Univerzita Pardubice

Fakulta ekonomicko-správní

# Pilier 3: Vplyv jazykovej zložitosti textu na správanie sa stakeholderov na webových stránkach komerčných bánk

Mgr. Ľubomír Benko, Ph.D.

Habilitačná práca

2024

**Prehlásenie**

Prehlasujem, že som túto prácu vypracoval samostatne. Všetky literárne pramene a informácie, ktoré som v práci použil, sú uvedené v zozname bibliografických odkazov.

Bol som oboznámený s tým, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., o autorskom zákone, o právach súvisiacich s autorským zákonom a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, najmä so skutočnosťou, že Univerzita Pardubice má právo na uzatvorenie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona, a s tým, že ak dôjde k použitiu práce mnou alebo bude poskytnutá licencia o využití inému subjektu, je Univerzita Pardubice oprávnená odo mňa požadovať primeraný príspevok na úhradu nákladov, ktoré na vytvorenie diela vynaložila, a to podľa okolností až do ich skutočnej výšky.

Beriem na vedomie, že v súlade s § 47b zákona č. 111/1998 Sb., o vysokých školách a o zmene a doplnení ďalších zákonov (zákon o vysokých školách), v znení neskorších predpisov a smernicou Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávaní, zverejňováni a formální úpravu závěrečných prací, v znení neskorších dodatkov, bude práca zverejnená prostredníctvom Digitálnej knižnice Univerzity Pardubice.

V Nitre dňa 20. 3. 2024

Mgr. Ľubomír Benko, Ph.D.

**Poďakovanie**

Chcel by som sa poďakovať mojim kolegom prof. RNDr. Michalovi Munkovi, PhD. a prof. RNDr. Daši Munkovej, PhD. za odborné rady a konzultácie počas môjho odborného napredovania. Poďakovať by som sa chcel aj spoluautorom odborných článkov, ktorých nápady a odborné znalosti viedli k môjmu napredovaniu v odbore.

Rád by som sa poďakoval svojmu domovskému pracovisku Katedre informatiky, FPVaI UKF v Nitre za vytvorenie priaznivého pracovného prostredia a podmienok na odborný rast. Taktiež by som sa chcel poďakovať Fakulte ekonomicko-správní Univerzity Pardubice za možnosť predložiť habilitačnú prácu.

Obzvlášť sa chcem poďakovať mojej manželke Mgr. Lucii Benkovej, PhD. za podporu a trpezlivosť počas tvorby práce.

**Anotácia**

Predkladaná habilitačná práca sa zameriava na analýzu zložitosti a čitateľnosti dokumentov týkajúcich sa povinne zverejňovaných informácií Pilier 3 na webovom portáli komerčných bánk a na skúmanie ich vplyvu na správanie sa stakeholderov. Výsledky experimentov ukázali, že zverejňovanie informácií Pilier 3 nie je potrebné počas celého roka, ale optimálne iba na začiatku, v prvých týždňoch roka. Záujem stakeholderov sa neobmedzuje len na povinne zverejňované informácie, ale zaujíma ich aj širší kontext informácií o banke. Analýza obsahu webu je realizovaná prostredníctvom metód spracovania prirodzeného jazyka s dôrazom na skúmanie vplyvu zložitosti a čitateľnosti dokumentov na správanie sa stakeholderov. Navrhnutá metodika je založená na kombinácií dát o používaní, štruktúre a obsahu webu, čo viedlo k návrhu ukazovateľov preferencií používateľov na webových portáloch komerčných bánk. Preukázal sa súvis medzi ukazovateľmi preferencií používateľov a metrikami zložitosti alebo čitateľnosti textu. Metriky základných charakteristík čitateľnosti textu ako početnosti tokenov, viet alebo znakov, dosiahla najvyššiu mieru závislosti s preferenciou používateľov. Navyše sa ukázalo, že stakeholderi preferujú rozsiahlejšie dokumenty pred krátkymi dokumentami, z dôvodu ich informačnej bohatosti.

**Kľúčové slová**

Web Usage Mining, Príprava dát, Reference Length, Data Mining, Web Content Mining, Spracovanie prirodzeného jazyka, Čitateľnosť textov, Zložitosť textov, Evalvácia strojového prekladu.

**Title**

Pillar 3: The influence of language complexity on the behaviour of stakeholders on the websites of commercial banks

**Annotation**

The presented habilitation thesis focuses on analyzing the complexity and readability of documents related to the Pillar 3 information disclosures on the web portal of commercial banks and examining their impacts on stakeholder behavior. The results of experiments have shown that Pillar 3 information disclosure is not necessary throughout the year but ideally in the first weeks of the year. The interest of stakeholders is not limited only to mandatory information, but also includes a broader context of bank information. The analysis of web

content is conducted through natural language processing methods with an emphasis on examining the impact of document complexity and readability on stakeholder behavior. The proposed methodology is based on the combination of data on usage, structure, and content of the website and has led to the design of user preference indicators on the web portals of commercial banks. A correlation has been demonstrated between user preference indicators and text complexity or readability metrics. Readability metrics like token, sentence or character frequency have achieved the highest degree of dependence on user preferences. Moreover, it has been shown that stakeholders prefer extensive documents over short documents, due to their information richness.

**Keywords**

Web Usage Mining, Data preparation, Data Mining, Web Content Mining, Natural Language Processing, Text readability, Text complexity, Evaluation of machine translation.

# OBSAH

# ZOZNAM OBRÁZKOV A TABULIEK

# ZOZNAM SKRATIEK

AWL – Academic Word List

BCBS – Basel Committee on Banking Supervision

CBMT – Corpus Based Machine Translation

CEE – Central and Eastern Europe

CTTR – Corrected Type-token ratio

EAWL – Economic Academic Word List

EBMT – Example-Based Machine Translation

GT – Google Translate

MT – Machine Translation

NDW – Number of Different Words

NGSL – New General Service List

NLP – Natural Language Processing

NMT – Neural Machine Translation

PER – Position-independent Error Rate

POS – Part-of-Speech

RBMT – Rule Based Machine Translation

RTTR – Root Type-token ratio

SMT – Statistical Machine Translation

TER – Translation Error Rate

TTR – Type-token ratio

UIH – User interest horizontal

UIV – User interest vertical

WCM – Web Content Mining

WER – Word Error Rate

WUM – Web Usage Mining

# ÚVOD

V každej ekonomike zohrávajú kľúčovú úlohu finančné inštitúcie a zvlášť, ak sú medzinárodne aktívne. Vzhľadom na ich veľký dopad v prípade zlyhania na národné ekonomiky ako aj globálny svet, Bazilejský výbor pre bankový dohľad (BCBS) vytvoril celú radu medzinárodne platných štandardov v oblasti regulácie týchto inštitúcií, známych ako Bazilejské rámce/dohody. Tie sa historicky vyvíjali od roku 1988 a dnes hovoríme už o Bazilej IV, resp. Bazilej V. Na tvorbe týchto dohôd sa významne podieľa Európska komisia, a to prostredníctvom svojej inštitúcie European Banking Authority (EBA). Európska únia sa totiž zaviazala plne implementovať štandardy vyplývajúce z Bazilejských rámcov. EBA podporuje BCBS s cieľom posilniť reguláciu, dohľad a manažovanie rizík bankového sektora. Od roku 2008 sú bazilejské štandardy budované na troch pilieroch: Pilier 1 – minimálne kapitálové požiadavky, Pilier 2 – proces kontroly vykonávaný bankovým dohľadom a Pilier 3 – trhová disciplína. S cieľom poskytnúť účastníkom trhu dostatok informácií a podporovať trhovú disciplínu, stanovil Bazilejský výbor komplexný súbor požiadaviek týkajúcich sa tejto oblasti. Rámec Pilier 3 poskytuje komplexný balík všetkých existujúcich požiadaviek na zverejňovanie nad rámec požiadaviek na kapitálové požiadavky. Pokiaľ nie je uvedené inak, rámec sa vzťahuje na všetky medzinárodne aktívne banky až na najvyššej konsolidovanej úrovni. Podobne ako bazilejské dohody aj Pilier 3 od uvedenia do účinnosti prechádzal viacerými revíziami. Revidované zverejnenia Pilier 3 vychádzajú z piatich hlavných princípov, ktorých základom sú skúsenosti získané z obdobia finančnej krízy v rokoch 2007 – 2009: zrozumiteľnosť, komplexnosť, zmysluplnosť/užitočnosť, konzistentnosť v čase a porovnateľnosť. Frekvencia zverejňovania údajov finančnými inštitúciami sa pohybuje medzi štvrťročnou, polročnou a ročnou frekvenciou, v závislosti od povahy požiadavky.

Komerčné banky v krajinách CEE (Central and Eastern Europe – stredná a východná Európa) majú množstvo špecifík, ktoré nie vždy sú adekvátne zohľadnené v príslušných reguláciách a štandardoch. Spomedzi nich je dôležité spomenúť prevládajúce vlastníctvo veľkými nadnárodnými skupinami, ich právna forma ako subjektov, s akciami, s ktorými sa neobchoduje na kapitálových trhoch (a z toho vyplývajúca aj iná požiadavka na informácie zo strany stakeholderov) alebo biznis modely založené na depozitných klientoch. Ich depozitný klienti – vkladatelia – predstavujú veľmi dôležitú skupinu zainteresovaných strán / stakeholderov. Avšak chýbajú empirické štúdie o správaní a záujmoch depozitných klientov pri využívaní informácií Pilier 3. Regulačné orgány nevedia, do akej miery sú existujúce pravidlá

zverejňovania zmysluplné, hodnotné pre používateľov a pomáhajú predchádzať zlyhaniu trhovej disciplíny, ako tomu bolo počas poslednej finančnej krízy. Pre dosiahnutie cieľov stanovených zo strany regulátorov je kľúčové, aby bol mechanizmus trhovej disciplíny účinný a využívaný v súlade s ich očakávaniami. Ako bolo spomenuté vyššie, v krajinách strednej a východnej Európy je nedostatok štúdií, ktoré by na základe relevantnosti obsahu pre kľúčové zainteresované strany komerčných bánk hodnotili zverejňovanie informácií podľa Pilier. Viaceré výskumy prezentované v tejto práci sú preto zamerané práve na analýzu záujmu o zverejňovanie informácií v rámci Pilier 3 – Trhová disciplína, komerčnými bankami, ktorých akcie nie sú verejne obchodované a kľúčovými stakeholdermi sú depozitní klienti. Význam tejto skupiny podporuje aj fakt, že napríklad na Slovensku takmer polovicu vkladov depozitných klientov tvoria nepoistené vklady a podobný stav možno očakávať aj v iných krajinách strednej a východnej Európy. A práve pre týchto klientov dostupnosť a zrozumiteľnosť informácií o finančnom a rizikovom profile ich finančnej inštitúcie by mali byť jednými z kľúčových faktorov pokiaľ chcú správne manažovať riziká vyplývajúce z umiestnenia ich vkladov do týchto inštitúcií.

## CIEĽ PRÁCE

Hlavný cieľ predkladanej habilitačnej práce vychádza z obsahu a používania webu, cieľom je návrh metodiky zameranej na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk ako aj na skúmanie ich vplyvu na správanie sa stakeholderov (medzi ktorých môžu patriť klienti, akcionári, regulačné orgány, ale aj široká verejnosť). Navrhovaná metodika kombinuje dáta o používaní, štruktúre a obsahu webu, čím prepája všetky domény webu. Experiment sa realizoval na dátach bankovej inštitúcie, ktoré boli získané za rok 2018 a dokumentami získanými z webového portálu (Príloha M). Naplnenie cieľa je založené na syntéze odbornej práce autora, ktorá je tvorená zjednocujúcim komentárom odkazujúcim na jednotlivé vedecko-výskumné práce zamerané na analýzu správania sa stakeholderov na webovom portáli a na analýzu dát o obsahu webu. Práce boli publikované v impaktovaných vedeckých časopisoch, na zahraničných vedeckých konferenciách a ako kapitoly v monografii. Publikované práce sú súčasťou prílohy habilitačnej práce (Príloha A - M).

Prvým čiastkovým cieľom je **skúmanie správania sa stakeholderov na webovom portáli bankovej inštitúcie pomocou rôznych prístupov**. Za týmto účelom sa realizovala séria experimentov zameraných na analýzu správania sa stakeholderov na webovom portáli.

V prvom rade bolo nutné realizovať predspracovanie dát v procese získavania znalostí z používania webu a zvolenie si vhodnej metodiky na získanie spoľahlivých dát o používaní webu (Príloha C a G). Ďalším krokom bola analýza správania sa stakeholderov v prostredí bankovej sféry (Príloha A, B, D, E a F).

Druhým čiastkovým cieľom je **analýza obsahu webu, vo forme kolekcie dokumentov, pomocou rôznych metód spracovania prirodzeného jazyka**. Viaceré odborné dokumenty a ich lokalizácia pre daný región je v dnešnej dobe vytváraná pomocou strojového prekladu. Z tohto dôvodu sa prvá séria experimentov zamerala na evalváciu strojového prekladu pomocou automatických metrík presnosti prekladu (Príloha J a K). Druhá časť experimentov sa zamerala na skúmanie zložitosti a čitateľnosti týchto dokumentov prostredníctvom rôznych automatických mier (Príloha H, I a L).

# 1 ZÍSKAVANIE ZNALOSTÍ Z WEBU

Pojem získavanie znalostí označuje proces, v ktorom sa pomocou jednej alebo viacerých data mining-ových techník hľadajú vo veľkých dátových zdrojoch vzory, ktoré slúžia na získanie užitočných informácií (Loshin 2013). Proces získavania znalostí si vyžaduje značné množstvo údajov, ktoré musia byť v spoľahlivom stave skôr ako budú podrobené samotnej analýze dát. Pod týmto procesom sa môžu rozumieť úlohy, ktoré zahŕňajú výber dát, predspracovanie dát, transformáciu dát, analýzu dát a interpretáciu výsledkov (Fayyad et al. 1996). Práve pomocou data minig-ových techník sa môžu analyzovať rôzne zdroje dát ako napríklad webové portály, texty alebo databázy a získať z nich rôzne zaujímavé znalosti.

Táto kapitola sa zameriava na dva zdroje dát, ktoré spolu úzko súvisia – dáta o používaní webu a dáta o obsahu webu. Informácie na webe sú najčastejšie uložené v textovej podobe, preto prepojenie oblastí získavania znalostí z webu a znalostí z textu môže priniesť zaujímavé vzory v skúmaní správania sa návštevníkov webových portálov.

## 1.1 ZÍSKAVANIE ZNALOSTÍ Z POUŽÍVANIA WEBU

Webové portály sú zdrojom informácií vyhľadávanými používateľmi. Web Usage Mining (WUM – získavanie znalostí z webu) v sebe zahŕňa porozumenie správania sa používateľov pri navštevovaní webových stránok. Podobnú filozofiu je možné použiť aj pre používateľov informačných systémov, ktorých správanie sa v systéme môže odhaliť prípadné chyby alebo prispieť k vylepšeniu systému. Na zaznamenávanie stôp, či už na webových portáloch alebo v informačných systémoch, slúžia logovacie súbory. Skúmanie logovacích súborov odhalí nielen správanie, ale aj návyky používateľov. Nakoľko sa v logovacích súboroch zaznamenávajú hlavne anonymné údaje, je nutné ich spracovať a pripraviť na analýzu pomocou metódy predspracovania dát. Predspracovanie dát je dôležitou súčasťou WUM, a pre tento účel bolo navrhnutých množstvo techník predspracovania. Cieľom získavania znalostí na základe používania webu je analýza správania sa používateľov pri prechádzaní webu (Srivastava et al. 2000; Romero et al. 2009). Dáta o používaní webu sa zaznamenávajú do logovacieho súboru webového servera, kde je možné z veľkého objemu dát získať informácie pre ich lepšie porozumenie.

Prvou fázou v procese získavania znalostí je porozumenie problematike. Cieľom tejto fázy je pochopiť ciele problému formulovaného z hľadiska modelovania dát. Medzi úlohy získavania znalostí patrí deskripcia a sumarizácia, segmentácia, deskripcia konceptov,

klasifikácia, predikcia a analýza závislostí (Liu 2011). V druhej fáze je cieľom získanie relevantných dát o používaní webu. Zdrojom sú dáta o používaní webu, prípadne informačných systémov a pod. Informačné systémy zväčša evidujú údaje o používaní systému vo vlastnej štruktúre, prevažne vo forme databázy. V prípade webových a proxy serverov sú dáta zaznamenávané v spoločnej štandardnej štruktúre v textovom formáte – v logovacom súbore. Logovací súbor v štandardnej štruktúre – Common Log File (W3C 1995) zaznamenáva informácie o IP adrese, čase a dátume návštevy a pristupovanom objekte. V prípade rozšírenej podoby (Extended Log File – ELF) dokáže zaznamenávať aj údaje o odkazovanom objekte a verzii prehliadača používateľa – User Agent (Liu 2011).

### 1.1.1 PREDSPRACOVANIE DÁT O POUŽÍVANÍ WEBU

Podmienkou dobrej analýzy sú kvalitné dáta. Logovacie súbory sú typické tým, že obsahujú značné množstvo nepodstatných údajov, ktoré by mohli analýzu dát ovplyvniť. V prípade skúmania správania sa používateľov alebo návštevníkov webového portálu je možné pre získanie logovacieho súboru použiť nasledovné metódy (Munk et al. 2010):

- výberové zisťovanie – zisťujú sa odpovede na konkrétne položky dotazníka a návštevník webu si je vedomý predmetu skúmania (Cerna a Poulova 2008),
- Web Usage Mining – analyzuje sa logovací súbor webového servera, ktorý obsahuje informácie o prístupoch na stránky webového portálu bez vedomosti návštevníka, pričom sú jeho údaje do istej miery anonymné (Cooley et al. 1999).

Predpokladom pre prácu s kvalitnými údajmi nie je len ich zber, ale aj príprava pre ďalšie analýzy (príprava dát sa v angličtine označuje pojmami data preprocessing alebo data preparation). Z dôvodu množstva irelevantných údajov v logovacích súboroch, ktoré treba odstrániť, je fáza prípravy dát nielen časovo najnáročnejšia, ale aj veľmi prácna. Losarwar a Joshi (2012) pri analyzovaní fázy prípravy dát vo WUM dospeli k záveru, že v oblasti analýzy webu je táto fáza veľmi dôležitá a vyžaduje si použitie nástrojov, ktoré sa zvyčajne na prípravu dát v iných doménach nepoužívajú. V prípade portálov virtuálneho vzdelávacieho prostredia, autori (Sael et al. 2013) prišli s vlastnou úpravou logovacieho súboru, čím minimalizovali nutnosť prípravy dát a rovno extrahovali všetky potrebné údaje pre analýzu. Napriek tomu, toto riešenie nie je použiteľné v prípade portálov s anonymným prístupom, kde je nutné postupovať klasickým procesom prípravy dát.

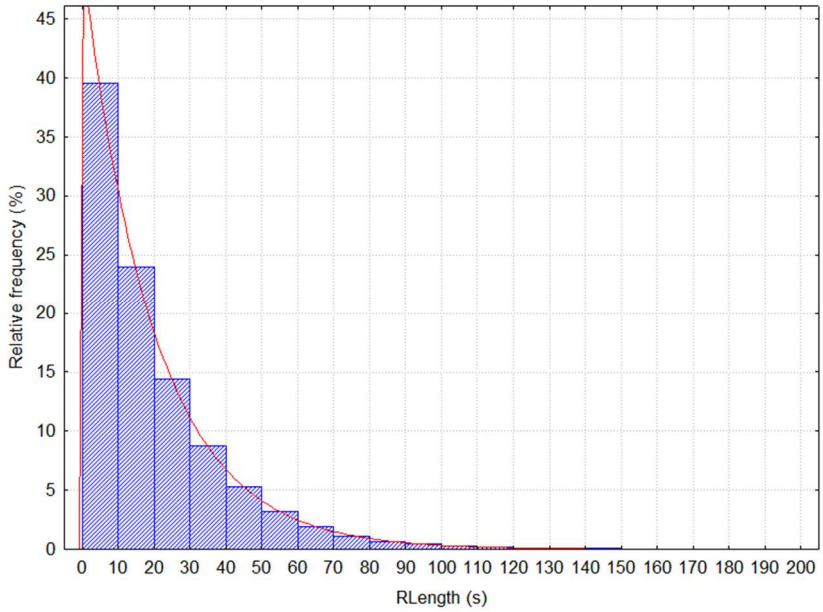
Príprava dát zahŕňa niekoľko krokov. Prvým krokom je čistenie dát od nepotrebných údajov, kde cieľom je odstránenie záznamov, resp. odkazov, ktoré sú irelevantné pre skúmanie správania sa používateľov (Cooley et al. 1999). Medzi takéto odkazy patria hlavne prístupy k obrázkom, flash videám, ikonám kurzora, javascriptom alebo štýlom. Postup identifikácie takýchto záznamov zvyčajne zahŕňa identifikáciu na základe prípony (*.jpg, *.jpeg, *.bmp, *.png, *.gif, *.css, *.js, *.flw, *.swf, *.cur, *.rss, *.ico, *.xml a podobne). Aj pri načítaní len jednej stránky sa všetky tieto požiadavky zapíšu do logovacieho súboru. Okrem požiadavky GET sa do logovacieho súboru zapisujú aj ďalšie požiadavky http protokolu, pričom je potrebné odstrániť aj návratové kódy 4xx/5xx, ktoré identifikujú chybu klienta/servera. Aye (2011) predstavil dva algoritmy pre získavanie dát z databáz. Jeho algoritmus pre čistenie dát v sebe navyše zahŕňal informácie o počte zmazaných údajov a počte unikátnych prístupov na skúmaný webový portál. Srivastava et al. (2015) predstavili algoritmus určený na čistenie logovacieho súboru od nepotrebných dát, v ktorom využívajú daný časový interval a taktiež dokážu vytvoriť sekvenciu záznamov. Nevýhodou predstaveného algoritmu je práca s veľkým objemom dát, kde v prípade čistenia väčších logovacích súborov dochádza k značnému spomaleniu. Spomínaní autori sa nezaoberali čistením logovacích súborov od prístupov robotov vyhľadávacích služieb.

Ďalším krokom čistenia dát je odstránenie prístupov robotov vyhľadávacích služieb ako napríklad Google, Yahoo, Bing a pod. Nakoľko roboty pristupujú k webovému portálu sekvenčne, tak nie je vhodné zahrnúť ich aktivitu do skúmania správania sa používateľov. Detekcia robotov prebieha buď na základe ich identifikácie v poli User Agent, alebo na základe IP adresy, ktorú je možné porovnať s databázou robotov (napr. www.robotstxt.org) (Cooley et al. 1999). Vellingiri a Chenthur Pandian (2011) sa sústredili na zlepšenie techník na čistenie dát, hlavne na odstraňovanie prístupov robotov. Okrem už vyššie spomínaných nepotrebných dát a prístupov robotov autori odstránili z logovacieho súboru všetky prístupy, ktoré mali dĺžku prístupu kratšiu ako dve sekundy.

V logovacom súbore sa prioritne zaznamenávajú anonymné údaje o používateľoch, čím vzniká problém s jednoznačnou identifikáciou návštevníka webu. Pri analýze nie je potrebné poznať konkrétnu identitu používateľa, ale rozlišovať medzi jednotlivými používateľmi. Avšak predpoklad, že na identifikáciu používateľa stačí IP adresa, je nesprávny, pretože za jednou IP adresou sa môže nachádzať viacero používateľov. Z toho dôvodu je nutné kombinovať viaceré metódy, ako napríklad využitie poľa Cookie (Pabarskaite a Raudys 2007), prípadne kombinácie IP adresy s poľom User Agent (Srivastava et al. 2000). Viaceré heuristické metódy

využívajú kombináciu IP adresy s poľom User Agent. Ak nastane zmena IP adresy, je zrejmé, že ide o nového používateľa. Ak je IP adresa rovnaká, porovnáva sa pole User Agent, ak nastane zmena, je identifikovaný nový používateľ, v opačnom prípade ide o toho istého používateľa (Srivastava et al. 2000). V prípade, že portál vyžaduje od používateľa registráciu, resp. prihlásenie, je identifikácia používateľov zjednodušená z dôvodu existencie záznamu v logovacom súbore. Používateľ môže navštíviť stránku viackrát, pričom v logovacom súbore sú zaznamenané viacnásobné sedenia (návštevy) pre každého používateľa. Cieľom identifikácie sedení je rozdeliť jednotlivé prístupy každého používateľa do oddelených relácií (Cooley et al. 1999). Sedenie môže byť definované ako postupnosť krokov, ktoré vedú k naplneniu určitej úlohy (Spiliopoulou a Faulstich 1999) alebo ako postupnosť krokov, ktoré vedú k dosiahnutiu určitého cieľa (Ming-Syan Chen et al. 1998). Na identifikáciu sedení sa používajú štruktúrovo-orientované heuristiky, časovo-orientované heuristiky (Liu 2011; Berendt et al. 2003), ako aj kombinácie týchto dvoch prístupov, ktoré predstavujú zaujímavý prístup (Munk a Kapusta 2014).

Metóda Reference Length patrí do skupiny heuristík, ktoré sú kombináciou štruktúrovo a časovo-orientovaných heuristík. Reference Length je založená na predpoklade, že dĺžka času stráveného používateľom na stránke je vo vzťahu s tým, či je stránka klasifikovaná ako obsahová alebo navigačná (Munk a Kapusta 2014; Kapusta et al. 2012a). Na obrázku (Obrázok 1) je znázornený histogram popisujúci rozdelenie premennej Length, ktorá slúži na reprezentáciu času stráveného na stránke webového portálu. Predpokladá sa, že ľavá strana grafu predstavuje navigačné stránky. Tie slúžia návštevníkom hlavne na rýchly prechod k obsahovým stránkam, ktoré sú ich cieľom. Z toho dôvodu pravú stranu tvoria stránky s obsahom, ktorých dĺžka stráveného času má väčší rozptyl.




**Obrázok 1** – Rozdelenie premennej RLength (Munk a Benko 2018)

Na základe predpokladu exponenciálneho rozdelenia premennej je možné vypočítať hraničný čas $C$, ktorý slúži na rozlíšenie navigačných stránok od obsahových. Premenná $RLength$ má exponenciálne rozdelenie

$$f(RLength) = \lambda e^{-\lambda RLength}, \tag{1}$$

$$F(RLength) = 1 - e^{-\lambda RLength}, \tag{2}$$

kde $RLength \geq 0$.

Ak je $p$ relatívna početnosť navigačných stránok, potom sa na odhad hraničného času $C$ využije kvantilová funkcia

$$F^{-1}(p, \lambda) = C = \frac{-\ln(1-p)}{\lambda}, \tag{3}$$

pre $0 \leq p < 1$. Maximálne vierohodný odhad parametra $\lambda$ (priemerná intenzita udalostí) je

$$\hat{\lambda} = \frac{1}{\overline{RLength}}, \tag{4}$$

kde $\overline{RLength}$ je pozorovaný priemer dĺžky návštev.

V okamihu, keď je odhadnutý hraničný čas, sedenie môže byť identifikované porovnaním každého času stráveného na stránke s hraničným časom. Práve hraničný čas rozdelí stránky na navigačné a obsahové podľa dĺžky času stráveného na konkrétnej stránke (Munk a Benko

2018; Munk et al. 2015). Následne, sedenie je sekvencia navštívených stránok s časovou známkou, pre ktorú platí:

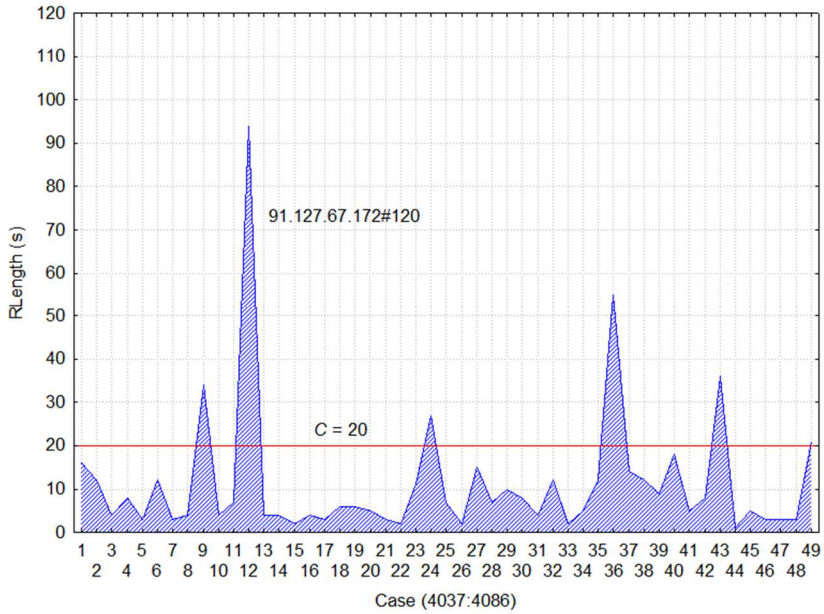
$$\langle USID, \langle URL_1, DTime_1, RLength_1 \rangle, \ldots, \langle URL_k, DTime_k, RLength_k \rangle \rangle, \quad (5)$$

$$RLength_i \leq C, \quad (6)$$

kde $1 \leq i < k$ a pre poslednú stránku sedenia platí:

$$RLength_k > C. \quad (7)$$

Podľa metódy Reference Length je nové sedenie definované od stránky s vlastnosťou (7), pričom prvých $k - 1$ stránok je klasifikovaných ako navigačné stránky a posledná $k$-tá stránka je klasifikovaná ako obsahová.




**Obrázok 2** – Metóda Reference Length (Munk a Benko 2018)

Na obrázku (Obrázok 2) je znázornená sekvencia navštívených stránok z danej IP adresy a agenta, ktorá je usporiadaná podľa času prístupu (os *x*) a času stráveného na stránke (os *y*). Hraničný čas bol 20 sekúnd, kde prvé sedenie je tvorené stránkami 1 až 9, prvých 8 je klasifikovaných ako navigačné stránky a posledná je obsahová stránka. Analogicky sa postupuje pri identifikácií ďalších sedení.

Ak premenná *length* nemá exponenciálne rozdelenie, tak autori (Munk et al. 2013; Kapusta et al. 2013; 2014; Munk et al. 2017a) sa prikláňajú k odhadom na základe kvartilového rozpätia, ktoré nie sú ovplyvnené odľahlými hodnotami, napr. $Q_{III} + 1{,}5Q$, kde $Q_{III}$ je horný kvartil (75. percentil) a $Q$ je kvartilové rozpätie (stredných 50 % hodnôt), t. j. ak je čas na stránke

považovaný za odľahlú hodnotu, začína sa nové sedenie. V opačnom prípade je lepšie použiť na identifikáciu sedení metódu Reference Length.

Dĺžka času stráveného na stránke je daná rozdielom prístupových časov súčasnej stránky a nasledujúcej, pričom sa nedá vypočítať čas poslednej stránky v sekvencii. Metóda Reference Length predpokladá, že každá posledná stránka je obsahová stránka. Môže sa však stať, že z dôvodu neočakávanej udalosti na strane návštevníka (napr. telefonát) je obsahová stránka klasifikovaná ako navigačná stránka. Rovnako je potrebné vziať do úvahy skutočnosť, že pre každého používateľa môže byť každá stránka rôzne klasifikovaná, pre jedného to môže byť navigačná stránka, ale pre druhého obsahová a naopak.

## 1.2 ZÍSKAVANIE ZNALOSTÍ Z OBSAHU WEBU A ZO SPRACOVANIA PRIRODZENÉHO JAZYKA

Návštevníci vyhľadávajú na webových portáloch informácie, ktoré tvoria ich obsah. Web Content Mining (WCM – získavanie znalostí z obsahu webu) extrahuje užitočné informácie alebo znalosti zo štruktúrovaných a neštruktúrovaných dát webu (Liu 2011). Obsah webu môže pozostávať z textu, obrázkov, zvuku, videa alebo štruktúrovaných dát, ako napr. tabuliek. Preto je WCM veľmi prepojený s oblasťou Text Mining (získavanie znalostí z textu). Medzi niektoré z riešených problémov WCM patrí extrahovanie kľúčových slov, zoskupovanie webových dokumentov a klasifikácia webových stránok (Srivastava et al. 2005). Klasifikácia webových stránok je proces priradenia kategórií webovým stránkam na základe predom stanovených kategórií. Z pohľadu návštevníkov webu predstavuje informatívne najdôležitejšiu časť obsahu práve text. V prípade webových portálov komerčných bánk predstavujú textové informácie nielen obsah webu, ale nachádzajú sa aj vo forme dokumentov. Návštevník webu na stránkach strávi určitý čas s cieľom získať informáciu, na tento čas môže mať vplyv viacero faktorov, medzi ktorými môže byť okrem štruktúry webu, aj zložitosť a čitateľnosť hľadaného textu. Na skúmanie textov sa používajú metódy spracovania prirodzeného jazyka (NLP – Natural Language Processing). Jednou z domén NLP je strojový preklad (MT – Machine Translation). Povinne zverejňované informácie na webových portáloch bankovej inštitúcie môžu byť lokalizované do rôznych jazykov. V prípade nedostupnosti lokalizovanej jazykovej verzie povinne zverejňovaných informácií sú s veľkou pravdepodobnosťou tieto informácie preložené pomocou strojového prekladu. Z toho dôvodu bolo nutné sa venovať kvalite strojového prekladu a predstaviť spôsoby evalvácie strojového prekladu. Existujú dva prístupy k procesu strojového prekladu, jeden je založený na pravidlách (Rule Based Machine Translation -

RBMT) a druhý na korpuse (Corpus Based Machine Translation - CBMT). RBMT stavia na lingvistických pravidlách a vyžaduje si širokú škálu gramatických pravidiel. Tento prístup si vyžaduje analýzu a reprezentáciu významu východiskového textu a syntézu (generovanie) jeho ekvivalentu v cieľovom jazyku. Dôležitým kritériom je vytvorenie abstraktnej reprezentácie textu, ktorá je po lexikálnej aj štrukturálnej stránke jednoznačná. Na druhej strane, prístup CBMT nie je založený na pravidlách a gramatike, ale ide o zarovnaný dvojjazyčný korpus. CBMT systém obsahuje v sebe korpusy v strojovo čitateľnej forme, ktoré sú písomného alebo hovoreného charakteru. CBMT si taktiež vyžaduje prekladové znalosti z veľkých dvojjazyčných korpusov. K systémom strojového prekladu založených na korpusoch patrí strojový preklad založený na príkladoch (Example-Based Machime Translation - EBMT) a štatistický strojový preklad (Statistical Machine Translation - SMT).

Štatistický strojový preklad je podobný strojovému prekladu založenému na príkladoch (EBMT). Napriek tomu, že oba prístupy vyžadujú rozsiahle bilingválne korpusy, systémy EBMT „sa učia“ na základe príkladov a systémy SMT na základe štatistiky.

SMT pozostáva z dvoch častí:

- v získaní translačného modelu, tak aj jazykového modelu cieľového textu,
- v dekódovaní východiskovej vety, t. j. v nájdení vhodného prekladu cieľovej vety ($c$), k východiskovej ($v$) s čo najväčšou pravdepodobnosťou.

Translačný model, charakterizujúci adekvátnosť prekladu, poskytuje informáciu o tom, aká je pravdepodobnosť, že reťazec (slovo/fráza) je adekvátnym prekladom iného reťazca, ktorý je natrénovaný na paralelných textoch bilingválneho korpusu. Jazykový model zasa charakterizuje plynulosť prekladu, poskytuje informáciu o tom, aká je pravdepodobnosť, že reťazec (slovo/fráza) je dobre formovaný (štruktúrovaný), pričom je natrénovaný na monolingválnom korpuse cieľového jazyka.

V dnešnej dobe je už bežný neurónový prístup ku strojovému prekladu, ktorý je založený na neurónových sieťach – neurónový strojový preklad (Neural Machine Translation, NMT). NMT dosiahol vynikajúce výkony pri veľkoobjemových prekladoch z angličtiny do francúzštiny (Luong et al. 2015), ako aj z angličtiny do nemčiny (Jean et al. 2015). Spoločnosti Google alebo Systran od konca roku 2016 postupne integrujú NMT do svojich prekladacích nástrojov. NMT chápe vetu ako celok a vytvára asociácie medzi frázami aj v dlhších vetách, t. j. prečíta si všetky východiskové slová až po koniec vety a až potom naraz

začne „vysielať“ jedno cieľové slovo. Výhodou NMT je schopnosť „priamo sa učiť“ (end-to-end), t. j. všetko sa učí ako jednu veľkú úlohu, pri učení neexistujú žiadne extra alebo medzikroky. SMT je založený na frázach, kde rozdeľuje východiskové segmenty (vety) na frázy (Koehn 2010). Počas trénovania SMT vytvára translačný a jazykový model. Počas prekladania vyberie dekóder preklad, ktorý je na základe týchto dvoch modelov najpravdepodobnejší. V princípe SMT dokáže produkovať veľmi dobré výsledky na úrovni fráz v zmysle adekvátneho transferu významu východiskového textu do cieľového jazyka. Avšak častokrát plynulosť prekladu v cieľovom jazyku (gramatika) nespĺňa požadovanú kvalitu. NMT je jedna rozsiahla neurónová sieť, ktorá je natrénovaná formou end-to-end, má malú pamäťovú stopu a schopnosť dobre generalizovať veľmi dlhé sekvencie slov. Dokáže spracovávať východiskové segmenty a transformovať ich na cieľové segmenty, pričom NMT prechádza celými vetami nielen frázami. Nepotrebuje uchovávať rozsiahle translačné (frázové tabuľky) a jazykové modely. Implementovanie NMT dekodéra je jednoduché v porovnaní s ostatnými strojovými prekladmi ako SMT. NMT využíva hlboké strojové učenie (deep learning), ktoré je reprezentované neurónovou sieťou (Bessenyei 2017). Výhodou NMT je, že sa vyhýba mnohým „veľmi krehkým“ prekladateľským návrhom v tradičnom strojovom preklade založenom na frázach. V praxi to znamená, že častokrát zlepšuje plynulosť prekladu na úkor adekvátnosti prekladu. NMT niekedy vyprodukuje vety, ktoré významovo nekorešpondujú s východiskovým textom, čo vyúsťuje do významových posunov a pomerne prekvapujúcich prekladov (hlavne pri natrénovaní na rozsiahlych textových dátach mimo domény).

### 1.2.1 EVALVÁCIA STROJOVÉHO PREKLADU

Napriek tomu, že manuálna evalvácia sa považuje za najspoľahlivejšiu, existujú problémy, s ktorými si nevie poradiť. Papineni et al. (2002) konštatovali, že metódy a metriky manuálnej evalvácie sú príliš pomalé a finančne náročné pre rozvoj systémov strojového prekladu, pre ktorý je dôležitá rýchla spätná väzba o kvalite prekladu. V snahe o zefektívnenie hodnotenia kvality prekladu sa začalo uvažovať o automatických metódach evalvácie strojového prekladu bez intervencie ľudského posudzovateľa. Bolo navrhnutých niekoľko automatických metrík hodnotenia MT kvality za účelom zníženia „času a ľudskej námahy“ počas evalvácie. Metriky automatickej evalvácie poskytujú vysokú efektívnosť a konzistentnosť pri pomerne nízkych nákladoch. Väčšinou sú založené na meraní podobnosti medzi strojovým prekladom, ktorý hodnotí – kandidátom a humánnym (ľudským) prekladom – referenčným. Metriky automatickej evalvácie môžu byť založené na štatistických princípoch (n-gramy alebo

vzdialenosť editovania) alebo na používaní lingvistických štruktúr na morfologickej, syntaktickej alebo sémantickej úrovni.

Medzi najpoužívanejšie metriky patria metriky založené na lexikálnej podobnosti, t. j. určujú podobnosť medzi kandidátom a referenciou na úrovni slov a znakov. Do tejto kategórie patria metriky ako BLEU, METEOR, WER, PER, TER, ROUGE alebo NIST. Lexikálna podobnosť medzi hypotézou (strojovým prekladom) a referenciou (reprezentovanou zväčša ľudským prekladom) sa dá určiť na základe Jaccardovej podobnosti, kosínusovej podobnosti alebo Levenstheinovej podobnosti.

Metrika BLEU (*Bilingual Evaluation Understudy*), ktorú navrhli Papineni a kol. (2002) poskytuje rýchly a lacný spôsob evalvácie modelov MT. Výpočet skóre metriky BLEU (8) je jednoduchý a nezávislý od jazyka. Napriek viacerým nedostatkom tejto metriky je stále vnímaná ako štandardná metrika a nové, navrhnuté metriky sa s ňou porovnávajú. Metrika bola navrhnutá na evalváciu strojového prekladu veľkých korpusov na úrovni n-gramov dĺžky 1-4. Častokrát je dĺžka kandidátskeho a referenčného prekladu rôzna. V takomto prípade sa na nájdenie najlepšieho prekladu používa parameter penalizácie krátkosti prekladu (*brevity penalty,* BP), pomocou ktorého sa zisťuje BP faktor (*multiplicative breavity penalty faktor).* BP faktor (10) nadobúda hodnoty medzi 0 a 1, pričom 1 znamená, že počet slov kandidátskeho a referenčného prekladu je rovnaký. Cieľom metriky je nájsť taký referenčný preklad, ktorý má čo najvyšší BP faktor. Následne sa penalizácia krátkosti prekladu vypočíta pre celý dokument:

$$BLEU(n) = BP\_factor \ * \ exp \sum_{i=1}^{n} Info(w_i) * \log precision_i \ , \qquad (8)$$

kde $Info(w_i)$ je váha *i*–tej premennej $precision_i$ a

$$Info(w_i) = \log_2 \frac{\sum_{r \epsilon R} \sum_{w_{i-1} \epsilon r} w_{i-1}}{\sum_{r \epsilon R} \sum_{w_i \epsilon r} w_i} \ , \qquad (9)$$

$$BP\_factor = min\left(1, \frac{length\_hyp}{length\_ref}\right) . \qquad (10)$$

BP je možné vypočítať pomocou rôznych počtov referenčných slov, čo vplýva aj na výpočet metriky BLEU, z toho dôvodu existujú rôzne verzie metriky BLEU, napríklad IBM BLEU používa vo svojom výpočte priemernú dĺžku referencií. Metrika BLEU patrí medzi tzv. metriky správnosti, čím sa skóre približuje k 1, tým je hypotéza podobnejšia referencii.

Alternatívny prístup ďalších metrík k hodnoteniu kvality strojového prekladu je z pohľadu vzdialenosti editovania (Levensteinova vzdialenosť). Cieľom automatických metrík je určenie miery chybovosti, nezhody, medzi hypotézou a referenciou. Medzi základné metriky chybovosti patria PER (Position – indipendent Error Rate), WER (Word Error Rate) a TER (Translation Error Rate).

Medzi novšie prístupy patria metriky založené na strojovom učení. Metriky evalvácie strojového prekladu založené na strojovom učení dosahujú oveľa lepšie výsledky ako štandardné metriky, najmä v prípade metrík založených na učení s učiteľom. Vyššiu koreláciu s ľudským hodnotením kvality je možné dosiahnuť použitím multilinguálnych a prispôsobiteľných modelov hodnotenia kvality MT. Tieto modely je možné použiť ako metriky na posúdenie kvality prekladu akéhokoľvek konkrétneho MT, čím sa proces evalvácie automatizuje a minimalizuje potrebu anotácie ľudským hodnotiteľom. Rei et al. (2020) navrhli neurónový rámec na trénovanie viacjazyčných modelov hodnotenia strojového prekladu (*Crosslingual Optimized Metric for Evaluation of Translation*, COMET). COMET sa učí pomocou neurónovej siete vyhodnotiť a predpovedať kvalitu MT pre viaceré rôzne jazyky na základe zhody medzi referenciou a hypotézou. Podobne ako pre BLEU, aj pre COMET platí, čím vyššie skóre, tým väčšia podobnosť medzi hypotézou a referenciou.

Základom rámca je medzijazykový kóder (Crosslingual Encoder) a združovacia vrstva (pooling layer). Ako medzijazykový kóder sa využíva vopred natrénovaný, krížovo-jazykový model (napr. BERT). Vďaka tréningu na údajoch z viacerých jazykov dosahuje dobré výsledky pri klasifikácii dokumentov ako aj pri zovšeobecňovaní pre neznáme jazyky (Rei et al. 2020). Rei et al. (2020) na základe očakávaných výsledkov rozlišuje dve možné architektúry pre metriku COMET:

- model odhadu (*Estimator Model*) – vstupom do modelu sú 3 segmenty: zdrojový text, referencia a hypotéza, ktoré sú nezávisle kódované pomocou predtrénovaného krížového jazykového kódovača so združovacou vrstvou. Vnorenia viet, ktoré sú výstupom združovacej vrstvy sa skombinujú a spoja do jedného vektora, ktorý predchádza do regresora s posuvným riadením (*Feed-Forward regressor*). Celý model je trénovaný na základe minimalizácie strednej kvadratickej chyby (*Mean Squared Error*).
- model hodnotenia prekladu (*Translation Ranking Model*) – vstupom do modelu sú 4 segmenty: zdrojový text, referencia, "lepšia" a "horšia" hypotéza, ktoré

sú kódované rovnako, ako v predchádzajúcom modeli. Model hodnotenia prekladu je trénovaný pomocou straty trojnásobnej marže (*Triplet Margin Loss*), pričom minimalizuje vzdialenosť medzi spomínanými segmentami.

### 1.2.2 ZLOŽITOSŤ A ČITATEĽNOSŤ TEXTU

Zložitosť textu zohráva dôležitú rolu v analýze textu. Pomocou rôznych metrík zložitosti textu dokážeme určiť, či dané texty sú vhodné pre cieľovú skupinu čitateľov alebo nie. Zásadnú rolu zohráva okrem štýlu textu aj jazyk, v ktorom je napísaný. Väčšina automatických metrík zložitosti textu je zameraná prevažne na anglické texty (Fisher et al. 2012). Vzdelávacia iniciatíva pripravujúca amerických študentov pre vzdelávanie a prax definuje vo svojich štandardoch (Common Core State Standards) zložitosť textu ako trojicu navzájom súvisiacich komponentov (Common Core State Standards Initiative 2023):

- Kvalitatívne dimenzie zložitosti textu: V štandardoch sa kvalitatívne dimenzie týkajú aspektov zložitosti textu, ktoré sú najlepšie merateľné alebo merateľné len vnímavým ľudským čitateľom, napr. úrovne významu alebo účelu, štruktúra, jazyková konvenčnosť alebo jasnosť, a požiadavky na znalosti;
- Kvantitatívne dimenzie zložitosti textu: vzťahujú sa na aspekty zložitosti textu, ako je dĺžka či frekvencia slova, dĺžka vety a kohéznosť (súdržnosť) textu, ktoré pre ľudského čitateľa sú ťažké, ak nie nemožné, efektívne vyhodnotiť, najmä v rozsiahlych textoch;
- Čitateľ a úloha: Zatiaľ čo predchádzajúce dva komponenty modelu sa zameriavajú na inherentné vlastnosti zložitosti textu, tretia dimenzia zložitosti textu sa zameriava na premenné špecifické pre konkrétnych čitateľov (ako je motivácia, znalosť a skúsenosť) a pre konkrétne úlohy (ako je účel a zložitosť úlohy). Napríklad, pri určovaní, či je text vhodný pre daného študenta, je potrebné zvážiť aj zadané otázky. Či na základe svojich doterajších vedomostí a skúseností vie odpovedať na danú otázku.

Pri analýze textu je okrem zložitosti dôležitá aj jeho čitateľnosť. Harris a Hodges (1995) definujú čitateľnosť ako jednoduchosť pochopenia textu pomocou štýlu písania, čím sa čitateľnosť textu rozširuje zo schopností čitateľa na analýzu štýlu písania. Naopak Collins a O'Brien (2003) definujú čitateľnosť ako kvalitu a zrozumiteľnosť písaného diela, pričom ide o text, ktorý je zrozumiteľný pre cieľového čitateľa. Čitateľnosť je možné chápať ako rovnováhu medzi schopnosťami čitateľa a samotným textom.

Kvantitatívne miery zložitosti textu sa zameriavajú hlavne na samotné charakteristiky slov a ich výskyt vo vetách a v odsekoch. Gunning (2003) predstavil viac ako sto metrík zložitosti textu, avšak iba zopár z nich sa v praxi používa. Analýza na úrovni slov patrí k prvej úrovni, na ktorú sa zameriava, pretože už samotná dĺžka slova (počet znakov) môže naznačovať do akej miery musí čitateľ slovo dekódovať, pričom jednoslabičné slová sú jednoduchšie na samotné čítanie a porozumenie textu ako viacslabičné. Početnosť slov však nie je možné chápať ako kompletnú metriku, pretože kontext, v ktorom sa slová nachádzajú môže zvyšovať zložitosť textu. Chall a Dale (1995) vytvorili zoznam slov, ktoré pomáhajú určiť zložitosť textu. Čím viac slov zo zoznamu sa v texte nenachádza, tým je daný text zložitejší. Ďalšou úrovňou v prípade kvantitatívnych mier čitateľnosti je dĺžka vety (počet slov) (Kintsch 1974) a s ňou súvisiace charakteristiky.

S ohľadom na plánovaný experiment a lepšiu prehľadnosť, bolo nutné rozdeliť metriky do viacerých kategórií. V prípade prekrytia niektorých kategórií (keď sa metriky nachádzali vo viacerých kategóriách) boli duplicitné metriky počítané pomocou iného nástroja. Zoznam skúmaných metrík a ich rozdelenie do kategórií je inšpirovaný (Lu 2012):

- **Charakteristiky textu** (Text characteristics) [char][1]: je skupina metrík, ktorá je zameraná na základné charakteristiky textu ako sú početnosti, priemer, medián (Gray a Leary 1935). Najčastejšie sa používa počet tokenov, počet viet alebo znakov ako aj počet unikátnych tokenov.
- **Čitateľnosť** (Readability) [read][1]: konvenčné metriky čitateľnosti vznikli hlavne z dôvodu nahradenia zastaralých metrík. Väčšina z nich vychádza z úrovne ročníka, ktorý študenti navštevujú, napríklad metrika Flesch-Kincaid grade level je založená na úrovni ročníkov v Spojených štátoch (Kincaid et al. 1975). Čitateľnosť textu môže byť vnímaná aj ako počet rokov učenia sa, potrebných na pochopenie daného textu: $Flesch - Kincaid = 0{,}39 * \left(\frac{total\ words}{total\ sentences}\right) + 11{,}8 * \left(\frac{total\ syllables}{total\ words}\right) - 15{,}59.$
  Metrika Flesch-Kincaid grade level je zameraná skôr na dĺžku vety, než na dĺžku slova. Ďalšou metrikou používanou v americkom školskom systéme je Gunning Fog Index, ktorá zároveň patrí medzi najpoužívanejšie metriky v súčasnej lingvistike (Spiers et al. 2017). Minimálnou požiadavkou pre výpočet metriky (indexu) je výber časti textu (okna), ktorý obsahuje aspoň sto slov, formálne zapísané: $index = 0{,}4 * \left[\left(\frac{words}{sentences}\right) +\right.$

[1] vektor skúmaných premenných [*x*], kde *x* označuje príslušnú kategóriu čitateľnosti alebo zložitosti textu

$100 * \left(\frac{complex\ words}{words}\right)\Big]$, pričom zložité slová pozostávajú z troch a viacslabičných slov. Texty, ktoré sú určené pre širšie publikum by mali mať index menší než 12. Medzi metriky čitateľnosti patria aj Coleman-Liau index (Coleman a Liau 1975), Automated Reliability Index (Senter a Smith 1967), SMOG (McLaughlin 1969), Flesch reading ease (Flesch 2016).

- **Lexikálna variácia** (Lexical variation) [lex_var][1]: sa vzťahuje na rozsah slovnej zásoby čitateľa v jeho jazykovom prejave (Malvern et al. 2004). Jednou zo základných metrík lexikálnej variácie je počet rôznych slov (number of different words – NDW), ktorá sa používa pri meraní jazykového vývoja dieťaťa (Klee 1992; Miller 1991). Nevýhodou tejto metriky je závislosť od dĺžky jazykovej vzorky, pretože nedokáže porovnať vzorky s rôznou dĺžkou. Možným riešením je skrátenie vzorky na jednotnú dĺžku na základe najkratšej vzorky (Thordardottir a Weismer 2001). Malvern et al. (2004) skracovanie vzoriek vnímajú ako mrhanie užitočnými dátami a preto navrhli dve metódy štandardizácie. V oboch prípadoch sa zo skúmanej vzorky náhodne vyberie súbor čiastkových vzoriek rovnakej dĺžky a následne sa spriemeruje ich NDW, aby sa aproximovala očakávaná hodnota NDW. V prvej metóde sa každá čiastková vzorka skladá zo štandardného počtu slov náhodne vybraných zo skúmanej vzorky. V druhej metóde obsahuje každá čiastková vzorka štandardný počet po sebe nasledujúcich slov zo skúmanej vzorky s náhodným počiatočným bodom. V súvislosti s lexikálnou variáciou počtu rôznych slov skúmal McClure (1991) rôzne pomery vybraných kontextových slovných druhov (počet slovies, podstatných mien, prídavných mien, prísloviek a ich modifikátorov, čo je kombinácia prídavných mien a prísloviek), s rovnakým menovateľom (počet lexikálnych slov).
- **Lexikálna bohatosť** (Lexical richness) [lex_rich][1]: sa vzťahuje na rozsah a rozmanitosť slovnej zásoby v skúmanom texte (McCarthy a Jarvis 2007). Používa sa v kombinácií s lexikálnou variáciou, hustotou a rozmanitosťou (diverzitou), a hovorí o počte rôznych výrazov v texte a rozmanitosti slovnej zásoby. Medzi najpoužívanejšie metriky lexikálnej bohatosti patrí Type-token ratio (TTR) určená vzťahom: $TTR = \frac{T}{N}$, kde $T$ je počet slovných druhov a $N$ celkový počet slov v skúmanom texte (Templin 1957). Nevýhodou tejto metriky je znižovanie pomeru v závislosti od zvyšovania skúmanej vzory (Arnaud 1992). Niektorí autori (Geeraerts et al. 1994; Jarvis 2002) uvádzajú, že lexikálna variácia a diverzita sú podobné vlastnosti, preto boli v prezentovanom experimente habilitačnej práce niektoré metriky uvedené v oboch

kategóriách, a vypočítané pomocou rôznych nástrojov. Ďalšie podobné metriky sú modifikáciou pôvodnej metriky TTR, ako napr. Root TTR (Guiraud 1960), Corrected TTR (Carroll 1964), Bilogarithmic TTR (LogTTR) (Herdan 1964), Uber Index (Dugast 1979) a normalizované TTR (zTTR) (Cvrček a Chlumská 2015).

- **Lexikálna rôznorodosť** (Lexical diversity) [lex_div][1]: je v podstate rozsah a rôznorodosť slovnej zásoby, ktorú v texte používa autor, pričom zohľadňuje kvalitu písania, znalosť slovnej zásoby, všeobecné charakteristiky a socioekonomický status (McCarthy a Jarvis 2007). Autori (Jarvis 2002; Geeraerts et al. 1994; Lu 2012) vnímajú lexikálnu diverzitu ako analógiu s lexikálnou variáciou alebo bohatosťou. V prezentovanom experimente habilitačnej práce boli vybrané metriky, ktoré sú primárne zamerané na lexikálnu diverzitu, t. j. rôznorodosť, aj napriek tomu, že väčšina zo skúmaných metrík je inšpirovaná metrikou TTR. Metrika Measure of textual lexical diversity (MTLD) rozdeľuje text do segmentov a pre každý sa počíta TTR skóre, kde dĺžka textu je premenná, ktorá závisí na hodnote TTR, ktorá hovorí o tom, ako sa rozširujú segmenty. Každý segment sa končí v momente, keď hodnota TTR dosiahne hodnotu 0,72 (McCarthy 2005). Ako ďalšie boli použité metriky Hypergeometric distribution diversity (HD-D) (McCarthy a Jarvis 2007), Herdans' lexical diversity measure (Herdan 1964), Dugasts' lexical diversity measure (Dugast 1979) a Maass' lexical diversity measure (McCarthy a Jarvis 2007).
- **Lexikálna sofistikovanosť** (Lexical sophistication) [lex_sop][1]: nazývaná aj lexikálna zriedkavosť (rareness) meria podiel relatívne neobvyklých alebo abstraktnejších slov v textoch. Linnarud (1986) a Hyltenstam (1988) použili na jej výpočet vzťah: $LS1 = \frac{N_{slex}}{N_{lex}}$, kde $N_{slex}$ je počet sofistikovaných lexikálnych slov a $N_{lex}$ je celkový počet lexikálnych slov v texte. Obaja autori metriku zamerali na študentov angličtiny ako druhého jazyka, Linnarud (1986) definoval sofistikované lexikálne slová ako anglické slová, ktoré sa naučia študenti od 9. ročníka a vyššie v švédskom školskom vzdelávacom systéme. Laufer (1994) vytvoril model Lexical Frequency Profile, ktorý sa zameriava na podiel slovných druhov v texte v kombinácií so zoznamom prvých 1000 najfrekventovanejších slov, druhých 1000 najfrekventovanejších slov a zoznamu univerzitných slov (Xue a Nation 1984). Model ponúka aj metriku lexikálnej sofistikovanosti, pre ktorú platí vzťah: $LS2 = \frac{T_s}{T}$, kde $T_s$ je počet sofistikovaných slovných druhov a $T$ je celkový počet slovných druhov v texte (Wolfe-Quintero et al. 1998). Ďalším z prístupov k lexikálnej sofistikovanosti bola metrika slovesnej

sofistikovanosti (verb sophistication), ktorá sa vypočíta nasledovne: $VS1 = \frac{T_{sverb}}{N_{verb}}$, kde $T_{sverb}$ je počet sofistikovaných slovesných druhov a $N_{verb}$ je celkový počet slovies v texte (Harley a King 1989). Alternatívnym prístupom je upravená slovesná sofistikovanosť (corrected verb sophistication): $CVS1 = \frac{T_{sverb}}{\sqrt{2*N_{verb}}}$, ktorú navrhol Wolfe-Quintero et al. (1998), aby zredukoval efekt veľkosti vzorky. Chaudron a Parker (1990) zvolili analogický prístup k úprave, avšak jej umocnením: $VS2 = \frac{T^2_{sverb}}{N_{verb}}$.

- **Expertné metriky** (Expert metrics) [expert][1]: metrika LIX sa zaraďuje medzi metriky čitateľnosti (Björnsson 1968), ktorá vznikla prioritne pre švédske texty, avšak úspešne bola aplikovaná nezávisle od skúmaného jazyka. Na jej výpočet sa používajú štandardné charakteristiky textu: $LIX = \frac{words}{periods} + \frac{long\ words*100}{words}$, kde *words* je počet slov, *periods* reprezentuje počet interpunkčných znamienok definovaných ako bodka, dvojbodka alebo prvé veľké písmeno a *long words* je počet dlhých slov (viac ako 6 znakov). Výsledné skóre reprezentuje podľa tabuľky (Anderson 1983) stupeň vzdelávania, pričom hodnota väčšia ako 55 indikuje vysoko odborné texty vhodné pre študentov a absolventov vysokých škôl. Anderson (1983) prišiel s optimalizáciou metriky LIX a nazval ju RIX (Rate Index). Definoval ju ako: $RIX = \frac{long\ words}{sentences}$, pričom skóre vyššie než 7,2 reprezentuje rovnako ako pri LIX vysokú zložitosť textu. Anderson (1983) preukázal, že LIX a RIX medzi sebou korelujú takmer dokonale ($r$ = 0,99). O'Hayre (1966) navrhol metriku LINSEAR Write, ktorá je založená na výpočte slabík. Výpočet skóre metriky je nasledovný: $LW = \frac{\frac{easy\ words}{sentences} + 2*hard\ words}{sentences}$, kde *hard words* je počet slov, ktorí obsahujú viac ako dve slabiky a naopak *easy words* sú slová, ktoré pozostávajú z dvoch a menej slabík. Výsledné skóre reprezentuje stupeň školy, pre ktorú je text určený, hodnota 13-16 reprezentuje študenta vysokej školy, zatiaľ čo 17+ reprezentuje absolventa vysokej školy. Všetky tieto metriky sa používajú aj v skúmaní ekonomických textov a preto boli zaradené do spoločnej kategórie. Gunning Fog Index, ktorý sa tiež často používa ako relevantná metrika pre ekonomické texty, bol zaradený do metrík čitateľnosti, na overenie rozdielu medzi týmito metrikami.
- **Slabiky** (Syllable) [syl][1]: ide o podobné metriky, ako v prípade skupiny charakteristík textu, avšak v tomto prípade metriky zachytávajú iba vlastnosti súvisiace s počtom slabík. Nevýhodou týchto metrík je nefunkčnosť, resp. nevyužiteľnosť pre niektoré jazyky.

- **Podiel slovných druhov** (Part of speech ratio) [pos_ratio][1]: tagovanie alebo morfologická anotácia je priradenie lemy a tagu (morfologickej značky) každému tokenu nachádzajúcemu sa v texte. Každý tag pozostáva zo súboru písmen latinskej abecedy, číslic a symbolov. V prípade skúmania čitateľnosti textu stačí identifikovať slovný druh tokenu. Nakoľko je tagovanie časovo náročný proces, v prezentovanom experimente habilitačnej práce bol použitý automatický nástroj Stanza (Qi et al. 2020), ktorý extrahuje podiely pre: číslovky, medzery, podstatné mená, adpozície, determinanty, vlastné podstatné mená, prídavné mená, slovesá, súradnicové spojky, interpunkčné znamienka, príslovky, pomocné výrazy, častice, zámená, podraďovacie spojky, citoslovcia, symboly a iné značky. Nástroj bol vybraný na základe dosiahnutých výsledkov v experimente na anglických textoch, kde dosahoval úspešnosť určenia slovného druhu viac ako 99% (Qi et al. 2018).
- **Ostatné charakteristiky** (Other characteristics) [other]: existuje veľké množstvo metrík čitateľnosti a zložitosti textu, pričom každá používa odlišnú škálu. Z toho dôvodu nebolo možné niektoré metriky zaradiť do vyššie uvedených. Vznikla samostatná kategória obsahujúca rôzne metriky ako napríklad veľkosť súboru v kB (Size of file in kB), ktorá sa pri ekonomických textoch ukazuje byť indikátorom zložitosti textu; Reading time (Demberg a Keller 2008), čo je v experimente podstatná metrika reprezentujúca dĺžku čítania textu v sekundách. Cvrček et al. (2020) vyvinuli nástroj QuitaUp, ktorý je určený na kvantitatívnu stylometrickú analýzu textu. Do nástroja implementovali viaceré už spomínané metriky, ale aj rôzne ďalšie, ktoré boli zaradené do tejto kategórie: h-point, frekvencia hapaxov, entropia, vzdialenosť slovesa, aktivita, deskriptivita, priemerná dĺžka tokenov, tematická koncentrácia, sekundárna tematická koncentrácia (Cvrček et al. 2020). Medzi ostatné metriky boli zaradené aj dve vlastné navrhnuté metriky EAWL a EAWL_unique.

Navrhnutá bola metrika zložitosti textov EAWL, primárne pre aplikovanie na ekonomické texty. Metrika vychádza zo zoznamu slov Economic Academic Word List (EAWL), pozostávajúci z 887 slov, ktoré sa najčastejšie nachádzajú v ekonomických textoch (O'Flynn 2019). EAWL zoznam je podobný Academic Word List (AWL) zoznamu, avšak je vhodnejší pre ekonomické texty, pretože je novší a vznikol ako nadstavba New General Service List (NGSL) zoznamu, ktorý vznikol v roku 2013. Rozdielom medzi nimi je, že EAWL obsahuje menej slovných foriem ako AWL. EAWL obsahuje iba skloňované tvary alebo varianty hláskovania slov, a nie celé skupiny slov, čo znamená, že hoci má viac hesiel ako AWL (887 v porovnaní s 570), celkovo má menej slovných tvarov (1763 v porovnaní s 3112). Skóre

navrhnutej metriky EAWL sa vypočíta ako: $EAWL = \frac{eawl\ words}{words}$, kde $eawl\ words$ je počet všetkých slov textu, ktoré sa nachádzajú v slovníku EAWL a $words$ je celkový počet slov v texte. Ako alternatíva bola navrhnutá optimalizovaná metrika, kde sa skúmali iba jedinečné slová: $EAWL\_unique = \frac{eawl\ words_{unique}}{words_{unique}}$, kde $eawl\ words_{unique}$ je počet unikátnych slov textu, ktoré sa nachádzajú v slovníku EAWL a $words_{unique}$ je celkový počet jedinečných slov v texte.

# 2 VÝSLEDKY VÝSKUMU ANALÝZY SPRÁVANIA SA STAKEHOLDEROV NA PORTÁLI KOMERČNEJ BANKY

Prvým čiastkovým cieľom je **skúmanie správania stakeholderov na webovom portáli bankovej inštitúcie pomocou rôznych prístupov**. Za týmto účelom sa realizovala séria experimentov zameraných na analýzu správania sa stakeholderov na webovom portáli.

Zdrojom dát v experimentoch (Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c) boli logovacie súbory webových portálov dvoch bankových inštitúcii (logovací súbor prvej bankovej inštitúcie je z obdobia rokov 2009 – 2012, logovací súbor druhej bankovej inštitúcie je z obdobia rokov 2013 – 2018). Oba webové portály mali podobnú štruktúru a pri ich predspracovaní sa postupovalo na základe podobnej metodiky. Podrobný popis logovacieho súboru webového portálu bankovej inštitúcie sa nachádza v článku (Munk, Pilkova, Benko et al. 2021b) (Príloha C). V článku je popísaná aj fáza transformácie dát, v ktorej boli vytvorené nezávislé premenné (prediktory). Premenná *week* bola vytvorená na základe štandardu ISO 8601 a reprezentuje týždne roka. Podobne boli vytvorené premenné *quartal, year* a *year quartal*. Skúmanou premennou bola závislá premenná *category*, ktorá bola vytvorená zlúčením navštívených webových častí do súvisiacich skupín – širších kategórii obsahu webu. Prezentované výskumy sú zamerané na webové časti súvisiace s Pilier 3.

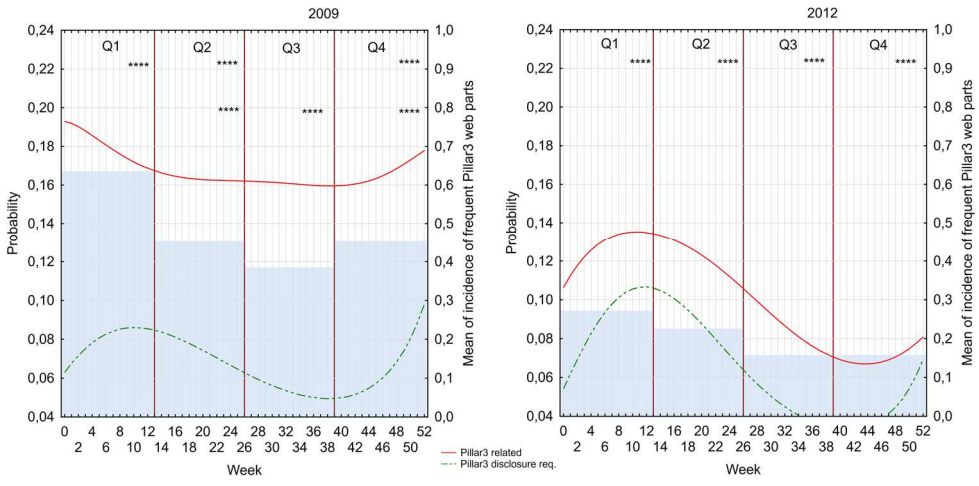
Chyby v predspracovaní údajov z logovacieho súboru môžu zásadne ovplyvniť výsledky analýzy dát a dosiahnuté závery. Z toho dôvodu je nutné napriek dodržiavaniu štandardných postupov prípravy dát vyhodnotiť získané znalosti. V experimente (Svec, Benko et al. 2020) (Príloha G) sa skúmal vplyv prípravy dát na získavanie nových znalostí vo fáze predikcie. Počas fázy vyhodnotenia výsledkov boli objavené prístupy automatizovaných nástrojov na webový portál, ktoré významne ovplyvnili identifikované znalosti, vo forme odhadu pravdepodobností prístupov na webové kategórie. Boli identifikované rozdiely v empirických a teoretických početnostiach prístupov na web a logitoch (Svec, Benko et al. 2020), kde počas 4. hodiny ráno, bola predikcia modelu pre väčšinu kategórií výrazne nadhodnotená, resp. podhodnotená. Na základe evalvácie dát bolo možné identifikovať chybu spôsobenú automatizovaným procesom, ktorý sa pravdepodobne staral o zálohovanie, kontrolu obsahu stránky na prítomnosť vírusov a podobne. Prítomnosť daného procesu významne ovplyvnila získané znalosti, či už na základe typu závislosti alebo vytváraní závislostí na miestach, kde žiadna nebola. Po odstránení týchto prístupov a zopakovaní analýzy dát, boli získané relevantné znalosti, ktoré potvrdila aj opätovná evalvácia dát. Výskum (Svec, Benko et al. 2020) slúžil ako praktická

ukážka ovplyvňovania získaných znalostí v prípade nedôkladného predspracovania dát logovacieho súboru. Podstatnú úlohu v tomto prípade zohrávala aj evalvácia dát, ktorá na všetkých troch skúmaných úrovniach poukazovala na chybu navrhnutého modelu pre konkrétnu kategóriu a čas.

Primárne smerovanie výskumu zameraného na dáta o používaní webu bolo na zhodnotenie správania sa a záujmov stakeholderov na webovom portáli komerčnej banky, ktorých akcie nie sú verejne obchodované. Kľúčovými stakeholdermi sú depozitní klienti, ktorí sa zaujímali o povinné zverejňované informácie v čase turbulentných zmien (Munk, Pilkova, Benko et al. 2021c) (Príloha A) ako aj po nich (Pilková, Munk, Benko et al. 2021a) (Príloha B). Experiment (Munk, Pilkova, Benko et al. 2021c) (Príloha A) bol zameraný na dátovú analýzu počas krízy a po nej, s cieľom identifikovať kľúčové typy informácií, ktoré zaujímajú stakeholderov, ako aj na možnosti optimalizácie politiky zverejňovania daných informácií. V experimente sa pracovalo s logovacím súborom popísaným v článku (Munk, Pilkova, Benko et al. 2021b) a na základe metodiky publikovanej v článku (Munk, Pilkova, Benko et al. 2021a) (Príloha D). V rámci metodiky boli podrobne popísané jednotlivé fázy spracovania údajov získaných z logovacích súborov webového portálu bankovej inštitúcie. Fáza modelovania dát bola zameraná na odhad pravdepodobností prístupov stakeholderov na webové kategórie súvisiace s Pilier 3. Skúmané bolo správanie sa návštevníkov počas obdobia viacerých rokov (2009-2012). Za roky krízy boli zvolené roky 2009 a 2010, pričom roky 2011 a 2012 boli považované za roky, kedy kríza začínala odznievať. Podstatný rozdiel oproti predchádzajúcim výskumom využívajúcim metodiku s multinomiálnym logitovým modelom bol v tom, že v predchádzajúcom prípade sa skúmali hodiny v rámci dňa a ďalšie umelé premenné rozlišujúce skúmané roky. To poskytovalo pohľad na správanie sa stakeholderov počas dní, avšak pre vytvorenie väčšieho obrazu o správaní sa návštevníkov webového portálu bolo potrebné zamerať sa aj na iné časové premenné. V prípade experimentu (Munk, Pilkova, Benko et al. 2021c) boli modelované pravdepodobnosti prístupu na webové kategórie na základe týždňov v rámci roka a umelá premenná identifikujúca obdobie finančnej krízy. Výskum (Pilková, Munk, Benko et al. 2021a) (Príloha B) je pokračovaním predchádzajúceho výskumu, pričom sa skúmalo akým spôsobom sa zmenilo správanie stakeholderov na webovom portáli inej rovnako dôležitej bankovej inštitúcie na Slovensku počas rokov 2016-2018. Predpokladal sa pokračujúci klesajúci trend záujmu o informácie súvisiace s Pilier 3. Napriek tomu, že logovacie súbory pochádzali z dvoch rôznych webových portálov, pomocou transformácie dát bola navrhnutá taxonómia webového portálu, ktorá tvorí

prienik pre webový obsah súvisiaci s informáciami Pilier 3. Vďaka tomu bolo možné sledovať časový trend a sezónnosť v správaní sa stakeholderov vo vzťahu k informáciám Pilier 3.

Cieľom výskumu (Pilková, Munk, Blažeková, Benko 2021b) (Príloha E) bolo zhodnotiť záujem stakeholderov o dve skupiny zverejňovaných informácií: *Pillar3 disclosure requirements* a *Pillar 3 related* počas obdobia rokov 2009-2012. Rovnako bolo cieľom zhodnotiť robustnosť overením výsledkov pomocou dvoch prístupov založených na rôznych časových premenných: týždeň a kvartál počas rokov 2009-2012. Prvý prístup bol založený na metodike popísanej v článku (Munk, Pilkova, Benko et al. 2021a). Druhý prístup sa zaoberal získavaním vzorov správania sa návštevníkov webu počas kvartálov. Výsledky boli spracované pomocou asociačnej analýzy s cieľom extrahovať frekventované položkové množiny s minimálnou podporou 1%. Predpokladom boli podobné výsledky, pričom týždenná analýza poskytuje podrobnejší prehľad správania sa návštevníkov webu v porovnaní s analýzou kvartálov. Grafy (Obrázok 3) vizualizujú pravdepodobnosti prístupov na skúmané webové kategórie súvisiace s trhovou disciplínou počas rokov 2009-2012. Rok 2009 (Obrázok 3) bol označený ako rok finančnej krízy, pričom najväčší záujem o webové kategórie bol na prelome rokov. Počas roka záujem o tieto kategórie klesá a koncom roka opäť začína stúpať. Hviezdičky nachádzajúce sa v grafoch označujú homogénne skupiny výskytu frekventovaných položkových množín. V roku 2009 ich najviac bolo identifikovaných v prvom kvartáli a najmenej v treťom. V roku 2012 (Obrázok 3) dochádza k výraznému prepadu záujmu o dané kategórie v priebehu celého roka. Napriek tomu je najväčší záujem hlavne v prvom kvartáli na začiatku roka.




**Obrázok 3** – Vizualizácia pravdepodobností kategórií súvisiacich s trhovou disciplínou počas rokov 2009 a 2012 (Pilková et al. 2021b)

Podrobná analýza, pravdepodobností prístupov stakeholderov počas týždňov na webový portál s povinne zverejňovanými informáciami komerčnej banky, ukázala, že výsledky korešpondujú s výsledkami analýzy kvartálov (Munk et al. 2017b). Najväčší záujem stakeholderov o informácie súvisiace s Pilier 3 bol počas prvého kvartálu, hlavne v období okolo 10. týždňa. 10. týždeň patril k obdobiu s najväčším záujmom o dané webové kategórie. Na základe výsledkov bolo možné konštatovať, že frekvencia povinného štvrťročného zverejňovania výsledkov nie je pre trhovú disciplínu nutná. Predpoklad, že aj po rokoch krízy (2012-2015) bude záujem o informácie Pilier 3 naďalej klesať sa potvrdil. Rovnako sa potvrdilo, že správanie sa stakeholderov vo vzťahu k zverejňovaným informáciám Pilier 3 už viac nepreukazuje žiadny trend alebo sezónnosť. Zvýšený záujem o dané informácie bol iba v čase krízy a jej následným odznievaním (2009-2011). Nasledujúce roky (2012-2018) boli charakteristické nízkym záujmom o tieto informácie, čo potvrdili dáta dvoch webových portálov bankových inštitúcií. Predchádzajúce revízie Pilier 3 regulátormi nepriniesli významný vplyv na zvyšovanie záujmu o tieto informácie.

Výskum v kapitole (Blažeková, Benko et al. 2021) (Príloha F) bol zameraný na skúmanie času stráveného na skúmaných webových stránkach v kontexte obsahu webových kategórií súvisiacich so zverejňovanými informáciami Pilier 3. Výsledky ukázali, že nadpriemerný čas bol stakeholdermi strávený na webovej stránke výročných správ (*annual reports*). Najviac času strávili návštevníci na stránkach poskytujúcich všeobecné informácie o banke a súvisiace informácie Pilier 3. Druhou najmenej navštevovanou kategóriou bola *Pillar3 Q-terly Information*, ktorá súvisí s povinne zverejňovanými informáciami. Podrobnejšia analýza tejto kategórie ukázala, že obsahuje niektoré webové časti, ktoré majú vysoký čas strávený návštevníkmi webu. To môže indikovať buď dôležitý obsah, ktorý návštevníci webu hľadajú alebo naopak, príliš veľa informácií na stránkach. Dosiahnuté výsledky potvrdzujú predchádzajúce výsledky (Munk et al. 2017b; Benko et al. 2020), ktoré ukázali, že návštevníci webu sa nezaujímajú o samotné informácie súvisiace s informáciami Pilier 3, ale skôr spolu s výročnými správami alebo informáciami o banke.

Na základe dosiahnutých výsledkov prezentovaných experimentov boli identifikované zaujímavé oblasti, ktoré môžu zvýšiť záujem stakeholderov o informácie Pilier 3 zhrnuté do nasledujúcich odporúčaní (Pilková, Munk, Blažeková, Benko 2021b):

- Zlepšiť štandardizáciu, teda harmonizáciu zverejňovania informácií vnútroštátnymi orgánmi požiadaviek a požiadaviek na zverejňovanie informácií na úrovni EÚ (Pilier 3 a národné požiadavky).
- Zvýšiť porovnateľnosť zverejňovaných informácií vytvorením jednej spoločnej šablóny (vizuálne predpísaných tabuliek), ktorú by v ideálnom prípade vytvorili regulačné orgány, s cieľom zaviesť jednotnosť.
- Znížiť frekvenciu zverejňovania informácií Pilier 3 vzhľadom na nízky záujem zainteresovaných strán o štvrťročné zverejňovanie informácií.
- Odlíšiť ročné zverejňovanie (formou zvýšenia objemu informácií) v porovnaní so štvrťročným, v prípade, ak sa štvrťročné zverejňovanie informácií použije na zníženie objemu zverejňovaných informácií.
- Zahrnúť informačné oblasti (povinne alebo dobrovoľne), o ktoré sa zainteresované strany zaujímajú (obchodné správanie inštitúcie, stratégia, reputácia, štruktúra, vlastníctvo, poslanie, hodnoty) a ovplyvňujú rizikovú pozíciu inštitúcie.
- Zabezpečiť dodržiavanie povinných požadovaných informácií zo strany regulačných orgánov - najmä obmedziť vynechávanie požadovaných informácií inštitúciami bez uvedenia dôvodu.
- Uložiť pravidlá pre umiestnenie zverejnených dokumentov, ktoré by mali byť na identifikovateľnom mieste časti webovej stránky.
- Povinnosť používať anglický jazyk ako spoločný jazyk pre zverejňovanie informácií.

# 3 VÝSLEDKY VÝSKUMU ANALÝZY DÁT Z OBSAHU WEBU

Druhým čiastkovým cieľom je **analýza obsahu webu vo forme dokumentov pomocou rôznych metód spracovania prirodzeného jazyka**. Analýza dokumentov pochádzajúcich z webu má podstatný vplyv na ďalší náš výskum, ktorý sa zameral na jazykovú zložitosť a čitateľnosť textu. Predpokladom je, že zložitosť textu, konkrétne jeho obsah, ktorý je zverejňovaný na stránkach komerčných bánk v súvislosti s Pilier 3 ovplyvňuje správanie sa stakeholderov. Na základe doterajších výsledkov našich štúdií a stanovených odporúčaní o používaní anglického jazyka ako univerzálneho jazyka pre zverejňovanie informácií, sme sa rozhodli skúmať iba anglické texty obsahujúce informácie Pilier 3. Výsledky tohto experimentu sú prezentované v predkladanej habilitačnej práci. Nakoľko pracujeme s dátami z roku 2018, tak nie ku všetkým slovenským dokumentom existovali oficiálne preklady do angličtiny. Z toho dôvodu bolo nutné niektoré dokumenty strojovo preložiť do angličtiny. Čo vyvolalo otázku kvality strojového prekladu a sekundárny výskum zameraný na hodnotenie kvality strojového prekladu. Výsledky prezentovaných výskumov vznikali v rámci projektov zameraných na evalváciu strojového prekladu dokumentov rôzneho štýlu extrahovaných z rôznych webových portálov.

Hodnotením kvality strojového prekladu sa zaoberali výskumy Benko et al. (2022) (Príloha J) a Benko et al. (2024a) (Príloha K). Benko et al. (2022) prepojili manuálnu evalváciu strojového prekladu s automatickými metrikami, kde pomocou analýzy chýb sa hľadali asociácie medzi kategóriami chýb a automatickými metrikami evalvácie strojového prekladu založenými na lexikálnej podobnosti. Výsledky štúdií indikujú, že nie všetky automatické metriky založené na n-gramoch (lexikálnej podobnosti) alebo vzdialenosti editovania by mali byť implementované do modelu hodnotenia kvality MT strojových prekladov z angličtiny do flektívnej slovenčiny. Pri určovaní kvality strojového prekladu vzhľadom na syntakticko-sémantickú korelatívnosť (plynulosť a adekvátnosť) stačí brať do úvahy metriky BLEU-4, NIST a CharacTER, pričom výsledky by mohli byť aplikovateľné aj pre iné flektívne jazyky. Následne bol prostredníctvom analýzy rezíduí porovnaný štatistický strojový preklad s neurónovým strojovým prekladom. Skúmalo sa, či zmena paradigmy vplýva na kvalitu strojového prekladu (Benko et al. 2024a). Porovnané boli dva prístupy k strojovému prekladu (SMT a NMT) použitím dvoch odlišných systémov (Google translátora a MT nástroja Európskej komisie pre preklad), pričom na evalváciu MT boli použité automatické metriky chybovosti. Predmetom skúmania boli publicistické dokumenty extrahované z webu pre

jazykový pár angličtina-slovenčina a nemčina-slovenčina. Výsledky analýz preukázali, že neurónové MT dosahovali štatisticky významne vyššiu kvalitu ako štatistické MT bez ohľadu na to, ktorý nástroj na preklad bol použitý. Neurónové MT generované nástrojom Google Translátorom (GT) dosahovali štatisticky významne najnižšiu chybovosť. Na druhej strane štatistické MT generované nástrojom mt@ec dosahovali štatisticky významne najvyššiu chybovosť. Predpoklad o vyššej kvalite neurónového MT v porovnaní so štatistickým MT sa potvrdil bez ohľadu na jazykový pár ako aj nástroj MT.

Výskum v článku Benko et al. (2024b) (Príloha L) bol zameraný na ďalšiu úlohu spracovania prirodzeného jazyka, konkrétne na porovnanie automatických nástrojov pre morfologickú anotáciu slovenského jazyka (POS taggers). Cieľom výskumu bolo navrhnúť metodiku na porovnávanie taggerov pre flektívne jazyky a nízko-zdrojové jazyky, a zároveň nájsť najefektívnejší automatický nástroj pre morfologickú anotáciu, tzn. ktorý dosahuje najlepší výkon na základe presnosti. Výber efektívneho (v zmysle presnosti) nástroja má veľký vplyv na metriky zložitosti textu, ktoré sú definované na základe tokenizácie ako napríklad pomer slovných druhov v texte (podiel podstatných miest, prídavných mien, slovies atď.). Článok sa zameriava na porovnanie najznámejších nástrojov TreeTagger (Schmid et al. 2007), RNNTagger (Schmid 2019), MorphoDita (online verzia a desktopová aplikácia) (Straka a Straková 2014), UDPipe2 (Straka a Straková 2017) a Stanza (Qi et al. 2020). Na tento účel bol zvolený pomerne jednoduchý a krátky ručne anotovaný subkorpus Slovenského korpusu závislostí (SDC) (Gajdošová a Šimková 2016). Nevýhodou použitého datasetu sú hlavne krátke vety (Benko a Benková 2022). Texty boli rozdelené na umelecké a náučné texty. Výsledky výskumu (Benko et al. 2024b) ukázali využiteľnosť POS taggerov v prípade málo zdrojového slovenského jazyka. Štyri zo šiestich skúmaných nástrojov dosiahli vysoký výkon vzhľadom na presnosť určenia, pričom RNNTagger sa ukázal najpresnejší pre oba typy textov. Prínos dosiahnutých výsledkov z hľadiska cieľa prezentovanej habilitačnej práce spočíva vo výbere a následnom aplikovaní pri metrikách zložitosti textu, ktoré sú založené na identifikácií slovných druhov v textoch.

Kvalita prekladu zohráva významnú úlohu v porozumení obsahu textu a extrahovaní dôležitých informácií, ktoré sú poskytované čitateľom. Práve na kvalitu prekladu sa zamerali výskumy v článkoch Munkova, Munk, Benko et al. (2021b) (Príloha H) a Benko et al. (2023) (Príloha I). Munkova, Munk, Benko et al. (2021b) skúmali vplyv kvality strojového prekladu na jazykovú zložitosť na úrovni slova a vetnej štruktúry. Cieľom štúdie bolo nájsť a overiť nový prístup k hodnoteniu kvality strojového prekladu na základe čitateľnosti a zložitosti textu.

Navrhovaná metodika bola založená na evalvácii frekventovaných tagsetoch strojového a post-editovaného strojového prekladu ako aj na základe frekventovaných POS tagsetoch a sumarizačných pravidiel. Prínos navrhovanej metodiky spočíva v identifikácií systematických a nie náhodných chýb. Munkova, Munk, Benko et al. (2021b) taktiež preukázali, že pre technické texty, MT systémy produkujú preklad s prijateľnou úrovňou kvality. Bol navrhnutý originálny a doteraz nepoužívaný unikátny prístup využívajúci miery zložitosti textu. Na dosiahnuté výsledky nadviazal výskum v článku Benko et al. (2023), v ktorom sa autori snažili prepojiť lexikálnu rôznorodosť vyjadrenú metrikami zložitosti textu s typmi (kategóriami) chýb strojového prekladu. Výsledky štúdie ukázali, že chyby vznikajúce v neurónovom MT súvisia s lingvistickými vlastnosťami založenými na početnostiach. Zaujímavým zistením štúdie je, že nie všetky metriky lexikálnej rôznorodosti súvisia s frekvenciou každého typu chýb. Štatisticky významnú závislosť s frekvenciou chýb v oblasti syntakticko-sémantickej korelatívnosti, súvetnej syntaxi a lexikálnej sémantike, vykázali iba metriky RTTR a CTTR. Limitáciou výskumu bolo obmedzenie sa iba na publicistické texty, ktoré sú pomerne ľahko čitateľné. Ukázalo sa, že čitateľnosť textu nezávisí od frekvencie chýb, čo umožnilo použiť strojový preklad v skúmaní ekonomických textov, konkrétne v oblasti bankovníctva.

Dosiahnuté výsledky výskumu analýzy dát z obsahu webu indikujú, že v prípade absentujúcich jazykových lokalizácií dokumentov, je možné vychádzať z ich strojových prekladov.

# 4 VPLYV JAZYKOVEJ ZLOŽITOSTI NA SPRÁVANIE SA STAKEHOLDEROV NA WEBOVÝCH STRÁNKACH KOMERČNÝCH BÁNK

Primárny cieľ prezentovanej habilitačnej práce vychádza z obsahu a používania webu. Cieľom práce je návrh metodiky zameranej na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk a skúmanie ich vplyvu na správanie sa stakeholderov. Súčasťou tohto procesu je vytvorenie ukazovateľov preferencií používateľov na webovom portáli bankovej inštitúcie a overenie vzťahu preferencií používateľov k zložitosti zverejňovaných textov. Experiment vychádza z dát bankovej inštitúcie získaných za rok 2018 a z dokumentov získaných z webového portálu (Benko et al. 2024c) (Príloha M). Motiváciou prezentovaného výskumu bolo nadviazať na dosiahnuté výsledky predchádzajúcich výskumov (Pilková et al. 2021a; Munk et al. 2021c; 2015; 2017b; Benko et al. 2020; Pilková et al. 2021b; Blažeková et al. 2021) a podrobne preskúmať obsah povinne zverejňovaných informácií v kontexte zložitosti a čitateľnosti textu.

Zložitosť a čitateľnosť odborných textov je vyhodnocovaná pomocou rôznych automatických mier navrhnutých viacerými autormi. Gunning (2003) predstavuje viac ako sto metrík zložitosti textu, avšak iba niekoľko z nich sa používa. Väčšina z nich sa používa na zisťovanie základných charakteristík textu ako dĺžka vety, počet slovných druhov a pod. (Sadeek Quaderi a Varathan 2024; Awan et al. 2021). Texty, ktoré sa analyzujú sú primárne určené pre učenie sa anglického jazyka ako druhého jazyka, čo smeruje k skúmaniu skôr edukačných ako odborných textov (Maqsood et al. 2022). Ehara (2022) sa zameral na skúmanie čitateľnosti úvodných textov informatiky. Vo svojom výskume porovnával BERT klasifikáciu s konvenčnými metrikami čitateľnosti ako sú Flesch-Kincaid Grade Level (Kincaid et al. 1975), ARI (Senter a Smith 1967), Coleman-Liau Index (Coleman a Liau 1975), Flesch Reading Ease (Flesch 2016), Gunning Fog Index (Gunning 2003), LIX (Björnsson 1968), SMOG Index (McLaughlin 1969), RIX (Anderson 1983) a Dale-Chall Index (Chall a Dale 1995). Ehara (2022) analyzoval prioritne texty extrahované z GitHub-u (návody k softvéru) a abstrakty vedeckých článkov zverejnených v rámci ACL Anthology a PubMed. Navrhol metriku čitateľnosti založenú na slovníku, ktorú porovnal s konvenčnými metrikami. Výsledky výskumu ukázali vyššiu koreláciu než konvenčné metriky, konkrétne, že vedecké texty sú nečitateľné pre stredne pokročilých študentov a naopak návody k softvéru sú študentov väčšinou čitateľné. Podobný výskum Ehara (2021) realizoval s ekonomickými (spravodajskými) textami, avšak zameral

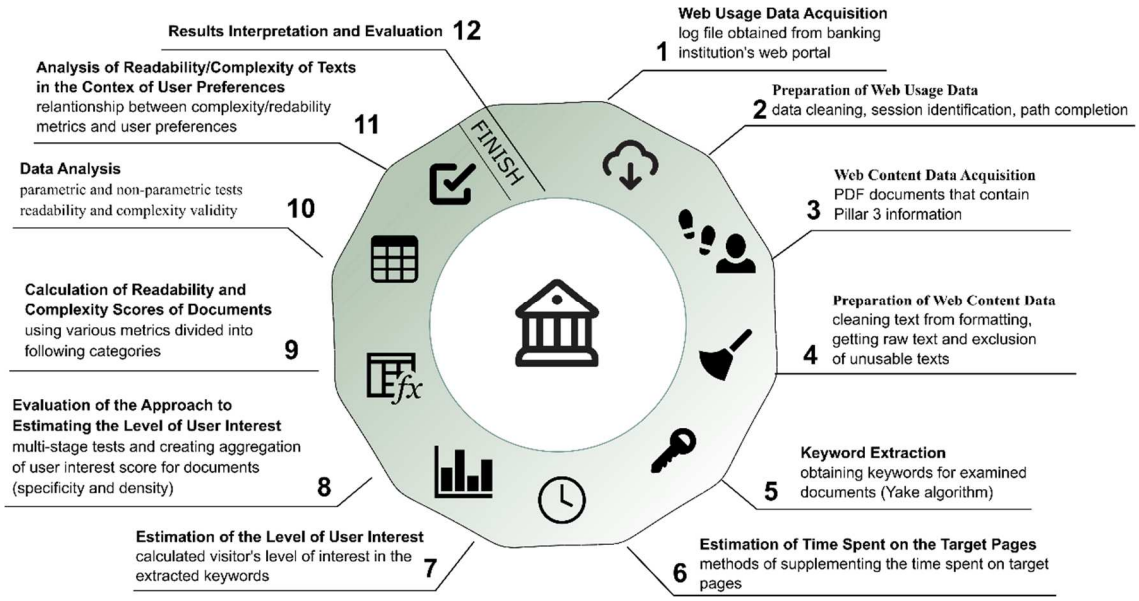
sa iba na porovnanie prístupu na báze BERT a slovníku. Výsledky ukázali, že väčšina textov bola čitateľná pre stredne pokročilých študentov, pričom 2,4% z nich neboli zrozumiteľné pre študentov.

Guay et al. (2015) sa zamerali na skúmanie zložitosti finančných výkazov (*financial statement*), v prípade čitateľnosti použili metriku ReadIndex, ktorá pozostávala zo skóre metrík: Flesch-Kincaid Grade Level, LIX, RIX, Gunning Fog Index, ARI a SMOG. Výsledky analýzy ukázali, že medzi všetkými šiestimi metrikami čitateľnosti je vysoká miera korelácie ako aj s ReadIndex-om. Prínos ich výskumu spočíva v tom, že napriek tomu, že zložité finančné výkazy negatívne ovplyvňujú informačné prostredie, tak niektoré firmy sa pokúšajú tieto vplyvy zmierniť dobrovoľným zverejňovaním ďalších informácií. Moreno a Casasola (2016) sa zamerali na analýzu čitateľnosti výročných správ v španielčine prostredníctvom upravenej metriky Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, LIX a RIX. Zistili, že výročné správy v španielčine vykazujú známky ťažšej čitateľnosti, čím potvrdili zistenia štúdií, ktoré boli zamerané na anglické výročné správy. Toerien a du Toit (2024) sa zamerali na zverejňované správy ohľadom rizika v Južnej Afrike. Analyzovali sa výročné správy (*annual reports*) zozbierané za roky 2005-2021, počas obdobia, kedy boli v krajine zavádzané viaceré štandardy týkajúce sa zverejňovania informácií vychádzajúce z praxe v EÚ. Skúmali texty pomocou metrík čitateľnosti Flesch-Kincaid a Gunning Fog Index ako aj metrík, ktoré sa zameriavajú na počet slov, dĺžku vety a podobne. Výsledky naznačujú (Toerien a du Toit 2024), že zverejňované správy majú nízku čitateľnosť. Zavedenie štandardov, ktoré majú zvyšovať čitateľnosť a znižovať zložitosť zverejňovaných správ, sa ukázali ako neefektívne, pričom je potrebné prehodnotiť formu informácií, aby zaujali širšiu verejnosť. Za limitáciu výskumu považujú nielen problémy s určením zložitých slov, vetnej štruktúry a jej dĺžky, ale aj nedostatok kontextu. Toerien a du Toit (2024) odporúčajú rozšíriť metriky o ďalšie, ktoré by mohli lepšie vysvetľovať zložitosť zverejňovaných informácií.

Na základe preštudovanej literatúry bolo smerovanie prezentovaného výskumu zamerané na skúmanie zložitosti textu súvisiaceho so zverejňovanými informáciami Pilier 3 na stránkach komerčných bánk a jeho vplyvu na správanie sa stakeholderov.

## 4.1 METODIKA VÝSKUMU

Metodika výskumu (Obrázok 4) bola inšpirovaná viacerými výskumami (Munk et al. 2021c; 2021a; Munkova et al. 2021a; Pilková et al. 2021a; Yao et al. 2017) a podrobnejšie je popísaná v nasledujúcich podkapitolách a v článku (Benko et al. 2024c) (Príloha M).




**Obrázok 4** – Metodika výskumu zameraného na jazykovú zložitosť a čitateľnosť textu (Benko et al. 2024c)

### 4.1.1 ZÍSKAVANIE A PRÍPRAVA DÁT O POUŽÍVANÍ WEBU

Výskum vychádza z dvoch dátových zdrojov. Prvým zdrojom je logovací súbor získaný z webového portálu bankovej inštitúcie a druhým zdrojom sú dokumenty extrahované z webového portálu na základe logovacieho súboru. Logovací súbor obsahoval prístupy na webový portál počas celého roku 2018 a prešiel fázou prípravy dát, ktorá pozostávala z čistenia dát, identifikácie používateľov/sedení a dopĺňania ciest (Munk et al. 2021a). Počas fázy predspracovania dát boli vytvorené ďalšie potrebné premenné ako *Category* a *Subcategory*, ktoré slúžili na prepojenie webových častí portálu, na ktoré pristupovali návštevníci. Taxonómia webového portálu, na základe ktorej boli webové časti rozdelené, sa nachádza v Tabuľke 1. Prioritou boli informácie súvisiace s informáciami Pilier 3, preto bol logovací súbor zredukovaný na sedenia, ktoré obsahovali aspoň jeden prístup na jednu zo skúmaných kategórií. Takto upravený logovací súbor pozostával z 265 216 záznamov.

**Tabuľka 1** – Taxonómia webového portálu komerčnej banky

| Kategória (Category) | Podkategória (Subcategory) |
|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ |
| /Pilar3 disclosure requirements/ | /information_about_bank/ |
| /Pillar3 related/ | /annual_reports/ |
| /Pillar3 related/ | /financial_reports/ |
| /Pillar3 related/ | /covered_bonds/ |
| /Pillar3 related/ | /information_for_investors_except_shareholders/ |
| /Pillar3 related/ | /information_for_shareholders/ |

### 4.1.2 ZÍSKAVANIE DÁT Z OBSAHU WEBU

Pre skúmanie vplyvu zložitosti textov na návštevnosť webového portálu, bolo nutné extrahovať texty, ku ktorým stakeholderi pristupovali. Skúmaný webový portál bankovej inštitúcie zverejňuje všetky informácie týkajúce sa Pilier 3 vo formáte PDF dokumentov. Z logovacieho súboru boli extrahované priame odkazy na dané dokumenty a pomocou webového crawlera bola získaná väčšina dokumentov. Nakoľko išlo o logovací súbor z roku 2018, nie všetky dokumenty boli prístupné, avšak podarilo sa extrahovať viac ako 90 % navštívených dokumentov. Viac ako polovica dokumentov bola v anglickom jazyku spolu s ich oficiálnym prekladom do slovenčiny. Dokumenty, ktoré obsahovali iba obrázky, tabuľky alebo grafy a duplikované jazykové verzie boli odstránené z kolekcie dokumentov (celkovo bolo odstránených 178 dokumentov). Niektoré slovenské dokumenty nemali sprístupnenú anglickú verziu, preto bol ich anglický preklad vytvorený pomocou MT systému Google Translátora (preložených bolo 97 dokumentov). Na analýzu textu bolo použitých 226 dokumentov rozdelených do jednotlivých kategórií podľa taxonómie (Tabuľka 1). Z týchto dokumentov bol pomocou PDF OCR nástroja extrahovaný čistý text, ktorý sa následne použil na analýzu zložitosti textu a hľadanie záujmu používateľov.

### 4.1.3 EXTRAKCIA KĽÚČOVÝCH SLOV

Pre potreby analýzy správania sa návštevníkov webu bolo potrebné z jednotlivých dokumentov extrahovať kľúčové slová. Bol použitý algoritmus Yake (Campos et al. 2020), ktorý dosahuje lepšie výsledky než štandardné techniky na identifikáciu kľúčových slov, ako napríklad TF-IDF, TextRank, KP-Miner alebo Rake (Campos et al. 2020). TF-IDF je častokrát používaná štatistická miera, ktorá určuje význam kľúčového slova vzhľadom na význam v jednom dokumente naprieč všetkými dokumentami celého korpusu. Experimenty preukázali, že v prípade odborných textov, nástroje ako Yake, KEA alebo KP-Miner dosahujú lepšie

výsledky než TF-IDF (Sarwar et al. 2021; Sarwar a Noor 2021). Nakoľko je prezentovaný experiment habilitačnej práce zameraný na dokumenty, ktoré sú odborného charakteru z oblasti bankovníctva, bol na extrakciu kľúčových slov zvolený nástroj Yake, ktorý bol implementovaný v jazyku Python. Algoritmus pozostával z nasledujúcich krokov (Campos et al. 2020):

- Predspracovanie textu a identifikácia kandidátnych pojmov – predbežné spracovanie dokumentu do strojovo čitateľného formátu s cieľom identifikovať potenciálne kandidátne pojmy, vďaka čomu sa zlepšuje účinnosť algoritmu,
- Extrakcia charakteristík – vstupom je zoznam jednotlivých pojmov reprezentovaných súborom štatistických znakov,
- Odhad skóre slov – spája vlastnosti do jedného skóre, ktoré odráža dôležitosť pojmov,
- Generovanie n-gramov a odhad skóre kandidátnych kľúčových slov – vygeneruje kandidátne kľúčové slová (prostredníctvom n-gramovej konštrukčnej metodológie) a priradí im skóre na základe ich dôležitosti,
- Deduplikácia a ranking – porovnáva podobné kľúčové slová pomocou miery podobnosti vzdialenosti deduplikácie. Zoznam konečných kľúčových slov je následne zoradený podľa ich skóre relevantnosti.

Pre každý skúmaný dokument bolo extrahovaných 100 kľúčových slov. Na základe dosiahnutého skóre boli vybrané kľúčové slová, ktoré spĺňali zvolenú hranicu pre každú kategóriu dokumentov. Hranica bola zvolená na základe priemerného skóre získaného pre dokumenty z danej kategórie. Kľúčové slová, ktoré mali vyššie skóre ako priemer, boli z ďalšej analýzy vylúčené (nižšie skóre znamená zaujímavejšie kľúčové slovo). V tabuľke 2 sa nachádza početnosť extrahovaných kľúčových slov pre jednotlivé kategórie ako aj s počtom skúmaných dokumentov. Táto tabuľka obsahuje súčet všetkých kľúčových slov za dokumenty vrátane duplicitných kľúčových slov, ktoré boli rovnaké pre niektoré dokumenty z tej istej kategórie.

**Tabuľka 2** – Počet dokumentov a kľúčových slov extrahovaných pre skúmané kategórie a podkategórie webového portálu

| Kategória | Podkategória | Počet dokumentov | Počet kľúčových slov |
|---|---|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ | 54 | 1367 |
| /Pilar3 disclosure requirements/ | /information_about_bank/ | 45 | 1554 |
| /Pillar3 related/ | /annual_reports/ | 17 | 1034 |
| /Pillar3 related/ | /financial_reports/ | 18 | 1134 |
| /Pillar3 related/ | /covered_bonds/ | 85 | 5939 |
| /Pillar3 related/ | /information_for_investors_ except_shareholders/ | 2 | 120 |
| /Pillar3 related/ | /information_for_shareholders/ | 5 | 312 |

### 4.1.4 ODHAD ČASU STRÁVENÉHO NA CIEĽOVÝCH STRÁNKACH

Dôležitým parametrom pri skúmaní čitateľnosti textu je odhad času tzv. cieľovej (obsahovej) stránky. Vo fáze identifikácie sedení pomocou metódy Reference Length (Kapusta et al. 2012b; Munk et al. 2015) sa stránky webového portálu v sedení rozdelia na navigačné a obsahové (cieľové) stránky. Cieľovými stránkami v prezentovanom experimente sú skúmané dokumenty, avšak pre ne nebolo možné určiť presný čas, ktorý na nich návštevníci strávili, a preto boli navrhnuté rôzne spôsoby odhadu tohto času. Nakoľko časové okno bolo vo fáze prípravy dát stanovené na 3 600 sekúnd (60 minút), jeden zo spôsobov vychádzal práve z daného časového okna. Ďalším spôsobom bolo použiť metriku čitateľnosti textu prostredníctvom času čítania (*reading_time*) jednotlivých dokumentov, doplnením priemerného času čítania dokumentu danej podkategórie alebo doplnením konkrétneho času čítania pre každý dokument. Navrhli sa štyri spôsoby doplnenia času stráveného na obsahových stránkach a vytvorili nové premenné v dátovom súbore:

- štandardný prístup bez doplnenia času (*length*),
- doplnenie 3 601 sekúnd pre všetky obsahové stránky v sedeniach (*length3601*),
- doplnenie priemerného času čítania pre dokumenty zo skúmaných podkategórií (*lengthRT_cat*),
- doplnenie konkrétneho času čítania pre jednotlivé dokumenty (*lengthRT_doc*).

Vzhľadom k tomu, že dĺžka času stráveného na stránke vstupovala do celého postupu odhadu úrovne záujmu používateľov, bol celý postup vykonaný pre všetky štyri prístupy doplnenia času obsahových stránok.

### 4.1.5 ODHAD ÚROVNE ZÁUJMU POUŽÍVATEĽOV

Úroveň záujmu návštevníka o extrahované kľúčové slová bola vypočítaná z času stráveného na stránke (Yao et al. 2017):

$$time_u(c_j, k_i) \begin{cases} \frac{length_u(c_j)}{m}, & \text{ak } k_i \text{ je v } c_j \\ 0, & \text{ak } k_i \text{ nie je v } c_j \end{cases},$$

kde $length_u(c_j)$ označuje dĺžku času, ktorú návštevník $u$ strávi na webovej časti kategórie $c_j$, ktorá obsahuje extrahované kľúčové slová $\{K_1, K_2, \dots, K_i\}$, pre celkový počet kľúčových slov $m$ vo všetkých kategóriách.

Celkový čas $sum_u$, ktorý návštevník stránky strávi nad určitým kľúčovým slovom $K_i$ v $KT_u$, je vypočítaný nasledovne:

$$sum_u(c_j, k_i) = \begin{cases} \sum_{i=j}^{f} time_u(c_j, k_i), & \text{ak } c_j \text{ je v } KT_u \\ 0, & \text{ak } c_j \text{ nie je v } KT_u \end{cases}.$$

Cieľom daného postupu je predpovedať prístup k určitému kľúčovému slovu, ktorý je založený na informáciách o navigácií návštevníkov. Ak má používateľ slovo, ktoré je pre neho zaujímavé či podstatné, opakovane navštevuje niektoré stránky, na ktorých s vysokou pravdepodobnosťou strávi viac času ako na iných stránkach. Na predpovedanie prístupu návštevníka webu k určitému kľúčovému slovu bol použitý model UISM (User interest structure model) (Yao et al. 2017), ktorý kombinuje dáta o obsahu, štruktúre a používaní webu a prepája všetky domény webu. UISM model je definovaný ako:

- súbor stavov: $Q = \{q_1, q_2, \dots, q_n\}$, počiatočný stav sa začína v $q_1$, každé $q_i$ reprezentuje webovú kategóriu,
- súbor kľúčových slov: $K = \{k_1, k_2, \dots, k_n\}$, $K$ obsahuje všetky kľúčové slová zo všetkých webových kategórií Q,
- pravdepodobnosť prechodu stavu $P_1(q \rightarrow q')$ medzi dvomi kategóriami, ktorá je definovaná nasledovne:

$$P_1(q \rightarrow q') = \frac{count\ (q \rightarrow q')}{count\ (q)},$$

kde $(q \rightarrow q')$ označuje cestu používateľa, ktorý najprv navštívil kategóriu $q$ a následne kategóriu $q'$. Yao et al. (2017) predstavili v rámci UISM dva rôzne prístupy k výpočtu $P_1$ na základe štruktúry webového portálu. Rozlišovali tzv. vertikálnu a horizontálnu štruktúru. V prípade vertikálnej štruktúry $count\ (q \rightarrow q')$ reprezentuje počet sedení, v ktorých

používateľ navštívil kategóriu $q'$ hneď po kategórií $q$. Teda obe kategórie sa musia nielen nachádzať v tom istom sedení, ale musia byť navštívené priamo po sebe. Na druhej strane, v prípade prístupu horizontálnej štruktúry, $count\ (q \rightarrow q')$ reprezentuje počet sedení, v ktorých sa nachádzajú obe kategórie $q$, $q'$, avšak nie sú obmedzené tým, že musia po sebe priamo nasledovať.

Vo všetkých stavoch $q$ existuje pravdepodobnosť rozdelenia $P_2\ (k_i|q)$ pre každé kľúčové slovo $k_i$ z $K$:

$$P_2\ (k_i|q) = \frac{\sum_{u=1}^{N} sum_u(q, k_i)}{\sum_{u=1}^{N}(\sum_{m=1}^{M} sum_u(q, k_i))},$$

ktorá sa označuje ako pravdepodobnosť symbolu pozorovania skrytého Markovovho modelu. Počíta sa na základe celkového času stráveného používateľom na danom kľúčovom slove na danej webovej kategórie. Predstavuje pravdepodobnosť záujmu o dané kľúčové slovo.

Pre sedenie $S^l$ ($l$ reprezentuje dĺžku sedenia) a záujem používateľa, je možné vyjadriť úroveň záujmu používateľa o dané kľúčové slovo $R\ (k|S_l)$, ktorého výpočet je nasledovný:

$$R\ (k|S^l) = P_1(q_{start} \rightarrow q_1) \times P_2(k|q_1) \times P_1(q_1 \rightarrow q_2) \times P_2(k|q_2) \times \ldots$$
$$\times P_1(q_{l-1} \rightarrow q_l) \times P_2(k|q_l)\ .$$

Ak je $R(k|S^l)$ väčšie alebo sa rovná $C$ – čo je hraničná hodnota spoľahlivosti, potom $R(k|S^l)$ je zaujímavé sedenie, pretože používatelia s rovnakým záujmom môžu pristupovať ku kategóriám sedenia. Hraničná hodnota spoľahlivosti bola nastavená podľa Yao et al. (2017) v intervale $10^{-3}$ až $10^{-7}$. V prípade klasického skrytého Markovovho modelu sa hraničná hodnota pohybuje v intervale $10^{-5}$ až $10^{-10}$.

### 4.1.6 ZHODNOTENIE PRÍSTUPU K ODHADU ÚROVNI ZÁUJMU POUŽÍVATEĽOV

Na základe vyššie spomenutého postupu bola vytvorená dátová matica, ktorá obsahovala horizontálnu/vertikálnu úroveň záujmu používateľov (UIH a UIV) pre jednotlivé extrahované kľúčové slová zo skúmaných dokumentov. Cieľom výskumu je prepojiť zložitosť textov so záujmom o navštevované webové kategórie. Z tohto dôvodu bolo nutné identifikovať vhodnú hraničnú hodnotu spoľahlivosti, prístup k úrovni záujmu používateľa (horizontálny alebo vertikálny) a prístup k odhadu času stráveného na cieľovej stránke.

Výsledky popisnej štatistiky ukázali, že dopĺňanie odhadu času stráveného na obsahových stránkach nemá vplyv na záujem používateľov. Z toho dôvodu bol zvolený prístup s doplnením

času na základe času čítania, tzn., ak používateľ hľadal daný dokument, tak pravdepodobne na dokumente strávil určitý čas, ktorý môže reprezentovať čas čítania daného dokumentu. Výsledky tiež ukázali, že vertikálny prístup k odhadu záujmu používateľov prináša najmenej užitočných informácií, pretože skúmané dokumenty sa v sedeniach nachádzajú v sekvencií za sebou zriedkavo. Z toho dôvodu bol do ďalšej analýzy vybraný iba horizontálny prístup, kedy sa dokumenty nachádzajú v sedeniach nezávisle od poradia návštevy.

Keďže navrhnutý postup odhadu záujmu používateľov autorov Yao et al. (2017) bol založený na kľúčových slovách a ich záujmu na základe hraničnej hodnoty spoľahlivosti, nebolo možné vyhodnotiť záujem používateľov o konkrétne dokumenty. Z toho dôvodu boli navrhnuté agregácie, ktoré mali reprezentovať záujem používateľov o jednotlivé dokumenty odhliadnuc od hraničnej hodnoty spoľahlivosti. Na základe výsledkov bol zvolený horizontálny prístup, ktorý vykazoval lepšiu rozlišovaciu schopnosť ako vertikálny prístup. Analogický postup odhadu úrovne záujmu používateľov je možné aplikovať nielen na dokumenty, ale aj na webové podkategórie alebo iné skúmané webové časti. Každé kľúčové slovo charakterizuje daný dokument iným spôsobom, a preto navrhnuté agregácie sa snažia zohľadniť váhu pre daný dokument. Agregácie boli navrhnuté na základe nasledovných vlastností s vytvorenými váhami:

- Záujem používateľa na základe špecificity kľúčových slov v dokumente: $UIH_{s_i} = \sum_{k=1}^{K_i} specificity_k * UIH_k$, kde $UIH$ je odhad úrovne záujmu používateľa o dané kľúčové slovo $k$ a $specificity_k = \ln \frac{n_i}{N}, i = 1, \ldots, N$, kde $N$ je celkový počet skúmaných dokumentov a $n$ je počet dokumentov, ktoré obsahujú dané kľúčové slovo $k$.
- Záujem používateľa na základe hustoty kľúčových slov v dokumente: $UIH_{d_i} = \sum_{k=1}^{K_i} density_k * UIH_k$, kde $UIH$ je odhad úrovne záujmu používateľa o dané kľúčové slovo $k$ a $density_k = \frac{n_i}{N}, i = 1, \ldots, N$, kde $N$ je celkový počet skúmaných dokumentov a $n$ je počet dokumentov, ktoré obsahujú dané kľúčové slovo $k$.

Napríklad kľúčové slovo „client“ sa nachádza v 83 dokumentoch z 226 skúmaných dokumentov. Prirodzeným logaritmom podielu týchto čísiel sa získa hodnota váhy -1,00169. Odhadnutá hodnota záujmu používateľov UIH pre toto kľúčové slovo je 0,002064. Vynásobením týchto dvoch hodnôt sa dosiahne hodnota -0,0020675, ktorá určuje špecificitu pre skúmané kľúčové slovo. Hodnota váhy v absolútnych číslach je bližšie k nule, ak je kľúčové

slovo všeobecnejšie. Analogicky sa postupuje aj v prípade hustoty, čo v tom prípade je hodnota 0,000758. Hodnota váhy je vyššia, čím kľúčové slovo je všeobecnejšie, tzn. nachádza sa vo viacerých dokumentoch. Na základe absolútnych čísiel špecifickosti je možné vidieť, že výsledky sú podobné, ale skóre sú dva opačné extrémy. Týmto spôsobom sa získali váhy pre jednotlivé kľúčové slová v dokumentoch, ktoré sa následne agregovali pre každý dokument.

Ako ďalšie ukazovatele preferencií používateľov sa vypočítali podpora (*support*), entropia (Shannon 1948) a počet sedení, v ktorých je dokument cieľová stránka. Vychádzalo sa z identifikovaných sedení pomocou metódy Reference Length (Kapusta et al. 2012b; Munk et al. 2015) a podpory daných dokumentov v identifikovaných sedeniach. Podpora vyjadruje záujem o navštevované dokumenty. Slúži ako referencia v prezentovanom experimente, v ktorom je snahou identifikovať ďalšie ukazovatele, ktoré by mali korelovať s validným kritériom. Jedným z takýchto ukazovateľov je entropia sedení (*entropy*) s dôrazom na zloženie jednotlivých sedení. Neusporiadanosť je charakterizovaná rôznorodosťou návštevností používateľa rôznych kategórií webového portálu počas sedenia. Pričom pri výpočte entropie sedenia sa vychádzalo z entropie definovanej (Shannon 1948): $entropy_s = -\sum_{x \in X} p(x) \log_n p(x)$, kde $n$ je počet stránok v sedení $s$ a $p(x)$ je pravdepodobnosť výskytu stránky $x$ v sedení. Ak sa entropia rovná 1, sedenia obsahovali webové stránky z rôznych kategórií. Ak sa entropia rovná 0, potom všetky stránky v sedení pochádzali z jednej kategórie, pričom používateľ hľadal cielene informáciu z danej kategórie. Entropia bola určená pre každé sedenie a pre každý dokument pričom výslednou entropiou bola priemerná hodnota pre každý dokument.

Druhým ukazovateľom bola premenná *target*, ktorá reprezentovala počet sedení, v ktorých bol dokument cieľovou (obsahovou) stránkou. Cieľová stránka je stránka, ktorej čas strávený používateľom na stránke je väčší než hraničný čas (Kapusta et al. 2012b), čo indikuje záujem používateľa o obsah a cieľ jeho hľadania. V prípade prezentovaného experimentu, cieľom návštevníkov bol obsah skúmaných dokumentov.

### 4.1.7 VÝPOČET SKÓRE ČITATEĽNOSTI A ZLOŽITOSTI DOKUMENTOV

V prezentovanom experimente habilitačnej práce sa aplikovalo niekoľko metrík zložitosti a čitateľnosti s motiváciou identifikovať tie, ktoré budú najlepšie charakterizovať zložitosť a čitateľnosť dokumentov z povinne zverejňovaných informácií súvisiacich s Pilier 3. Metriky boli rozdelené do viacerých kategórií pre lepšiu prehľadnosť. V prípade niektorých kategórii dochádza k prekrytiu (spôsobené prekrytím kategórií), preto boli niektoré metriky použité

duplicitne, avšak v takom prípade boli počítané pomocou iného nástroja. Skúmané metriky boli implementované pomocou jazyka Python alebo boli použité externé nástroje (Cvrček et al. 2020; Lu 2012; 2011; 2010; Lu a Ai 2015). Podrobnejší popis jednotlivých kategórií bol uvedený v kapitole 1.2.2, pre zvýšenie prehľadnosti metodiky sa na tomto mieste uvádza iba zoznam skupín metrík čitateľnosti a zložitosti textu: Charakteristiky textu [char][2], Čitateľnosť [read][2], Lexikálna variácia [lex_var][2], Lexikálna bohatosť [lex_rich][2], Lexikálna rôznorodosť [lex_div][2], Lexikálna sofistikovanosť [lex_sop][2], Expertné metriky [expert][2], Slabiky [syl][2], Podiel slovných druhov [pos_ratio][2], Ostatné charakteristiky [other][2].

Po realizovaní všetkých spomenutých krokoch bol výsledkom dátový súbor, ktorý obsahoval skúmané dokumenty a k nim odhad úrovne záujmu o dokumenty, časové charakteristiky, počet sedení, v ktorých je dokument cieľovou stránkou, entropiu sedení a charakteristiky na základe zložitosti a čitateľnosti textu.

## 4.2 VÝSLEDKY EXPERIMENTU

Analýza závislosti medzi úrovňou záujmu používateľov a čitateľnosťou/zložitosťou textu pozostávala z viacerých krokov (Benko et al. 2024c) (Príloha M). V prvom kroku bolo nutné vyhodnotiť, ktorá z agregácii úrovne záujmu používateľov najlepšie vystihuje skúmané dokumenty v kombinácií s časovými charakteristikami a ostatnými používateľsky zameranými charakteristikami (počet sedení, v ktorých je dokument cieľovou stránkou a entropiou sedení) a podporou.

*Bol stanovený predpoklad, že ukazovatele preferencií používateľov (počet sedení, v ktorých je dokument cieľová stránka, entropia sedení a úroveň záujmu používateľa zohľadňujúci špecificitu a hustotu kľúčového slova v dokumente) budú relevantné, v zmysle rozlišovacej schopnosti a miery vysvetlenia návštevnosti.*

V ďalšom kroku sa porovnávala agregácia s premennou podpory (support), ktorá vyjadruje záujem o navštívené dokumenty a slúži ako validné kritérium. Výsledky agregácií (Benko et al. 2024c) ukázali, že počet sedení, v ktorých je dokument cieľová stránka; entropia sedení a úroveň záujmu používateľa zohľadňujúci hustotu kľúčového slova v dokumente, sú relevantné. Nové ukazovatele navrhnuté v experimente, reprezentujúce preferenciu používateľa (počet sedení, v ktorých je dokument cieľová stránka a entropia sedení) dosiahli rovnakú rozlišovaciu schopnosť, pričom v oboch prípadoch bola dosiahnutá veľmi veľká

[2] vektor skúmaných premenných [*x*], kde *x* označuje príslušnú kategóriu čitateľnosti alebo zložitosti textu

štatisticky významná korelácia. Oba ukazovatele sa ukázali ako zaujímavé aj z hľadiska miery vysvetlenia návštevnosti. Počet sedení, v ktorých je dokument cieľová stránka vysvetľuje 96 % variability podpory a entropia sedení vysvetľuje 79 % variability podpory.

Zaujímavý výsledok priniesli neparametrické odhady v prípade počtu relácií, v ktorých je dokument cieľovou stránkou a v prípade entropie sedení, kde podkategória výročných správ (*annual reports*) dosiahla najvyšší priemer poradí (mean rank) v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a najnižší priemer poradí v prípade entropie sedení (na základe absolútnych výsledkov špecifickosti možno vidieť, že výsledky sú podobné, ale skóre je opačné voči počtu sedení). V prípade ostatných skúmaných ukazovateľov, nedosahovala podkategória výročných správ také vysoké hodnoty.

*Bol stanovený predpoklad, že najvýkonnejší (v zmysle rozlišovacej schopnosti a miery vysvetlenia návštevnosti) ukazovateľ preferencií používateľov z hľadiska času stráveného na stránke bude čas strávený na stránke zohľadňujúci čas čítania dokumentu.*

Globálna nulová hypotéza sa zamieta na hladine významnosti 0,001 (lengthRT_doc_mean: $F(6, 219) = 3{,}538$, $p < 0{,}001$; $H(6, N = 226) = 122{,}387$, $p < 0{,}001$), ktorá tvrdí, že neexistuje žiadny štatisticky významný rozdiel v preferenciách používateľa vyjadrený časom stráveným na stránke pri zohľadnení času čítania dokumentu medzi skúmanými podkategóriami obsahu. To znamená, že medzi skúmanými podkategóriami obsahu boli identifikované rozdiely v časových hodnotách. V prípade času stráveného na stránke pri zohľadnení času čítania dokumentu bola v prípade podkategórie *annual reports* dosiahnutá štatisticky významne najväčšia preferencia používateľov ($p < 0{,}05$). V prípade ďalších časových premenných sa globálna nulová hypotéza nezamieta ($p > 0{,}05$).

### 4.2.1 VALIDITA ZLOŽITOSTI A ČITATEĽNOSTI TEXTU

Nasledujúcim krokom bolo zameranie sa na validitu metrík čitateľnosti a zložitosti. Kvôli prehľadnosti boli skúmané metriky rozdelené do desiatich kategórií podľa spoločných vlastností. Vo výskumoch zameraných na ekonomické texty sa používajú rôzne metriky čitateľnosti, pričom časť autorov preferuje metriku Gunning Fog Index a druhá skupina autorov preferuje metriku LIX alebo RIX (Ebaid 2023; Guay et al. 2015; Moreno a Casasola 2016). Z toho dôvodu bola v experimente vytvorená kategória expertných metrík LIX, RIX a LinsearWrite. Metrika Gunning Fog Index bola zaradená medzi metriky čitateľnosti, s ktorými sa často kombinuje. Bola vypočítaná korelácia medzi vektormi jednotlivých skupín

metrík a skupiny expertných metrík, ktorých vektor premenných [expert] predstavuje validné kritérium. Výsledky poukazujú na fakt, že všetky kategórie sú štatisticky významné, avšak najvyššiu mieru závislosti s expertnými metrikami dosahujú metriky čitateľnosti (skupina obsahuje aj metriku Gunning Fog Index) a metriky základných charakteristík textu. Metriky čitateľnosti a expertné metriky dokážu spoločne odhaliť podobné znaky zložitosti a čitateľnosti finančných textov. Z toho dôvodu metrika Gunning Fog Index nebola zaradená medzi expertné metriky, čo nemalo vplyv na výsledky experimentu. Za validné kritérium tak mohli byť zvolené akékoľvek metriky čitateľnosti.

Posledná skupina metrík (ostatné) obsahovala metriky zložitosti textu rôznorodého charakteru, preto k nej nebolo možné pristupovať ako k vektoru konzistentných metrík. Porovnanie bolo realizované pomocou viacnásobnej analýzy v kombinácií premenná vs. vektor expertných metrík. Vo všetkých prípadoch bol viacnásobný korelačný koeficient medzi jednotlivými premennými a vektorom expertných metrík štatisticky významný na hladine významnosti 0,001, okrem poslednej (*other_Size_in_kB*), ktorá bola na hladine významnosti 0,01.

Výsledkom validity metrík čitateľnosti a zložitosti textu je, že všetky skupiny skúmaných metrík sú použiteľné. Navrhnutá metrika *other_eawl* založená na slovníku ekonomických slov dosiahla štatisticky významnú strednú mieru závislosti. Od nej odvodená metrika *other_eawl_unique*, ktorá brala do úvahy iba jedinečné slová dosiahla dokonca štatisticky významnú veľkú mieru závislosti voči vektoru expertných metrík. Ukázalo sa, že tieto metriky majú potenciál charakterizovať čitateľnosť ekonomických textov.

### 4.2.2 ANALÝZA ČITATEĽNOSTI/ZLOŽITOSTI TEXTOV V KONTEXTE PREFERENCIÍ POUŽÍVATEĽA

Hlavným cieľom výskumu bolo zistiť, aký je vzťah medzi metrikami zložitosti/čitateľnosti a preferenciami používateľa (počet sedení, v ktorých je dokument cieľová stránka, entropia sedení, čas strávený na stránke zohľadňujúci čas čítania dokumentu a úroveň záujmu používateľa na základe hustoty kľúčových slov v dokumente). Analýza závislostí bola vykonaná pre každú skupinu metrík zvlášť, v kombinácií s ukazovateľmi preferencií používateľa. Viacnásobná analýza sa aplikovala pre skupinu metrík a následne bola vypočítaná jednorozmerná analýza pre jednotlivé metriky danej skupiny v kombinácií s ukazovateľmi preferencií používateľa.

Podrobné výsledky analýzy pre všetky skupiny metrík sa nachádzajú v článku (Benko et al. 2024c) (Príloha M). Na základe dosiahnutých výsledkov boli identifikované skupiny metrík, ktoré dosiahli najvýznamnejší vzťah so skúmanými ukazovateľmi preferencií používateľa.

**Úroveň záujmu používateľa vs. čitateľnosť/zložitosť textu**

Metriky skupiny [pos_ratio] dosiahli najvyššiu mieru závislosti s úrovňou záujmu používateľa na základe hustoty kľúčových slov v dokumente (*Multiple R* = 0,848; *Multiple R2* = 0,720; *Adjusted R2* = 0,697), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Zaujímavé výsledky dosiahli v tejto skupine metriky popisujúce podiel vlastných mien a slovies. Zo skupiny ostatných metrík [other] sa ukázalo, že viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,01 pre metriku *eawl_unique* s úrovňou záujmu používateľa na základe hustoty kľúčových slov v dokumente. Výsledky ukázali, že úroveň záujmu používateľa na základe hustoty kľúčových slov v dokumente súvisí hlavne s podielom čísloviek (*numerals*), podielom vlastných mien (*proper noun*), podielom slovies (*verbs*) a podielom unikátnych ekonomických slov (*eawl_unique*). Dá sa predpokladať, že vyšší počet vlastných mien ($r$ = 0,4) a slovies ($r$ = 0,2) môže znamenať vyššiu úroveň záujmu používateľa. Naopak nižší počet čísloviek ($r$ = -0,2) v texte naznačuje vyššiu úroveň záujmu používateľa. Rovnako, navrhnutá metrika podielu unikátnych ekonomických slov dokáže zachytiť úroveň záujmu používateľa. Dôvodom môže byť fakt, že generované kľúčové slová sú prevažne vlastné mená, ktoré sa nachádzajú aj v zozname ekonomických slov. Z hľadiska interpretácie metrík zložitosti, je vhodné používať metriku *eawl_unique* z viacerých dôvodov:

- nevyžaduje si použitie nástroja tretej strany, pomocou ktorého je nutné vykonať časovo náročnú morfologickú anotáciu na identifikáciu počtu slovných druhov;
- zoznam ekonomických slov sa môže rozširovať, a tým neustále zlepšovať presnosť metriky;
- ekonomické slová sú zrozumiteľné pre odborníkov na finančníctvo, na druhej strane môžu byť menej zrozumiteľné pre bežných stakeholderov. Ukázalo sa, že vyšší podiel týchto slov značí vyšší záujem používateľov ($r$ = 0,2), čo môže naznačovať, že o dané dokumenty majú záujem hlavne odborníci na finančníctvo.

Napríklad v prípade podkategórie *information-for-shareholders-not-investors* dokument „slovakiavub_presentation_for_investors_062018“ mal nadpriemernú úroveň záujmu 0,03029

a *eawl_unique* hodnotu 0,08, t. j. 8% unikátnych ekonomických slov v dokumente. Na druhej strane dokument podkategórie *annual-reports* s názvom „vubannualreport14“ mal podpriemernú úroveň záujmu 0,00001 a *eawl_unique* hodnotu 0,02, čo znamená, že nižšie percento ekonomických slov môže znižovať záujem o daný dokument medzi expertami na danú oblasť.

**Čas strávený na stránke vs. čitateľnosť/zložitosť textu**

Metriky skupiny [char] a [lex_rich] dosiahli najvyššiu mieru závislosti s časom stráveným na stránke zohľadňujúcim čas čítania dokumentu ([char]: *Multiple R* = 0,566; *Multiple R2* = 0,321; *Adjusted R2* = 0,269; [lex_rich]: *Multiple R* = 0,558; *Multiple R2* = 0,311; *Adjusted R2* = 0,258), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že čas strávený na stránke zohľadňujúci čas čítania dokumentu značne súvisí s početnosťou slovných jednotiek textu, ako je počet znakov ($r = 0,3$), tokenov ($r = 0,2$) a jedinečných tokenov ($r = 0,2$). Potvrdili to aj výsledky skupiny lexikálnej bohatosti, kde obe metriky, zachytávajúce mieru rôznych slovných druhov v texte, sa ukázali ako zaujímavé ($r = -0,2$). V prípade metrík zo skupiny [other] je pozitívne, že metrika času čítania *other_RT* ($r = 0,3$) súvisí s časom stráveným na stránke zohľadňujúcim čas čítania dokumentu. Výsledky tiež ukázali, že vyšší čas strávený na stránke súvisí s väčším rozsahom (väčšou dĺžkou) textov, čo potvrdzujú aj metriky súvisiace s časom čítania a veľkosťou daného dokumentu ($r = 0,3$). Zaujímavým zistením bolo, že nižšia miera rôznych slovných druhov ($r = -0,2$) v textoch zvyšuje čas strávený na stránkach.

Napríklad v prípade podkategórie *annual-reports* mal dokument „ar_2017_en_final _web“ nadpriemernú úroveň priemerného času stráveného na stránkach 188 sekúnd a hodnoty metrík pre daný dokument dosahovali nadpriemerné skóre: počet tokenov = 106317, počet viet = 3378, čas čítania = 7484 sekúnd.

**Počet sedení, v ktorých je dokument cieľová stránka vs. čitateľnosť/zložitosť textu**

Metriky skupiny [char] dosiahli najvyššiu mieru závislosti s počtom cieľových stránok (*Multiple R* = 0,628; *Multiple R2* = 0,394; *Adjusted R2* = 0,348), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že počet sedení, v ktorých je dokument cieľová stránka, podobne ako čas strávený na stránke, súvisí hlavne s početnosťou slovných jednotiek textu, ako počet znakov ($r = 0,4$), tokenov ($r = 0,4$) a jedinečných tokenov ($r = 0,4$). Väčšia dĺžka textu reprezentovaná dĺžkou vety ($r = 0,3$) a vyšší počet znakov, tokenov a aj jedinečných tokenov v dokumente má vplyv

na väčší počet sedení, v ktorých je dokument cieľová stránka. Tieto výsledky potvrdzujú aj metriky zo skupiny [other], či už veľkosť súboru ($r = 0,4$) alebo čas čítania ($r = 0,4$). Podobne aj metrika h-point (v prípade väčších textov ($r = 0,3$)) a entropia textu ($r = 0,2$) potvrdzujú dosiahnuté výsledky: čím väčšia diverzita slovníka je v dokumente identifikovaná, tým je väčší počet sedení, v ktorých je dokument cieľová stránka.

Napríklad v prípade podkategórie *information-about-bank* mal dokument „pillar-iii_15_12_en“ nadpriemernú úroveň počtu sedení (35), v ktorých je dokument cieľová stránka. Hodnoty metrík pre daný dokument dosahovali nadpriemerné skóre: počet tokenov = 34367, počet viet = 3952, čas čítania = 2526 sekúnd, h-point = 54 a entropia textu = 9,59.

**Entropia sedení vs. čitateľnosť/zložitosť textu**

Metriky skupín [char] a [pos_ratio] dosiahli najvyššiu mieru závislosti s entropiou sedení ([char]: *Multiple R* = 0,646; *Multiple R2* = 0,418; *Adjusted R2* = 0,373; [pos_ratio]: *Multiple R* = 0,630; *Multiple R2* = 0,397; *Adjusted R2* = 0,348), pričom viacnásobné koeficienty korelácie sú štatisticky významné na hladine významnosti 0,001. Výsledky ukázali, že entropia sedení taktiež súvisí s početnosťou znakov ($r = -0,5$), tokenov ($r = -0,4$) a jedinečných tokenov ($r = -0,4$) v dokumente. Dokazujú to metriky zo skupiny [pos_ratio], kde vyšší podiel vlastných mien ($r = -0,4$) v dokumente znižuje entropiu sedení, t. j. návštevník hľadá v sedeniach súvisiace informácie. Výsledky tiež ukázali, že nižšia entropia sedení zodpovedá väčšej dĺžke textov. Dokazujú to metriky času čítania ($r = -0,5$) a veľkosti súboru ($r = -0,2$), ako aj väčšia entropia dokumentu ($r = -0,3$), ktorá reprezentuje mieru rôznych slovných druhov v dokumente.

Napríklad v prípade podkategórie *financial-reports* mal dokument „polrocna-financna-sprava-za-rok-2012“ nadpriemernú úroveň entropie sedení = 0,99990 a hodnoty metrík pre daný dokument dosahovali podpriemerné skóre: počet tokenov = 30354, počet viet = 396, čas čítania = 1188 sekúnd, h-point = 44 a entropia textu = 8,64.

# ZÁVER

Habilitačná práca bola zameraná na prepojenie zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3 zverejňovanými na stránkach komerčných bánk a správanie sa stakeholderov na webových stránkach. Realizované experimenty boli smerované na skúmanie správania sa stakeholderov na webovom portály bankovej inštitúcie. Ukázalo sa, že príprava dát má vplyv na kvalitu analýzy dát a získavanie znalostí z webu (Munk, Pilkova, Benko et al. 2021b; Svec, Benko et al. 2020), čo viedlo k vytvoreniu metodiky, ktorá bola základom pre všetky vykonané experimenty (Munk, Pilkova, Benko et al. 2021a). Výsledky viacerých experimentov potvrdili (Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c; Pilková, Munk, Blažeková, Benko 2021b), že zverejňovanie povinných informácií Pilier 3, nie je potrebné v priebehu celého roka. Preukázali to výsledky na báze rôznych časových premenných, či už týždňov alebo kvartálov. Kombináciou rôznych metód realizovaných v experimentoch bolo dosiahnuté zlepšenie výsledkov získaných z údajov z logovacieho súboru webového portálu, čo prispelo k lepšiemu pochopeniu správania sa stakeholderov v prípade informácií súvisiacich s Pilier 3. Analýza zameraná na skúmanie času stráveného na webových stránkach (Blažeková, Benko et al. 2021) ukázala, že záujem iba o povinne zverejňované informácie nie je až taký vysoký. Stakeholderov zaujíma širší kontext informácií o banke a výročných správach. Na základe dosiahnutých výsledkov boli stanovené odporúčania, ktoré môžu zvýšiť záujem stakeholderov o informácie Pilier 3.

Skúmanie obsahu webu vo forme dokumentov a práca s nimi má podstatný vplyv na ďalší výskum, ktorý sa zameral na jazykovú zložitosť a čitateľnosť textov. Na základe stanovených odporúčaní o používaní anglického jazyka ako jednotného jazyka pre zverejňovanie informácií, bolo predefinované, že v prezentovanom experimente habilitačnej práce sa budú skúmať iba anglické texty obsahujúce informácie Pilier 3. Z toho dôvodu bolo ďalšie smerovanie zamerané na skúmanie kvality strojového prekladu, keďže v súčasnosti veľké korporácie používajú na lokalizáciu svojich produktov pre daný región strojový a nie humánny preklad. Výsledky experimentu (Benko et al. 2022) indikujú, že nie všetky automatické metriky založené na lexikálnej podobnosti (n-gramoch alebo vzdialenosti editácie) by mali byť implementované do modelu určovania kvality MT, či už ekonomických alebo iných typov textov prekladaných z anglického jazyka do flektívnej slovenčiny. V ďalších experimentoch sa pracovalo prevažne s metrikami, ktoré boli vhodné pre jazykový pár slovenčina-angličtina. Ďalším prínosom pri analýze chybovosti strojových prekladov bolo prepojenie analýzy rezíduí na identifikáciu

konkrétnych segmentov alebo textov, v ktorých systémy strojového prekladu dosahujú vyššiu chybovosť (Benko et al. 2024a). Významným prínosom je schopnosť identifikovať segmenty a lingvisticky charakterizovať segmenty, v ktorých strojový preklad vykazoval chybovosť v zmysle adekvátnosti a plynulosti do slovenčiny. Výsledky výskumu (Benko et al. 2024b) ukázali, že používanie POS taggerov by mohlo byť v prípade slovenského jazyka prínosné. Štyri nástroje zo šiestich skúmaných taggerov dosiahli vysoký výkon, v zmysle presnosti, pri lingvistickom anotovaní do 15-pozičného tagsetu. Použitie RNNTagger (najefektívnejšieho nástroja) by malo byť preferované pre generovanie morfologických značiek pre slovenský jazyk. Dosiahnuté výsledky v oblasti spracovania textu poukázali na skutočnosť, že v prípade absentujúcich jazykových mutácií (lokalizácií), je možné vychádzať z výstupov systémov strojového prekladu a použiť ich strojové preklady. Zložitosť textov sa ukázala ako jeden z atribútov, ktorý môže tiež poukázať na chybovosť strojového prekladu (Benko et al. 2023). Experiment (Munkova, Munk, Benko et al. 2021b) bol zameraný na vplyv jazykovej zložitosti na úrovni slov a vetnej štruktúry, pričom za hlavný prínos je možné považovať navrhnutú metodiku, ktorá zohľadňuje miery jazykovej zložitosti, slovných druhov, frekventovaných tagsetov, asociačných pravidiel a ich sumarizácie.

Cieľom prezentovaného experimentu (Benko et al. 2024c) bolo navrhnúť metodiku zameranú na analýzu zložitosti a čitateľnosti textu súvisiaceho s informáciami Pilier 3, ktorý sa podarilo splniť. Hlavným prínosom habilitačnej práce do odboru je navrhnutie metodiky prepájajúcej zdroje dát o používaní, obsahu a štruktúre webu. Metodika viedla k návrhu ukazovateľov preferencií používateľov na webových portáloch komerčných bánk a umožnila skúmať, či zložitosť a čitateľnosť povinne zverejňovaných informácií má vplyv na vytvorenie preferencie používateľov. Počas realizovania experimentu boli navrhnuté dve metriky zložitosti textu *eawl* a *eawl_unique*, ktoré sú vhodné pre ekonomické texty. Za hlavný prínos experimentu je možné považovať ukazovatele preferencií používateľov z hľadiska návštevnosti: entropia sedení, počet sedení, v ktorých dokument je cieľová stránka a úrovne záujmu používateľa na základe hustoty kľúčových slov v dokumente, a z hľadiska času: čas strávený na stránke zohľadňujúci čas čítania dokumentu. Ukazovatele orientované na návštevnosť boli porovnané s podporou, ktorá vystupovala ako referencia. Preukázal sa súvis medzi ukazovateľmi preferencií používateľov a metrikami zložitosti alebo čitateľnosti. Skupina metrík základných charakteristík, ako sú rôzne početnosti tokenov, viet, znakov dosiahla najvyššiu mieru závislosti s preferenciou používateľov na základe entropie sedení; počtu sedení, v ktorých je dokument cieľová stránka a času stráveného na stránke zohľadňujúceho čas čítania dokumentu.

Najvýznamnejší prínos habilitačnej práce spočíva v prepojení oblasti spracovania prirodzeného jazyka a používateľských preferencií, ktoré reprezentujú doménu webu.

Dĺžka textu hrá podstatnú úlohu v jeho zložitosti a čitateľnosti. Rozsiahlejšie dokumenty obsahujú väčšie množstvo informácií, a preto sú podľa dosiahnutých výsledkov pre stakeholderov zaujímavejšie a preferujú ich pred krátkymi dokumentami. Z hľadiska ukazovateľa úrovne záujmu používateľa na základe hustoty kľúčových slov v dokumente sa preukázala ako zaujímavá navrhnutá metrika *eawl_unique*, ktorá bola navrhnutá za účelom charakterizovať zložitosť ekonomických textov.

Zaujímavý výsledok priniesli neparametrické odhady v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a v prípade entropie sedení, kde podkategória výročných správ dosiahla najvyšší priemer poradí v prípade počtu sedení, v ktorých je dokument cieľovou stránkou a najnižší priemer poradí v prípade entropie sedení. Potvrdzujú to aj výsledky výskumu správania stakeholderov na portáli skúmanej bankovej inštitúcie (Blažeková, Benko et al. 2021; Pilková, Munk, Blažeková, Benko 2021b; Pilková, Munk, Benko et al. 2021a; Munk, Pilkova, Benko et al. 2021c).

Z hľadiska čitateľnosti a zložitosti textu sa ukázalo, že všetky kategórie metrík čitateľnosti/zložitosti sú štatisticky významné, pričom najvyššiu mieru závislosti s expertnými metrikami dosahujú metriky čitateľnosti a metriky základných charakteristík textu. Metriky čitateľnosti a expertné metriky dokážu spoločne popísať podobné znaky zložitosti a čitateľnosti finančných textov. Viac ako polovica skúmaných dokumentov (151 z 226) dosiahla skóre Gunning Fog Index, ktoré reprezentuje úroveň absolventov vysokej školy (*index* > 17), čo naznačuje vysoko odborné dokumenty. Podobné výsledky boli dosiahnuté aj pre metriku LIX (*index* > 56), ktorá viac ako polovicu dokumentov (157 z 226) identifikovala ako odborné texty. Na druhej strane v kombinácií metrík čitateľnosti/zložitosti s ukazovateľmi preferencií používateľov, v prípade metriky Gunning Fog Index sú viacnásobné koeficienty korelácie štatisticky významné na hladine významnosti 0,01 len v spojení s entropiou sedení. V prípade ostatných ukazovateľov sú viacnásobné koeficienty korelácie štatisticky nevýznamné. V experimente bolo implementovaných 110 metrík zložitosti/čitateľnosti zaradených do 10 skupín. Výsledky analýzy identifikovali iba jednu metriku (*pos_ratio_PROPN*), pre ktorú boli koeficienty korelácie štatisticky významné v prípade všetkých ukazovateľov preferencií používateľov. Znamená to, že z hľadiska preferencií používateľov, podiel vlastných mien v dokumentoch zvyšuje záujem o tieto dokumenty zo strany stakeholderov.

Dosiahnuté výsledky potvrdzujú výsledky predchádzajúcich štúdií (Pilková et al. 2021a; Munk et al. 2021c). Prezentovaný výskum nadviazal na predchádzajúce zistenia veľmi nízkeho záujmu o informácie Pilier 3 zo strany investorov komerčných bánk pôsobiacich v strednej a východnej Európe. Z doterajších výsledkov štúdií možno vyvodiť, že skupina klientov, ktorí sa o tieto informácie zaujímajú v tomto type komerčných bánk, sú tí, ktorí majú v banke nepoistené vklady (právnické osoby, fyzické osoby s vkladmi nad 100 tis. EUR). Ako však ukazujú výsledky prezentovaného výskumu, títo klienti majú väčší záujem o menej náročné a čitateľnejšie texty, ako sú výročné správy. Keďže informácie Pilier 3 a ďalšie dokumenty Pilier sú zložitejšie a ťažšie čitateľné, majú o ne menší záujem. To vedie k dôležitému záveru pre regulátorov: zvýšenie záujmu o tieto informácie v tomto type bánk si vyžaduje nájsť spôsoby ich prezentácie, aby boli menej zložité a čitateľnejšie.

Limitácií hlavného výskumu je niekoľko. Prvou je počet extrahovaných dokumentov, ktorých bolo 226, z toho v niektorých skúmaných podkategóriách bolo možné extrahovať iba niekoľko dokumentov. Skúmané dáta o používaní webu a obsahu webu pochádzali z roku 2018, ktorý je považovaný za najstabilnejšie obdobie, keďže nešlo o turbulentné obdobie. Nevýhoda skúmania starších dát spočíva v nedostupnosti všetkých dát a v zmene štruktúry webového portálu. Získavanie obsahu webu je možné cez archívne záznamy webu, ktoré ukladajú webovú stopu z daného obdobia. Avšak ani to nemusí garantovať použiteľnosť extrahovaných dokumentov.

Prínos habilitačnej práce nie je iba do oblasti bankovníctva, ale aj do vzdelávania. V rámci predmetov magisterského štúdia Objavovanie znalostí a Hĺbkováanalýza dát vyučovanými autorom práce, študenti pracujú s rôznymi zdrojmi dát. Študenti sa učia ako funguje proces získavania znalostí práve pomocou domény webu, pretože táto oblasť ponúka najlepšie zdroje dát – štruktúrované aj neštruktúrované. Úlohou študentov je pochopiť procesy prípravy dát a ich následné spracovanie pre potreby analýzy dát. Pri riešení semestrálnych projektov postupujú na základe metodiky (Munk, Pilkova, Benko et al. 2021a) a skúmajú okrem dát o používaní webu, aj dáta o obsahu webu, kde v textoch získaných z webových portálov hľadajú znalosti a skúmajú zložitosť textu.

Ďalšie smerovanie výskumu sa zameria na porovnanie ukazovateľov preferencií používateľov a čitateľnosti dokumentov naprieč obdobím viacerých rokov s ohľadom na turbulentné obdobia (globálna finančná kríza, pandémia a pod.) a revíziu zverejňovaných informácií.

# ZOZNAM POUŽITEJ LITERATÚRY


ANDERSON, Jonathan, 1983. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*. **26**(6), 490–496.

ARNAUD, Pierre J.L., 1992. Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In: Pierre J.L. ARNAUD a H. BÉJOINT, ed. *Vocabulary and Applied Linguistics*. London, UK: Palgrave Macmillan.

AWAN, Malik Daler Ali, Nadeem Iqbal KAJLA, Amnah FIRDOUS, Mujtaba HUSNAIN a Malik Muhammad Saad MISSEN, 2021. Event classification from the Urdu language text on social media. *PeerJ Computer Science* [online]. **7**, e775. ISSN 2376-5992. doi:10.7717/peerj-cs.775

AYE, Theint Theint, 2011. Web log cleaning for mining of web usage patterns. *2011 3rd International Conference on Computer Research and Development* [online]. **2**, 490–494. doi:10.1109/ICCRD.2011.5764181

BENKO, Ľubomír a Lucia BENKOVÁ, 2022. Comparison of Novel Approach to Part-Of-Speech Tagging of Slovak Language. In: *DIVAI 2022 – The 14th international scientific conference on Distance Learning in Applied Informatics*. Štúrovo, Slovakia: Wolters Kluwer, s. 327–333.

BENKO, Ľubomír, Lucia BENKOVA, Dasa MUNKOVA, Michal MUNK a Danylo SHULZENKO, 2022. Error Classification Using Automatic Measures Based on n-grams and Edit Distance. In: *Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022* [online]. Springer, Cham, s. 345–356. doi:10.1007/978-3-031-20319-0_26

BENKO, Ľubomír, Petra BLAŽEKOVÁ, Michal MUNK a Anna PILKOVÁ, 2020. Time Spent on Web Page as an Indicator of Interest. In: *DIVAI 2020 The 13 th international scientific conference on Distance Learning in Applied Informatics*. s. 489–497. ISBN 9788075988416.

BENKO, Ľubomír, Dasa MUNKOVÁ a Michal MUNK, 2023. Relationship Between Linguistic Complexity and MT Errors in the Context of Inflectional Languages. In: *Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2023* [online]. Springer, Cham, s. 546–557. doi:10.1007/978-3-031-42430-4_45

BENKO, Ľubomír, Dasa MUNKOVA, Michal MUNK, Lucia BENKOVA a Petr HAJEK, 2024a. The use of residual analysis to improve the error rate accuracy of machine translation. *Scientific Reports (v recenznom konaní)*.

BENKO, Ľubomír, Dasa MUNKOVA, Mária PAPPOVÁ a Michal MUNK, 2024b. Comparison of various approaches to tagging for the inflectional Slovak language. *PeerJ Computer Science (v recenznom konaní)*.

BENKO, Ľubomir, Anna PILKOVA, Michal MUNK a Slavka ELEY, 2024c. Pillar 3: The impact of language complexity on the preferences of commercial bank website users. *Expert Systems with Applications (v recenznom konaní)*.

BERENDT, Bettina, Bamshad MOBASHER, Miki NAKAGAWA a Myra SPILIOPOULOU, 2003. The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. *Lecture Notes in Computer Science* [online]. **2703**, 159–179. doi:10.1007/978-3-540-39663-5_10

BESSENYEI, Gabor, 2017. Neural Machine Translation: The Rising Star. *Memsource* [online] [vid. 2022-10-04]. Dostupné z: https://www.memsource.com/blog/2017/09/19/neural-machine-translation-the-rising-star/?utm_source=mailchimp&utm_medium=email&utm_content=blog_article

BJÖRNSSON, Carl Hugo, 1968. *Lasbarhet*. Stockholm, Sweden: Bokforlaget Liber.

BLAŽEKOVÁ, Petra, Ľubomír BENKO, Anna PILKOVÁ a Michal MUNK, 2021. Is Pillar 3 a Good Tool for Stakeholders in CEE Commercial Banks? In: *Studies in Systems, Decision and Control* [online]. Springer, s. 421–440. doi:10.1007/978-3-030-76632-0_15

CAMPOS, Ricardo, Vítor MANGARAVITE, Arian PASQUALI, Alípio JORGE, Célia NUNES a Adam JATOWT, 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* [online]. **509**, 257–289. ISSN 00200255. doi:10.1016/j.ins.2019.09.013

CARROLL, John Bissell, 1964. *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

CERNA, Miloslava a Petra POULOVA, 2008. VISIT RATE OF INTERNET PORTALS AND UTILIZATION OF THEIR TOOLS AND SERVICES. *E & M EKONOMIE A MANAGEMENT*. **11**(4), 132–143. ISSN 1212-3609.

COLEMAN, Meri a Ta Lin LIAU, 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*. **60**, 283–284.

COLLINS, John William a Nancy P. O'BRIEN, 2003. *The Greenwood Dictionary of Education*. Greenwood Press.

COMMON CORE STATE STANDARDS INITIATIVE, 2023. *English Language Arts Standards* [online]. [vid. 2024-02-02]. Dostupné z: https://corestandards.org/wp-content/uploads/2023/09/ELA_Standards1.pdf

COOLEY, R, B MOBASHER, J SRIVASTAVA a OTHERS, 1999. Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*. **1**(1), 5–32.

CVRČEK, Václav, Radek ČECH a Miroslav KUBÁT, 2020. QuitaUp – nástroj pro kvantitativní stylometrickou analýzu. *Czech National Corpus and University of Ostrava* [online] [vid. 2023-07-21]. Dostupné z: https://korpus.cz/quitaup/

CVRČEK, Václav a Lucie CHLUMSKÁ, 2015. Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics* [online]. **39**(3), 309–325. ISSN 0304-3487. doi:10.1007/s11185-015-9151-8

DEMBERG, Vera a Frank KELLER, 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* [online]. **109**(2), 193–210. ISSN 00100277. doi:10.1016/j.cognition.2008.07.008

DUGAST, Daniel, 1979. *Vocabulaire et stylistique: Théâtre et dialogue*. Geneva, Switzerland: Slatkine-Champion.

EBAID, Ibrahim El-Sayed, 2023. IFRS adoption and the readability of corporate annual reports: evidence from an emerging market. *Future Business Journal* [online]. **9**(1), 80. ISSN 2314-7210. doi:10.1186/s43093-023-00244-x

EHARA, Yo, 2021. To What Extent Can English-as-a-Second Language Learners Read Economic News Texts? In: *Proceedings of the Third Workshop on Economics and Natural Language Processing* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 62–68. doi:10.18653/v1/2021.econlp-1.9

EHARA, Yo, 2022. Neural Language Model-based Readability Assessment of Computer Science Introductory Texts for English-as-a-Second Language Learners. In: *Proceedings of the*

*44th Annual Conference of the Cognitive Science Society*. Toronto, Canada: eScholarship, s. 1698–1704.

FAYYAD, Usama M., Gregory PIATETSKY-SHAPIRO a Padhraic SMYTH, 1996. From Data Mining to Knowledge Discovery in Databases [online]. **17**(3), 37–54. ISSN 0738-4602. doi:10.1609/AIMAG.V17I3.1230

FISHER, Douglas, Nancy FREY a Diane LAPP, 2012. *Text Complexity: Raising Rigor in Reading*. International Reading Association.

FLESCH, Rudolf, 2016. How to Write Plain English. *University of Cantenbury* [online] [vid. 2024-01-21]. Dostupné z: https://web.archive.org/web/20160712094308/http:/www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

GAJDOŠOVÁ, Katarína a Mária ŠIMKOVÁ, 2016. Slovak Dependency Treebank. *http://hdl.handle.net/11234/1-1822*.

GEERAERTS, Dirk, Stefan GRONDELAERS a Peter BAKEMA, 1994. *The Structure of Lexical Variation* [online]. DE GRUYTER MOUTON. ISBN 978-3-11-014387-4. doi:10.1515/9783110873061

GRAY, William Scott a Bernice Elizabeth LEARY, 1935. *What Makes a Book Readable: With Special Reference to Adults of Limited Reading Ability*. Chicago, IL: University of Chicago Press.

GUAY, Wayne R., Delphine SAMUELS a Daniel J. TAYLOR, 2015. Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure. *SSRN Electronic Journal* [online]. ISSN 1556-5068. doi:10.2139/ssrn.2564350

GUIRAUD, Pierre, 1960. *Probl`emes et m´ethodes de la statistique linguistique*. Dordrecht, Netherlands: Springer.

GUNNING, Thomas G., 2003. The Role of Readability in Today's Classrooms. *Topics in Language Disorders*. **23**(3), 175–189.

HARLEY, Brigit a Mary Lou KING, 1989. Verb Lexis in the Written Composition of Young L2 Learners. *Studies in Second Language Acquisition*. **11**(4), 415–439.

HARRIS, Theodore L. a Richard E. HODGES, 1995. *The Literacy Dictionary: The Vocabulary of Reading and Writing*. International Reading Association. ISBN 9780872071384.

HERDAN, Gustav, 1964. *Quantitative linguistics*. London, UK: Butterworths.

HYLTENSTAM, Kenneth, 1988. Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development* [online]. **9**(1–2), 67–84. ISSN 0143-4632. doi:10.1080/01434632.1988.9994320

CHALL, Jeanne Sternlicht a Edgar DALE, 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

CHAUDRON, Craig a Kate PARKER, 1990. Discourse Markedness and Structural Markedness: The Acquisition of English Noun Phrases. *Studies in Second Language Acquisition*. **12**, 43–64.

JARVIS, Scott, 2002. Short texts, best-tting curves and new measures of lexical diversity. *Language Testing*. **19**(1), 57–84.

JEAN, Sébastien, Kyunghyun CHO, Roland MEMISEVIC a Yoshua BENGIO, 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [online]. Beijing, China: Association for Computational Linguistics, s. 1–10. doi:10.3115/v1/P15-1001

KAPUSTA, Jozef, Michal MUNK a Martin DRLÍK, 2012a. Cut-off time calculation for user session identification by reference length. In: *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*.

KAPUSTA, Jozef, Michal MUNK a Martin DRLÍK, 2012b. Cut-off time calculation for user session identification by reference length. In: *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings* [online]. 2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012. ISBN 9781467317405. doi:10.1109/ICAICT.2012.6398500

KAPUSTA, Jozef, Michal MUNK, Peter SVEC a Anna PILKOVA, 2014. Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal. *Procedia Computer Science*. **29**, 1779–1790.

KAPUSTA, Jozef, Anna PILKOVA, Michal MUNK a Peter SVEC, 2013. Data pre-processing for web log mining: Case study of commercial bank website usage analysis. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. **61**(4), 973–979.

KINCAID, Peter J., Robert P. FISHBURNE JR., Richard L. ROGERS a Brad S. CHISSOM, 1975. *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel*.

KINTSCH, Walter, 1974. *The Representation of Meaning in Memory* [online]. Lawrence Erlbaum. ISBN 9781317744894. Dostupné z: doi:10.4324/9781315794563

KLEE, Thomas, 1992. Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*. **12**, 28–41.

KOEHN, Philipp, 2010. *Statistical Machine Translation*. Cambridge University Press. ISBN 0521874157, 9780521874151.

LAUFER, Batia, 1994. The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal* [online]. **25**(2), 21–33. ISSN 0033-6882. doi:10.1177/003368829402500202

LINNARUD, Moira, 1986. *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund, Sweden: CWK Gleerup.

LIU, Bing, 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* [online]. ISBN 978-3-642-19459-7. doi:10.1007/978-3-642-19460-3

LOSARWAR, Vijayashiri a Madhuri JOSHI, 2012. Data Preprocessing in Web Usage Mining. In: *International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore*. s. 1–5.

LOSHIN, David, 2013. Knowledge Discovery and Data Mining for Predictive Analytics. In: *Business Intelligence* [online]. Elsevier, s. 271–286. doi:10.1016/B978-0-12-385889-4.00017-X

LU, Xiaofei, 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*. **15**(4), 474–496.

LU, Xiaofei, 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. *TESOL Quaterly*. **45**(1), 36–62.

LU, Xiaofei, 2012. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal* [online]. **96**(2), 190–208. ISSN 00267902. doi:10.1111/j.1540-4781.2011.01232.x

LU, Xiaofei a Haiyang AI, 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*. **29**, 16–27.

LUONG, Minh-Thang, Ilya SUTSKEVER, Quoc V. LE, Oriol VINYALS a Wojciech ZAREMBA, 2015. Addressing the Rare Word Problem in Neural Machine Translation. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* [online]. Beijing, China: Association for Computational Linguistics, s. 11–19. doi:10.3115/v1/P15-1002

MALVERN, David, Brian RICHARDS, Ngoni CHIPERE a Pilar DURÁN, 2004. *Lexical Diversity and Language Development* [online]. London: Palgrave Macmillan UK. ISBN 978-1-4039-0232-0. doi:10.1057/9780230511804

MAQSOOD, Shazia, Abdul SHAHID, Muhammad TANVIR AFZAL, Muhammad ROMAN, Zahid KHAN, Zubair NAWAZ a Muhammad Haris AZIZ, 2022. Assessing English language sentences readability using machine learning models. *PeerJ Computer Science* [online]. **7**, e818. ISSN 2376-5992. doi:10.7717/peerj-cs.818

MCCARTHY, Philip M., 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Memphis, TN. PhD. Thesis. The University of Memphis.

MCCARTHY, Philip M. a Scott JARVIS, 2007. vocd: A theoretical and empirical evaluation. *Language Testing* [online]. **24**(4), 459–488. ISSN 0265-5322. doi:10.1177/0265532207080767

MCCLURE, Erica, 1991. A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*. **2**, 141–154.

MCLAUGHLIN, Harry G., 1969. SMOG Grading - a New Readability Formula. *Journal of Reading*. **12**(8), 639–646.

MILLER, Jon F, 1991. Quantifying productive language disorders. In: Jon F. MILLER, ed. *Research in child language disorders: A decade of progress*. Austin, TX: Pro-Ed, s. 211–220.

MING-SYAN CHEN, Ming-Syan, Jong Soo JONG SOO PARK a P.S. YU, 1998. Efficient data mining for path traversal patterns. *IEEE Transactions on Knowledge and Data Engineering* [online]. **10**(2), 209–221. ISSN 10414347. doi:10.1109/69.683753

MORENO, Alonso a Araceli CASASOLA, 2016. A Readability Evolution of Narratives in Annual Reports. *Journal of Business and Technical Communication* [online]. **30**(2), 202–235. ISSN 1050-6519. doi:10.1177/1050651915620233

MUNK, Michal a Lubomir BENKO, 2018. Using Entropy in Web Usage Data Preprocessing. *Entropy* [online]. **20**(1), 67. doi:10.3390/e20010067

MUNK, Michal, Ľubomír BENKO, Mikuláš GANGUR a Milan TURČÁNI, 2015. Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management* [online]. **18**(3), 144–159. doi:10.15240/tul/001/2015-3-013

MUNK, Michal, Martin DRLIK, Lubomir BENKO a Jaroslav REICHEL, 2017a. Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques. *IEEE Access* [online]. **5**, 8989–9004. ISSN 21693536. doi:10.1109/ACCESS.2017.2706302

MUNK, Michal a Jozef KAPUSTA, 2014. *Web Usage Mining: Príprava a modelovanie dát*. Nitra: Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-558-0692-1.

MUNK, Michal, Jozef KAPUSTA, Peter ŠVEC a Milan TURČÁNI, 2010. Data Advance Preparation Factors Affecting Results of Sequence Rule Analysis in Web Log Mining. *E+M Ekonomie a Management*. **13**(4), 143–160.

MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021a. Methodology of stakeholders' behaviour modelling based on time. *MethodsX* [online]. **8**, 101570. ISSN 22150161. doi:10.1016/j.mex.2021.101570

MUNK, Michal, Anna PILKOVA, Ľubomír BENKO, Petra BLAZEKOVA a Peter SVEC, 2021b. Pillar 3–Pre-processed web server log file dataset of the banking institution. *Data in Brief* [online]. **39**, 107672. ISSN 23523409. doi:10.1016/j.dib.2021.107672

MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021c. Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Systems with Applications* [online]. **185**, 115503. ISSN 09574174. doi:10.1016/j.eswa.2021.115503

MUNK, Michal, Anna PILKOVA, Lubomir BENKO a Petra BLAŽEKOVÁ, 2017b. Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management* [online]. **18**(5), 954–973. doi:10.3846/16111699.2017.1360388

MUNK, Michal, Anna PILKOVA, Jozef KAPUSTA, Peter SVEC a Martin DRLIK, 2013. Pillar 3 and Modelling of Stakeholders' Behaviour at the Commercial Bank Website during the Recent Financial Crisis. *Procedia Computer Science* [online]. **18**, 1747–1756. ISSN 18770509. doi:10.1016/j.procs.2013.05.343

MUNKOVA, Dasa, Michal MUNK, Ľubomír BENKO a Petr HAJEK, 2021a. The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science* [online]. **7**, e706. ISSN 2376-5992. doi:10.7717/peerj-cs.706

MUNKOVA, Dasa, Michal MUNK, Ľubomír BENKO a Jiri STASTNY, 2021b. MT Evaluation in the Context of Language Complexity. *Complexity* [online]. **2021**, 1–15. ISSN 1099-0526. doi:10.1155/2021/2806108

O'FLYNN, James Adam, 2019. An Economics Academic Word List (EAWL): Using online resources to develop a subject-specific word list and associated teaching-learning materials. *Journal of Academic Language and Learning*. **13**(1).

O'HAYRE, John, 1966. *Gobbledygook has gotta go*. U.S. Government Printing Office.

PABARSKAITE, Zidrina a Aistis RAUDYS, 2007. A process of knowledge discovery from web log data: Systematization and critical review. *Journal of Intelligent Information Systems* [online]. **28**(1), 79–104. ISSN 09259902. doi:10.1007/s10844-006-0004-1

PAPINENI, Kishore, Salim ROUKOS, Todd WARD a WeiJing ZHU, 2002. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. s. 311–318.

PILKOVÁ, Anna, Michal MUNK, Ľubomír BENKO, Petra BLAŽEKOVÁ a Jozef KAPUSTA, 2021a. Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? *PLOS ONE* [online]. **16**(10), e0258449. ISSN 1932-6203. doi:10.1371/journal.pone.0258449

PILKOVÁ, Anna, Michal MUNK, Petra BLAŽEKOVÁ a Ľubomír BENKO, 2021b. Web usage analysis: Pillar 3 information assessment in turbulent times. In: Mohammad Z. ABEDIN, Kabir HASSAN, Petr HAJEK a Mohammed M. UDDIN, ed. *The Essentials of Machine Learning in Finance and Accounting* [online]. Routledge, s. 24. ISBN 9780367480813. doi:10.4324/9781003037903

QI, Peng, Timothy DOZAT, Yuhao ZHANG a Christopher D. MANNING, 2018. Universal Dependency Parsing from Scratch. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 160–170. doi:10.18653/v1/K18-2016

QI, Peng, Yuhao ZHANG, Yuhui ZHANG, Jason BOLTON a Christopher D. MANNING, 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 101–108. doi:10.18653/v1/2020.acl-demos.14

REI, Ricardo, Craig STEWART, Ana C FARINHA a Alon LAVIE, 2020. COMET: A Neural Framework for MT Evaluation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 2685–2702. doi:10.18653/v1/2020.emnlp-main.213

ROMERO, Cristóbal, Sebastián VENTURA, Amelia ZAFRA a Paul de BRA, 2009. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers and Education*. **53**(3), 828–840.

SADEEK QUADERI, Shah Jafor a Kasturi Dewi VARATHAN, 2024. Identification of significant features and machine learning technique in predicting helpful reviews. *PeerJ Computer Science* [online]. **10**, e1745. ISSN 2376-5992. doi:10.7717/peerj-cs.1745

SAEL, N, A MARZAK a H BEHJA, 2013. Web Usage Mining data preprocessing and multi level analysis on Moodle. In: *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on* [online]. s. 1–7. ISSN 2161-5322. doi:10.1109/AICCSA.2013.6616427

SARWAR, Talha Bin a Noorhuzaimi Mohd NOOR, 2021. An Experimental Comparison of Unsupervised Keyphrase Extraction Techniques for Extracting Significant Information from Scientific Research Articles. In: *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* [online]. IEEE, s. 130–135. ISBN 978-1-6654-1407-4. doi:10.1109/ICSECS52883.2021.00031

SARWAR, Talha Bin, Noorhuzaimi Mohd NOOR, M. Saef Ullah MIAH, Mamunur RASHID, Fahmid Al FARID a Mohd Nizam HUSEN, 2021. Recommending Research Articles: A Multi-Level Chronological Learning-Based Approach Using Unsupervised Keyphrase Extraction and Lexical Similarity Calculation. *IEEE Access* [online]. **9**, 160797–160811. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3131470

SENTER, RJ a EA SMITH, 1967. *Automated Readability Index*.

SHANNON, C. E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* [online]. **27**(3), 379–423. ISSN 00058580. doi:10.1002/j.1538-7305.1948.tb01338.x

SCHMID, Helmut, 2019. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage* [online]. New York, NY, USA: ACM, s. 133–137. ISBN 9781450371940. doi:10.1145/3322905.3322915

SCHMID, Helmut, Marco BARONI, Eros ZANCHETTA a Achim STEIN, 2007. The Enriched TreeTagger System. In: *Proceedings of the EVALITA 2007 workshop*.

SPIERS, Harry, Nikul AMIN, Raj LAKHANI, Andrew J. MARTIN a Parag M. PATEL, 2017. Assessing Readability and Reliability of Online Patient Information Regarding Vestibular Schwannoma. *Otology & Neurotology* [online]. **38**(10), e470–e475. ISSN 1531-7129. doi:10.1097/MAO.0000000000001565

SPILIOPOULOU, Myra a Lukas C. FAULSTICH, 1999. WUM: A Tool for Web Utilization Analysis. In: *The World Wide Web and Databases* [online]. Springer Berlin Heidelberg, s. 184–203. doi:10.1007/10704656_12

SRIVASTAVA, Jaideep, Robert COOLEY, Mukund DESHPANDE a Pang-ning TAN, 2000. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *Text* [online]. **1**(2), 12–23. ISSN 19310145. doi:10.1145/846183.846188

SRIVASTAVA, Jaideep, Prasanna DESIKAN a Vipin KUMAR, 2005. Web Mining - Concepts, Applications, and Research Directions. In: *Foundations and Advances in Data Mining*. Springer, Berlin, Heidelberg, s. 275–307.

SRIVASTAVA, Mitali, Rakhi GARG a P. K. MISHRA, 2015. Analysis of Data Extraction and Data Cleaning in Web Usage Mining. In: *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET '15* [online]. New York, New York, USA: ACM Press, s. 1–6. ISBN 9781450334419. doi:10.1145/2743065.2743078

STRAKA, Milan a Jana STRAKOVÁ, 2014. MorphoDiTa: Morphological Dictionary and Tagger. *http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0*.

STRAKA, Milan a Jana STRAKOVÁ, 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* [online]. Stroudsburg, PA, USA: Association for Computational Linguistics, s. 88–99. doi:10.18653/v1/K17-3009

SVEC, Peter, Lubomir BENKO, Miroslav KADLECIK, Jan KRATOCHVIL a Michal MUNK, 2020. Web Usage Mining: Data Pre-processing Impact on Found Knowledge in Predictive Modelling. *Procedia Computer Science* [online]. **171**, 168–178. ISSN 18770509. doi:10.1016/j.procs.2020.04.018

TEMPLIN, Mildred, 1957. *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

THORDARDOTTIR, Elin T. a Susan Ellis WEISMER, 2001. High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders* [online]. **36**(2), 221–244. ISSN 1368-2822. doi:10.1080/13682820118239

TOERIEN, Franz Eduard a Elda DU TOIT, 2024. Fighting through the Flesch and Fog: the readability of risk disclosures. *Accounting Research Journal* [online]. **37**(1), 39–56. ISSN 1030-9616. doi:10.1108/ARJ-03-2023-0094

VELLINGIRI, J. a S. CHENTHUR PANDIAN, 2011. A novel technique for web log mining with better data cleaning and transaction identification. *Journal of Computer Science* [online]. **7**(5), 683–689. ISSN 15493636. doi:10.3844/jcssp.2011.683.689

W3C, 1995. *Configuration File of W3C httpd* [online] [vid. 2022-01-23]. Dostupné z: https://www.w3.org/Daemon/User/Config/Logging.html

WOLFE-QUINTERO, Kate, Shunji INAGAKI a Hae-Young KIM, 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Honolulu, US: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

XUE, Guo-yi a Paul NATION, 1984. A university word list. *Language Learning and Communication*. **3**, 215–229.

YAO, Zheng, X. WANG a J. LUAN, 2017. Using Hidden Markov Model to Predict the Web Users' Linkage. *Journal of Residuals Science & Technology* [online]. **14**(3), 554–565. doi:10.14355/jrst.2017.1403.053

# PRÍLOHY: ZOZNAM POUŽITÝCH PUBLIKOVANÝCH PRÁC

Príloha A: MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Systems with Applications*. **185**, 115503. doi:10.1016/j.eswa.2021.115503 (**Web of Science, 2021IF: 8.665, Q1**) [WoS: 2, Scopus: 1]

Príloha B: PILKOVÁ, Anna, Michal MUNK, Ľubomír BENKO, Petra BLAŽEKOVÁ a Jozef KAPUSTA, 2021. Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? *PLOS ONE*. **16**(10), e0258449. doi:10.1371/journal.pone.0258449 (**Web of Science, 2021IF: 3.752, Q2**) [WoS: 0, Scopus: 0]

Príloha C: MUNK, Michal, Anna PILKOVA, Ľubomír BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Pillar 3–Pre-processed web server log file dataset of the banking institution. *Data in Brief*. **39**, 107672. doi:10.1016/j.dib.2021.107672 (**Web of Science; Scopus**) [WoS: 1, Scopus: 0]

Príloha D: MUNK, Michal, Anna PILKOVA, Lubomir BENKO, Petra BLAZEKOVA a Peter SVEC, 2021. Methodology of stakeholders' behaviour modelling based on time. *MethodsX*. **8**, 101570. doi:10.1016/j.mex.2021.101570 (**Web of Science; Scopus**) [WoS: 0, Scopus: 1]

Príloha E: PILKOVÁ, Anna, Michal MUNK, Petra BLAŽEKOVÁ a Ľubomír BENKO, 2021. Web usage analysis: Pillar 3 information assessment in turbulent times. In: Mohammad Z. ABEDIN, Kabir HASSAN, Petr HAJEK a Mohammed M. UDDIN, ed. *The Essentials of Machine Learning in Finance and Accounting*. Routledge, s. 24. doi:10.4324/9781003037903 (**Scopus**) [Scopus: 1]

Príloha F: BLAŽEKOVÁ, Petra, Ľubomír BENKO, Anna PILKOVÁ a Michal MUNK, 2021. Is Pillar 3 a Good Tool for Stakeholders in CEE Commercial Banks? In: *Studies in Systems, Decision and Control*. Springer, s. 421–440. doi:10.1007/978-3-030-76632-0_15 (**Scopus**) [Scopus: 0]

Príloha G: SVEC, Peter, Lubomir BENKO, Miroslav KADLECIK, Jan KRATOCHVIL a Michal MUNK, 2020. Web Usage Mining: Data Pre-processing Impact on Found

Knowledge in Predictive Modelling. *Procedia Computer Science*. **171**, 168–178. doi:10.1016/j.procs.2020.04.018 (**Scopus**) [Scopus: 10]

Príloha H: MUNKOVA, Dasa, Michal MUNK, ĽUBOMÍR BENKO a Jiri STASTNY, 2021. MT Evaluation in the Context of Language Complexity. *Complexity*. **2021**, 1–15. doi:10.1155/2021/2806108 (**Web of Science, 2021IF: 2.121, Q2**) [WoS: 2, Scopus: 0]

Príloha I: BENKO, Ľubomír, Dasa MUNKOVÁ a Michal MUNK, 2023. Relationship Between Linguistic Complexity and MT Errors in the Context of Inflectional Languages. In: *Recent Challenges in Intelligent Information and Database Systems. ACIIDS 2023*. Springer, Cham, s. 546–557. doi:10.1007/978-3-031-42430-4_45 (**Scopus**) [Scopus: 0]

Príloha J: BENKO, Ľubomír, Lucia BENKOVA, Dasa MUNKOVA, Michal MUNK a Danylo SHULZENKO, 2022. Error Classification Using Automatic Measures Based on n-grams and Edit Distance. In: *Advanced Research in Technologies, Information, Innovation and Sustainability. ARTIIS 2022*. Springer, Cham, s. 345–356. doi:10.1007/978-3-031-20319-0_26 (**Web Of Science**, **Scopus**) [WoS:0, Scopus: 0]

Príloha K: BENKO, Ľubomír, Dasa MUNKOVA, Michal MUNK, Lucia BENKOVA a Petr HAJEK, 2024. The use of residual analysis to improve the error rate accuracy of machine translation. *Scientific Reports* (v recenznom konaní od 2023, 3. kolo) (**Web of Science, 2022IF: 4.6, Q2**)

Príloha L: BENKO, Ľubomír, Dasa MUNKOVA, Mária PAPPOVÁ a Michal MUNK, 2024. Comparison of various approaches to tagging for the inflectional Slovak language. *PeerJ Computer Science* (v recenznom konaní od 2023, 2. kolo) (**Web of Science, 2022IF: 3.8, Q2**)

Príloha M: BENKO, Ľubomír, Anna PILKOVA, Michal MUNK a Slavka ELEY, 2024. Pillar 3: The impact of language complexity on the preferences of commercial bank website users. *Expert Systems with Applications* (v recenznom konaní od 2024, 1. kolo) (**Web of Science, 2022IF: 8.5, Q1**)

PRÍLOHA A: MUNK, MICHAL, ANNA PILKOVA, LUBOMIR BENKO, PETRA BLAZEKOVA A PETER SVEC, 2021C. WEB USAGE ANALYSIS OF PILLAR 3 DISCLOSED INFORMATION BY DEPOSIT CUSTOMERS IN TURBULENT TIMES. *EXPERT SYSTEMS WITH APPLICATIONS*. 185, 115503. DOI:10.1016/J.ESWA.2021.115503 (**WEB OF SCIENCE, 2021IF: 8.665, Q1**) [WOS: 2, SCOPUS: 1]

Expert Systems With Applications 185 (2021) 115503

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times

Michal Munk [a,c,1], Anna Pilkova [b,2], Lubomir Benko [a,3,*], Petra Blazekova [b,4], Peter Svec [a,5]

[a] Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, SK 949 01 Nitra, Slovakia
[b] Comenius University in Bratislava, Odbojárov 10, SK 820 05 Bratislava, Slovakia
[c] Tomas Bata University in Zlín, nám. T. G. Masaryka 5555, CZ 760 01 Zlín, Czechia

ARTICLE INFO

*Keywords:*
Risk management
Market discipline
Data pre-processing
Financial regulation
Web usage mining

ABSTRACT

Market discipline has been a scrutinized area since the last financial crisis in 2008. Regulators strengthened their role particularly through Pillar 3 in Basel III. However, there are still some aspects of market discipline that deserve special attention to avoid future failures. This study focuses on the analysis of the interest and behaviour of deposit stakeholders based on website data dedicated to disclosures of commercial bank in Slovakia during and after turbulent times (period 2009–2012). The data consists of log files, and web mining techniques were applied (the modelling of web user behaviour in dependence on time - based on the proposals of the authors). The results show that also in turbulent times, stakeholders' interest in Pillar 3 disclosures is low (in line with (Munk, Pilkova, Benko, & Blažeková, 2017)) and the highest interest was identified for the Pricing List category. After turbulent times, Pillar 3 categories (Pillar 3 related information and Pillar 3 disclosures) have weak interest, with peaks at the beginning of the year, and the highest increase was in the Business Conditions category. The results suggest that the enhancement of interest of key stakeholders in disclosures inevitably requires changes to deliver sufficient disclosure data structures and to design a disclosure policy that fulfils regulatory expectations.

## 1. Introduction

Currently, we live in a world of fast changes in products, technologies, companies, markets and industries etc. These changes are sources of economic turbulence which banking sectors also cannot avoid. The last financial crisis proved and highlighted the weaknesses of the global financial regulatory system which was not able to avoid the failures and losses generated by turbulence in the banking industry. Regulators, policy makers, and academics learnt many lessons from this period and tried to fix identified weaknesses. Market discipline is also one of the areas on which regulators have focused. In regulation, it has been included since Basel II's introduction as a Pillar 3 component. Pillar 3 complements the minimum risk-based capital requirements and the other quantitative requirements (Pillar 1), and the supervisory review process (Pillar 2). It aims to promote market discipline by providing meaningful regulatory information to market participants consistently and to be able to assess banks' risk appetite, risk exposure, and level of risk management. Market discipline, in its broadest terms, can be understood as a mechanism via which market participants monitor, assess, and discipline risk-taking by financial institutions. In the studies of Bliss and Flannery (Bliss & Flannery, 2002), a market discipline is defined by its two distinguishing aspects into market monitoring – market participants' assessment of banks' conditions, which are to be reflected in banks' security prices and deposit rates; market influence – banks' reaction brought on by market monitoring or to counteract adverse changes in banks' conditions.

Pillar 3 has been discussed and reviewed by key market participants and concluded with the obligatory regulatory standards for information disclosure for banks. The first obligatory standard for information disclosure for banks was launched as a revised version of the Pillar 3

* Corresponding author.
*E-mail addresses:* mmunk@ukf.sk (M. Munk), anna.pilkova@fm.uniba.sk (A. Pilkova), lbenko@ukf.sk (L. Benko), psvec@ukf.sk (P. Svec).
[1] ORCID: 0000-0002-9913-3596.
[2] ORCID: 0000-0002-4296-4823.
[3] ORCID: 0000-0002-1657-395X.
[4] ORCID: 0000-0001-9545-9436.
[5] ORCID: 0000-0002-1713-6444.

https://doi.org/10.1016/j.eswa.2021.115503
Received 5 January 2020; Received in revised form 3 June 2021; Accepted 26 June 2021
Available online 1 July 2021
0957-4174/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

framework in December 2015, which is a background to promote market discipline through regulatory disclosure requirements. This document has been reviewed several times in consultation processes. Firstly, the Pillar 3 review and the Pillar disclosure requirements consultative document was disclosed in March 2016, and available for comment until March 2017. Consequently, the Basel Committee on Banking Supervision (BCBS) issued the Pillar 3 disclosure requirements - updated the single Pillar 3 framework in February 2018 (available for comment until May 2018). Moreover, the European Banking Authority (EBA) also published the final guidelines on regulatory disclosure requirements in December 2016, the goal of which is the consistency and comparability of institutions' disclosures and to ensure market discipline. It is important to note that changes and updates to the regulatory disclosure documents should reflect the requirements of key market participants, which is the objective of regulation and also of its consultation processes. Therefore, the complexity of the finalisation process of the disclosure standards, and the very high amount of changes in the disclosed forms suggest that the goal of the standard has not been reached during the implementation of its versions (enhancement of market discipline), also from the regulator's point of view (BCBS - Basel Committee on Banking Supervision, 2015; BCBS - Basel Committee on Banking Supervision, 2016; BCBS - Basel Committee on Banking Supervision, 2018; 2016; 2017). Consequently, the standard and its forms (disclosure tables) should cover the interests of key market participants on the information disclosed based on their behaviour.

All of these discovered and confirmed processes covering requirements of key market participants required changes to find the most effective structure of information. It also has been found that the incentive to monitor the risk that a bank takes is higher for depositors' investments, which are neither guaranteed by deposit insurance nor by other government regulation (Fonseca & González, 2010; Nier & Baumann, 2006). Therefore, in terms of enhancement of market discipline, one specific group of stakeholders is uninsured depositors, who are the source of market discipline in monitoring the bank's risk. Moreover, depositors in Europe actively monitor the riskiness (interest rate on deposits) of banks and prefer safer banks, which can repay deposits anytime and pay higher interest rates (Fueda & Konishi, 2007; Hori, Ito, & Murata, 2009; Martinez Peria & Schmukler, 2001). However, the Guillemin study (Guillemin, 2017) focusing on the effects of disclosure on depositors' behaviour suggests that disclosures negatively influence deposit levels in European banks. Furthermore, according to Bourgain, Pieretti, and Zanaj (2012), disclosures in the European Union (EU) have a positive impact on deposit levels in Middle Eastern and North African (MENA) countries and also in Russia (Wu & Bowe, 2012). Nevertheless, different behaviours among depositors in Europe serve as the background for our research together with a lack of research in the field of analysis of the behaviour of key stakeholders to disclosures in Central and Eastern European (CEE) countries.

### 1.1. Motivation and research objectives

Commercial banks in CEE countries have numerous specifics. Among them, the most important is prevailing ownership by large international groups and focus on deposit collection. Their depositors represent a very important group of stakeholders; however, empirical studies on their behaviour and interests in using Pillar 3 disclosures are missing. Regulators do not know to what extent existing disclosure rules are meaningful and add value for this type of users and help avoid market discipline failure similar to during the last financial crisis. However, it is crucial for the regulators' goal achievement that market discipline mechanism should be effective and used according to the regulators' expectations. In CEE countries, there is a lack of studies assessing Pillar 3 information disclosures based on the content relevancy to key commercial banks' stakeholders.

Therefore, our study analyses the interest in information disclosures aimed at a specific type of stakeholders (uninsured depositors) in foreign-owned bank not traded on capital markets. The importance of this group is further supported by the fact that nearly half of the deposits in bank accounts in Slovakia are uninsured deposits, and a similar status can be expected in other CEE countries. Moreover, we agree with Kuranchie-Pong, Bokpin, and Andoh (2016) that stakeholders are expected to contribute to effective risk management in the banking industry through market discipline, and they need a sufficient disclosure of risk-related information to assess the risk profiles of banks. In terms of disclosing information, Goldstein and Leitner (2015) stress the necessity of disclosures in preventing market breakdown but they also point out that in terms of sufficient disclosures, there is a potential threat in disclosing too much information. However, according to Bouaiss, Refait-Alexandre, and Alexandre (2017), an increase in disclosure enhances transparency and efficient market discipline by supervising excessive risk-taking and also improves stockholders' monitoring and sensitivity to risk-taking (Goldstein & Sapra, 2014). On the other hand, transparency can have a positive impact on bank performance and stability but only up to a certain point (Iren, Reichert, & Gramlich, 2014). This is due to its conflicting effects: more transparency decreases efficient liquidity, increases rollover risk, and has a negative impact on banks' stock (Bouvard, Chaigneau, & de Motta, 2015). Moreover, Andrievskaya & Semenova conclude that the concentration of the markets is lower with stricter public disclosure requirements, and there is a reduction of competition due to stricter disclosure requirements, which depends on bank credit risks (Andrievskaya & Semenova, 2016). Additionally, they also highlight the marginal positive effects of additional regulation, mostly in developed countries, where regulation is high. These are general findings derived mostly from developed countries. It is important to take into consideration all these factors in the process of designing an optimal disclosure policy.

To sum up, based on the findings stated above, the main goal of this study is twofold: Firstly, to assess the behaviour and interest of the key stakeholders (primarily depositors) in a foreign-owned commercial bank (not traded on capital markets) on disclosed information during and after turbulent times in a country that belongs to the CEE region and subsequently, based on website data analysis during and after crises to identify the key types of information, which are in the particular interest of the key stakeholders as a component for the design of optimal disclosure policy. This study is also aims to contribute to fulfilling a significant research gap in the area of the content relevancy of Pillar 3 disclosures to web users of commercial banks in CEE countries.

The structure of this study is as follows: The first section contains an introduction and the main goal of the study, followed by a section that offers the current status of empirical research and theories connected with market discipline and web usage analysis. The methodology of the modelling of web user behaviour dependent on time, and research results are included in the third and fourth sections. The last section is the discussion and conclusion.

## 2. Related work

The theory of market discipline is related to the efficient markets hypothesis (Fama, 1970) and explains that depositors (and similar money instruments investors) can rein in the risk taken by banks through market-based mechanisms (Garten, 1986). The theory of market discipline has been scrutinised in the years following the last financial crisis in 2008. Common outcomes of the numerous studies on why market discipline failed during the crisis are twofold (Min, 2015): a/ market discipline very much relies on investors in money instruments who are relatively insensitive to the risk and b/ neglect very risk-sensitive investors who might encourage even greater risk. However, even before the last financial crisis, some authors dispute the ability of, in particular, retail depositors to monitor and change the risky behaviour of banks (Nagarajan & Sealey, 1997; Garten, 1986). The reason, why retail depositors are unsophisticated, poorly equipped to receive risk-related information about their banks and are likely to misinterpret

such information lies in a lack of financial literacy (Semenova, 2012). Nevertheless, post-crisis regulators' efforts tend to enhance market discipline. Enhancement of market discipline as an issue is connected to factors, which support or discourage disciplined behaviour. These factors are studied by researchers from different perspectives. One of these perspectives is the effectiveness of market discipline and disclosure is regarded as one of the most effective tools for the enhancement of market discipline (Fonseca & González, 2010). According to Hamid and Yunus (Hamid & Yunus, 2017), disclosures are effective market discipline tools when the market is more concentrated. On the other hand, Andrievskaya and Semenova (2016) conclude that the concentration of markets is lower with stricter public disclosure requirements and that increased levels of banking disclosures stimulate banks' competition and positively affect banks' market share by attracting more retail deposits (Andrievskaya & Raschupkin, 2015). However, Guillemin (2017) states that depositors in EU countries actively monitor banks and the decline of the deposit base is connected with higher levels of disclosure, which is the case for more transparent banks. Moreover, Hamid's research showed that foreign banks are subject to market disciplining effects when disclosure is taken into account. This enhances the supervisory need to impose stricter disclosure requirements in a more pertinent and timely manner.

However, what about commercial banks for which customers are retail depositors who are risk insensitive? Several studies are focused on the impact of the increase of information disclosures on these banks. According to Bouaiss et al. (2017), the increase in disclosure levels enhances transparency and efficient market discipline to supervise excessive risk-taking. Moreover, according to Guillemin and Semenova (2018), broader and more transparent disclosures increase the ability to attract interbank funding, while the larger and riskier banks (179 largest Russian banks in 2004–2013) tend to have broader and more transparent disclosures. Moreover, increases in banks' disclosures positively influence investor's attitude to bank's risk profiles, actively increase bank's value (Zer, 2015), boosts depositors' sensitivity to equity levels (Kozłowski, 2016), and improve stockholders' monitoring and sensitivity to risk-taking (Goldstein & Sapra, 2014). On the other hand, Goldstein and Leitner (2015) stress the necessity of disclosing information in preventing market breakdown but also point out the potential threat of disclosing too much information, which destroys risk-sharing opportunities.

Nevertheless, an increase in disclosures, which is also connected with an increase in transparency, is outlined by interpretability, availability of information disclosures, and improvement of information transition mechanisms in the stimulation of market discipline. However, studies are lacking which would analyse the behaviour of customers of commercial banks as far as the usage of disclosed information during and after a period of crisis. Therefore, we focus on this issue to explore sufficient disclosures and transparency to enhance commercial bank depositors' incentives to monitor banks, which is an issue in market discipline practice.

Improving transparency is a component of post-crisis regulators' efforts to enhance market discipline (Min, 2015). Researchers appreciated transparency as an effective factor in incentives to market disciplinary effects that lead to greater market efficiency (Gandrud & Hallerberg, 2014). These authors point out that even from a regulatory point of view current EU bank transparency is insufficient. Furthermore, Moreno and Takalo (2016) suggest that only an intermediate level of transparency is socially optimal and effective (Bouvard et al., 2015) to balance its conflicting effects: more transparency decreases efficient liquidity and increases rollover risk. Parwada, Lau, and Ruenzi (2015) also support this view claiming that the increases in transparency by reporting Pillar 3 and its disclosures have a negative impact on banks' stock and the cost of trading corporate bonds decreased, which is in conjunction with Iren et al. (2014) stating that broader and greater disclosures increase bank performance and stability only up to a certain point.

Referring to market discipline theory failures due to the risk-insensitive behaviour of depositors, we studied papers in which authors concentrate on the analysis of depositors' behaviours in connection with the timing of market turbulences (Li, Liu, Siganos, & Zhou, 2016; Arnold, Größl, & Koziol, 2015; Li & Wang, 2014). Arnold et al. (2015) state that strong market discipline is observed during and after turbulent times (sample of German depositors in 2003–2012) and it strengthened after the crisis, with no exemptions for banks with government support (Antzoulatos & Karanastasis, 2016). This is in line with Kaffenberger's study (Kaffenberger, 2015), which states that market discipline has increased and is stronger in the European Union (decrease in deposit growth rates) for fiscally weak countries in the aftermath of the Cypriot bail out. According to Andrievskaya and Semenova (2016), the timing of market turbulence influences the behaviour of depositors, who prefer more sensitive, less risky, and more reliable banks (with strong capital adequacy and liquidity) connected with higher growth in the volume and the share of deposits.

Those results provide evidence that the timing of market turbulence influences the behaviour of depositors and their perception of banks' risk. Depositors in Europe show different disciplining behaviour in terms of transparency and disclosures, and research in this field is rather rare. Nahar, Azim, and Anne Jubb (2016) also conclude that disclosure in general offers lower costs of capital and helps investors to maintain information about the bank's risk and its management. This is valid for an investor in securities or capital instruments, whose risk behaviour, and according to Min (2015), was neglected even before the 2008 financial crisis. The lower risk disclosure occurs in the case of high performing banks, and it generates ambiguity for potential stakeholders. At this point, we have to agree with those authors who point out that further research is needed and inevitable as far as studying the behaviour of different stakeholders' groups of commercial banks. Our study contributes to cover the existing literature gap in depositors' behaviour towards information disclosure during and after turbulent times in the CEE region. Our intent is also to contribute to design and optimise disclosure policy that would prevent future market discipline failures. Moreover, the analysis of interest in disclosed information by depositors' behaviour helps regulators and other interested parties to assess the effectiveness of the implementation of the regulations and their goals, which are enhancements to an effective market disciplining mechanism, relevant, meaningful, and sufficient disclosures.

### 2.1. Weekly-based web usage analysis literature review

The use of web portals is mainly represented by time data. The variable time occurs often only when extracting sequential rules but only when determining the order of visited web parts. There is no time-based modelling of website users' behaviour in the application area. Nevertheless, several authors tried to track behaviour in other ways. Arbelaitz et al. (2013) focused on analysing and creating navigational profiles of visitors to tourist web portal. They designed a system that can achieve profiles that correspond to real visitors' search sequences with a success rate of 60%. The authors aimed to use navigational profiles to better personalize the web portal for visitors. They used basic principles of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to create segments of users with common interests. Other authors (Anitha, 2010; Bhawsar, Pathak, & Patidar, 2012; Vojíř, Zeman, Kuchař, & Kliegr, 2018; Yin & Guo, 2013) also focused on the effort to detect, respectively, to predict the next step of visitors to the web portal. Makkar, Gulati, and Sharma (2010) used Petri Nets with information obtained from the log file and the structure of the web portal to predict the behaviour of users on the web portal. Carmona et al. (2012) focused on developing a methodology for e-commerce web portals using Google Analytics, not sequentially accessing the data. Subsequently, the extracted data was processed by aggregation, association analysis, and subgroup discovery. Based on the results, they identified recommendations and problematic areas of the surveyed web portal for the portal

management team. Van Nguyen, Zhou, Chong, Li, and Pu (2019) dealt with a data mining prediction approach in their paper. The authors aimed to obtain a prediction model for remanufactured products and described the non-linear effect of online market factors as predictors of customer demands. The results of the research showed that the most accurate results were obtained for the ensemble regression tree model.

Inspired by a time-based weekly survey of site users, similar research from other areas will be analysed. Dabrowska-Zielinska, Kogan, Ciolkosz, Gruszczynska, and Kowalik (2002) focused on the development of a model for studying the conditions of crops in various regions of Poland, where the remote exploration of the Earth was used. Based on satellite imagery, they calculated two indices for each week for 14 years. Then they monitored the measured values each week and evaluated the most suitable planting and cropping period during the year. All the results were supported by meteorological observations. The authors used regression analysis relating to yield deviation from the mean. The model was used for yield prediction for the year 1998, where the results of the prediction were compared with results published by the Central Statistical Office. Raffi et al. (2006) focused on examining the impact of the application of antiretroviral treatment to the twelfth week of the disease and the prediction of treatment over the next few weeks. To model the results of the critical weeks of treatment 24, 48, and 96, they used a prediction based on the observed treatment effects during the first twelve weeks. Logistic regression analysis was used to compare patients. The 12 week border was used as a baseline and was compared with the following critical weeks. On the basis of the results, after twelve weeks of treatment, the patient may be advised to continue or discontinue the treatment. The long-term predictive value of early treatment was analysed and the results of continuous treatment validated the assumptions of the authors. Even in the modelling of river basins, Verdhen, Chahar, and Sharma (2014) made predictions focusing on the Himalayan snow melting pattern during the spring. In their research, they used data from 2008 to create a prediction model for the years 2003 and 1983 back-to-back years, examined weekly during the spring periods of those years. The authors focused on the Nash and Sutcliffe efficiency coefficient and the linear regression coefficient to verify the efficiency of computations concerning the observed data. The variability in prediction from temperature index or energy balance models with observations was evaluated. The accuracy of temperature index or energy balance algorithms was determined in terms of probability. In addition to the research results, they also found that their input data was not reliable enough to achieve an effective simulation coefficient for the year 1983. They want to remove this in future research by adding additional parameters to help them to improve the simulation quality.

## 3. Materials and methods

In this paper, data related to Pillar 3 were gathered from bank web server log files (Munk, Pilková, Drlík, Kapusta, & Švec, 2012). The log files contain information about visitors. This information can be used for further analysis of visitors' behaviour. However, on the other hand, the web server log file also contains irrelevant and unnecessary data as well as inaccurate and incomplete information. Therefore, it is necessary to pre-process the data obtained from the web server log files. Data preparation of the log files consists of data cleaning, session identification, and path completion (Kapusta, Munk, Svec, & Pilkova, 2014; Kapusta, Pilková, Munk, & Švec, 2013). Data preparation was completed according to (Kapusta et al., 2014; Kapusta et al., 2013), where it was needed to pre-process log files from multiple servers that are used as load balancers. After data preparation, the data sample comprised 2 071 235 accesses. Subsequently, the variables were created. The dependent variable *category* was created from URL address, and its levels represent the examined web categories of the web portal (*Pricing List, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements,* and *We support*). The timestamp served to create the independent variables-predictors *week* (0–53) and *crisis* (0: 2011–2012, 1: 2009–2010), that define the time and period of the access to the examined web categories of the web portal. The dummy variable *internal* (0, 1) was created from the IP address, and it divides the accesses from inside and outside of the organization network. Modelling the probabilities of the accesses to the examined web categories of the web portal of the bank depending on time was done using the multinomial logit model, which is a part of generalized linear models. The estimation of the models' parameters was done by maximizing the logarithm of the multinomial likelihood function. Subsequently, the logits were estimated and used to estimate the probabilities of accesses to the examined web portal categories. The models were evaluated by comparing the observed and expected values at the level of frequencies, probabilities and logits. The created research methodology was inspired by Munk, Benko, Gangur, and Turčáni (2015), Munk et al. (2017), Munk and Drlk (2014), Munk, Kapusta, and Švec (2010) and Munk, Kapusta, Švec, and Turčáni (2010).

## 4. Results

The probabilities of the accesses to the examined web portal categories were modelled (*category*) based on time, where time was represented by the variable week (*week*). The web server log file from multiple servers was used as load balancers. The source data came from a domestic significant commercial bank operating in Slovakia. The experiment was conducted on the sample of 2 071 235 log accesses obtained after data pre-processing, which is comprised of data cleaning, session identification, and path completion. Web users' behaviour was monitored over several years (2009–2012). The influence of other factors was also analysed, whether it was significant to distinguish the internal and external accesses (*internal*) and whether it was significant to distinguish the years of financial crisis and years after the financial crisis (*crisis*). In the case of the dummy variable *internal,* a trivial degree of dependency with the variable *category* (*Contingency coefficient* $C = 0.077$; *Cramer* $V = 0.077$) was identified. The contingency coefficient can have values from 0 (represents no dependence between variables) to 1 (represents the perfect dependence between variables).

In the case of the dummy variable *crisis,* a small degree of dependency with the variable *category* (*Chi-square* $= 81\,455.210$; $df = 5$; $p = 0.000$; *Contingency coefficient* $C = 0.224$; *Cramer* $V = 0.230$) was identified. The contingency coefficient is statistically significant.

Based on these results, a model for all accesses was created (the internal and external accesses were not distinguished), and a dummy variable *crisis* was implemented into the model as a predictor.

Although the methodology was based on the research of Munk, Drlik, and Vrabelova (2011), it was not clear, by week, whether the model would be a polynomial model of the second, third or fourth degree. Using the Likelihood-ratio (LR) test (Tables 1–3), it was possible to compare estimates of the theoretical counts of accesses with the empirical counts of accesses. On the other hand, the disadvantage of the LR test is in the case of not sufficiently large expected counts where the condition of the LR test usage is violated. In that case, alternative techniques such as visualization of the difference in empirical and theoretical counts, and extreme values identification are used. The results of the LR test for all three polynomial models (second, third and fourth-degree) showed that the value of the LR test is small, so all models can be taken as appropriate. The value of the Pearson Chi-square is roughly equal to 1, which also indicates the suitability of the models.

The maximum of the logarithm of the likelihood function was appropriate for a comparison of the models. The smaller the value, the

**Table 1**
Evaluation of the second-degree polynomial model.

| | df | Stat | Stat/df |
|---|---|---|---|
| Deviance | 7,684,410 | 4,529,670 | 0.589462 |
| Pearson Chi-square | 7,684,410 | 7,794,534 | 1.014331 |
| Log-likelihood | | −2264835 | |

**Table 2**
Evaluation of the third-degree polynomial model.

| | df | Stat | Stat/df |
|---|---|---|---|
| Deviance | 7,684,405 | 4,510,212 | 0.586931 |
| Pearson Chi-square | 7,684,405 | 7,784,416 | 1.013015 |
| Log-likelihood | | −2255106 | |

**Table 3**
Evaluation of the fourth-degree polynomial model.

| | df | Stat | Stat/df |
|---|---|---|---|
| Deviance | 7,684,400 | 4,503,695 | 0.586083 |
| Pearson Chi-square | 7,684,400 | 7,829,313 | 1.018858 |
| Log-likelihood | | −2251847 | |

more appropriate the model. The highest value of log-likelihood (Tables 1–3) was identified for the fourth-degree polynomial model that is not in favour of this model. However, further evaluation of the model by visualizing empirical and theoretical logits allowed us to choose the appropriate degree of the polynomial for the examined model.

Parameter estimation for individual data was done using *STATISTICA Generalized Linear/Nonlinear Models*. The significance of the parameters was tested using the Wald test. The probability of access to the web portal categories was modelled depending on the time-week of access and crisis period. Time was represented by the variable *week* and its transformation based on the degree of the polynomial (*week*$^2$, *week*$^3$, and *week*$^4$) and the dummy variable *crisis* that represents the years of the financial crisis.

Based on the results of the test of all effects (Table 4) for the third-degree polynomial model, the parameters of the model are statistically significant. In the created model the years of the crisis and after the crisis represent a statistically significant sign that is represented by the dummy variable *crisis*. The weeks of the year represented by the variables *week* and its transformation based on the degree of polynomial also showed statistically significant signs.

The estimated parameters for almost all categories (except the category *Pricing List*) were significantly dependent on the week of the access and also for its transformations (Table 5). The values of logits were significantly affected by the period of the crisis. The logit model provides a probability estimate of the output. The absolute size of the parameters reflects predictors with the highest influence on the examined variable. A high absolute value of the parameter refers to a large dependency. A negative value refers to indirectly proportional dependence.

Using the estimated parameters, it was possible to calculate the logits for each category $j$ in time $i$. The third-degree polynomial model

$$\widehat{\eta}_{ij} = \alpha_j + \beta_{1j} week_i + \beta_{2j} week_i^2 + \beta_{3j} week_i^3 + \gamma_j crisis_i, j = 1, 2, \cdots, J-1,\ i = 0, 1, 2, \cdots, 53$$

The parameters were analogically estimated for the polynomial models of the second and fourth-degree.

Calculated logits were used to estimate the probability for the referential category. The estimate calculation was denoted by $\widehat{\pi}_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\widehat{\eta}_{ij}}}$, where $\widehat{\eta}_{ij}$ are the logits for the web category $j$ in time $i$. Based

**Table 4**
Test of all effects for the third-degree polynomial model.

| | df | Wald Stat | p |
|---|---|---|---|
| **Intercept** | 5 | 52325.51 | 0.0000 |
| **week** | 5 | 10800.15 | 0.0000 |
| **week$^2$** | 5 | 18259.02 | 0.0000 |
| **week$^3$** | 5 | 18995.98 | 0.0000 |
| **crisis** | 5 | 68622.95 | 0.0000 |

**Table 5**
Parameter estimation of the third-degree polynomial model.

| | Category | Estimate | Std Dev | Wald Stat | p |
|---|---|---|---|---|---|
| **week** | Pricing List | 0.0233 | 0.0025 | 84.7536 | 0.0000 |
| **week$^2$** | Pricing List | 0.0001 | 0.0001 | 1.3018 | 0.2539 |
| **week$^3$** | Pricing List | 0.0000 | 0.0000 | 0.3505 | 0.5538 |
| **crisis** | Pricing List | −1.1730 | 0.0087 | 18003.3166 | 0.0000 |
| **week** | Reputation | 0.0507 | 0.0030 | 291.4300 | 0.0000 |
| **week$^2$** | Reputation | −0.0023 | 0.0001 | 267.4115 | 0.0000 |
| **week$^3$** | Reputation | 0.0000 | 0.0000 | 305.2583 | 0.0000 |
| **crisis** | Reputation | −0.9237 | 0.0101 | 8362.0226 | 0.0000 |
| **week** | Business Conditions | −0.0623 | 0.0028 | 494.4624 | 0.0000 |
| **week$^2$** | Business Conditions | 0.0056 | 0.0001 | 1814.5684 | 0.0000 |
| **week$^3$** | Business Conditions | −0.0001 | 0.0000 | 1721.9480 | 0.0000 |
| **crisis** | Business Conditions | −2.0717 | 0.0095 | 47714.1908 | 0.0000 |
| **week** | Pillar3 related | 0.0655 | 0.0027 | 574.4090 | 0.0000 |
| **week$^2$** | Pillar3 related | −0.0024 | 0.0001 | 343.8587 | 0.0000 |
| **week$^3$** | Pillar3 related | 0.0000 | 0.0000 | 301.2951 | 0.0000 |
| **crisis** | Pillar3 related | −0.8455 | 0.0093 | 8189.3534 | 0.0000 |
| **week** | Pillar3 disclosure requirements | 0.1703 | 0.0030 | 3182.4843 | 0.0000 |
| **week$^2$** | Pillar3 disclosure requirements | −0.0074 | 0.0001 | 2640.8711 | 0.0000 |
| **week$^3$** | Pillar3 disclosure requirements | 0.0001 | 0.0000 | 2284.0930 | 0.0000 |
| **crisis** | Pillar3 disclosure requirements | −0.9935 | 0.0100 | 9890.9264 | 0.0000 |

on the estimate of the probability of access to the referential web category and estimated logits, it was possible to estimate the probabilities of accesses to the other web categories $\widehat{\pi}_{ij} = e^{\widehat{\eta}_{ij}} \widehat{\pi}_{iJ}, j = 1, 2, \cdots, J-1$, where $\widehat{\eta}_{ij}$ are the logit estimates of the web category $j$ in time $i$ and $\widehat{\pi}_{iJ}$ is the estimate of the probability of access to the referential web category $J$ in time $i$.

To find the most suitable model, it was necessary to evaluate the results for each of the models and to visualize the probabilities of access to web categories during the weeks of the year. The evaluation of the model was conducted on three levels – firstly, the counts of accesses, then the probabilities, and finally, the logits. We used contingency tables to define empirical counts of accesses $y_{ij}$ for the web category $j$ in the time $i$. Based on the estimated probabilities of accesses of the visitors on the examined web categories, it was possible to estimate theoretical counts of accesses $\widehat{y}_{ij} = \widehat{\pi}_{ij} \sum_j y_{ij}$, where $\widehat{\pi}_{ij}$ are the estimates of probabilities of accesses and $y_{ij}$ are empirical counts of accesses to the web category $j$ in time $i$. Problematic parts in counts were identified using the visualization of differences of empirical and theoretical counts $d_{ij} = y_{ij} - \widehat{y}_{ij}$, where the extreme values are identified based on rule $2\sigma$.

The comparison of differences in counts for each model allowed us to determine the most suitable model. Calculated differences were similar for each model. The most considerable difference was identified for the category *Business Conditions* for the model of the second-degree polynomial (Fig. 1) in the 31st week (second-degree polynomial: 14 445.715; third-degree polynomial: 14 020.797; fourth-degree polynomial: 13 548.354). The plot (Fig. 1) visualizes the differences in empirical and theoretical counts of accesses of visitors in the years of the crisis. After applying rules, $2\sigma$ extreme values were identified. For the category *Business Conditions* in the case of the 31st week, the prediction was undervalued. All three models contained a similar count of extreme values and represent 5.55% of all cases.

The next step was to evaluate the probabilities. Based on the observed counts, the empirical relative counts of accesses $p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}}$ can be calculated. Following that, the distributions of the probabilities of empirical relative counts of accesses and the estimated probabilities of the selected web category $j$ in time $i$ were compared. The zero hypothesis was tested using the Wilcoxon pair test. The distribution of the pairs was symmetrical around zero $r_{ij} = p_{ij} - \pi_{ij}$, H0: $F(-r) = 1 - F(r)$. Only two

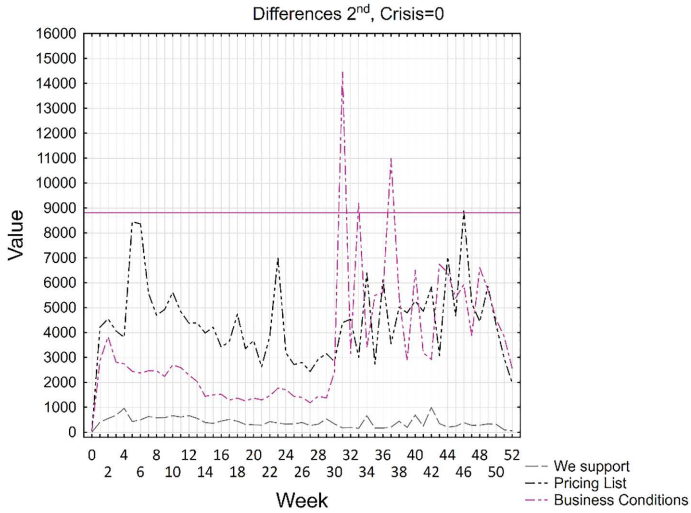



**Fig. 1.** Differences of counts of the second-degree polynomial model.

categories showed issues (Table 6). It was not clear which degree of the polynomial is the best for the model. During the years of crisis, an issue for the category *We support* (second-degree polynomial: $p = 0.000320$ was identified; third-degree polynomial: $p = 0.000299$; fourth-degree polynomial: $p = 0.000342$). In the case of years after the crisis, the category *Business Conditions* was discovered by the evaluation of differences of counts of accesses.

The last option, to decide which degree of the polynomial is the most suitable, was to evaluate theoretical and empirical logits. In this case, it was analysed whether the estimated theoretical logits fit (model) the empirical logits calculated from the empirical relative counts of accesses $h_{ij} = \ln\left(\frac{p_{ij}}{p_{iJ}}\right), j = 1,2,\cdots,J-1$, where $p_{ij}$ is the empirical relative count of access to the web category $j$ in time $i$ and $p_{iJ}$ is the empirical relative count of access to the referential web category $J$ in time $i$. The visualization of empirical and theoretical logits of each of the examined web categories (except the referential web category) can show how the theoretical logits model the empirical logits. Based on the visualization, it was seen that all the theoretical logits fit the empirical logits. However, in the case of the second-degree polynomial model (Fig. 2), it was observed that the theoretical logits were more similar to a linear function in the case of some categories, rather than quadratic. In the case of the theoretical logits, the third (Fig. 3) and fourth (Fig. 4) degree polynomials better fitted the empirical logits. However, the fourth-degree polynomial model (Fig. 4) captured the course of the empirical logits in too much detail. On the other hand, the third-degree polynomial model (Fig. 3) offered the same fitting but did not adjust to the data at the expense of the trend. The suitability of the choice of the polynomial degree is shown on plots of the web category *Pillar3 related* (Figs. 2–4).

**Table 6**
Probability distribution for the third-degree polynomial.

| Category | Crisis | N | T | Z | p-value |
|---|---|---|---|---|---|
| **We support** | 1 | 53 | 307.0000 | 3.6164 | 0.0002 |
| **Pricing List** | 1 | 53 | 570.0000 | 1.2881 | 0.1977 |
| **Reputation** | 1 | 53 | 678.0000 | 0.3320 | 0.7399 |
| **Business Conditions** | 1 | 53 | 621.0000 | 0.8366 | 0.4028 |
| **Pillar3 related** | 1 | 53 | 659.0000 | 0.5002 | 0.6169 |
| **Pillar3 disclosure requirements** | 1 | 53 | 686.0000 | 0.2612 | 0.7940 |
| **We support** | 0 | 53 | 691.0000 | 0.2169 | 0.8283 |
| **Pricing List** | 0 | 53 | 696.0000 | 0.1726 | 0.8629 |
| **Reputation** | 0 | 53 | 597.0000 | 1.0490 | 0.2942 |
| **Business Conditions** | 0 | 53 | 456.0000 | 2.2973 | 0.0216 |
| **Pillar3 related** | 0 | 53 | 564.0000 | 1.3412 | 0.1799 |
| **Pillar3 disclosure requirements** | 0 | 53 | 698.0000 | 0.1549 | 0.8769 |

The plots (Figs. 5 and 6) show the visualization of probabilities of access to each of the examined web categories during the years of the financial crisis. The highest access during the crisis was estimated for the web category *Pricing List*, where the highest access probability was during the weeks at the end of the year (the 50th week has the value of 0.481). The lowest estimated values were identified for the weeks at the beginning of the year (the 10th week has a value of 0.381). The second most accessed web category for almost half the year was the web category *Pillar3 related*, where the highest visit rate for this web category was achieved during the first quarter of the year (the 10th week has the value of 0.214). In the second half of the year, access to this web category decreased but significantly increased again in the last four weeks of the year. From the 33rd week to the 50th week of the year, the web category *Business Conditions* was, based on the estimate of the probability of access, the second most visited web category. The highest value of 0.210 was identified in the 42nd week of the year. On the other hand, in the first half of the year, the web category *Business Conditions* was the least visited web category where the lowest value of 0.053 was identified in the 10th week. The probabilities of access to the other web categories are around the value of 0.10. The small rise in the probability of access to the web category *Pillar3 disclosure requirements* is also interesting with the highest value of 0.145 in the 14th week but which afterwards decreases to a value of under 0.10.

The plots (Figs. 7 and 8) show the probabilities of accesses to each examined web category during the years after the financial crisis. The highest access rate was estimated for the web category *Pricing List* in the period of years after the financial crisis. Except for the period of the 36th – 46th week where the web category *Business Conditions* had the highest probability of access. The biggest difference in comparison to the years

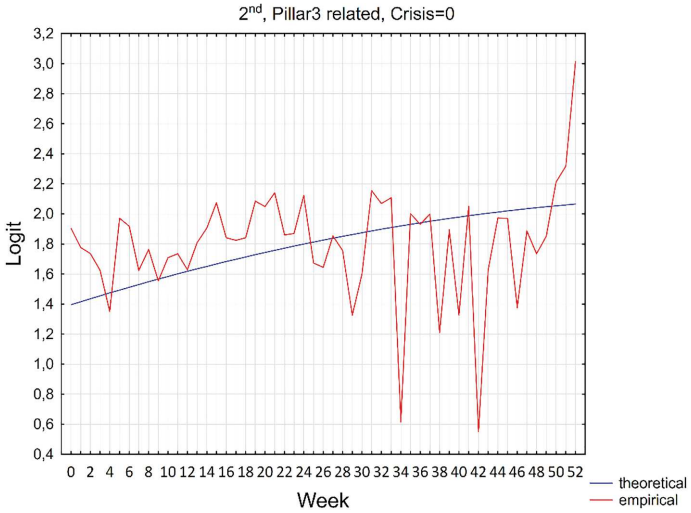



**Fig. 2.** Logit visualisation of the second-degree polynomial model of the Pillar3 related category.

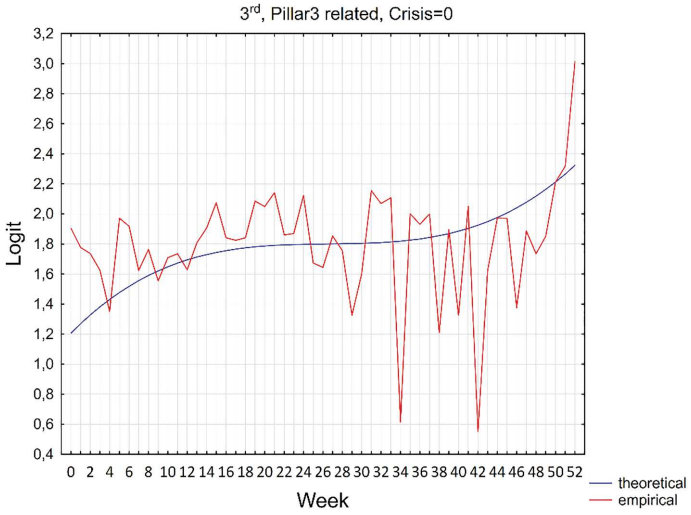



**Fig. 3.** Logit visualisation of the third-degree polynomial model of the Pillar3 related category.

of the crisis was identified for the web category *Business Conditions,* where the access rate almost doubled. On the other hand, access to the web category *We support* decreased in the years after the crisis. In the case of the web category *Reputation*, *Pillar3 related* and *Pillar3 disclosure requirements,* the probability of access decreased, but the behaviour of the visitors was the same as in the years of the financial crisis. The visitors had higher interest at the beginning of the year and during the year interest decreased but at the end of the year, it began to increase again.

From the detailed weekly analysis (of the years 2009–2012) of users' behaviours on the web portal of published financial and risk information by a commercial bank, it was discovered that the results of the analysis correspond to the results of previous quarterly analysis (Munk et al., 2017). The stakeholders had the highest interest in the mandatory and supplementary Pillar 3 information during the first quarter, where the period around the 10th week could be specified (that week displayed the highest interest in these web categories). On this basis, it can be concluded that the required quarterly publication frequency of the results is not necessary for market discipline. It would be sufficient to publish this information annually, ideally in the early weeks of the year.

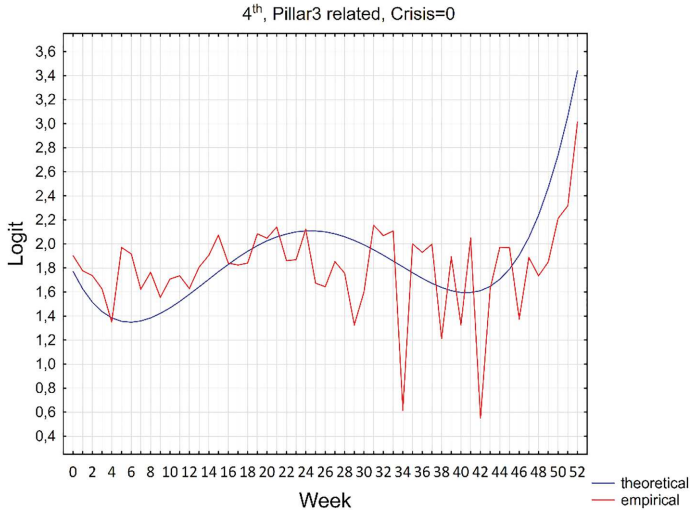



**Fig. 4.** Logit visualization of the fourth-degree polynomial model of the Pillar3 related category.

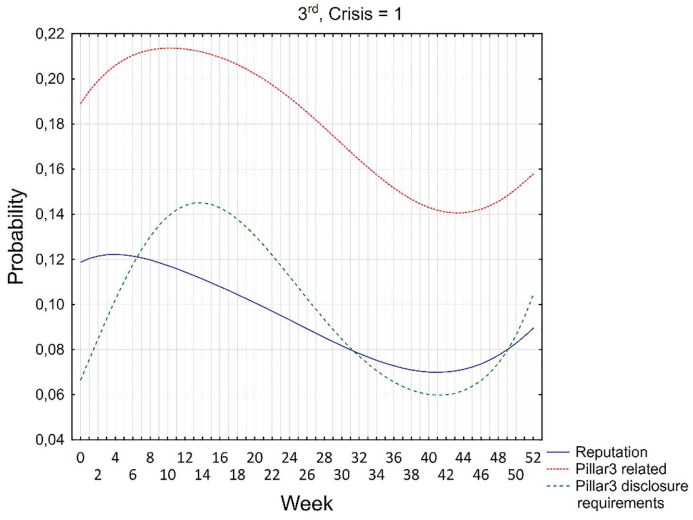



**Fig. 5.** Probability visualization of market discipline related categories during the years of the financial crisis.

## 5. Discussion and conclusion

The process of enhancement of market discipline has been implemented by Pillar 3 requirements since the introduction of the Basel II regulations, and it has been identified as its key objective. The Pillar 3 information disclosure document objective is to provide meaningful, comparable, and relevant disclosures. The document has been revised several times to achieve consolidated disclosures. It offers obligatory forms, flexible forms, and new dashboards of key metrics. Moreover, to achieve more consolidation, EBA guidelines on disclosure requirements have been published to enhance the consistency and comparability of institutions' disclosures. The guidelines aim to ensure market discipline, and they are based on the update of the Pillar 3 requirements by the Basel Committee in January 2015.

All of these ongoing changes suggest that the main goal of Pillar 3 information disclosures requires the delivery of relevant information to key market participants and consequent enhancement of market discipline mechanisms. Therefore, any analysis of the current structure and the content of disclosures is inevitable, and the design of the relevant disclosure models is important for delivering sufficient information to

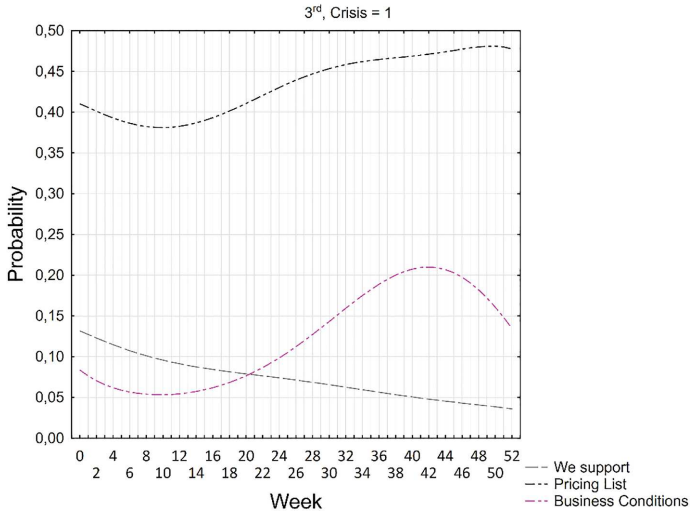



**Fig. 6.** Probability visualization of other categories during the years of the financial crisis.

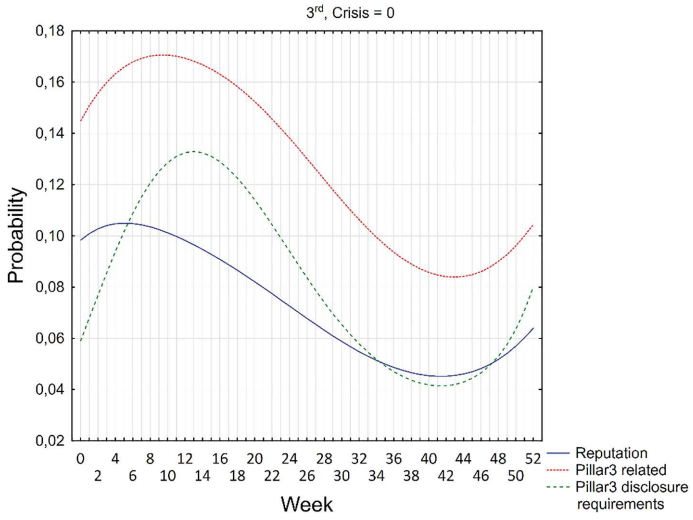



**Fig. 7.** Probability visualization of market discipline related categories during the years after the financial crisis.

stakeholders, and for their ability to assess bank risk-taking. Consequently, based on the analysis of the depositor's interest in disclosures after turbulent times and on the assessment of the current regulatory disclosure standards, it creates the opportunity to design a data structure model, which would help to achieve this meaningful goal and ensure relevant disclosures to avoid market discipline failure.

In this study, key findings have been identified. Firstly, in turbulent times depositors' interest was highest in the *Pricing List* web category (lowest in the first weeks of the year), which is followed by the *Pillar 3 related* web category, which was the highest in the first two quarters of the year. Peaks for the Pillar 3 categories (*Pillar 3 related, Pillar3 disclosure requirements*) were identified in the first weeks of the year. In the second half of the year, *Business Conditions* has the highest interest (lowest interest in the first half of the year). It is important to note that *Pillar 3 disclosure requirements* recorded a minor increase in the first quarter. After turbulent times, the highest interest is identified in the *Pricing List* web category. It was replaced by *Business Conditions* in only a few weeks in the third quarter. This category has been identified as having twice as high levels of interest in comparison to turbulent times. Lastly, the web categories *Reputation*, *Pillar 3 related,* and *Pillar 3*

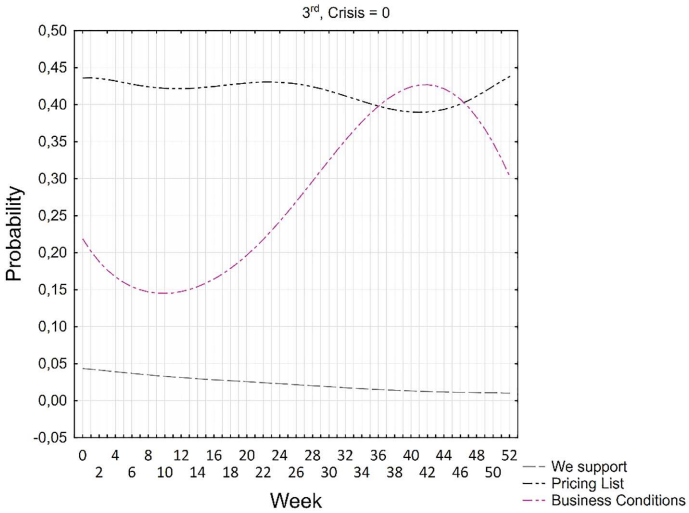



**Fig. 8.** Probability visualization of other categories during the years after the financial crisis.

*disclosure requirements* have displayed declining interest, which is in conjunction with Munk et al. (2017) (in turbulent times). Moreover, the behaviour of depositors' in these categories is also high at the beginning and the end of the year, and during the year, a decline in interest is identified. These findings also suggest that depositors do not show interest in *Pillar 3 related* and *Pillar 3 disclosure requirements* but instead in the current status of the commercial bank, especially in the *Business Conditions* and *Pricing List* categories.

Generally, these results are in line with De Araujo and Leyshon (2016) that conclude that stakeholders have a higher interest in other types of information than Pillar 3 disclosures or financial risk disclosures (Giner, Allini, & Zampella, 2020). It should be noted that our results related to the analysis of the interest of stakeholders in Pillar 3 disclosures during turbulent times are only in conjunction with our previous studies because the research volume within this topic is low.

These findings suggest that changes in commercial banking disclosures in favour of achieving Pillar 3 goals are inevitable, even after years of crisis. Consequently, changes in banks' disclosures are insufficient and we agree with Kuranchie-Pong et al. (2016) that disclosures should add value to key market participants to assess the risk profiles of banks to enhance effective market discipline, which is also in line with regulators' expectations. The removal of redundant and addition of useful information is essential to deliver relevant and sufficient disclosures to key market participants. This serves as a background for the identification of a disclosure model which is attractive and relevant for key market participants. The challenges for this topic are the low interest of depositors in officially requiring Pillar 3 disclosures, as the main issue is also the structure and identification of a relevant model based on disclosed information, which would enhance the interest of key market participants. Mainly, the existence of some limitations in this research encourages further study of this topic, which is inevitable.

## 6. Implications and limitations

The goal of the Pillar 3 regulation framework is the promotion of transparency and enhancement of public disclosures by financial institutions to reinforce market discipline. Therefore, the outcomes of the study, such as the identification of the nature of parts, content and specific categories of Pillar 3 disclosures, which are important to stakeholders, along with those types of information viewed as efficient are crucial to implementing effective market discipline as a supervisory goal. There are a few implications from our results. Firstly, the disclosures are most attractive at the beginning and the end of the year. During turbulent times, strategic issues like business conditions have the highest peak in attractiveness. The other implications are related to depositors' incentives to discipline commercial banks during turbulent times, which are lower for banks with good business conditions and pricing compared to riskier banks, because interest in Pillar 3 related disclosures is low. Moreover, this information is most important at the beginning and end of the year. During the year, there is an implication that their interest is concentrated on managerial actions and strategic issues, which covers current business conditions. These results can be firstly used by regulators in the process of designing new Pillar 3 regulations as they propose practical changes to the design of the disclosure framework. Secondly, commercial banks and financial institutions can use these outcomes when designing their Pillar 3 disclosures on their websites and provide voluntary information disclosures to attract their depositors. Accordingly, enhancement of market discipline through effective disclosures brings a range of benefits to market participants, such as stability on financial markets through potential avoidance of market discipline failure in turbulent times, and to society in general. Finally, we have identified the analysis of the categories of Pillar 3 disclosures in more detail and in other countries as topics for future research.

Our research has few limitations, mainly in the data analysis. There are limited options in obtaining historical data from banking and financial institutions. Storage of such data is difficult as web server records much information about each access to the web portal page. The institutions must have large data storage capabilities to preserve all the historical data, and that is why they often store only a fragment of historical data (e.g. only one year, month, etc.), based on the available data storage. Another limitation is also connected to working with historical data. In this case, it is about the structure of the web portal. As time passes, the web portal evolves with the needs of both visitors and administrators. During data pre-processing phases, various methods require a site map. Because the web portal has changed over time on

several occasions, it is rather difficult to obtain an accurate and complete site map of a web portal. It is possible to obtain an incomplete site map from the log file, but still, it is an issue that can result in generating inaccurate data. The last limitation is connected to the evaluation of the data analysis results as standard evaluation methods could not be used. This was because the examined variable has many categories (web pages). Novel evaluation methods were designed to evaluate the model across various levels (counts of accesses, probabilities, and logits).

## CRediT authorship contribution statement

**Michal Munk:** Conceptualization, Methodology, Formal analysis, Writing - review & editing. **Anna Pilkova:** Conceptualization, Writing - review & editing, Supervision. **Lubomir Benko:** Methodology, Investigation, Writing - original draft. **Petra Blazekova:** Conceptualization, Writing - original draft. **Peter Svec:** Data curation, Resources, Writing - original draft.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0776/18 and VEGA-1/0821/21.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2021.115503.

## References

Andrievskaya, I., & Raschupkin, M. (2015). Is it Worth Being Transparent? Evidence from the Russian Banking System. In Higher School of Economics Research Paper No. WP BRP 51/FE/2015, 28. 10.2139/ssrn.2700621.

Andrievskaya, I., & Semenova, M. (2016). Does banking system transparency enhance bank competition? Cross-country evidence. *Journal of Financial Stability, 23*, 33–50. https://doi.org/10.1016/j.jfs.2016.01.003

Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications, 8*(11), 7–10.

Antzoulatos, A. A., & Karanastasis, D. (2016). Expected government support of banks and market discipline by shareholders. *SSRN Electronic Journal, 31*. https://doi.org/10.2139/ssrn.2719781

Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M., & Perona, I. (2013). Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. *Expert Systems with Applications, 40*(18), 7478–7491. https://doi.org/10.1016/j.eswa.2013.07.040

Arnold, E. A., Größl, I., & Koziol, P. (2015). Market discipline across bank governance models: Empirical evidence from German depositors. Discussion Papers. Retrieved from <https://ideas.repec.org/p/zbw/bubdps/132015.html>.

BCBS - Basel Committee on Banking Supervision. (2015). Standards for Revised Pillar 3. Disclosure Requirements. [online], [cited 1 April 2021]. Available from Internet: https://www.bis.org/bcbs/publ/d309.pdf.

BCBS - Basel Committee on Banking Supervision. (2016). Consultative Document Pillar 3. Disclosure requirements – consolidated and enhanced framework. [online], [cited 1 April 2021]. Available from Internet: http://www.bis.org/bcbs/publ/d356.pdf.

BCBS - Basel Committee on Banking Supervision. (2018). The Basel Committee consults on revisions to the Pillar 3 dislocsure framework. [online], [cited 1 April 2021]. Available from Internet: https://www.bis.org/PRESS/P180227.HTM.

Bhawsar, S., Pathak, K., & Patidar, V. (2012). New framework for web access prediction. *International Journal of Computer Technology and Electronics Engineering, 2*(1), 48–53.

Bliss, R. R., & Flannery, M. J. (2002). Market DISCIPLINE in the Governance of U.S. Bank holding Companies: Monitoring vs influencing. *Review of Finance, 6*(3), 361–396. https://doi.org/10.1023/A:1022021430852

Bouaiss, K., Refait-Alexandre, C., & Alexandre, H. (2017). Will Bank Transparency really Help Financial Markets and Regulators? Retrieved from <https://hal.archives-ouvertes.fr/hal-01637917>.

Bourgain, A., Pieretti, P., & Zanaj, S. (2012). Financial openness, disclosure and bank risk-taking in MENA countries. *Emerging Markets Review, 13*(3), 283–300. https://doi.org/10.1016/J.EMEMAR.2012.01.002

Bouvard, M., Chaigneau, P., & de Motta, A. (2015). Transparency in the financial system: Rollover risk and crises. *The Journal of Finance, 70*(4), 1805–1837. https://doi.org/10.1111/jofi.12270

Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., & García, S. (2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. *Expert Systems with Applications, 39*(12), 11243–11249. https://doi.org/10.1016/j.eswa.2012.03.046

Dabrowska-Zielinska, K., Kogan, F., Ciolkosz, A., Gruszczynska, M., & Kowalik, W. (2002). Modelling of crop growth conditions and crop yield in Poland using AVHRR-based indices. *International Journal of Remote Sensing, 23*(6), 1109–1123. https://doi.org/10.1080/01431160110070744

De Araujo, P., & Leyshon, K. I. (2016). The impact of international information disclosure requirements on market discipline. *Applied Economics, 49*(10), 954–971. https://doi.org/10.1080/00036846.2016.1208361

EBA. (2016). Guidelines on disclosure requirements under Part Eight of Regulation (EU). [cited 1 April 2021]. Available from Internet: https://eba.europa.eu/regulation-and-policy/transparency-and-pillar-3/guidelines-on-disclosure-requirements-under-part-eight-of-regulation-eu-.

EBA. (2017). Guidelines on disclosure requirements under Part Eight of Regulation (EU) Updated version[cited 1 April 2021]. Available from Internet: https://eba.europa.eu/sites/default/documents/files/documents/10180/1918833/8daeb580-5f64-418e-bf10-.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417.

Fonseca, A. R., & González, F. (2010). How bank capital buffers vary across countries: The influence of cost of deposits, market power and bank regulation. *Journal of Banking & Finance, 34*(4), 892–902. https://doi.org/10.1016/J.JBANKFIN.2009.09.020

Fueda, I., & Konishi, M. (2007). Depositors' response to deposit insurance reforms: Evidence from Japan, 1990–2005. *Journal of Financial Services Research, 31*(2–3), 101–122. https://doi.org/10.1007/s10693-007-0010-7

Garten, H. A. (1986). Banking on the market: Relying on depositors to control bank risks. *Yale Journal on Regulation, 4*, 129–172.

Giner, B., Allini, A., & Zampella, A. (2020). The value relevance of risk disclosure: An analysis of the banking sector. *Accounting in Europe, 17*(2), 129–157. https://doi.org/10.1080/17449480.2020.173092

Gandrud, C., & Hallerberg, M. (2014). Supervisory transparency in the European banking union. In Bruegel Policy Contribution, (2014/01). Retrieved from <https://www.econstor.eu/handle/10419/106314>.

Goldstein, I., & Leitner, Y. (2015). Stress tests and information disclosure. No 15-10, Working Papers. Federal Reserve Bank of Philadelphia. Retrieved from <https://econpapers.repec.org/paper/fipfedpwp/15-10.htm>.

Goldstein, I., & Sapra, H. (2014). Should banks' stress test results be disclosed? An analysis of the costs and benefits. *Foundations and Trends® Finance, 8*(1), 1–54. https://doi.org/10.1561/0500000038

Guillemin, F., & Semenova, M. (2018). Transparency and Market Discipline: Evidence from the Russian Interbank Market. In Higher School of Economics Research Paper No. WP BRP 67/FE/2018, 32. https://doi.org/10.2139/ssrn.3225061.

Guillemin, F. (2017). Disclosure, ambiguity and depositors' discipline in European banking system. *SSRN Electronic Journal, 27*. https://doi.org/10.2139/ssrn.2845194

Hamid, F. S., & Yunus, N. M. (2017). Market discipline and bank risk taking: Evidence from the East Asian Banking Sector. *East Asian Economic Review, 21*(1), 29–57. https://doi.org/10.2139/ssrn.2945792

Hori, M., Ito, Y., & Murata, K. (2009). Do depositors respond rationally to bank risks? Evidence from Japanese banks during crises. *Pacific Economic Review, 14*(5), 581–592. https://doi.org/10.1111/j.1468-0106.2009.00470.x

Iren, P., Reichert, A. K., & Gramlich, D. (2014). Information disclosure, bank performance and bank stability. *International Journal Banking, Accounting and Finance, 5*(4), 39.

Kaffenberger, B. (2015). Depositor market discipline in the EMU - The effect of the cypriot bail-in on European peripheral countries. *SSRN Electronic Journal, 38*. https://doi.org/10.2139/ssrn.2698773

Kapusta, J., Munk, M., Svec, P., & Pilkova, A. (2014). Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal. In *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2014.05.163

Kapusta, J., Pilková, A., Munk, M., & Švec, P. (2013). Data pre-processing for web log mining: Case study of commercial bank website usage analysis. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 61*(4), 973–979.

Kozłowski, Ł. (2016). Cooperative banks, the internet and market discipline. *Journal of Co-Operative Organization and Management, 4*(2), 76–84. https://doi.org/10.13140/RG.2.1.3768.6809

Kuranchie-Pong, L., Bokpin, G. A., & Andoh, C. (2016). Empirical evidence on disclosure and risk-taking of banks in Ghana. *Journal of Financial Regulation and Compliance, 24*(2), 197–212. https://doi.org/10.1108/JFRC-05-2015-0025

Li, H., Liu, H., Siganos, A., & Zhou, M. (2016). Bank regulation, financial crisis, and the announcement effects of seasoned equity offerings of US commercial banks. *Journal of Financial Stability, 25*, 37–46. https://doi.org/10.1016/j.jfs.2016.06.007

Li, S., & Wang, S. (2014). A financial early warning logit model and its efficiency verification approach. *Knowledge-Based Systems, 70*, 78–87. https://doi.org/10.1016/j.knosys.2014.03.017

Makkar, P., Gulati, P., & Sharma, A. (2010). A novel approach for predicting user behavior for improving web performance. *International Journal on Computer Science and Engineering, 2*(4), 1233–1236.

Martinez Peria, M. S., & Schmukler, S. L. (2001). Do depositors punish banks for bad behavior? Market discipline, deposit insurance, and banking crises. *The Journal of Finance, 56*(3), 1029–1051. https://doi.org/10.1111/0022-1082.00354

Min, D. (2015). Understanding failures the failures of market discipline. *Washington University Law Review, 92*(6), 1421–1501.

Moreno, D., & Takalo, T. (2016). Optimal bank transparency. *Journal of Money, Credit and Banking, 48*(1), 203–231. https://doi.org/10.1111/jmcb.12295

Munk, M., Pilková, A., Drlík, M., Kapusta, J., & Švec, P. (2012). Verification of the fulfilment of the purposes of basel ii, pillar 3 through application of the web log mining methods. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 60*(2), 217–222.

Munk, M., Drlík, M., & Vrabelova, M. (2011). Probability modelling of accesses to the course activities in the web-based educational system. In *Computational Science And Its Applications - Iccsa 2011, Pt V* (pp. 485–499).

Munk, M., Benko, Ľ., Gangur, M., & Turčáni, M. (2015). Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management, 18*(3), 144–159.

Munk, M., & Drlk, M. (2014). Analysis of stakeholders' behaviour depending on time in virtual learning environment. *Applied Mathematics and Information Sciences, 8*(2), 773–785.

Munk, M., Kapusta, J., & Švec, P. (2010). Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. *Procedia Computer Science, 1*(1), 2273–2280. https://doi.org/10.1016/j.procs.2010.04.255

Munk, M., Kapusta, J., Švec, P., & Turčáni, M. (2010). Data advance preparation factors affecting results of sequence rule analysis in web log mining. *E+M Ekonomie a Management, 13*(4), 143–160.

Munk, M., Pilkova, A., Benko, L., & Blaželková, P. (2017). Pillar 3: Market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management, 18*(5), 954–973. https://doi.org/10.3846/16111699.2017.1360388

Nagarajan, Sri, & Sealey, C. William (1997). Market discipline, moral hazard and bank regulation. *Federal Reserve Bank of Chicago*.

Nahar, S., Azim, M., & Anne Jubb, C. (2016). Risk disclosure, cost of capital and bank performance. *International Journal of Accounting & Information Management, 24*(4), 476–494. https://doi.org/10.1108/IJAIM-02-2016-0016

Nier, E., & Baumann, U. (2006). Market discipline, disclosure and moral hazard in banking. *Journal of Financial Intermediation, 15*(3), 332–361. https://doi.org/10.1016/j.jfi.2006.03.001

Parwada, J. T., Lau, K., & Ruenzi, S. (2015). The Impact of Pillar 3 Disclosures on Asymmetric Information and Liquidity in Bank Stocks: Multi-Country Evidence. CIFR Paper No. 82/2015, 27. https://doi.org/10.2139/ssrn.2670403.

Raffi, F., Katlama, C., Saag, M., Wilkinson, M., Chung, J., Smiley, L., & Salgo, M. (2006). Week-12 response to therapy as a predictor of week 24, 48, and 96 outcome in patients receiving the HIV Fusion inhibitor enfuvirtide in the T-20 versus optimized regimen only (TORO) trials. *Clinical Infectious Diseases, 42*(6), 870–877. https://doi.org/10.1086/500206

Semenova, Maria (2012). Market discipline and banking system transparency: Do we need more information? *Journal of banking regulation, 13*(3), 241–248.

Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2019). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2019.08.015

Verdhen, A., Chahar, B. R., & Sharma, O. P. (2014). Snowmelt modelling approaches in watershed models: Computation and comparison of efficiencies under varying climatic conditions. *Water Resources Management, 28*(11), 3439–3453. https://doi.org/10.1007/s11269-014-0662-7

Vojíř, S., Zeman, V., Kuchař, J., & Kliegr, T. (2018). EasyMiner.eu: Web framework for interpretable machine learning based on rules and frequent itemsets. *Knowledge-Based Systems, 150*, 111–115. https://doi.org/10.1016/j.knosys.2018.03.006

Wu, Y., & Bowe, M. (2012). Information disclosure and depositor discipline in the Chinese banking sector. *Journal of International Financial Markets, Institutions and Money, 22*(4), 855–878. https://doi.org/10.1016/j.intfin.2012.05.004

Yin, P. Y., & Guo, Y. M. (2013). Optimization of multi-criteria website structure based on enhanced tabu search and web usage mining. *Applied Mathematics and Computation, 219*(24), 11082–11095. https://doi.org/10.1016/j.amc.2013.05.033

Zer, I. (2015). Information disclosures, default risk, and bank value. *Finance and Economics Discussion Series, 2015*(104), 1–43. https://doi.org/10.17016/FEDS.2015.104

PRÍLOHA B: PILKOVÁ, ANNA, MICHAL MUNK, ĽUBOMÍR BENKO, PETRA BLAŽEKOVÁ A JOZEF KAPUSTA, 2021A. PILLAR 3: DOES BANKING REGULATION SUPPORT STAKEHOLDERS' INTEREST IN BANKS FINANCIAL AND RISK PROFILE? *PLOS ONE*. 16(10), E0258449. DOI:10.1371/JOURNAL.PONE.0258449 (**WEB OF SCIENCE, 2021IF: 3.752, Q2**) [WOS: 0, SCOPUS: 0]

PLOS ONE

RESEARCH ARTICLE

# Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile?

**Anna Pilková[1], Michal Munk[2,3], Ľubomír Benko[2]*, Petra Blažeková[1], Jozef Kapusta[2,4]**

**1** Comenius University in Bratislava, Bratislava, Slovakia, **2** Constantine the Philosopher University in Nitra, Nitra, Slovakia, **3** University of Pardubice, Pardubice, Czech Republic, **4** Pedagogical University of Cracow, Kraków, Poland

* lbenko@ukf.sk

Check for updates

OPEN ACCESS

**Citation:** Pilková A, Munk M, Benko Ľ, Blažeková P, Kapusta J (2021) Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? PLoS ONE 16(10): e0258449. https://doi.org/10.1371/journal.pone.0258449

**Editor:** László Vasa, Szechenyi Istvan University: Szechenyi Istvan Egyetem, HUNGARY

**Received:** June 18, 2021

**Accepted:** September 27, 2021

**Published:** October 27, 2021

**Copyright:** © 2021 Pilková et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data underlying this study are available on Mendeley (DOI: 10.17632/8tkxfrdcc9.1).

**Funding:** This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0821/21 (MM), also by the scientific research project of the Czech Sciences Foundation Grant No. 19-15498S (MM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

The paper examines the interest of the commercial banks' stakeholders in Pillar 3 disclosures and their behaviour during the timing of serious market turbulence. The aim is to discover to which extent current banking regulation supports stakeholders' interest in the information required by regulators to be disclosed. The examined data consists of log files that were pre-processed using web mining techniques and from which were extracted frequent item sets by quarters and evaluated in terms of quantity. The authors have proposed a methodology to evaluate frequent item sets of web parts over a dedicated time. Based on the verification of applied methodology on two commercial banks, the results show that stakeholders' interest in disclosures is highest in the first quarter at each year and after turbulent times in 2009 their interests decreased. Moreover, the results suggest that stakeholders expressed higher interest than in regulatory required Pillar 3 information in the following group of information: Pillar3 related information, Annual reports, Information on Group. Following our results, the paper contributes to cover the gap in the research by analysing Pillar 3 disclosures and their compliance with regulatory requirements, which also increase the interest of the relevant stakeholders to conduce them as an effective market discipline tool.

## Introduction

In the recent post-financial crisis years, significant changes have occurred in financial markets regulation and supervision. The financial institutions face challenges as reduction of risk-taking in line with tightening of regulatory requirements and the increase of regulatory directives and guidelines published by the European authorities. These directives cover primarily risk areas related to capital, liquidity, credit, counterparty exposures, market, operational and securitisation. They aim to eliminate systemic risk through supervision and more importantly to reinforce market discipline [1].

Market discipline as one of the three pillars of the stable financial market (the other two are regulation and supervision) has been in regulatory interest since 2001. However, the issuance of Basel II and Basel III regulations has built in market discipline into the regulatory

**Competing interests:** The authors have declared that no competing interests exist.

framework through Pillar 3 standards. The process of final Pillar 3 tuning is still ongoing, and its main focus is focused on the methodological changes, feedback from related parties and broad discussions aimed to reach enhancement and reinforcement of the market discipline [2,3].

Topics of Pillar 3 and disclosures have also attracted researchers. They have focused mainly on weaknesses of the disclosures (mainly huge costs, ineffective implementation) and less on benefits [4–6]. Research on stakeholders' interest and comments towards Pillar 3 disclosures is very limited. The situation is even worse in commercial banks operating in Central and Eastern European countries where their owners are big international financial groups located in developed countries. Evidence is in the paper Munk et al. [7] according to which the relevance of the published Pillar 3 information is not in particular interest of stakeholders, especially information related to mandatory Pillar 3 disclosures. All in all, both from regulatory status and research focus is evident that Pillar 3 framework as a supervisory market discipline tool still does not fulfil its main goal to bring adequate information to key market participants to effectively enhance market discipline. Therefore, consolidation of further research, theory and regulation is crucial for its reinforcement. Following these results our research focuses on studying stakeholders' behaviour and how and to which extent stakeholders of the commercial banks use Pillar 3 disclosed information.

### Research objective and assumptions

The paper aims to examine the website data dedicated to Pillar 3 disclosures of commercial banks operating in CEE country to study the behaviour of stakeholders concerning the timing of serious market turbulence.

The article summarizes the results of previous research focused on stakeholders' interest in Pillar 3 information during the years 2009–2015. In the current research, we focus on verifying our latest findings, especially in connection with verifying the continuation of the established trend characterized by the low interest of stakeholders in Pillar 3 information.

We have made the following assumptions:

We assume the continuation of the established trend after the gradual fading of the consequences of the financial crisis from 2012–2015, which is characterized by a lack of interest in Pillar 3 information.

We assume that the behaviour of stakeholders to the published Pillar 3 information will no longer be reflected in any time trend or seasonality.

We will verify the assumptions on data of the accesses of the website of another equally important bank operating in Slovakia during the years 2016–2018.

The rest of the paper is structured as follows. The next section describes the related work in the field of market discipline and Pillar 3 disclosures. The third section summarizes the applied methodology to obtain the Results. The fourth section is focused on the results of the web usage analysis and interpretation of the results. Subsequently, the last section provides the conclusion and contribution of the research to improve Pillar 3 regulatory aims and increase stakeholders' interests in Pillar 3 regulatory information.

## Literature review

Pillar 3 represents the regulators' tool for market discipline enhancement by meaningful, relevant, and transparent disclosures. In the last two decades, the regulators' goal is to conduce them as an effective market discipline tool, which fulfils regulators' expectations in the areas of standardization, consistency, comparability, and transparency. To bring the stability of the financial markets, the coherence among market discipline, regulatory requirements and

supervision is required, as this consistency reduces the potential for regulatory arbitrage. However, there are some depressive effects of implementing regulation, such as reduction in competition, innovation, and profitability and increase in regulatory change costs [8]. On the contrary, these extensive reporting requirements imposed after the global financial crisis support a more efficient, stable, inclusive financial system and are safer than before the financial crisis [9,10]. Based on our reviewed studies, which assess compliance of the Pillar 3 disclosures with regulatory requirements we obtained the following findings. First of all, there is a lack of standardization of the disclosures, which is important due to its impact on the banks' performance indicators. Disclosures were found general, providing no further useful information, infrequent, annual disclosures were repetitive and symbolic [11], there is an imbalance among some disclosure categories and lack of standardization due to data scarcity [12]. Furthermore, disclosures in their prescriptive nature do not provide an adequate risk picture as they are predominantly backwards-looking. The delivery of the understanding and relevant information to stakeholders is impeded by a dynamic of the risk events [13]. Thus, we agree with the authors that there is an opportunity for directors as stakeholders to decide the provision of additional information to supplement Pillar 3 disclosures and support the market discipline mechanism. Moreover, Savvides and Savvidou [14] point out that the level of harmonization is far from regulators' intentions because the banks do not disclose risk information in a consistent, easily accessible, or usable manner. Supervisors need to advocate informative risk disclosures, which should be reviewed and updated regularly.

Moreover, the level of the quality of the risk disclosures influences also Pillar 3 content efficiency. Additionally, the overall quality of risk disclosures is poor based on Kabir and Sobhani's study [15], which analysed the level of risk disclosures of the 46 banks between 2009 and 2013. This is in conjunction with Barakat and Hussainey's study [16], which concludes that the quality of disclosures also depends on the ownership structure of the bank. These authors propose to achieve better risk reporting quality through various channels: more active audit committee, lower executive ownership, operating under regulations promoting bank competition. Moreover, Khalil and Alam [17] propose an enhancement of market discipline by the introduction of the independent risk disclosure council (composing of representatives of regulators, banks, investors). The council could determine the minimum standard for risk disclosure and the utilization of the information disclosed in the periodic accounts and annual reports, which also reduces the cost of supervision.

Importantly, a balance in transparency should be held as the aggressive decrease in the risk-taking by increasing disclosures can hinder economic growth. This negative effect of market discipline implementation should be closely monitored, as a market discipline plays a vital part in risk-taking behaviour of commercial banks [18]. The role of market discipline is important in evading financial crisis and has come out even stronger in this endeavour because information asymmetry led to moral hazard in commercial banks. Accordingly, there is an assumption that banks do not regard the content of the disclosures in turbulent times. The results of Kabir and Sobhani's [15] study (period from 2009 to 2013) revealed mimicking tendency of the banks to disclose in the same standard as competitors, in a conventional prescribed format of reporting risk with no explanation of significant risk factors.

Substantially, the beneficial effects of Pillar 3 disclosures are supported (revealed) by loads of research studies. Pillar 3 improves the safety of the banking system [19], offers banks to raise cheaper capital [20], decreases information asymmetry [21–23], quarterly reporting is useful to investors [22] and improves the ex-ante risk sharing provided by financial intermediation [24]. Additionally, the increased risk disclosures brought liquidity benefits when Pillar 3 became effective and its compliance is enforced by the banking regulator [25]. The level of supervisory power positively influences the disclosure of Pillar 3 requirements, specifically on

loan loss provisions [26]. Moreover, in the assessment of efficiency of Pillar 3 disclosures the interest of key parties in the content of the disclosures is important. In a few studies, the authors highlight these categories with high content relevancy for investors, which are: credit risk and liquidity risk [27], market risk [28]. De Araujo and Leyshon's study [29] reveals a window in content relevancy of disclosures as depositors and creditors are most responsive to information such as the bank's assets, off-balance sheet items, and ratings for other banking activities. Stakeholders value the quantitative information more than the qualitative one and can positively influence bank risk-taking [30]. Additionally, website disclosures are a timely disclosure medium and rich form of communication available to a broad range of stakeholder groups [31]. Therefore, the sensitive response of the stakeholders to the negative disclosures can also trigger runs on an inefficient bank [32] and can lower the distance to default [33]. Furthermore, Oliveira et al. [34] found that stakeholder perceptions in case of manipulation of risk disclosures may be influenced by the reputational benefit of the bank's management. While investors and customers welcome increased disclosure, care must be taken to ensure that adequate value is being derived from the disclosed information.

In addition, there are rare studies assessing Pillar 3 information disclosures in CEE countries from stakeholders' perspectives and usage information by them. This type of research is crucial for effective supervisory market discipline implementation. There are arising issues about the specific factors concerning Pillar 3 disclosures, such as content relevancy, timing, and the compliance of the Pillar 3 disclosures with regulatory requirements to serve as an effective market discipline tool [35–37]. Additionally, according to a few studies that analyse the interest of the stakeholders in Pillar 3 disclosures, the most visited parts of the disclosures were not solo Pillar 3 information required by regulators but information on Group, which owns the bank and Pillar 3 information together with Annual reports [7] or Emitent Prospects [38].

The study contributes to cover the research gap by deep analysis of the interest of the stakeholders in Pillar 3 disclosures in relation to market turbulences based on website data of commercial banks in CEE countries. The study is also dedicated to analyzing Pillar 3 disclosures' compliance with regulatory requirements, which also increase the interest of the relevant stakeholders to serve as an effective market discipline tool.

## Materials and methods

The used methodology is based on the procedures and results of the research project VEGA 1/0776/18 (Optimizing the content and structure of Pillar 3 disclosures based on modelling their use by commercial bank's stakeholders), where particular attention is on web usage mining. Since Basel regulations are complex systems, detecting the problems of their lack of efficiency requires the most accurate analysis and application of non-traditional approaches such as web mining. This research project serves as a basis for the methodology of this study, and we try to identify potential problems in the usage of bank portals by analysing behaviour patterns of web users following the required information from Pillar 3.

### Methodology

The research methodology was inspired by Munk et al. [39] and it was done the same way as in Munk et al. [7] where it was used to evaluate the frequent item sets in terms of quantity. In this web usage analysis were data related to Pillar 3 obtained from the web server log files of domestic significant commercial banks operating in Slovakia. In comparison to the previous research, another bank was used with the idea to validate the results of previous research of Munk et al. [7]. The webserver log files keep information about visitors, which can be used for the analysis of visitor's behaviour. Data pre-processing of their usage consists of data cleaning,


**Fig 1. The approach used to process the log files.**

https://doi.org/10.1371/journal.pone.0258449.g001

integration, transformation, session identification, path completion and data reduction [40,41]. Data preparation was done based on Munk et al. [39] where it was needed to pre-process log files from multiple servers that are used as load balancers and the used methods were inspired by Pamutha et al. [42] and Spiliopoulou et al. [43]. The web usage analysis was realised on a sample of log accesses which were obtained after the data pre-processing included data cleaning, session identification and path completion. The applied methodology is depicted in Fig 1 and is based on Munk et al. [7] research.

### Dataset composition

This article is based on the results of the research by Munk et al. [7] that dealt with the analysis of website data dedicated to Pillar 3 disclosures of commercial banks. Therefore, two web server log files of domestic significant commercial banks operating in Slovakia were used. The first log file used in the research of Munk et al. [7] contained more than 10 000 000 log accesses that were obtained after data pre-processing. These accesses represent the five years 2009–2015 where the year 2009 represents the year of the financial crisis. On the other hand, the years 2010–2015 represent the years after the financial crisis. The second log file used in this research comes from a different web server. The log file consisted of more than 35 000 000 log accesses that were obtained after data pre-processing. The examined period was of years 2016–2018. During the data pre-processing phase were created various variables. This way time variables were created for both log files that represented the quarter of the specific year. Three variables were added to the dataset—Quartal, Year and YearQuartal which represents the number of quartal of the year. Also, a variable Category was created to unite the accessed the web parts of web portal into web categories. The composition of both created datasets from the log files are described in Table 1 that contains information about the analysed web categories. The analysed banks provided different levels of content based on the website taxonomy of the surveyed banks (Table 1). Although the web categories of the surveyed banks are at different levels of detail, in the case of both banks we clearly identified web parts with Pillar 3 content (in accordance with the regulator's requirements). This allowed us to monitor the time trend and seasonality in stakeholder behaviour in relation to Pillar 3 information.

## Results

The experiment analysed website data dedicated to Pillar 3 disclosures of two commercial banks operating in Slovakia and studied the behaviour of stakeholders in relation to the timing

**Table 1. Website taxonomy.**

| | | |
|---|---|---|
| **First analysed bank 2009–2015** | | |
| /About bank/ | /Pillar 3 disclosure requirements/ | /Pillar3 Q-terly Info/ |
| /About bank/ | /Pillar 3 disclosure requirements/ | /Pillar3 Semiannualy Info/ |
| /About bank/ | /Pillar 3 related/ | /Rating/ |
| /About bank/ | /Pillar 3 related/ | /Annual Reports/ |
| /About bank/ | /Pillar 3 related/ | /Group/ |
| /About bank/ | /Pillar 3 related/ | /Information for Banks/ |
| /About bank/ | /Pillar 3 related/ | /Emitent Prospects/ |
| /About bank/ | /Pillar 3 related/ | /General Shareholder Meeting/ |
| /About bank/ | /Pillar 3 related/ | /Financial Reports/ |
| /About bank/ | /Other/ | |
| **Second analysed bank 2016–2018** | | |
| /Pillar 3/ | /Pillar 3 disclosure requirements/ | |
| /Pillar 3/ | /Pillar 3 related/ | |
| /Other/ | /Products and services for customers/ | |
| /Other/ | /Information service/ | |
| /Other/ | /About us/ | |
| /Other/ | /Press centre/ | |
| /Other/ | /Social responsibility/ | |
| /Other/ | /Documents/ | |

https://doi.org/10.1371/journal.pone.0258449.t001

of serious market turbulence. The aim is to find out whether current banking regulation supports stakeholders' interest in the information required by the regulator to be disclosed and confirm the research assumptions. The first subsection is a summary of the previous research done by Munk et al. [7]. A log file from the years 2009–2015 was analysed and based on the results were established the hypotheses for further research. The second subsection is focused on website data collected during the years 2016–2018. The data is dedicated to Pillar 3 disclosures of the second commercial bank operating in Slovakia and examined was the behaviour of stakeholders to their interest in Pillar 3 disclosures.

## Analysis of the stakeholders' behaviour and their interest in pillar 3 disclosure during turbulent times (2009–2015) in the Slovak commercial bank

Munk et al. [7] analysed accesses to the website of a commercial bank operating in Slovakia that was focused on Pillar 3 disclosures. The authors recommend further changes in commercial bank's information disclosure for obtaining an effective market discipline mechanism. The results served as a stepping-stone for our further research presented in the following subsection. For that reason, the most important conclusions were described in this subsection. The web portal was divided into the following parts, which were subsequently analysed: */Group/*, */Annual Reports/*, */Rating/*, */Emitent Prospects/*, */General Shareholder Meeting/*, */Financial Reports/*, */Information for Banks/*, */Pillar 3 Semiannually Information/*, */Pillar 3 Quarterly Information/*. The analysis was done on five-year data 2009–2015 and key findings are summarized according to the years below.

In the first quarter of 2009 (Table 2), the most visited were web parts */Group/*, */Pillar3 Q-terly Info/*, */Rating/*, */Annual Reports/*, */Information for Banks/*, */Pillar3 Semiannualy Info/*, */Emitent Prospects/*. The pairs identified (Table 3) with high interest were (*/Group/*, */Pillar3 Q-terly Info/*), (*/Annual Reports/*, */Pillar3 Q-terly Info/*) and others (Table 3). Positive correlation was identified for (*/Emitent Prospects/*, */Pillar3 Semiannualy Info/*). In the first quarter of 2010 (Table 2), the web parts */Group/*, */Pillar3 Q-terly Info/*, */Annual Reports/*, */Rating/* were the most visited parts. The most visited pair (Table 3) of the web parts was (*/Pillar3 Q-terly Info/*, */Annual Reports/*). Web parts pairs (*/Group/*, */Rating/*) and (*/Pillar3 Q-terly Info/*, */Rating/*) were pairs with less interest. The highest degree of positive correlation was identified for (*/General Shareholder Meeting/*, */Pillar3 Q-terly Info/*), (*/Pillar3 Semiannualy Info/*, */Pillar3 Q-terly Info/*) and (*/Pillar3 Q-terly Info/*, */Annual Reports/*).

**Table 2. Support: Web part traffic rate.**

| ***min support = 1%*** | **09Q1** | **10Q1** | **11Q1** | **12Q1** | **13Q1** | **14Q1** | **15Q1** |
|---|---|---|---|---|---|---|---|
| */Group/* | 58.48 | 38.89 | 50.00 | 45.83 | 48.08 | 45.76 | 46.23 |
| */Pillar3 Q-terly Info/* | 36.77 | 37.04 | 28.57 | 20.83 | 25.00 | 24.26 | 23.68 |
| */Rating/* | 30.12 | 22.22 | 17.86 | 16.67 | 17.31 | 16.32 | 16.98 |
| */Annual Reports/* | 25.87 | 31.48 | 21.43 | 16.67 | 19.23 | 17.43 | 15.99 |
| */Information for Banks/* | 23.89 | 11.11 | 7.14 | < 1 | 3.85 | < 1 | < 1 |
| */Pillar3 Semiannualy Info/* | 17.42 | 3.70 | < 1 | 8.33 | 3.85 | < 1 | < 1 |
| */Emitent Prospects/* | 15.15 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| */General Shareholder Meeting/* | < 1 | 1.85 | < 1 | 12.50 | 5.77 | 5.64 | < 1 |
| */Financial Reports/* | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |

Note: Marked supports are > 15%.

https://doi.org/10.1371/journal.pone.0258449.t002

**Table 3. Support: Web part pair traffic rate.**

| ***min support* = 1%** | **09Q1** | **10Q1** | **11Q1** | **12Q1** | **13Q1** | **14Q1** | **15Q1** |
|---|---|---|---|---|---|---|---|
| (*/Group/*, */Pillar3 Q-terly Info/*) | 18.90 | 3.70 | < 1 | 8.33 | 3.85 | < 1 | < 1 |
| (*/Annual Reports/*, */Pillar3 Q-terly Info/*) | 18.80 | 22.22 | 17.86 | 8.33 | 13.46 | 13.78 | 14.38 |
| (*/Rating/*, */Group/*) | 16.66 | 7.41 | 3.57 | < 1 | < 1 | < 1 | < 1 |
| (*/Pillar3 Q-terly Info/*, */Pillar3 Semiannualy Info/*) | 16.32 | 3.70 | < 1 | 4.17 | < 1 | < 1 | < 1 |
| (*/Rating/*, */Information for Banks/*) | 16.27 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Rating/*, */Annual Reports/*) | 15.93 | 1.85 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Group/*, */Information for Banks/*) | 15.78 | < 1 | 3.57 | < 1 | < 1 | < 1 | < 1 |
| (*/Rating/*, */Pillar3 Q-terly Info/*) | 15.54 | 5.56 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Annual Reports/*, */Information for Banks/*) | 15.00 | 1.85 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Annual Reports/*, */Group/*) | 14.59 | 3.70 | < 1 | 4.17 | < 1 | < 1 | < 1 |
| (*/Pillar3 Q-terly Info/*, */Emitent Prospects/*) | 14.57 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Pillar3 Q-terly Info/*, */Information for Banks/*) | 14.18 | 1.85 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Annual Reports/*, */Pillar3 Semiannualy Info/*) | 14.05 | 1.85 | < 1 | 4.17 | < 1 | < 1 | < 1 |
| (*/Annual Reports/*, */Emitent Prospects/*) | 13.64 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Group/*, */Pillar3 Semiannualy Info/*) | 13.62 | < 1 | < 1 | 4.17 | < 1 | < 1 | < 1 |
| (*/Group/*, */Emitent Prospects/*) | 13.56 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Rating/*, */Pillar3 Semiannualy Info/*) | 13.45 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Rating/*, */Emitent Prospects/*) | 13.36 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Emitent Prospects/*, */Pillar3 Semiannualy Info/*) | 13.24 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Information for Banks/*, */Emitent Prospects/*) | 13.23 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Information for Banks/*, */Pillar3 Semiannualy Info/*) | 12.98 | < 1 | < 1 | < 1 | < 1 | < 1 | < 1 |
| (*/Pillar3 Q-terly Info/*, */General Shareholder Meeting/*) | < 1 | 1.85 | < 1 | < 1 | < 1 | < 1 | < 1 |

Note: Marked supports are > 5%.

https://doi.org/10.1371/journal.pone.0258449.t003

In the first quarter of 2011 (Table 2), the web part */Group/* was the most visited web part with the highest interest. The other web parts with high interest were */Pillar 3 Q- terly Info/*, */Annual Reports/*, */Rating/*. The pair (Table 3) with the high level of interest was (*/Pillar3 Q-terly Info/*, */Annual reports/*) and a positive correlation was detected for this pair too. Similar behaviour was identified for the year 2012 (Table 2) as in the previous years. The highest support was detected for */Group/*, */Pillar3 Q-terly Info/*, */Rating/*, */Annual Reports/*. The pairs (Table 3) with the highest visits were (*/Pillar3 Q-terly Info/*, */Group/*) and (*/Pillar3 Q-terly Info/*, */Annual Reports/*). The highest correlation was detected for (*/Annual Reports/*, */Pillar 3 Semiannualy Info/*), (*/Pillar3 Q-terly Info/*, */Annual Reports/*) and (*/Pillar 3 Semiannualy Info/*, */Pillar3 Q-terly Info/*).

In 2013 (Table 2), the most visited web part was */Group/*. The other very interesting parts were */Pillar3 Q-terly Inf/*, */Annual Reports/* and */Rating/*. The pair (Table 3) identified with high interest was (*/Pillar3 Q-terly Info/*, */Annual Reports/*). The highest correlation was detected for (*/Annual Reports/*, */Pillar3 Q-terly Info/*) and (*/Pillar3 Q-terly Info/*, */Group/*). In the first quarter of 2014 and 2015 (Table 2), the most visited web parts were */Group/*, */Pillar3 Q-terly Info/*, */Annual Reports/*, */Rating/*. The pair (Table 3) identified with high interest was (*/Pillar3 Q-terly Info/*, */Annual Reports/*) and also positive correlation was detected for this pair.

**Results summary.** The results highlighted that stakeholders were interested in Pillar 3 disclosures (regulatory and accounting) mainly in the first quarter. Stakeholders' interest was not in Pillar 3 information as a solo, but together with *Annual reports* and *Information on Group*. Moreover, the interest in the disclosed information decreased after turbulent times in 2009.

The global financial crisis has had a significant impact on the increased interest in disclosure and related information. This manifested itself in seasonality, i.e. in the increased interest in Pillar 3 information in the first, respectively in the second quarter in the first years (2009–2011) after the end of the financial crisis (2009: $Q = 8.258$, $df = 3$, $p = 0.0410$; 2010: $Q = 12.581$, $df = 3$, $p = 0.0056$; 2011: $Q = 11.539$, $df = 3$, $p = 0.0091$). In the following years (2012–2015) not only did the interest in Pillar 3 information decrease but there was no time trend ($p > 0.05$) and no seasonality in the interest of stakeholders in this information (2012: $Q = 4.154$, $df = 3$, $p = 0.2453$; 2013: $Q = 3.255$, $df = 3$, $p = 0.3539$; 2014: $Q = 4.565$, $df = 3$, $p = 0.2066$; 2015: $Q = 3.000$, $df = 3$, $p = 0.3916$) [7].

## Analysis of the stakeholders' behaviour and their interest in pillar 3 disclosures in 2016–2018

In this section, we focus on verifying the continuation of the established trend in the years after the crisis of 2012–2015. The aim of this part is to analyse the behaviour of stakeholders in the period of 2016–2018 in relation to the published Pillar 3 information, in terms of verifying null statistical hypotheses, resulting from established research assumptions that the behaviour of stakeholders in relation to published Pillar 3 information is not influenced by time trend or seasonality. In this part of our research, we analysed website data dedicated to Pillar 3 disclosures of the second commercial bank operating in Slovakia and studied the behaviour of stakeholders in relation to their interest in Pillar 3 disclosures. The website data were divided into eight categories: */Pillar 3 disclosure requirements/*, */Pillar 3 related/*, */Other—About us/*, */Other—Press centre/*, */Other—Information service/*, */Other—Social responsibility/*, */Other—Products and services for customers/*, */Other—Documents/*. During the first quarter of 2018 (Fig 2), the web part */Other—Products and services for customers/* was one of the most visited web parts with 94% *support*. Web parts */Other—Information service/* and */Other—Press Center/* occurred in identified sessions with a probability of about 4%. Web Parts */Other—About us/* and */Pillar3 related/* were less popular with a probability of less than 0.3%. Web parts */Other—Social responsibility/* and */Pillar3 disclosure requirements/* occurred in identified sessions with the *support* of less than 0.1%. The web part */Other—Documents/* did not meet the minimum support, i.e. the probability of occurrence in the identified sessions is less than 0.01%.

In the first quarter of 2018 (Fig 2), pairs (*/Other—Products and services for customers/*, */Other—Information service/*) with approximately 2% *support* were among the most visited pairs of web parts. Other pairs (*/Other—Products and services for customers/*, */Other—Press center/*), (*/Other—Products and services for customers/*, */Other—About us/*) and (*/Other—Information service/*, */Other—About us/*) reached a probability of about 0.14%. Other pairs of web parts achieved a probability of less than 0.08%. The highest level of positive correlation (*lift* $> 1$) was obtained for the pair (*/Pillar3 related/*, */Pillar3 disclosure requirements/*) with *lift* = *235*. The high level was also reached by the pair (*/Other—About us/*, */Other—Social responsibility/*) and (*/Other—Social responsibility/*, */Pillar3 related/*) with *lift* between 100–109. High level of positive correlation was also achieved by positive pair (*/Other—About us/*, */Pillar3 related/*) with *lift* = 86. Positive correlation with *lift* $< 10$ reached a pair (*/Other—Information service/*, */Other—About us/*), (*/Other—Social responsibility/*, */Other—Information service/*), (*/Pillar3 related/*, */Other—Information service/*), (*/Pillar3 related/*, */Other—Press center/*), (*/Other—Press center/*, */Other—Social responsibility/*), (*/Other—Press center/*, */Pillar3 disclosure requirements/*) and (*/Other—About us/*, */Other—Press center/*). Negative correlation (*lift* $< 1$) was identified and web parts occur more often separately than together in identified sessions for pairs (*/Other—Press center/*, */Other—Information service/*), (*/Other—About us/*, */Other—Products and services for customers/*), (*/Other—Social responsibility/*, */Other—Products and*

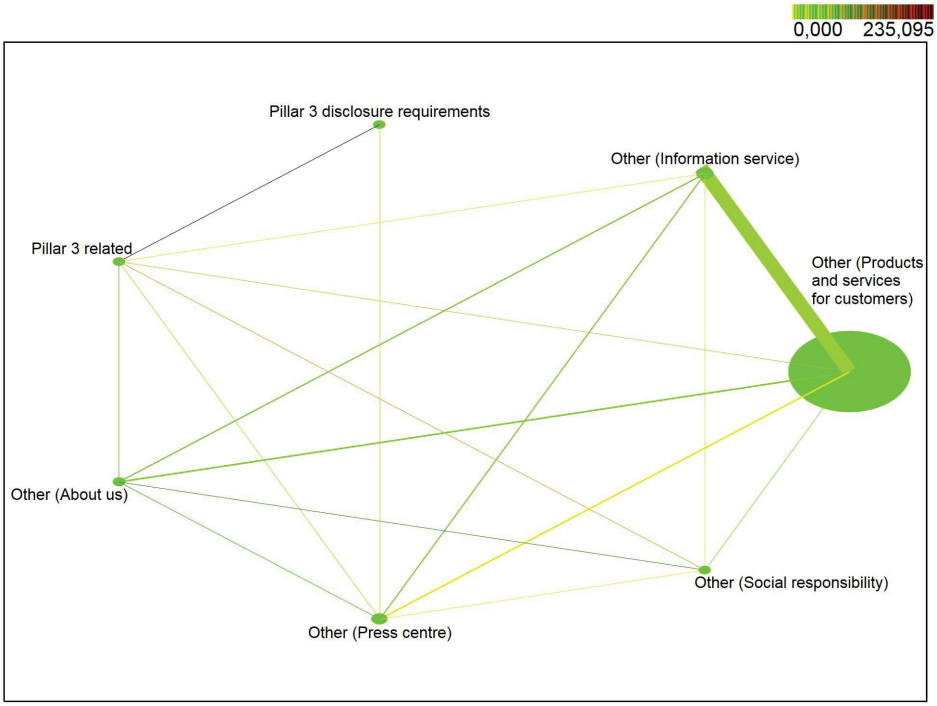



**Fig 2. Visualization of frequented web parts of the first quarter of 2018.**

https://doi.org/10.1371/journal.pone.0258449.g002

*services for customers/*), (*/Other—Information service/*, */Other—Products and services for customers/*), (*/Pillar3 related/*, */Other—Products and services for customers/*) and (*/Other—Press center/*, */Other—Products and services for customers/*).

*Note*: *NODE SIZE—relative support of each web part, LINE THICKNESS—relative joint support of two web parts, COLOR DARKNESS OF LINE—relative lift of two web parts.*

The results of the segmentation (Fig 3) confirm the results of the association analysis for the pairs of web parts (*/Pillar3 disclosure requirements/*, */Pillar3 related/*) and (*/Other—About us/*, */Other—Social responsibility/*), where a positive correlation was identified (*lift* > 1). Part of the sessions was characterized by a visit to both Pillar 3 web parts, where the highest level of interest was achieved (*lift* = 235).

In the second quarter of 2018 (Fig 4), the web part */Other—Products and services for customers/* belonged to the most visited web part with the *support* of 92%. The web part */Other—Information service/* occurred in the identified sessions with a probability of more than 6% and the web part */Other—Press centre/* with a probability of about 3%. The web part */Other—About us/* also belonged to the less visited web parts with a probability of just over 1%. The web part */Pillar3 related/* was one of the least visited with *support* of about 0.2%. Other web parts */Other—Social responsibility/*, */Pillar3 disclosure requirements/* and */Other—Documents/* occurred in the identified sessions with a probability of less than 0.1%.

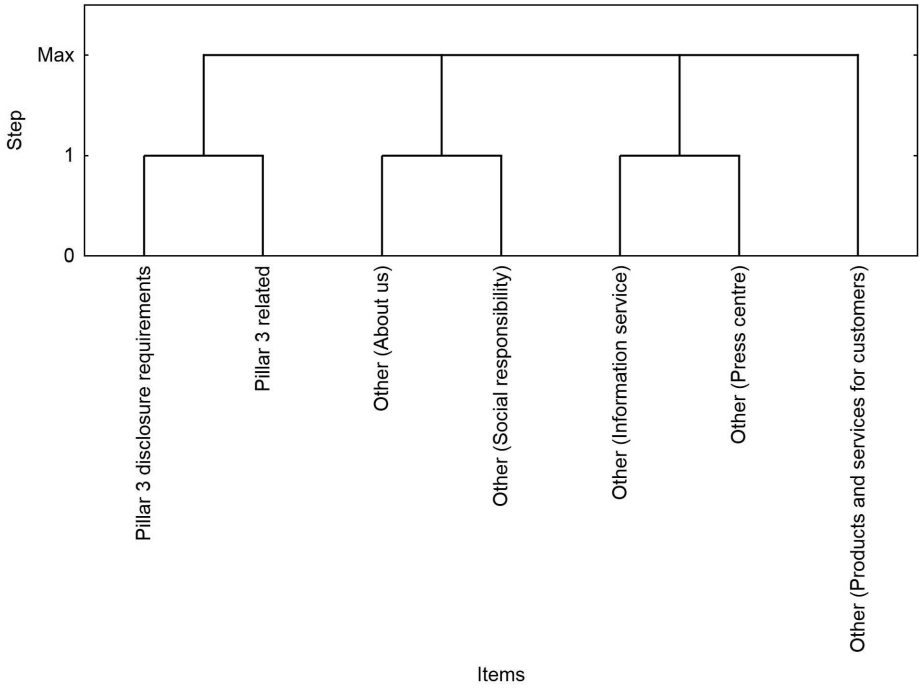



**Fig 3. Visualization of segments in the first quarter.**

https://doi.org/10.1371/journal.pone.0258449.g003

In the second quarter of 2018 (Fig 4), the pair (*/Other—Products and services for customers/*, */Other—Information service/*) with the *support* of almost 2% was among the most visited pairs of web parts. Pairs of web parts (*/Other—Information service/*, */Other—Press center/*), (*/Other—Products and services for customers/*, */Other—About us/*), (*/Other—Information service/*, */Other—About us/*) and (*/Other—Products and services for customers/*, */Other—Press center/*) achieved *support* between 0.2–0.4%. The remaining pairs of web parts achieved the *support* of less than 0.06% but met the requirement for minimum support of at least 0.01%. The highest level of interest was achieved by a pair (*/Pillar3 disclosure requirements/*, */Pillar3 related/*) with *lift* = 196. A high degree of positive correlation was also achieved by a pair of web parts (*/Pillar3 related/*, */Other—Social responsibility/*) with *lift* = 74. The higher interest was also achieved by pairs (*/Other—Social responsibility/*, */Other—About us/*) and (*/Pillar3 related/*, */Other—About us/*) with a lift between 19 and 27. A positive correlation (*lift* > 1) was identified and web parts occur more often together than separately in the identified sessions for (*/Other—Documents/*, */Other—Information service/*), (*/Other—Social responsibility/*, */Other—Information service/*), (*/Pillar3 related/*, */Other—Information service/*), (*/Other—Press center/*, */Other—Social responsibility/*), (*/Pillar3 related/*, */Other—Press center/*), (*/Other—Information service/*, */Other—About us/*) and (*/Other—Press center/*, */Other—Information service/*), which reached a value between 1 and 7. On the other hand, for pairs (*/Other—About us/*, */Other—Products and services for customers/*), (*/Other—Information service/*, */Other—Products and services for*

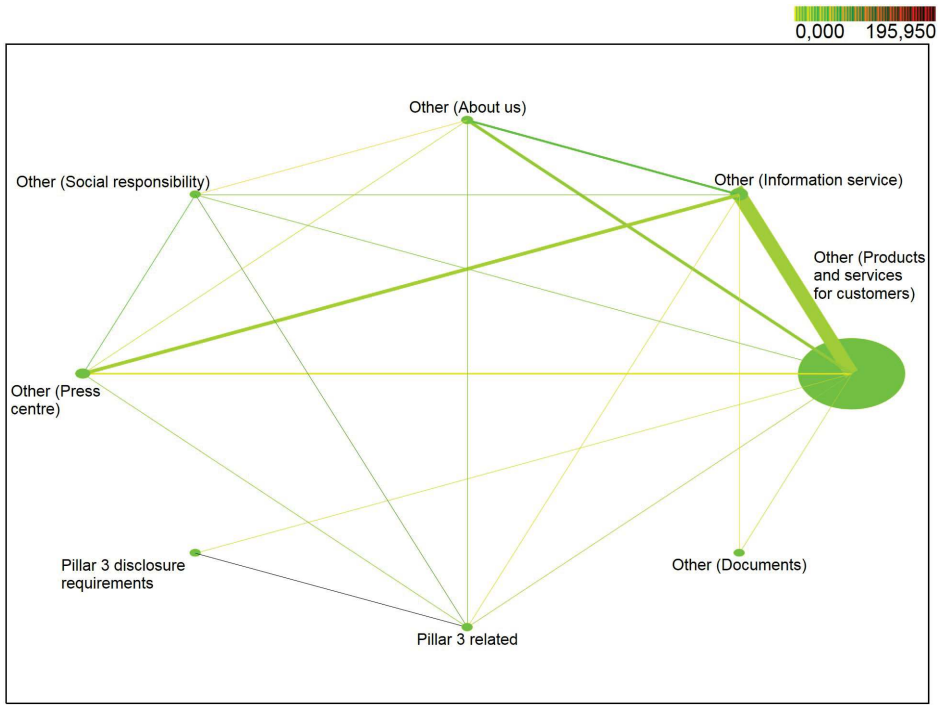



**Fig 4. Visualization of frequented web parts of the second quarter of 2018.**

https://doi.org/10.1371/journal.pone.0258449.g004

*customers/*) a (*/Other—Press center/*, */Other—Products and services for customers/*) a negative correlation was identified (*lift* is less than 1).

*Note: NODE SIZE—relative support of each web part, LINE THICKNESS—relative joint support of two web parts, COLOR DARKNESS OF LINE—relative lift of two web parts.*

The segmentation outcome (Fig 5) confirms the results of the association analysis for the web part pairs (*/Pillar3 disclosure requirements/*, */Pillar3 related/*), (*/Other—About us/*, */Other—Social responsibility/*) and (*/Other—Information service/*, */Other—Documents/*), where a positive correlation was identified (*lift* > 1). Part of the sessions was characterized by a visit to both Pillar 3 web parts, where the highest level of interest was achieved (*lift* = 196).

In the third quarter of 2018 (Fig 6), one of the most visited web parts was the web part */Other—Products and services for customers/* with 66% *support*. The web part */Other—Information service/* with *support* of more than 22% was also interesting for visitors. Web parts */Other—Press center/* and */Other—About* us/ were among the less visited with the *support* of 11% and 7%. Less traffic was identified for the web parts */Other—Documents/*, */Pillar3 related/*, */Other—Information Service/*, */Other—Social responsibility/* and */Pillar3 disclosure requirements/* with *support* between 0.2%–0.6%.

In the third quarter of 2018 (Fig 6), the pair (*/Other—Products and services for customers/*, */Other—Information Service/*) was among the most visited web parts with *support* of more

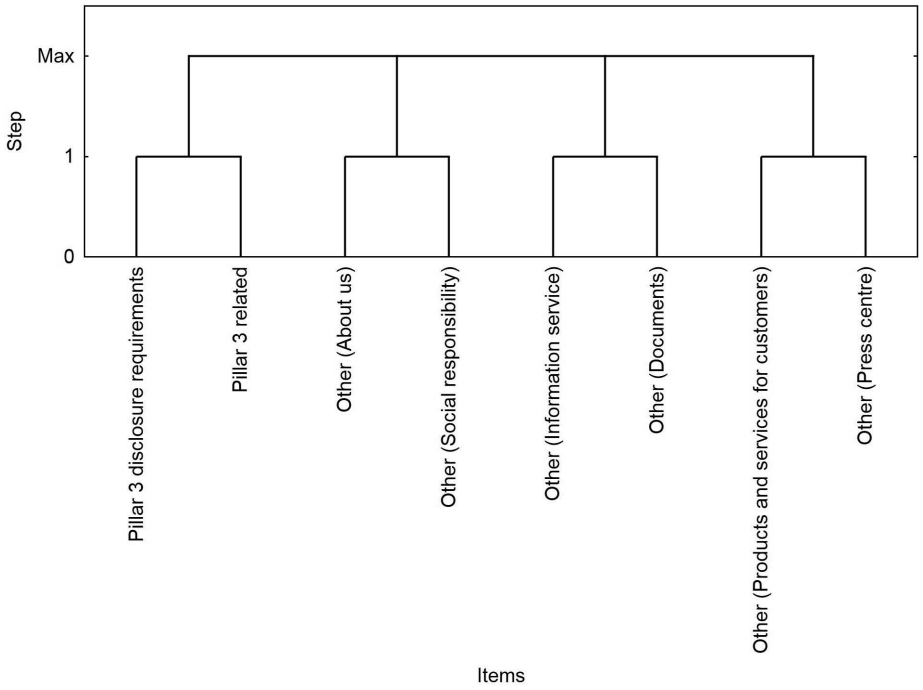



**Fig 5. Visualization of segments in the second quarter.**

https://doi.org/10.1371/journal.pone.0258449.g005

than 4%. The pairs of web parts (*/Other—Information Service/*, */Other—About us/*) and (*/Other—Products and services for customers/*, */Other—About us/*) achieved the *support* of around 1%. Less visited pairs (/Other—Information Service/, */Other—Press center/*), (*/Other—Products and services for customers/*, */Other—Documents/*) and (*/Other—Products and services for customers/*, */Other—Press center/*) achieved *support* between 0.4–0.6%. The other pairs (Fig 6) achieved *support* between 0.01–0.2%. The highest level of interest was achieved by the pair (*/Pillar3 disclosure requirements/*, */Pillar3 related/*) with *lift* = 92. A significant degree of interest was also identified for the pairs of web parts (*/Other—Social responsibility/*, */Pillar3 disclosure requirements/*) and (*/Other—Social responsibility/*, */Pillar3 related/*), with a *lift* between 30–37. Positive correlation was identified for pairs (*/Other—Social responsibility/*, */Other—About us/*), (*/Pillar3 disclosure requirements/*, */Other—About us/*), (*/Pillar3 related/*, */Other—About us/*), (*/Other—Social responsibility/*, */Other—Information Service/*), (*/Other—Social responsibility/*, */Other—Press center/*), (*/Other—Documents/*, */Other—Information Service/*) and (*/Other—Documents/*, */Other—Products and services for customers/*) with a *lift* between 1 and 6. A negative correlation was identified for the other pairs of web parts, that is, they occurred more frequently in the identified sessions separately than together (*lift* <1).

*Note*: *NODE SIZE—relative support of each web part, LINE THICKNESS—relative joint support of two web parts, COLOR DARKNESS OF LINE—relative lift of two web parts.*

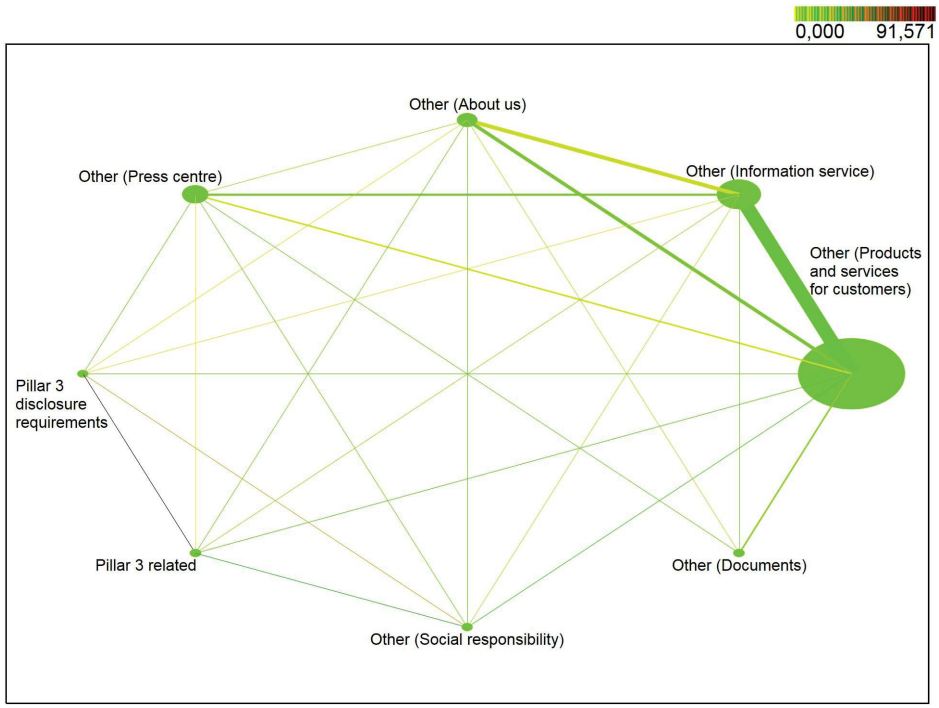



**Fig 6. Visualization of frequented web parts of the third quarter of 2018.**

https://doi.org/10.1371/journal.pone.0258449.g006

The segmentation results (Fig 7) confirm the results of the association analysis for the web part pairs (*/Pillar3 disclosure requirements/*, */Pillar3 related/*), (*/Other—About us/*, */Other—Social responsibility/*) and (*/Other—Information Service/*, */Other—Documents/*), where a positive correlation was identified (*lift* > 1). Part of the sessions was characterized by a visit to both Pillar 3 web parts, where the highest level of interest was achieved (*lift* = 92).

In the fourth quarter of 2018 (Fig 8), the same behaviour was observed as in the third quarter of the year under study. The most visited web parts included the web part */Other—Products and services for customers/* with 64% *support*. The web part */Other—Information Service/* with *support* of more than 24% was also interesting for visitors. Web parts */Other—Press center/* and */Other—About us/* were among the less visited with the *support* of 11% and 7%. Less traffic was identified for the web parts /Other—*Documents/*, */Pillar3 related/*, */Other—Information Service/*, */Other—Social responsibility/* and /Pillar3 disclosure requirements/ with *support* between 0.2%–0.6%.

A similar trend was also in the case of pairs of web parts, where the results identified in the fourth quarter (Fig 8) are like the third quarter of 2018. The pair (*/Other—Products and services for customers/*, */Other—Information Service/*) was one of the most visited web parts with more *support* of 4%. The pairs of web parts (*/Other—Information Service/*, */Other—About us/*) and (*/Other—Products and services for customers/*, */Other—About us/*) achieved the *support* of

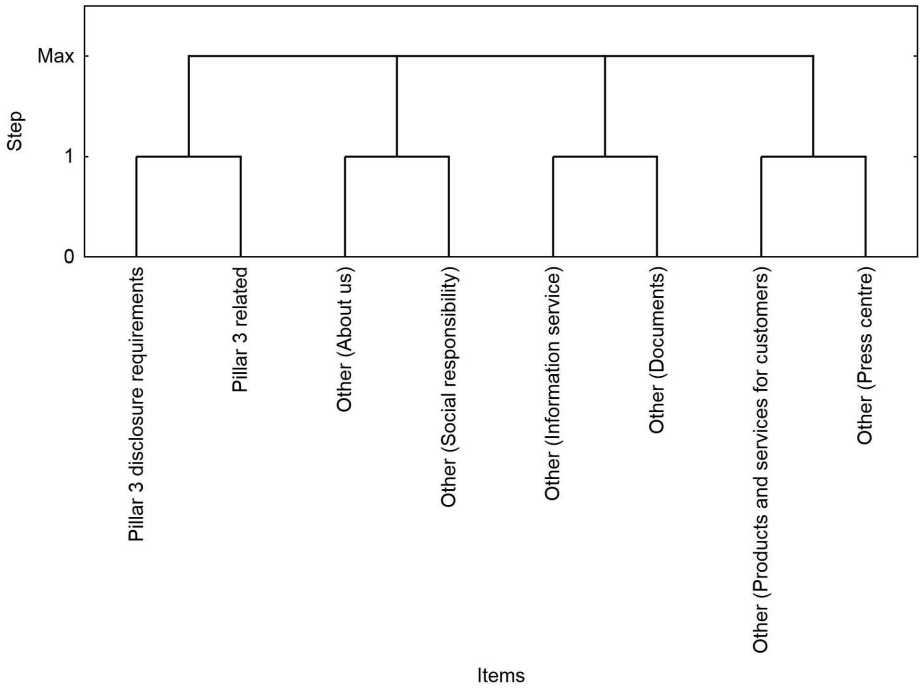



**Fig 7. Visualization of seating segments in the third quarter.**

https://doi.org/10.1371/journal.pone.0258449.g007

around 1%. Less visited pairs (*/Other—Information Service/*, */Other—Press center/*), (*/Other—Products and services for customers/*, */Other—Documents/*) and (*/Other—Products and services for customers/*, */Other—Press center/*) achieved *support* between 0.4–0.6%. The other pairs (Fig 8) achieved *support* between 0.01–0.2%.

*Note*: *NODE SIZE—relative support of each web part, LINE THICKNESS—relative joint support of two web parts, COLOR DARKNESS OF LINE—relative lift of two web parts*.

The highest level of interest in the fourth quarter was achieved by the pair (*/Pillar3 disclosure requirements/*, */Pillar3 related/*) with *lift* = 59. A significant degree of interest was also identified for the pairs of web parts (*/Other—Social responsibility/*, */Pillar3 disclosure requirements/*) and (*/Other—Social responsibility/*, */Pillar3 related/*), with *lift* around 18. Positive correlation was identified for pairs (*/Other—Social responsibility/*, */Other—About us/*), (*/Pillar3 disclosure requirements/*, */Other—About us/*), (*/Pillar3 related/*, */Other—About us/*), (*/Other—Social responsibility/*, */Other—Information Service/*), (*/Other—Social responsibility/*, */Other—Press center/*), (*/Other—Documents/*, */Other—Information Service/*) and (*/Other—Documents/*, */Other—Products and services for customers/*) with a *lift* between 1 and 5. A negative correlation was identified for the other pairs of web parts, i.e. they occurred more frequently in the identified sessions separately than together (*lift* $< 1$). These results correspond to the behaviour of stakeholders in the third quarter, but in the fourth quarter, a lower level of interest was achieved for the examined pairs of web parts.

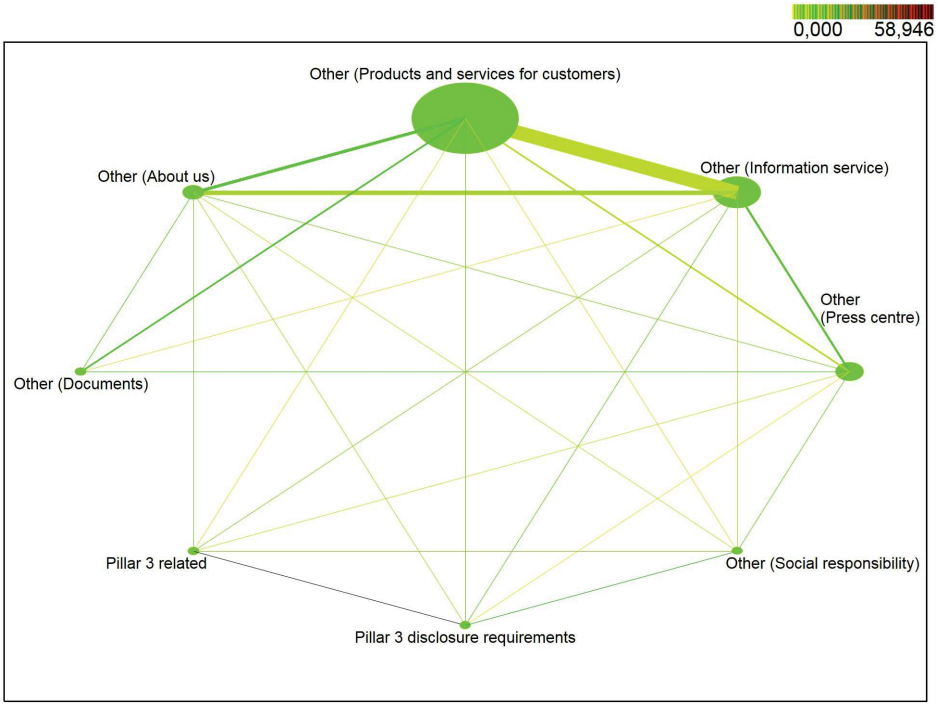



**Fig 8. Visualization of frequented web parts of the fourth quarter of 2018.**

https://doi.org/10.1371/journal.pone.0258449.g008

The segmentation results (Fig 9) confirm the results of the association analysis for the web part pairs (*/Pillar3 disclosure requirements/*, */Pillar3 related/*), (*/Other—About* us/, */Other—Social responsibility/)* and (*/Other—Information Service/*, */Other—Documents/*), where a positive correlation was identified (*lift* > 1). Part of the sessions was characterized by a visit to both Pillar 3 web parts, where the highest level of interest was achieved (*lift* = 59).

**Results summary.** The results reveal that during the year 2018 the most visited web parts pair was (*/Other—Products and services for customers/*, */Other—Information Service/*). The highest level of positive correlation was achieved by a pair (*/Pillar3 disclosure requirements/*, */Pillar3 related/*). The results revealed that stakeholders showed higher interest in Pillar 3 related information, Annual reports, Information on Group than in Pillar 3 disclosure requirements. This could imply that the stakeholders found important content in these categories that are not published in Pillar 3 disclosures.

We evaluated the patterns of stakeholder behaviour in terms of time (years and quarters) to verify the presence or absence of a time trend and seasonality in the behaviour of stakeholders in relation to Pillar 3 information. For this purpose, we evaluated the proportion of occurrence of patterns with Pillar 3 information from all possible combinations of the examined web parts.

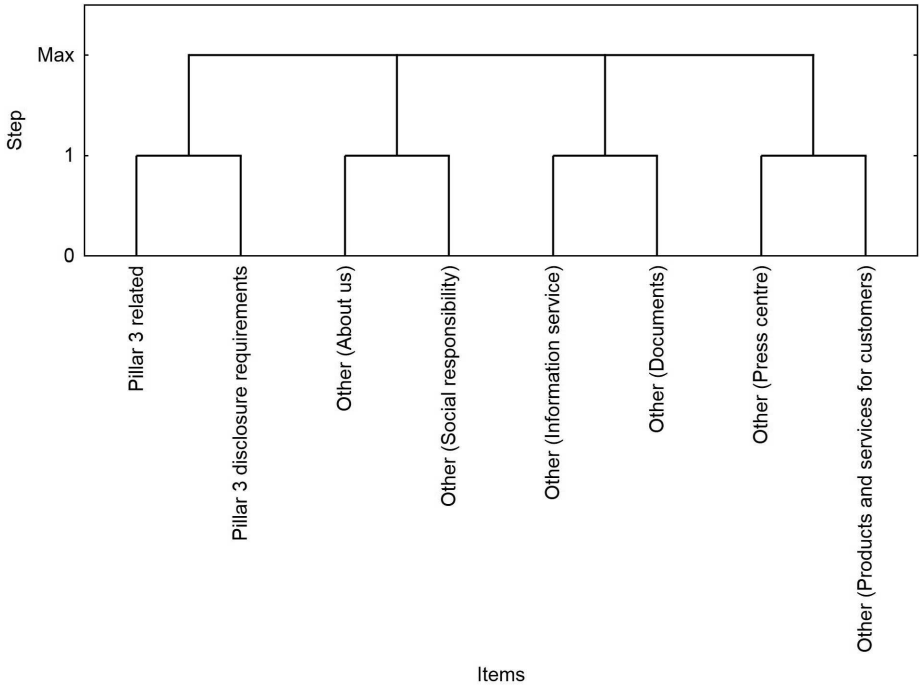



**Fig 9. Visualization of segments in the fourth quarter.**

https://doi.org/10.1371/journal.pone.0258449.g009

In the case of years (Table 4a–4d) for all quarters (Q1: $Q = 4.666667$, $df = 2$, $p = 0.096974$; Q2: $Q = 3.600000$, $df = 2$, $p = 0.165301$; Q3: $Q = 2.000000$, $df = 2$, $p = 0.367881$; Q4: $Q = 0.000000$, $df = 2$, $p = 1.000000$), the null hypotheses are not rejected, i.e. the occurrences of frequent item sets of web parts related to Pillar 3 do not depend on the year.

One homogeneous group (16QX, 17QX, 18QX) was identified for each quarter (Table 4a–4d) based on the average occurrence of the found frequent item sets of web parts related to Pillar 3 information ($p > 0.05$). The results correspond to our previous findings [7] on the absence of a time trend in the years after the crisis.

In the case of the examined web parts related to Pillar 3 (Table 5c), no statistically significant differences (2018: $Q = 7.714$, $df = 3$, $p = 0.0523$) were identified between the individual quarters of 2018. The most frequent item sets related to Pillar 3 in 2018 (Table 5c) were identified in the third and fourth quarters (< 40%), the least in the first and second quarters (< 30%). Similarly, seasonality in the interest of Pillar 3 information (Table 5a and 5b) was not identified in 2016 and 2017 (2016: $Q = 1.737$, $df = 3$, $p = 0.6288$; 2017: $Q = 0.857$, $df = 3$, $p = 0.8358$), which confirms the established trend identified in the years after the crisis 2012–2015 ($p > 0.05$) [7], i.e. in the years after the crisis, no seasonality was identified in accesses to Pillar 3 information.

**Table 4. Homogeneous groups for the occurrence of frequented item sets of web parts related to Pillar 3 for the quarter a) Q1, b) Q2, c) Q3 and d) Q4.**

| Year | Percent | 1 |
|---|---|---|
| 18Q1 | 25.00% | **** |
| 17Q1 | 30.56% | **** |
| 16Q1 | 33.33% | **** |
| 17Q3 | 33.33% | **** |
| 18Q3 | 36.11% | **** |
| 16Q3 | 38.89% | **** |
| 18Q2 | 25.00% | **** |
| 17Q2 | 33.33% | **** |
| 16Q2 | 33.33% | **** |
| 17Q4 | 36.11% | **** |
| 18Q4 | 36.11% | **** |
| 16Q4 | 36.11% | **** |

Note:

****—homogeneous groups ($p > 0.05$).

https://doi.org/10.1371/journal.pone.0258449.t004

## Conclusion and contribution of the work

The disclosure of commercial banks is of particular importance for their stakeholders. In the process of the search for meaningful Pillar 3 disclosures as a direct channel for market discipline implementation, it is important to take into account a lot of factors, which influence their quality. These factors are transparency, accuracy, timing, which serves as a background for meaningful, comparable, and sufficient disclosures to deliver relevant information to key market participants. However, if market discipline should be effective it requires the relevant interest of key market participants in regulatory required Pillar 3 disclosed information. In our research, we studied these interests in commercial banks operating in CEE country. We found

**Table 5. Homogeneous groups for the occurrence of frequented item sets of web parts related to Pillar 3 for the year a) 2016, b) 2017 and c) 2018.**

| Quarter | Percent | 1 |
|---|---|---|
| 16Q1 | 33.33% | **** |
| 16Q2 | 33.33% | **** |
| 16Q4 | 36.11% | **** |
| 16Q3 | 38.89% | **** |
| 17Q1 | 30.56% | **** |
| 17Q3 | 33.33% | **** |
| 17Q2 | 33.33% | **** |
| 17Q4 | 36.11% | **** |
| 18Q2 | 25.00% | **** |
| 18Q1 | 25.00% | **** |
| 18Q4 | 36.11% | **** |
| 18Q3 | 36.11% | **** |

Note:

****—homogeneous groups ($p > 0.05$).

https://doi.org/10.1371/journal.pone.0258449.t005

out that in studied banks single interest in regulatory required Pillar 3 information was very low and not equally intensive interest in regulatory required timing—quarters of the year.

In the first analysed bank, from 2009 to 2015, the most visited were web parts */Group/* and */Pillar3 Q-terly Info/*. The pairs identified with high interest were */Group/*, */Pillar3 Q-terly Info/* and */Annual Reports/*, */Pillar3 Q-terly Info/* in these years. Generally, stakeholders were interested in Pillar 3 disclosures (regulatory and accounting) mainly in the first quarter of the year. It is important to note, that after turbulent times in 2009 the interest of the stakeholders in disclosed information steadily decreased in the analysed first commercial bank in Slovakia.

In the second analysed bank, the results show that in all four quarters of 2018 the pair (*/Other—Products and services for customers/*, */Other—Information Service/*) was the most visited web parts with the highest *support*. Additionally, the highest level of positive correlation was achieved by a pair (*/Pillar3 disclosure requirements/*, */Pillar3 related/*). Moreover, the results of these analyses suggest that stakeholders expressed higher interest in Pillar 3 related information, Annual reports, Information on Group (financial reports, annual reports, information about the group, rating, general shareholder meeting, emitent prospects) than in Pillar 3 disclosure requirements. It can also indicate that these categories contain important content for web users, which is not published in Pillar 3 disclosures. Similar findings were achieved for 2016 and 2017.

The assumption about the continuation of the established trend after the gradual disappearance of the consequences of the financial crisis from 2012–2015 [7], which is characterized by a lack of interest in Pillar 3 information, was confirmed. It was further confirmed that the behaviour of stakeholders in relation to the published Pillar 3 information no longer shows any time trend or seasonality.

We saw an increased interest in this information only in the years of the global financial crisis and in the immediate aftermath of the crisis (2009–2011). The following years (2012–2018) were characterized by an only low interest in this information and based on data on the use of the web portals of two commercial banks, we did not identify any time trends or seasonality in the behaviour of stakeholders in relation to Pillar 3 information. Previous revisions of Pillar 3 by regulatory authorities have not had a significant impact on increasing interest in this information.

The results suggest that market discipline mechanisms should start to be ready and operate efficiently, particularly during turbulent times. If the commercial banking institutions owned by foreign shareholders in CEE countries are not ready to disclose information satisfactorily and on time it can cause losses related to the reputational risk of the bank. These results also suggest that changes in the Pillar 3 disclosures are inevitable to bring relevant and meaningful information to stakeholders which are key for this type of institution. The stability of the financial market is the most important goal of the regulators since the severe market turbulence, which influenced the whole financial system. Therefore, market discipline mechanism implementation should take into account balancing between its economic costs and benefits and should adequately respond to its current challenges. It is important to note, that the weaknesses of the Pillar 3 disclosures are present, but their advantages are undeniable due to the focus of the majority of theoretical studies on their positive aspects, as a beneficial tool for the enhancement of effective market discipline mechanism.

## Contribution to the improvement of the effective market discipline in commercial banks operating in CEE countries

In our research, we have identified areas of improvement, which can increase the interest of the stakeholders in the Pillar 3 disclosures, and we summarize them as the following recommendations:

- To improve standardisation, meaning harmonisation of national authorities' disclosure requirements and disclosure requirements on EU level (Pillar 3 and national requirements).
- To increase comparability of disclosures by creating one common template (visually prescribed tables) ideally created by regulators, to implement uniformity.
- To decrease the frequency of the Pillar 3 disclosures, due to the low interests of the stakeholders in quarterly disclosures.
- If quarterly disclosures are applied to decrease the amount of the disclosed information could be beneficial and to differentiate extend of the annual disclosures (more information) in comparison to a quarterly one.
- To include information areas (either on an obligatory or voluntary basis), in which stakeholders are interested (business behaviour of the institution, strategy, rumours, structure, ownership, mission, values), which influence the risk position of the institution.
- Regulators should assure compliance with the obligatory required information—mainly to restrict the institutions' omission of the required information without any indication of the reasons.
- To impose rules for the location of the disclosed documents, which should be in an identifiable section of the web page.
- Obligation to use English as a unified language for disclosures.

The stability of the financial market is the most important goal of the regulators since the severe market turbulences influenced the whole financial system. Therefore, market discipline mechanism implementation should take into account balancing between its economic costs and benefits and should adequately respond to its current challenges. It is important to note, that the weaknesses of the Pillar 3 disclosures are present, but their advantages are undeniable as a beneficial tool for enhancement of effective market discipline mechanism.

The future work will be focused on the verification of our findings particularly concerning the effectiveness of published and revised information in the period of the COVID-19 pandemic crisis.

### Research limitations

Our research has a few limitations. Importantly, these limitations are concerning the nature of the data and its characteristics. The extraction of the data also has deficiencies in the redundancy of the data, which means that the data source has also been aimed at customers of the bank web portal, who do not preferentially look for Pillar 3 data to assess the bank's risk profile. Importantly, customers of the bank are also an important group of targets of market discipline implementation, but they are not a direct source of market discipline enhancement. Although these deficiencies in the data might have caused some statistical deviations, we consider them as minor deviations with weak influence on overall research results as customers of the bank are also part of the market discipline.

## Author Contributions

**Conceptualization:** Anna Pilková, Michal Munk.

**Data curation:** Jozef Kapusta.

**Formal analysis:** Anna Pilková, Petra Blažeková.

**Investigation:** Anna Pilková, Michal Munk.

**Methodology:** Michal Munk, Ľubomír Benko.

**Resources:** Michal Munk, Ľubomír Benko, Jozef Kapusta.

**Supervision:** Anna Pilková.

**Validation:** Anna Pilková, Petra Blažeková.

**Visualization:** Michal Munk.

**Writing – original draft:** Anna Pilková, Petra Blažeková.

**Writing – review & editing:** Michal Munk, Ľubomír Benko, Jozef Kapusta.

## References

1. EBA. EBA notes enhanced consistency on institutions' Pillar 3 disclosures but calls for improvements to reinforce market discipline. 2020. https://www.eba.europa.eu/eba-notes-enhanced-consistency-institutions'-pillar-3-disclosures-calls-improvements-reinforce.
2. BIS. The Basel Committee consults on revisions to the Pillar 3 disclosure framework. 2018. https://www.bis.org/press/p180227.htm.
3. EBA. EBA launches consultation on comprehensive Pillar 3 disclosures. 2019. https://www.eba.europa.eu/eba-launches-consultation-on-comprehensive-pillar-3-disclosures.
4. Benli VF. Basel's Forgotten Pillar: The Myth of Market Discipline on the Forefront of Basel III. Financ Internet Q. 2015; 11: 70–91.
5. Biljanovska B. Aligning Market Discipline and Financial Stability: A More Gradual Shift from Contingent Convertible Capital to Bail-in Measures. Eur Bus Organ Law Rev. 2016; 17: 105–135. https://doi.org/10.1007/s40804-016-0028-0
6. Wilms W. The dark side of the Basel Committee's Pillar 3 framework. In: National Bank of Belgium [Internet]. 2014 p. 17. http://www.eurofiling.info/201411/presentations/20141125TheDarkSideOfTheBasselCommitteesWilfriedWilms.pdf.
7. Munk M, Pilkova A, Benko L, Blažeková P. Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. J Bus Econ Manag. 2017; 18: 954–973. https://doi.org/10.3846/16111699.2017.1360388
8. Anagnostopoulos Y, Kabeega J. Insider perspectives on European banking challenges in the post-crisis regulation environment. J Bank Regul. 2019; 20: 136–158.
9. Buckley R, Arner D, Zetzsche D, Weber R. The road to RegTech: the (astonishing) example of the European Union. J Bank Regul. 2020; 21: 26–36.
10. Basten M, Sanchez Serrano A. European banks after the global financial crisis: a new landscape. J Bank Regul. 2019; 20: 51–73.
11. Weekes-Marshall D. A developing country's commercial banking risk governance disclosures: Post-financial crisis. Int J Financ Econ. 2020; https://doi.org/10.1002/ijfe.2320
12. Ivan Moreno A, Caminero T. Application of Text Mining to the Analysis of Climate-Related Disclosures. Banco Espana Work Pap No 2035. 2020; 43. https://doi.org/10.2139/ssrn.3738629
13. Linsley PM., Shrives PJ. Transparency and the disclosure of risk information in the banking sector. J Financ Regul Compliance. 2005; 13. https://doi.org/10.1108/13581980510622063
14. Savvides SC, Savvidou N. Market risk disclosures of banks: a cross-country study. Int J Organ Anal. 2012; 20. https://doi.org/10.1108/19348831211268599
15. Kabir MR., Sobhani FA. Risk Disclosures in Bank's Annual Report: Bangladesh Perspective. Aust Acad Account Financ Rev. 2017; 3: 11–20.
16. Barakat A, Hussainey K. Bank governance, regulation, supervision, and risk reporting: Evidence from operational risk disclosures in European banks. Int Rev Financ Anal. 2013; 30. https://doi.org/10.1016/j.irfa.2013.07.002
17. Khalil F, Alam H.M. Evidence of market discipline through operational risk disclosures in commercial banking sector of Pakistan. Pakistan Bus Rev. 2018; 20: 768–782.
18. Naz M, Ayub H. Impact of Risk-Related Disclosure on the Risk-Taking Behavior of Commercial Banks in Pakistan. J Indep Stud Res Soc Sci Econ. 2017; 15. https://doi.org/10.31384/jisrmsse/2017.15.2.9

19. Vauhkonen J. The Impact of Pillar 3 Disclosure Requirements on Bank Safety. J Financ Serv Res. 2012; 41: 37–49. https://doi.org/10.1007/s10693-011-0107-x

20. Frolov M. Why do we need mandated rules of public disclosure for banks? J Bank Regul. 2007; 8: 177–191. https://doi.org/10.1057/palgrave.jbr.2350045

21. Niessen-Ruenzi A, Parwada JT, Ruenzi S. Information Effects of the Basel Bank Capital and Risk Pillar 3 Disclosures on Equity Analyst Research An Exploratory Examination. SSRN Electron J. 2015. https://doi.org/10.2139/ssrn.2670418

22. Parwada JT, Ruenzi S, Sahgal S. Market Discipline and Basel Pillar 3 Reporting. SSRN Electron J. 2013. https://doi.org/10.2139/ssrn.2443189

23. Shair F, Sun N, Shaorong S, Atta F, Hussain M. Impacts of risk and competition on the profitability of banks: Empirical evidence from Pakistan. PLoS One. 2019; 14. https://doi.org/10.1371/journal.pone.0224378 PMID: 31710614

24. Chen Y, Du K. The role of information disclosure in financial intermediation with investment risk. J Financ Stab. 2020; 46: 100720. https://doi.org/10.1016/j.jfs.2019.100720

25. Bischof J, Daske H, Elfers F, Hail L. A Tale of Two Supervisors: Compliance with Risk Disclosure Regulation in the Banking Sector. Contemp Account Res Forthcom. 2021. https://ssrn.com/abstract=3881110.

26. Albuquerque D, Isabel Morais A, Pinto I. The role of banking supervision in credit risk disclosures and loan loss provisions. Rev Bras Gest Negócios. 2020; 22: 932–948.

27. Giner B, Allini A, Zampella A. The Value Relevance of Risk Disclosure: An Analysis of the Banking Sector. Account Eur. 2020. https://doi.org/10.1080/17449480.2020.1730921

28. Bischof J, Daske H, Elfers F, Hail L. A Tale of Two Regulators: Risk Disclosures, Liquidity, and Enforcement in the Banking Sector. SSRN Electron J. 2016. https://doi.org/10.2139/ssrn.2580569

29. de Araujo P, Leyshon KI. The impact of international information disclosure requirements on market discipline. Appl Econ. 2016; 49: 954–971. https://doi.org/10.1080/00036846.2016.1208361

30. Fernandes C, Farinha J, Vitorino Martins F, Mateus C. The impact of board characteristics and CEO power on banks' risk-taking: stable versus crisis periods. J Bank Regul. 2021. https://doi.org/10.1057/s41261-021-00146-4

31. Akbar S, Deegan C. Analysis of corporate social disclosures of the apparel industry following crisis: an institutional approach. Account Financ. 2021; 61: 3565–3600. https://doi.org/10.1111/acfi.12712

32. Faria-e-Castro M, Martinez J, Philippon T. Runs versus Lemons: Information Disclosure and Fiscal Capacity. Cambridge, MA; 2017 May. https://doi.org/10.3386/w21201

33. Del Gaudio BL, Megaravalli A V., Sampagnaro G, Verdoliva V. Mandatory disclosure tone and bank risk-taking: Evidence from Europe. Econ Lett. 2020; 186: 108531. https://doi.org/10.1016/j.econlet.2019.108531

34. Oliveira J, Lima Rodrigues L, Craig R. Voluntary risk reporting to enhance institutional and organizational legitimacy. J Financ Regul Compliance. 2011; 19. https://doi.org/10.1108/13581981111147892

35. Arsov S, Bucevska V. Determinants of transparency and disclosure—evidence from post-transition economies. Econ Res Istraživanja. 2017; 30: 745–760. https://doi.org/10.1080/1331677X.2017.1314818

36. Bartulovic M, Pervan I. Comparative analysis of voluntary internet financial reporting for selected CEE countries. Recent Res Appl Econ Manag. 2012; 1: 296–301.

37. Farvaque E, Refait-Alexandre C. Implementing Basel Framework's Transparency Requirements in Emerging Countries: Bane or Boon? SSRN Electron J. 2013. https://doi.org/10.2139/ssrn.2381867

38. Blažeková P, Benko Ľ, Pilková A, Munk M. Is Pillar 3 a good tool? Days of Management and Economics Students 2020. Bratislava, Slovakia; 2020.

39. Munk M, Benko Ľ, Gangur M, Turčáni M. Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. E+M Ekon a Manag. 2015; 18: 144–159.

40. Cooley R, Mobasher B, Srivastava J. Data preparation for mining world wide web browsing patterns. Knowl Inf Syst. 1999; 1: 5–32.

41. Liu B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Computer Knowledge and Technology (Academic . . .. 2011. https://doi.org/10.1007/978-3-642-19460-3

42. Pamutha T, Chimphlee S, Kimpan C, Sanguansat P. Data Preprocessing on Web Server Log Files for Mining Users Access Patterns. Int J Res Rev Wirel Commun. 2012; 2: 92–98. Available: http://www.sciacademypublisher.com/journals/index.php/IJRRWC/article/view/835.

43. Spiliopoulou M, Mobasher B, Berendt B, Nakagawa M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. INFORMS J Comput. 2003; 15: 171–190. https://doi.org/10.1287/ijoc.15.2.171.14445

PRÍLOHA C: MUNK, MICHAL, ANNA PILKOVA, ĽUBOMÍR BENKO, PETRA BLAZEKOVA A PETER SVEC, 2021B. PILLAR 3–PRE-PROCESSED WEB SERVER LOG FILE DATASET OF THE BANKING INSTITUTION. *DATA IN BRIEF*. 39, 107672. DOI:10.1016/J.DIB.2021.107672 (**WEB OF SCIENCE; SCOPUS**) [WOS: 1, SCOPUS: 0]

Data in Brief 39 (2021) 107672

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Pillar 3–Pre-processed web server log file dataset of the banking institution

Michal Munk[a], Anna Pilkova[b], Ľubomír Benko[a,*], Petra Blazekova[b], Peter Svec[a]

[a] Constantine the Philosopher University in Nitra, Slovakia
[b] Comenius University in Bratislava, Slovakia

ARTICLE INFO

*Article history:*
Received 13 July 2021
Revised 29 November 2021
Accepted 30 November 2021
Available online 3 December 2021

*Keywords:*
Pillar 3
Basel II
Web usage mining
Data modeling
Multinomial Logit model

ABSTRACT

The dataset presented in this article represents the pre-processed web server log file of the commercial bank. The source of data is the web server of the bank and keeps access of web users starting the year 2009 till 2012. It contains accesses to the bank website during and after the financial crisis. Unnecessary data saved by the web server was removed to keep the focus only on the textual content of the website. Many variables were added to the original log file to make the analysis workable. To keep the privacy of website users, sensitive information in the log file were anonymized. The dataset offers a way to understand the behavior of stakeholders during and after the crisis and how they comply with the Basel regulations. The behavior of users can be modeled using the multinomial logit model, which is in detail described in the research article [1] related to this data article.

© 2021 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

DOI of original article: 10.1016/j.eswa.2021.115503
* Corresponding author.
*E-mail address:* lbenko@ukf.sk (Ľ. Benko).

https://doi.org/10.1016/j.dib.2021.107672
2352-3409/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

**Specifications Table**

| | |
|---|---|
| Subject | Computer Science (General) |
| Specific subject area | Modeling user behavior based on accesses to bank website content mandatory since Basel II Pillar 3 regulations |
| Type of data | Pre-processed log file |
| How data were acquired | Data were acquired from the web servers log files. Log files are stored as text files on the webserver and were raw copied for analysis. Using a simple script, the data were inserted/converted into an SQL database. Each line in the log file represents one row in the database table. After that, data were exported to a common CSV text file using the SQL query for exporting data. |
| Data format | Pre-processed data (.csv) |
| Parameters for data collection | Unnecessary data from the log file (access to graphic files, style sheets, java scripts and search engines robots) was removed. We used the Linux bash script to filter out lines containing specified keywords, e.g. png, jpg for images, css for stylesheets, js for javascript. To filter out access of search engines robots we searched for specific keywords for robot identification (e.g. BingBot, Googlebot, AOL, Baiduspider etc.), specific IP addresses based on a public list of robots IP addresses and access to robots.txt. |
| Description of data collection | The web server log file is created automatically by the web server when a user visits the website, and then the log file is copied for pre-processing and further analysis. |
| Data source location | Data related to Pillar 3 was obtained from the web server log files of domestic significant commercial banks operating in Slovakia. The log files are usually generated by the web server application as the text files or database entries. The exact form is depended on the technology the bank is using, which is, however, irrelevant for data analysis. The analysed data were obtained based on a request to the web server system administrator and were given as the text file. |
| Data accessibility | Munk, Michal; Pilkova, Anna; Benko, Ľubomír; Blazekova, Petra; Svec, Peter (2021), "Pillar 3: Pre-processed web server log file dataset of the banking institution", Mendeley Data, V1, doi:10.17632/5bvkm76sdc.1 |
| Related research article | M. Munk, A. Pilkova, Ľ. Benko, P. Blaželková, P. Svec, Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times, Expert Systems with Applications 185 (2021) 115503. doi:10.1016/j.eswa.2021.115503 |

**Value of the Data**

- The dataset contains data of web server log file of significant domestic commercial bank operating in Slovakia during the financial crisis and after the crisis and provides an option to analyse the stakeholders' behavior according to EU regulations.
- Researchers focused on the analysis and modeling of user behavior in the online environment can benefit from this data where the focus is on the time with examining the behavior of users in time.
- A reusable dataset for modeling the probabilities of the accesses to the web portal of the bank depending on time using the multinomial logit model which is a part of generalized linear models can be used for further insights and development of experiments.
- Dataset compromise Pillar 3 disclosure requirements quarterly and semi-annually; Pillar 3 related information; other information and describe market discipline categories regarding Basel II Pillar 3 regulations.
- Data pre-processing methods applied to this dataset can be used also in the field of educational data mining and learning analytics.

## 1. Data Description

The dataset is pre-processed standard web server log files. It contains only useful information for the analysis of user behavior. In the dataset, there are identified user sessions to distinct

visits of stakeholders. Pairs of visited part of the website and its referrer offer the way to restore the sitemap of the portal. As the source of the dataset in the web server of the commercial bank, the sensitive information is anonymized, but the researcher is still able to understand the meaning of this sensitive information. The dataset contains variables connected to the Basel II, Pillar 3 regulations. The Pillar 3 are specific regulatory disclosures requirements set out since the Basel II framework and incorporated into EU law and subsequent laws of the member states. Those regulations order the bank to publish various information and stakeholders can better understand the risk. The multinomial logit model offers the possibility to model bank website visitor behavior.

## 2. Experimental Design, Materials, and Methods

The research (in detail in Expert Systems with Applications [1]) is focused on modeling the bank visitors' behavior using a multinomial logit model [2]. The web usage analysis was done on log files obtained from the web servers of a bank institution. The log files consisted of a sample of 2 071 235 logged accesses that were obtained after data preparation. The analysis was oriented on examining the behavior of visitors over an extended period of time (2009-2012). The applied methodology to obtain and pre-process the web server log file is based on [3]. After obtaining log files from multiple servers, the data preparation was done in multiple steps.

The first step is **data cleaning**. All client requests on multimedia files, cascade styles sheet, java scripts and other non-content file types were removed using basic Linux commands. We are using a simple Linux bash script based on the grep command to filter our lines containing unnecessary keywords, e.g. png or jpg for images, js for JavaScript, css for stylesheets etc. All those requests form the web page design and are not valuable for the content analysis. The access of search engines robots was removed, too. Search engines robots are identified based on the *userAgent* or special access to robots.txt file on the webserver. We used the Linux bash script based on grep again, but we searched for robot identification, e.g. Google robots are identified as GoogleBot in the *userAgent* field, Microsoft robot is identified as Bingbot, etc. Some robots do not identify themselves but access directly to robots.txt. We search for access to robots.txt and marked down the IP of the origin to clean out all lines with this IP. We also used the public list of known IPs of search engines robots to filter them out. After the data cleaning phase, the raw data file only with accesses to html or pdf files of the portal is available. The associated file contains an already cleaned log. The algorithm for filtering data is as follows:

*Foreach record from the logfile {*
*Parse record;*
*Foreach keyword from keywords filter_out (record, keyword);*
*Foreach IP from IPS filter_out(record, IP);*
*}*

The second step in data preparation is **user/session identification**. To identify which lines of the log file, belong to the same user (visitor) and thus form a session it is needed to add fields into the log file. It was also needed to add a record to the log file. Some lines are missing because the user hits the back button in the browser. In this case, there is no request to the web server and the browser shows the content from its cache. To add missing lines the Path completion [4] was used. Visitors can be distinguished based on their IP address and the type of browser. While various users can share the same IP address or computer, the Reference Length method [5–8] for user/session was used. The published dataset does not contain a full IP address but only the first two bytes of the IP address. To keep the privacy of users according to GDPR the *anonIP* variable was created. The *anonIP* is a salted hash of the full IP. Researchers can analyze the dataset based on this hash but are not able to reveal the original IP. The first two bytes *ipPart* are enough to distinguish access of bank employees and other users. The bank employees connected from the internal bank network use the private network 10.238.x.x. To mark those users the *internal* variable was created in the dataset.

**Table 1**
Extract of the pre-processed log file.

| Frame | Agent | ipPart | anonIP | length |
|---|---|---|---|---|
| 883526 | 37885 | 81.83 | 9ad5c1b20796f7215dd25f597367df6e8a1a834c | –2 |
| 1005099 | 4528 | 85.248 | 1026980e36670a17d16758a71b6921cc304c5edd | 8 |
| 819284 | 3416 | 78.98 | 211d24985f0162c0f2d6586db18e6bf09eeb2ecd | –1 |
| 1005099 | 4528 | 85.248 | 1026980e36670a17d16758a71b6921cc304c5edd | –1 |
| 603493 | 23 | 213.151 | 19190cdb4218cd39c9505622b43a5fde8f88cdf5 | –2 |
| 73781 | 3530 | 10.238 | 141fa2a4792f00d8f4049b9cdd50417ad3dece5f | 4 |
| 73781 | 3530 | 10.238 | 141fa2a4792f00d8f4049b9cdd50417ad3dece5f | 7 |
| 73781 | 3530 | 10.238 | 141fa2a4792f00d8f4049b9cdd50417ad3dece5f | 33 |
| 519356 | 18346 | 212.5 | fd6504078435cf655f76cdb81cb920cdad3ef8b8 | –1 |
| 877098 | 376288 | 80.94 | 40be9376d4efbc23b7d06c7561164ed1d9d09832 | 99 |

The values of internal can be

- 0, if the user is located outside the bank internal network,
- 1, if the user is in the bank internal network,
- 2, if the user is from the network of IT company that maintain and develop the web site.

To use the Reference Length Method, the *anonIP, Agent,* and *unixTime* are needed. The Agent is referred to as the type of web browser and the *unixTime* represents the standard unix time in seconds starting on January 1, 1970. The result of the Reference Length Method is the session represented by the *Frame-Agent-anonIP and* the *Length.*

The *Length* represents the session length calculated using the Reference Length method and its value represents the time in seconds a user spent on the website except for

- -1, which represents the end of the session caused by overrun of the session time threshold,
- -2, which represents the end of the session caused by the change of IP address,
- 0 represents robot access (crawling the website was too quick).

Table 1 shows the *Agent* variable which is created based on the *User_agent*. We kept both variables in the dataset. The session is a combination of *Frame-Agent-anonIP*.

The third step is **variables determination** for the user behavior analysis. A bank expert was asked to categorize every webpage according to the terminology used in the bank environment. The result of this process is the 19 different parts of the web represented by the *webPart* field each attached to 6 different categories – the variable *category* (Pricing list, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements, We support..). Pillar 3 related information are those which are contained in the bank annual reports, minutes from general assembly meetings, prospect of emitent, information about group and information for banks. The Pillar 3 Disclosure Requirements consists of information on the Bank (organizational chart, information about employees or bank activities), financial information (financial statement information, information on asset quality, information on liquidity); information on risk management (risk strategies, policies, credit risk management). The webpage taxonomy is depicted in Table 2 and the detailed frequency of accesses to the web categories can be seen in Figs. 1 and 2. The algorithm for assigning variables is as follows:

*Foreach record from the logfile {*
*Parse record;*
*Get IP, Agent, Unixtime from record;*
*Calculate Internal Access from record;*
*Calculate Length from record;*
*Get category from URL;*
*If not exist category {*
*Get category from Referrer;*
*}*
*If exist category {*

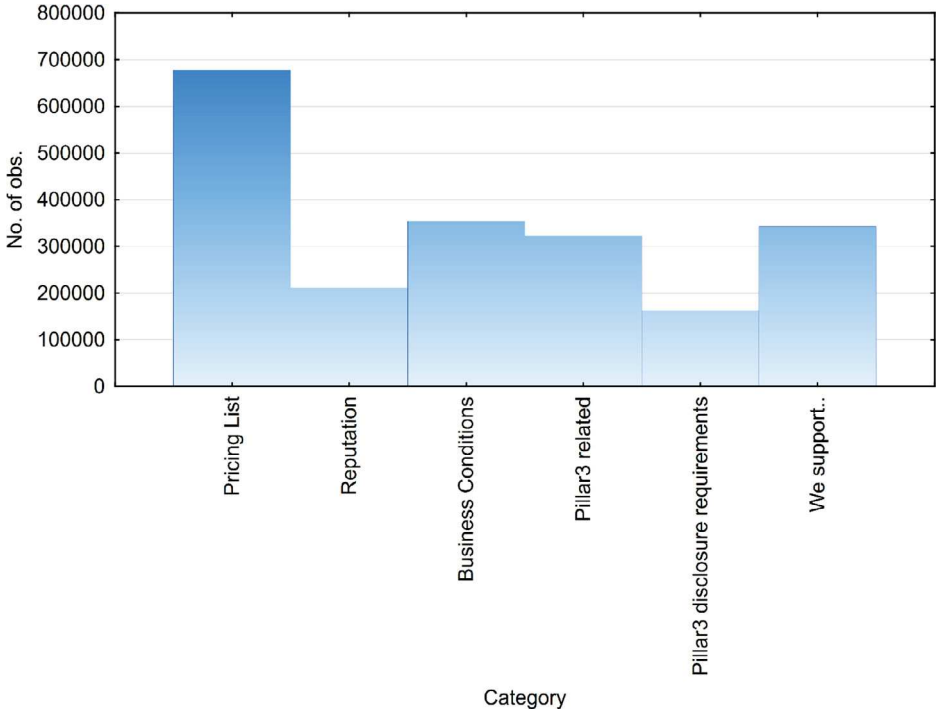



**Fig. 1.** Number of accesses to the examined web categories.

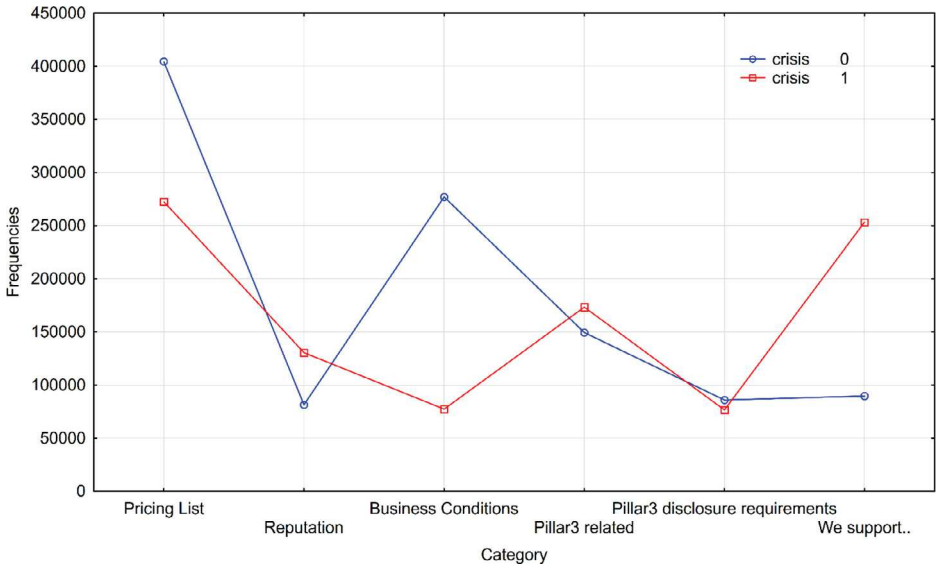



**Fig. 2.** Interaction plot for the accesses of the examined web categories.

**Table 2**
Taxonomy of the examined web portal content.

| | |
|---|---|
| /Pillar 3 disclosure requirements/ | /Pillar3 Q-terly Info/ |
| /Pillar 3 disclosure requirements/ | /Pillar3 Semi-annually Info/ |
| /Pillar 3 related/ | /Rating/ |
| /Pillar 3 related/ | /Annual reports/ |
| /Pillar 3 related/ | /Group/ |
| /Pillar 3 related/ | /Information for banks/ |
| /Pillar 3 related/ | /Emitent prospects/ |
| /Pillar 3 related/ | /General shareholder meeting/ |
| /Pillar 3 related/ | /Financial reports/ |
| /Other/ | |

*Assign webPart based on category;*
*Calculate Quartal from Timestamp;*
*}*
*}*

The independent variables-predictors were identified as the next step. Field *Week* based on ISO 8601 with values of 0-53 was created. The value of 0 represents the case a week begins in the previous year. Another three fields were added to the dataset - *Quartal, Year* and *Year Quartal* which represents the number of quartal of the year.

To distinguish the time during the financial crisis and after the financial crisis, a field was added to the dataset. The *crisis* variable identifies the period of years during a financial crisis (the value is 1) and after a financial crisis (the value is 0).

To be able to restore the sitemap from the dataset, the URL and Referrer were kept. To keep privacy, only the salted hash of URL and the salted hash of Referrer is published. The same URL has the same hash *(urlHash)* and the same Referrer has the same hash *(referrerHash).* As the hash hides the extension of the file and it is not possible to distinguish html or pdf files or just the directory part of the URL the type of content was added before the anonymization took place. The fields *urlExt* and *referrefExt* were added to the dataset.

The obtained data file after the data preparation was used for web usage analysis. The investigated categorical dependent variable category that represents a group of web parts that deal with a similar issue was used. The analysis was oriented on examining the behavior of visitors over an extended period of time (2009-2012). The years 2009-2010 represent the years of the financial crisis. On the other hand, the years 2011-2012 represent the years after the financial crisis. The time independent variable was chosen as the variable week that was created from the date of the access of the visitor. The whole methodology of the experiment is described in detail in the MethodsX article [9]. Similar data pre-processing methods can be used also in other fields such as educational data mining or learning analytics where you work with data stored in log files from virtual learning environments [10,11]

## Ethics Statement

The authors declare that the data presented in this article did not involve any use of human subjects, animal experiments nor data collected from social media platforms.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## CRediT Author Statement

**Michal Munk:** Conceptualization, Methodology, Formal analysis, Writing – review & editing; **Anna Pilkova:** Conceptualization, Writing – review & editing, Supervision; **Ľubomír Benko:** Methodology, Investigation, Writing – original draft; **Petra Blazekova:** Conceptualization, Writing – original draft; **Peter Svec:** Data curation, Resources, Writing – original draft.

## Acknowledgments

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under contract VEGA-1/0776/18 and VEGA-1/0821/21.

## References

[1] M. Munk, A. Pilkova, Ľ. Benko, P. Blažeková, P. Svec, Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times, Expert Syst. Appl. 185 (2021) 115503, doi:10.1016/j.eswa.2021.115503.
[2] M. Munk, M. Drlik, M. Vrabelova, Probability modeling of accesses to the course activities in the web-based educational system, Computational Science and Its Applications - ICCSA (2011) 485–499 Pt V.
[3] M. Munk, M. Drlík, Analysis of stakeholders' behavior depending on time in virtual learning environment, Appl. Math. Inf. Sci. 8 (2014) 773–785.
[4] J. Kapusta, M. Munk, P. Svec, A. Pilkova, Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal, Procedia Comput. Sci. 29 (2014) 1779–1790.
[5] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowl. Inf. Syst. 1 (1999) 5–32.
[6] R. Cooley, B. Mobasher, J. Srivastava, Grouping Web page references into transactions for mining World Wide Web browsing patterns, in: Proc. 1997 IEEE Knowl. Data Eng. Exch. Work., 1997, doi:10.1109/KDEX.1997.629824.
[7] J. Kapusta, M. Munk, M. Drlík, Cut-off time calculation for user session identification by reference length, in: 2012 6th Int. Conf. Appl. Inf. Commun. Technol. AICT 2012 - Proc., 2012, doi:10.1109/ICAICT.2012.6398500.
[8] J. Kapusta, M. Munk, M. Drlik, User Session Identification Using Reference Length, in: Capay, M and Mesarosova, M and Palmarova, V (Ed.), DIVAI 2012 9TH Int. Sci. Conf. DISTANCE Learn. Appl. INFORMATICS Conf. Proc, 2012, pp. 175–184.
[9] M. Munk, A. Pilkova, Ľ. Benko, P. Blažeková, P. Svec, Methodology of stakeholders' behavior modeling based on time, MethodsX 8 (2021) 101570, doi:10.1016/j.mex.2021.101570.
[10] M. Munk, M. Drlík, Ľ. Benko, J. Reichel, Quantitative and qualitative evaluation of sequence patterns found by application of different educational data preprocessing techniques, IEEE Access 5 (2017) 8989–9004, doi:10.1109/ACCESS.2017.2706302.
[11] M. Drlík, M. Munk, Understanding time-based trends in stakeholders' choice of learning activity type using predictive models, IEEE Access 7 (2018) 3106–3121, doi:10.1109/ACCESS.2018.2887057.

PRÍLOHA D: MUNK, MICHAL, ANNA PILKOVA, LUBOMIR BENKO, PETRA BLAZEKOVA A PETER SVEC, 2021A. METHODOLOGY OF STAKEHOLDERS' BEHAVIOUR MODELLING BASED ON TIME. *METHODSX*. 8, 101570. DOI:10.1016/J.MEX.2021.101570 (**WEB OF SCIENCE; SCOPUS**) [WOS: 0, SCOPUS: 1]

MethodsX 8 (2021) 101570

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

# Methodology of stakeholders' behaviour modelling based on time

Michal Munk[a], Anna Pilkova[b], Lubomir Benko[a,*], Petra Blazekova[b], Peter Svec[a]

[a] Constantine the Philosopher University in Nitra, Slovakia
[b] Comenius University in Bratislava, Slovakia

A B S T R A C T

The methods presented in this article were created to model and describe the behaviour of the web users of a bank institution web portal. The source dataset is represented by a log file of the commercial bank web server. The analysis is oriented on examining the behaviour of visitors over an extended period (2009-2012). The years 2009-2010 represent the years of the financial crisis, and the years 2011-2012 represent the years after the financial crisis. The following method describes the sequence of steps necessary to pre-process the raw log file and model the web user behaviour using the multinomial logit model. The introduced methods can be used also for other domains in the case of appropriate data preparation.

- Data preparation- data cleaning, user/session identification, path completion, variables determination;
- Data analysis- model definition, parameters estimation, logits estimation, probabilities estimation;
- Results evaluation- comparison of empirical and theoretical values in term of counts, probabilities and logits.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

A R T I C L E I N F O
*Method name:* Methodology of Stakeholders' Behaviour Modelling Based on Time
*Keywords:* Data pre-processing, Web usage mining, Multinomial logit model
*Article history:* Received 6 July 2021; Accepted 31 October 2021; Available online 2 November 2021

## Specification table

| | |
|---|---|
| Subject Area | Computer Science |
| More specific subject area | *Web Usage Mining* |
| Method name | *Methodology of Stakeholders' Behaviour Modelling Based on Time* |
| Name and reference of original method | *Web usage data pre-processing and analysis* [1,2] |
| Resource availability | *The pre-processed log file is located in Data in Brief* [3]. |

DOI of original article: 10.1016/j.eswa.2021.115503
* Corresponding author.
*E-mail address:* lbenko@ukf.sk (L. Benko).

https://doi.org/10.1016/j.mex.2021.101570
2215-0161/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Methods described in this article were created to analyse and model the behaviour of the visitors of a web portal – bank web portal for a research article [4]. The data source was log files obtained from the webservers and it contained the visitors' accesses to the web parts of the web portal. A detailed description of the log files is located in Data in Brief [3]. Data related to Pillar 3 were gathered from bank webserver log files [5]. The research methodology was inspired by [6–8]. The model of bank visitors' behaviour was created based on a multinomial logit model [9]. The web usage analysis was done based on a sample of 2 071 235 logged accesses that were obtained after data preparation. The investigated categorical dependent variable is a variable *category* that represents a group of web parts that deal with a similar issue. The variable contains these categories: *Pricing List, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements* and *We support*. The analysis is oriented on examining the behaviour of visitors over an extended period of time (2009-2012). The years 2009-2010 represent the years of financial crisis (variable *crisis=1*). On the other hand, the years 2011-2012 represent the years after the financial crisis (variable *crisis=0*). The time independent variable was chosen as the variable *week* that was created from the date of the access of the visitor. The variable was created based on the standard ISO 8601 and the variable acquires the values 0-53. If the week number equals 0, it means that the given date belongs to the preceding year. The applied methodology is based on [10] and is as follows:

1. Obtaining log files from multiple servers.
2. Data preparation involving the following multiple tasks:
   a. Data cleaning – removing the unnecessary data from the log files (requests for pictures, styles and so on; and also accesses of robots of search engines) which leads to raw data of only accesses to the web portal.
   b. User/session identification – the visitors were identified based on the variables IP address and user agent; and sessions were identified based on the Reference Length method.
   c. Path completion –used to complete the records of the users' path that the user followed using the Back button in the web browser (these visited pages are not recorded in the log file since they have already been stored on the client side under the previous steps).
   d. Variables determination – the log file contains the variables in a typical Extended Log Format (ELF), so a transformation and variable definition are needed for the user behaviour analysis of the examined web portal. A dependent variable *category* is created, and it represents the web parts of the web portal. In case the web parts have low traffic, it is appropriate to create wider categories based on their relevance to the content [9]. The variable *category* will in the case of the examined web portal of the bank institution contain the following web categories: *Pricing List, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements,* and *We support*. It is also necessary to identify independent variables- predictors that represent the time variables created from the timestamp of the access to the web category. In case of the weeks of the year, it is the variable *week* that was created based on ISO 8601 and will have values of 0-53. The variable will be 0 in case it is a week that begins in the previous year. The next predictor will be the dummy variable *crisis* that identifies the period of years during a financial crisis and after a financial crisis. Next a dummy variable *internal* was created for the identification of accesses from inside and outside of the organization's network. In this way the behaviour of users accessing from the inside/outside (internal/external access) of the organizations' network (the variable was created based on the sets of IP addresses) can be analysed.
3. Data analysis follows based on the presumption that the examined data consists of individual accesses to web portal parts:
   a. Model definition - probability distribution of accesses $Y_{ij}$ in time $i$ for the category $j$ with observations $y_{ij}$, if the count of accesses is given $n_i = \sum_j y_{ij}$ in time $i$ is multinomial $P[Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iJ} = y_{iJ}] = \frac{n_i!}{y_{i1}!y_{i2}!\ldots y_{ij}!}\pi_{i1}^{y_{i1}}\pi_{i2}^{y_{i2}}\ldots\pi_{ij}^{y_{ij}}$. Since $\sum_{j=1}^{J}\pi_{ij} = 1$ it is necessary to estimate $J-1$ of unknown probabilities. The estimates are calculated using the Maximum Likelihood method. In the logarithmic function of likelihood (without constants) $\sum_i\sum_{j=1}^{J} y_{ij}\ln\pi_{ij}$ (1) is denoted a logit transformation $\eta_{ij} = \ln\frac{\pi_{ij}}{\pi_{ij}}$, where the last category

is chosen as the reference category $\eta_{ij} = 0$, and it is assumed that the logits $\eta_{ij}$ are linear functions of the independent variables $\eta_{ij} = \alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. Using inverse transformation, it is denoted $\pi_{ij} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\eta_{ij}}}, \pi_{ij} = e^{\eta_{ij}} \pi_{ij}, j = 1,\ 2, \ldots,\ J-1$, respectively $\pi_{ij} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}_j}}, \pi_{ij} = \frac{e^{\alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}_j}}{1+\sum_{j=1}^{J-1} e^{\alpha_j + \boldsymbol{x}_i^T \boldsymbol{\beta}_j}}, j = 1,\ 2, \ldots, J-1$ (2). The logarithm function is a likelihood function with unknown parameters $\alpha_j$ and $\boldsymbol{\beta}_j$ $(j = 1,\ 2, \ldots,\ J-1)$ after substituting such expressed $\pi_{ij}$ (2) into (1). After determining the model, it is necessary to identify the type of dependence for determining the degree of polynomial and the selection of predictors, including dummy variables.

b. The estimation of the models' parameters $\alpha_j, \boldsymbol{\beta}_j$ by maximizing the logarithm of the multinomial likelihood function. The *STATISTICA Generalized Linear/Nonlinear Models* was used to estimate the parameters of individual values. The significance of parameters was tested using the Wald test H0 : $\alpha_j = 0$, H0 : $\beta_{kj} = 0$, where $k$ is the number of predictors. The estimated parameters are used to calculate estimates of logits and from logits can be calculated probabilities of selection of specific categories at a given time.
c. The estimation of logits $\eta_{ij}$ for all values of independent variables $\hat{\eta}_{ij} = a_j + \boldsymbol{x}_i^T \boldsymbol{b}_j, j = 1,\ 2, \ldots, J-1$.
d. Probability estimation of accesses $\pi_{ij}$ in time $i$ for reference web category $J$ $\hat{\pi}_{ij} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\hat{\eta}_{ij}}}$.
e. Probability estimation of accesses $\pi_{ij}$ in time $i$ for web category $\hat{\pi}_{ij} = e^{\hat{\eta}_{ij}} \hat{\pi}_{ij}, j = 1,\ 2, \ldots, J-1$.
f. Visualization of the probabilities of web category $j$ in time $i$, where $j = 1,\ 2, \ldots, J$.

4. Results evaluation:

Based on the assumption that the expected counts $\hat{y}_{ij} = n_i \hat{\pi}_{ij}$ are big enough (they are not zero and no more than 20% from $\hat{y}_{ij}$ is less than 5) to compare the actual model with the saturated model that is used to predict the probabilities independently for $i = 0,\ 1,\ \ldots,\ 53$, then the statistics $G^2$ (deviance, likelihood ratio) can be used

$$G^2 = LR\left(\hat{\pi}\right) = 2\left(L(p) - L\left(\hat{\pi}\right)\right) = 2\sum_{i=0}^{53}\sum_{j=1}^{J} y_{ij}\left(\ln p_{ij} - \ln \hat{\pi}_{ij}\right) = 2\sum_{i=0}^{53}\sum_{j=1}^{J} y_{ij} \ln \frac{p_{ij}}{\hat{\pi}_{ij}} = 2\sum_{i=0}^{53}\sum_{j=1}^{J} y_{ij} \ln \frac{y_{ij}}{n_i \hat{\pi}_{ij}}.$$

After the last edit it is following:

$$G^2 = 2\sum_{i=0}^{53}\sum_{j=1}^{J} y_{ij} \ln \frac{y_{ij}}{\hat{y}_{ij}}.$$

The hypothesis H0 : $\pi_{ij} = \hat{\pi}_{ij}$ can be testes using the LR test. LR test makes it possible to compare the estimates $\hat{y}_{ij}$ with $y_{ij}$. The saturated model has $54(J-1)$ free parameters and the current model $k(J-1)$, then the degrees of freedom $df$ are equal $(54-k)(J-1)$. Statistics $G^2$ has approximately $\chi^2(df)$ distribution. Pearson statistics can be used to compare the estimates $\hat{y}_{ij}$ with $y_{ij}$ either:

$$\chi^2 = \sum_{i=0}^{53}\sum_{j=1}^{J} r_{ij}^2,$$

Where $r_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{y}_{ij}}}$ is Pearson residue that has also the $\chi^2(df)$ distribution.

In the given application field is often the condition of using the LR test/ Pearson statistics violated. Usually, the examined variable has a considerable number of levels that are web parts of the portal or system (pages, content categories, activities, etc.). This results in the violation of using the LR test/Pearson statistics- the expected counts are not large enough. For this reason are used alternative methods to evaluate the model [9,11] - visualization of empirical and theoretical counts differences, extreme identification, comparison of the distribution of empirical relative counts of accesses and estimated probabilities of the examined web part $j$ in time $i$, and empirical and theoretical logit visualization of each web part, except the reference web part.

The created model was evaluated based on the following steps:

a. Empirical counts determination $y_{ij}$.
b. Theoretical counts estimation $\hat{y}_{ij} = \hat{\pi}_{ij} \sum_j y_{ij}$.
c. Visualization of differences in the empirical and theoretical counts of accesses $d_{ij} = y_{ij} - \hat{y}_{ij}$.
d. Extreme values identification $d_{ij}$, where $d_{ij} > \bar{d}_j + 2s_j$ represents an underestimated prediction and $d_{ij} < \bar{d}_j - 2s_j$ represents overestimated prediction where $s_j$ is standard deviation and $\bar{d}_j$ is the mean of differences of the category $j$.
e. Calculation of relative empirical counts of accesses $p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}}$.
f. Comparison of the distribution of the relative empirical counts of accesses with the estimated probabilities of the selected web category $j$ in time $i$. To test the zero hypothesis, dividing the pair differences is symmetric around zero $r_{ij} = p_{ij} - \pi_{ij}$, H0: $F(-r) = 1 - F(r)$, a Wilcoxon pair test will be used.
g. Calculation of empirical logits $h_{ij} = \ln(\frac{p_{ij}}{p_{ij}}), j = 1,\ 2, \ldots, J-1$.
h Visualization of empirical and theoretical logits for individual web categories except for the reference one.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0776/18 and VEGA-1/0821/21.

## References

[1] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowl. Inf. Syst. 1 (1999) 5–32.
[2] B. Liu, Web data mining: exploring hyperlinks, contents, and usage data, 2011. doi:10.1007/978-3-642-19460-3.
[3] M. Munk, A. Pilkova, Ľ. Benko, P. Blažeková, P. Svec, Pillar 3: pre-processed web server log file dataset of the banking institution, Data Br. (2021) Submitted for publication.
[4] M. Munk, A. Pilkova, Ľ. Benko, P. Blažeková, P. Svec, Web usage analysis of pillar3 disclosed information in turbulent times, Expert Syst. Appl. 185 (2021) 115503 (in press), doi:10.1016/j.eswa.2021.115503.
[5] M. Munk, A. Pilková, M. Drlík, J. Kapusta, P. Švec, Verification of the fulfilment of the purposes of basel ii, pillar 3 through application of the web log mining methods, Acta Univ. Agric. Silvic. Mendelianae Brun. 60 (2012).
[6] M. Munk, J. Kapusta, P. Švec, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor, Procedia Comput. Sci. (2010) 2273–2280, doi:10.1016/j.procs.2010.04.255.
[7] M. Munk, J. Kapusta, P. Švec, M. Turčáni, Data advance preparation factors affecting results of sequence rule analysis in web log mining, E a M Ekon, a Manag 13 (2010).
[8] M. Munk, Ľ. Benko, M. Gangur, M. Turčáni, Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining, E+M Ekon, a Manag 18 (2015) 144–159 doi:dx.doi.org/, doi:10.15240/tul/001/2015-3-013.
[9] M. Munk, M. Drlik, M. Vrabelova, Probability Modelling of Accesses to the Course Activities in the Web-Based Educational System, Comput. Sci. Its Appl. - Iccsa (2011) 485–499 Pt V, 2011.
[10] M. Munk, M. Drlík, Analysis of stakeholders' behaviour depending on time in virtual learning environment, Appl. Math. Inf. Sci. 8 (2014) 773–785.
[11] M. Munk, M. Vrábelová, J. Kapusta, Probability modeling of accesses to the web parts of portal, Procedia Comput. Sci. 3 (2011) 677–683, doi:10.1016/j.procs.2010.12.113.

PRÍLOHA E: PILKOVÁ, ANNA, MICHAL MUNK, PETRA BLAŽEKOVÁ A ĽUBOMÍR BENKO, 2021B. WEB USAGE ANALYSIS: PILLAR 3 INFORMATION ASSESSMENT IN TURBULENT TIMES. IN: MOHAMMAD Z. ABEDIN, KABIR HASSAN, PETR HAJEK A MOHAMMED M. UDDIN, ED. *THE ESSENTIALS OF MACHINE LEARNING IN FINANCE AND ACCOUNTING*. ROUTLEDGE, S. 24. DOI:10.4324/9781003037903 (**SCOPUS**) [SCOPUS: 1]

# Web Usage Analysis: Pillar 3 Information Assessment in Turbulent Times

Anna Pilkova[a], Michal Munk [b,c], Petra Blazekova[a] and Lubomir Benko[c]

[a]Comenius University in Bratislava, Address: Odbojárov 10, SK-820 05, Bratislava, Slovakia; email: anna.pilkova@fm.uniba.sk, pblazekova@vub.sk [b]University of Pardubice, Address: Studentská 84, CZ-532 10, Pardubice, Czech Republic; [c]Constantine the Philosopher University in Nitra, Address: Tr. A. Hlinku 1, SK-949 74, Nitra, Slovakia; email: mmunk@ukf.sk, lbenko@ukf.sk

**ARTICLE HISTORY**
Compiled October 22, 2020

**ABSTRACT**
This study analyses the interest in information disclosures aimed at a specific type of stakeholders in foreign owned commercial banks, not traded on the capital market. The main goal is a) to assess interests of depositors in the disclosed two groups of information: Pillar 3 disclosure requirements, Pillar 3 related information; b) to conduct robustness checks by verifying results applying two approaches based on different time variables. The chapter deals with a comparison of two different approaches to model the behaviour of web users. The results suggest that changes in information disclosures' design in commercial banks operating according to the analysed model are inevitable to enhance the efficiency of market discipline mechanisms and to add value to key stakeholders.

**KEYWORDS**
Risk Management; Market Discipline; Financial Regulation; Data Science in Finance; Web Usage Mining; Data Pre-processing; Association Rule Analysis; Multinomial Logit Model.

## 1. Introduction

The stability of the financial system is a long-term challenge for governments, regulators and academics. The 2007 – 9 financial crisis proved stability's importance and highlighted the weaknesses of the financial regulatory system around the world which was not able to avoid the failures and losses generated by the turbulence in the banking industry. Regulators, policy makers and academics learnt many lessons from this period and have attempted to fix the identified weaknesses. One of these areas is market discipline. There are different definitions of this term and one of them is a mechanism to used by market participants to discipline risk-taking by financial institutions. Basel supervisors stressed this mechanism's importance in the Basel II architecture (BCBS, 2006) by the introduction of Pillar 3 components that complemented the other two pillars: Pillar 1 – minimum risk-based capital requirements and other quantitative requirements; Pillar 2 – supervisory review processes. At that time, it aimed to provide

CONTACT Michal Munk. https://orcid.org/0000-0002-9913-3596. Email: mmunk@ukf.sk

meaningful regulatory information to market participants on a consistent basis and to be able to assess a bank's risk. However, the 2007 – 9 financial crisis additionally exposed some weaknesses in Basel II and also with Pillar 3. In Basel III (a revision of Basel II after the financial crisis) Pillar 3 started to become more detailed, structured and with information more frequently disclosed. Lengthy discussion among regulators and stakeholders on the new content, structure and frequency of information disclosure just confirmed some authors' opinions that it would have been more accurate to label Pillar 3 "information disclosure" rather than "market discipline" (Flannery & Bliss, 2019). Above all in Europe the ideas on global financial stability and cross border banking have to be achieved through centralization in the European Central Bank (Miklaszewska & Pawłowska, 2014). All in all, the current version of Pillar 3 is highly standardised with limited room for banks regarding flexibility of the content, structure and frequency of reporting. Due to high standardization and low differentiation disclosure reporting, according to Pillar 3, for some banks it can be a very costly process which does not incentivize their stakeholders to behave in such a manner that would discipline banks. It is the case with banks when insured deposits are major items in their liabilities. According to research findings (Flannery & Bliss, 2019), insured depositors have no incentive to spend resources on monitoring. In addition to that, they are probably not sophisticated enough to interpret bank results correctly. Other specific cases are also banks with majority ownership (many times more than 90%) by single foreign bank/financial group and when key stakeholders are uninsured depositors. These stakeholders have an interest in information about banks to be able to monitor the bank's risk. To cover their requirements, it is important to establish which information they are particularly interested in and furthermore to design the most effective structure regarding both content and time points of disclosed information. This study analyses the interest in information disclosures aimed at a specific type of stakeholders (depositors) in foreign owned commercial banks, not traded on the capital market. These types of banks can represent a "model" in CEE countries where many commercial banks have similar ownership and liability structures. Stakeholders' behaviour related to Pillar 3 disclosed information can be expected to be similar in these countries.

The main goal is as follows: firstly, to assess interests of depositors in the disclosed two groups of information: Pillar 3 disclosure requirements, Pillar 3 related information during the period 2009 – 2012 (year of crisis and subsequent years). Secondly, to conduct robustness checks by verifying results applying two approaches based on different time variables: week, quarter during 2009 – 2012.

The study has the following structure: the first section contains market discipline status of the research; methodology of the research is introduced in the second section; results are dealt with in the third section where the outcomes are compared based on different time frameworks; subsequently discussion and conclusion is dealt with in the last section.

## 2. Related Work

Market discipline mechanisms work via the prospect of failure and financial losses to minimise risk-taking, by which market participants discipline banks for their risk-taking behaviour. The market disciplining effect has been broadly identified by many studies and also by substantial evidence concerning market participants' reactions (Distinguin, 2008; Evanoff & Wall, 2000; Hadad, Agusman, Monroe, Gasbarro &

Zumwalt, 2011; Jagtiani, Kaufman & Lemieux, 1999; Jordan, Peek & Rosengren, 2000; Sironi, 2003). Market discipline as a concept is evaluated empirically and factors which influence its enhancement have also been analysed. This concept has been analysed from different perspectives, which also point out its weak points. It is important to note, that these frailties are important generally, but their impact on financial markets can be increased in turbulent times. Consequently, market discipline can be weakened by implicit government guarantees such as insured deposits (Distinguin, Rous & Tarazi, 2006), and by negative aspects of financial regulation highlighted by the crisis (Calomiris, 2009). Substantially, this involves disclosure standards, which reinforce an accurate reaction to turbulent times in the financial markets. On the other hand, Cubillas et al. (Cubillas, Fonseca & González, 2012) conclude that while market discipline is weakened by banking crises and policy implications, regulations and interventions strengthen market discipline. At this point, these outcomes of disclosures as a market disciplining tool can differ, they are important in the functioning of market discipline mechanisms and as a background for market discipline enhancement. Furthermore, recapitalization and forbearance have negative effects on market discipline, but less supervisory power and more private ownership and supervision of banks have opposite effects. Generally, market discipline's efficiency is important and mainly during crises, due to stronger risk-taking incentives which can be eliminated by market discipline (Nier & Baumann, 2006).

Market discipline framework is important in the concept of market discipline. This framework may represent a functional system and is a key component in modern banking regulation (Bartlett, 2012). However, this is true only in the case when the four following blocks are in perfect coherence (Stephanou, 2010): information disclosure, market participants, discipline mechanism and internal governance. Moreover, the interaction of these four blocks influences the market discipline effectiveness, which depends on the enhancement of accurate and timely financial disclosures (Jagtiani & Lemieux, 2001) and market disclosure of private information that penetrates the market (Berger & Davies, 1998).

There is no doubt that disclosures are one of the most effective tools for the enhancement of market discipline (Fonseca & González, 2010) and serve as a macro prudential tool in reducing uncertainty in the capital markets during a financial crisis (Ellahie, 2012; Peristiani, Morgan & Savino, 2010). Moreover, Sowerbutts (Sowerbutts, Zer & Zimmerman, 2013) conclude that disclosures' mechanism failure contributed to the last financial crisis, because of inadequate public disclosure that was followed by the inability of investors to judge risk and the withdrawal of lending in times of systemic stress. Therefore, a few studies have been reviewed in which the authors concentrate on the factors of disclosures, which contribute to the market discipline efficiency. Most of the authors concentrate on the impact of the increase of information disclosures on commercial banks. According to Bouaiss et al. (Bouaiss, Refait-Alexandre & Alexandre, 2017), the increase in disclosure levels enhances transparency and efficient market discipline and supervises excessive risk-taking. It positively influences investor's attitude to banks' risk profiles and actively increases banks' value (Zer, 2015). Furthermore, it increases the ability to attract interbank funding (Guillemin & Semenova, 2018), boosts depositors' sensitivity to equity levels (Kozłowski, 2016), improves sensitivity to risk-taking (Goldstein & Sapra, 2014) and prevents market breakdown. Additionally, an increase in disclosures is connected with a reduction of risk-taking by commercial banks (Naz & Ayub, 2017) and with a lower probability of default (Li, Li & Gao, 2020). However, it also implies the potential threat of disclosing too much information, which destroys risk-sharing opportunities (Goldstein & Leitner, 2015).

The nature of the disclosure content as a base for adequate, accurate and timely disclosures depends on the level of transparency, which is connected to the enhancement of market discipline. Accordingly, the bank stability and the probability of falling into crisis is influenced by transparency that increases accountability and leads to greater market efficiency (Gandrud & Hallerberg, 2014; Nier, 2005) and enables banks to raise cheaper capital (Frolov, 2007). However, according to Moreno and Takalo (Moreno & Takalo, 2016) conclude that only an intermediate level of transparency is socially optimal and effective (Bouvard, Chaigneau & Motta, 2015). Key explanation of this finding is statement that more transparency can decrease efficient liquidity and increase rollover risk. This development has negative impact on the bank's prices of shares and the cost of trading corporate bonds can decrease (Parwada, Lau & Ruenzi, 2015). This is in conjunction with Iren et al. (Iren, Reichert & Gramlich, 2014) that concludes that transparency can have a positive impact on bank performance only up to some level.

After the last financial crisis market discipline has become a background for stable financial markets and its implementation by regulatory requirements was preceded by a complex discussion process, which has led to significant changes and improvement of Pillar 3. The Pillar 3 disclosure requirements enhance the efficiency of market discipline in order to achieve a resilient banking system. Numerous studies evaluate Pillar 3 disclosures as a market disciplining tool and the majority of authors concentrate on the benefits of Pillar 3 disclosures as effective market disciplining tool. Firstly, Pillar 3 improves the safety of the banking system (Vauhkonen, 2012), decreases information asymmetry (Niessen-Ruenzi, Parwada & Ruenzi, 2015) and its quarterly reporting is useful to investors (Parwada, Ruenzi & Sahgal, 2013). Secondly, banks' adequate disclosures (Pillar 3 and annual reports) have a significant effect on market risk-taking behaviour and can minimize risks (Sarker & Sharif, 2020). Thirdly, the similarity of Pillar 3 regulation and COREP (Common Reporting Standards (COREP) used in Europe) reports leads to a positive relationship between these regulations and market discipline effectiveness (Yang & Koshiyama, 2019). On the other hand, a few authors concentrate on Pillar 3 weaknesses. Concretely, implementation of additional regulations is effective in the area of financial reporting quality, but there arise differences for smaller banks, which face disproportionately higher increases in the costs of compliance but big banks can be engaged in higher risk acceptance (Poshakwale, Aghanya & Agarwal, 2020). Correspondingly, Freixas and Laux (Freixas & Laux, 2011) in their study put doubt about the transparency of Pillar 3 reports, mainly thanks to stories that occurred during the financial crisis. Moreover, Pillar 3 has excessive superstructure and monitoring costs (Benli, 2015), which can bring a range of issues.

The content of Pillar 3 disclosures is also an important factor in any assessment of the efficiency of Pillar 3 as a market disciplining tool. According to Giner et. al. (Giner, Allini & Zampella, 2020), the most relevant categories of Pillar 3 disclosures are credit risk, liquidity risk and market risk category (Scannella, 2018). This is in conjunction with Bischof et al. (Bischof, Daske, Elfers & Hail, 2016), who conclude that the improved content of Pillar 3 disclosures translates into higher market liquidity. Scannella and Polizzi (Scannella & Polizzi, 2019) concentrated on improvement in the disclosure of derivatives and hedging strategies, which is important for the enhancement of interest of the market participants. Additionally, IT risk popped up as one of the key risk categories that is positively correlated with a firm's future stock price decrease (Song, Cavusoglu, Lee & Ma, 2020).

However, the related work review suggests a lack of studies assessing the interest of stakeholders to the content of Pillar 3 formation disclosures in commercial banks,

which is crucial in order to implement effective supervisory market discipline.

Firstly, the sensitivity of stakeholders to negative content in disclosures is validated by a few authors. It can trigger inefficient bank runs (Faria-e-Castro, Martinez & Philippon, 2017) and qualitative information disclosed in a negative way that contributes to explain the bank risk insolvency and increase probability of default (Del Gaudio, Megaravalli, Sampagnaro & Verdoliva, 2020). Secondly, Araujo and Leyshon (de Araujo & Leyshon, 2016) revealed a window in the content relevancy of disclosures in relation to banks' risk profile, because stakeholders are most responsive to information related to the value of the bank's assets, off-balance sheet, and ratings. These authors also highlight important factors, which influence Pillar 3 content efficiency, which are the overall quality of risk disclosures, and valuation of quantitative information more than qualitative information.

Research in the CEE region focused on the analysis of disclosures is rare. A few authors concentrate on their weaknesses (Bartulovic & Pervan, 2012). According to Arsov and Bucenska (Arsov & Bucevska, 2017), disclosures in CEE countries lack transparency and independent verification of the data in the presented reports (Habek, 2017). Moreover, for some financial institutions in Poland, there is a lack of resources strongly related to disclosures (Fijalkowska, Zyznarska-Dworczak & Garsztka, 2017). But it is clear that research in the field of Pillar 3 disclosures in CEE countries is even more rare. Despite Matuszaka's and Rozanska's interesting study (Matuszak & Rozanska, 2017), which points out the Pillar 3 benefits to financial the performance of commercial banks in Poland (positive relationship between Pillar 3 disclosure and banks profitability measured by ROA and ROE of commercial banks), research, which would evaluate its content relevancy, is insufficient.

As has been already stated in the introduction, depository markets are important in CEE countries. Research studies identify a few challenges these markets cope with. According to Distinguin et al. (Distinguin, Kouassi & Tarazi, 2012), it is high correlation between the level of interbank deposits and risk of the bank. Higher proportion of interbank deposits in the bank's balance sheet means lower levels of risk in this region. They also conclude that explicit deposit insurance (implemented in the 1990's) contributed to effective market discipline in CEE. Moreover, Karas et al. (Karas, Pyle & Schoors, 2013) discovered that in Russia the introduction of deposit insurance for households caused lower sensitivity for insured deposits flows than for uninsured ones. According to researchers the uninsured depositors have stronger market discipline than insured ones. On the other hand, Lapteacru (Lapteacru, 2019) suggests that external support has no impact on non-deposit funding at any type of banks according to ownership. Hasan et al. (Hasan, Jackowicz, Kowalewski & Kozłowski, 2013) presents a relevant finding for banks owned by foreign investors (which is the case in many CEE banks): a more positive correlation in interest to negative rumours on the banks' parent companies than to banks' disclosures. This is in conjunction with Accornero and Moscatelli (Accornero & Moscatelli, 2018), who concluded that depositors of threatened banks used to be more sensitive than other depositors to negative rumours. Substantially, it can be agreed with Berger and Bouwman (Berger & Bouwman, 2013) that particularly depository discipline research in Europe is not sufficient and according to the findings even more rare in CEE countries. At this point it can be agreed also with authors (Miklaszewska & Pawłowska, 2014) who questioned post-crisis regulatory architecture, especially for CEE banks in the competitive and unstable environment, which may produce negative effects on relatively stable CEE banks. Based on these findings this study aims to cover the gap in the field of adequate and relevant Pillar 3 disclosures, which would also contribute to higher interests of stakeholders to enhance

the efficiency of market discipline.

## 3. Research methodology

The methods described in this chapter were created to analyse the behaviour of the visitors to a web portal – a bank web portal. The source data comes from webservers that are set as load balancers and are saved in an extended log file format. It contains only useful information for the analysis of user behaviour. In the log file (Munk, Pilková, Drlik, Kapusta & Švec, 2012), are identified user sessions to distinct visits of stakeholders. The log file contains variables connected to the Basel II, Pillar 3 regulations. The Pillar 3 are exact regulatory disclosures requirements set out in the Basel II framework and incorporated into EU law and the subsequent laws of the EU states. Those regulations order the bank to publish various information and stakeholders can better understand the risk. The log file consists of around two million logged accesses that were obtained after data preparation. The data preparation is very important as was proved in previous experiments (Drlik & Munk, 2019; Munk, Kapusta, Švec & Turčáni, 2010; Munk, Drlik & Vrabelova, 2011; Munk, Benko, Gangur & Turčáni, 2015; Munk, Kapusta & Švec, 2010) where bad data preparation can lead to different results from the analysis.

The chapter deals with a comparison of two different approaches to model the behaviour of web users. Both approaches deal with the time variable as an indicator of the analysis. The first approach deals with the analysis of weekly accesses to web categories of banking portal. The multinomial logit model is used to analyse the data. The second approach deals with the evaluation of frequent itemsets based on quantity. The itemsets were evaluated based on quarters. Both approaches deal with the period 2009 – 2012. The year 2009 represents the year of the financial crisis. Contrary, the years 2010 – 2012 represent the years after the financial crisis. The investigated categorical dependent variable is a variable category that represents a group of web parts that deal with a similar issue. The variable contains these categories of the web content: *Business Conditions, Pricing List, Pillar3 related, Reputation, Pillar3 disclosure requirements* and *We support*. In this experiment, the focus will be on two of these categories that are related to the topic of Pillar3: *Pillar3 related* and *Pillar3 disclosure requirements*. The *Pillar3 related* category consists of parts: Rating, Group, Information for Banks, Annual Reports, General Shareholder Meeting, Financial Reports, and Emitent Prospects. The *Pillar3 disclosure requirements* consists of parts: Pillar3 Semiannually Info and Pillar3 Q-terly Info. Detailed analysis of web parts frequent itemsets of the Pillar3 category (*Pillar3 disclosure requirements, Pillar3 related*) in the respective quarters in the examined period was done in (Munk, Pilkova, Benko & Blažeková, 2017). The applied methodology has a similar data preparation phase for both approaches:

(1) retrieving log files.
(2) data pre-processing consisting of the following tasks:
(3) data cleansing – unnecessary data is removed from the log file (requests on pictures, fonts, styles, scripts, etc.). Search engines robots' access to the web portal are also removed from the log file. The result of this phase is raw data that contains only accesses to the web portal.
(4) user/session identification – Reference Length method was used to identify sessions and the visitors who accessed the web were identified using the fields IP

address and user agent (Cooley, Mobasher & Srivastava, 1999; Kapusta, Munk & Drlik, 2012a,b).

(5) path completion – based on the visitors usage of the *Back* button of the web browser, the records of his/her path on the web portal can be reconstructed (Munk et al., 2015).

(6) variables determination – the log file contains the variables in a typical Extended Log Format (ELF), so a transformation and variable definition are needed for a user behaviour analysis of the examined web portal. A dependent variable category is created, and it represents the web parts of the portal. In case the web parts have low traffic, it is appropriate to create wider categories based on their relevance to the content (Munk, Drlik et al., 2011). It is also necessary to identify independent variables - predictors that represent the time variables created from the timestamp of the access to the web category. In the case of the weeks of the year, it is the variable *week* that was created based on ISO 8601 and will have values of 0-53. The variable will be 0 in case it is a week that begins in the previous year. The next variable will be a nominal variable year representing individual years- 2009, 2010, 2011, and 2012. From the nominal variable dummy variables representing the examined years will be created by binarization. The next nominal variable *quarterYear* served to make the quarters and specific years distinct. Similarly, dummy variables representing the specific quarter and year (2009Q1, 2009Q2, ...) were created.

After data preparation, the experiment, divided into two approaches, is done. One method is focused on the time variable week and the other quarter. The first analysis is done using the multinomial logit model. After determining the model, it is required to identify the type of dependence for determining the degree of the polynomial and the selection of predictors, including dummy variables. The first approach was done as follows:

(7) the estimation of the models' parameters $\alpha_j, \beta_j$ by maximizing the logarithm of the likelihood function. The *STATISTICA Generalized Linear Models* was used to estimate the parameters of individual values.

(8) the estimation of logits $\eta_{ij}$ for all predictors values $\hat{\eta}_{ij} = a_j + x_i^T b_j, j = 1, 2, ..., J-1$.

(9) estimation of probability of accesses $\pi_{iJ}$ in time $i$ for reference web category $J$ $\hat{\pi}_{iJ} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\hat{\eta}_{iJ}}}$.

(10) probability estimation of accesses $\pi_{ij}$ in time $i$ for web category $\hat{\pi}_{ij} = e^{\hat{\pi}_{ij}} \hat{\pi}_{iJ}, j = 1, 2, ..., J-1$.

(11) visualization of the probabilities of web category $j$ in time $i$, where $j = 1, 2, ..., J$.

The model was afterwards evaluated using alternative methods to evaluate the model (Munk, Drlik et al., 2011; Munk, Vrábelová & Kapusta, 2011). The evaluation consisted of the visualization of observed and expected counts differences, extreme identification, comparison of the distribution of observed relative counts of accesses and estimated probabilities of the examined web part $j$ in time $i$, and observed and expected logit visualization of each web part, except the reference web part. If the model is suitable on all levels then it is a suitable model for the analysed data.

The second approach dealt with discovering the behaviour patterns of web users during quarters in the examined period. The results were processed by association rule analysis using *STATISTICA Sequence and Association Analysis*, which is an implementation of the algorithm using apriori algorithm together with a tree-structured

procedure that requires only one pass through the data (Hill & Lewicki, 2013). The aim was to extract frequent itemsets with the min support of 0.01 (Pilkova, Munk, Švec & Medo, 2015). After extracting web part frequent itemsets of web parts in the identified sessions the interest is in comparison to the proportion of incidence in the quarters of examined years. It can be summed up into the following:

(12) extracting the frequent itemsets.
(13) incidence matrix in the examined periods.
(14) assessment of seasonality in terms of quantity (occurrence rate) in the examined years.

The results of the two different approaches will be compared based on the specified time variable. It can be assumed that the results should be similar and that the weekly analysis should offer a more detailed look at the behaviour of the web users in comparison to the quarterly analysis.

## 4. Results

Frequent itemsets (a) were extracted and probabilities of the accesses (b) were examined of web portal categories (*category*) based on time, where time was represented by the variable (a) quarter and (b) week. The data saved in the log file originated from a significant domestic commercial bank operating in Slovakia. The examined log file was pre-processed, and variables were created that represented the analysed factors. Firstly, it was important to determine whether it was significant to distinguish the individual years (variable year). In the case of the nominal variable *year*, a moderate degree of dependency with the variable *category* (*Chi-square = 389 844.7; df = 15; p = 0.000; Contingency coefficient C = 0.4; Cramer V = 0.3*) was identified. The contingency coefficient can obtain values from 0 (represents no dependence between variables) to 1 (represents the perfect dependence between variables). The contingency coefficient is statistically significant. Based on these results dummy variables were created representing the examined years (*2009, 2010*, and *2011*). These variables gain only two values: 0 or 1 meaning whether the access was done in the specific year. The dummy variable for the year 2012 is not needed as the accesses from this year will have all other variables values set to 0.

Based on the Likelihood-ratio (LR) test (Table 9.1) estimates of the theoretical counts of accesses were compared with the empirical counts of accesses. The results of the LR test helped to identify the appropriate polynomial model of the third-degree for the time variable week. The value of the Pearson Chi-square approximates 1 and it means that the chosen model is suitable. Also, the maximum of the logarithm of the likelihood function helps to choose the appropriate model where the smallest value is the best.

**Table 9.1.** Evaluation of the model

| | **df** | **Stat** | **Stat/df** |
|---|---|---|---|
| Deviance | 10356140 | 6473096.45 | 0.625049 |
| Pearson Chi-square | 10356140 | 10445031.93 | 1.008584 |
| Log-likelihood | | -3236548 | |

The *STATISTICA Generalized Linear Models* was used to estimate parameters for individual data. The significance of the parameters was tested using the Wald test. The probability of access to the web portal categories has been modelled depending

on time- week of access and years. Time was represented by the predictor *week* and its transformation based on the degree of the polynomial (*week*$^2$ and *week*$^3$) and the dummy variables of the examined years (*2009, 2010*, and *2011*).

Based on the results (Table 9.2) of all effects test for the model, the parameters are statistically significant. In the created model, all of the years represent statistically significant features that are represented by the dummy variables. The weeks of the year represented by the variables week and its transformation based on the degree of polynomial showed also statistically significant features.

**Table 9.2.** All effects test for the model

| | df | Wald Statistic | p |
|---|---|---|---|
| **Intercept** | 5 | 98888.3 | <0.001 |
| **week** | 5 | 19440.9 | <0.001 |
| **week$^2$** | 5 | 25555.8 | <0.001 |
| **week$^3$** | 5 | 22826.6 | <0.001 |
| **2009** | 5 | 218115.7 | <0.001 |
| **2010** | 5 | 109030.4 | <0.001 |
| **2011** | 5 | 80968.1 | <0.001 |

The estimated parameters for both categories were significantly dependent on the week of access and for its transformations too (Table 9.3). The values of logits were significantly influenced by the examined years. The logit model provides a probability estimate at the output. The absolute size of the parameters reflects predictors with the highest influence on the examined variable. A high absolute value of the parameter refers to a large dependency. A negative value refers to indirectly proportional dependence.

**Table 9.3.** Estimate the parameters of the model

| | Category | Estimate | Standard Deviation | Wald Statistic | p |
|---|---|---|---|---|---|
| **week** | Pillar3 related | -0.07695 | 0.001659 | 2151.0 | <0.001 |
| **week$^2$** | Pillar3 related | 0.00324 | 0.000078 | 1707.9 | <0.001 |
| **week$^3$** | Pillar3 related | -0.00003 | 0.000001 | 968.8 | <0.001 |
| **2009** | Pillar3 related | -1.33753 | 0.008442 | 25100.1 | <0.001 |
| **2010** | Pillar3 related | -0.78993 | 0.008914 | 7852.6 | <0.001 |
| **2011** | Pillar3 related | 0.36939 | 0.009105 | 1645.9 | <0.001 |
| **week** | Pillar3 disclosure requirements | 0.00299 | 0.002050 | 2.1 | 0.1443 |
| **week$^2$** | Pillar3 disclosure requirements | -0.00106 | 0.000096 | 121.8 | <0.001 |
| **week$^3$** | Pillar3 disclosure requirements | 0.00002 | 0.000001 | 382.3 | <0.001 |
| **2009** | Pillar3 disclosure requirements | -1.78236 | 0.009811 | 33002.1 | <0.001 |
| **2010** | Pillar3 disclosure requirements | -1.10608 | 0.010141 | 11896.1 | <0.001 |
| **2011** | Pillar3 disclosure requirements | -0.60084 | 0.010140 | 3510.9 | <0.001 |

Using the estimated parameters, it was possible to estimate the logits for each category $j$ in time $i$. The third-degree polynomial model:

$$\hat{\eta}_{ij} = \alpha_j + \beta_{1j} week_i + \beta_{2j} week_i^2 + \beta_{3j} week_i^3 + \gamma_j year_i, i = 0, 1, 2, ..., 53, j = 1, 2, ..., J-1.$$

Evaluation of the suitability of the model was done. The importance of thorough data preparation can be shown in the following example. The log file contained a big sample of unnecessary data that was not discovered during the data preparation phase. The evaluation of theoretical and empirical counts of accesses helped to identify this issue. During a specific week in 2012 a systematic error was identified as having occurred. It was an automated script that could be related to maintenance, backup, etc. This was identified by examining the extreme values of differences between the theoretical and empirical counts of accesses. As can be seen in Figure 9.1 during the 21$^{st}$ week of the year 2012 there was a high extreme value (extreme value border is

depicted using the dashed line). This led to a more detailed analysis of the analysed log file and finding the issue.

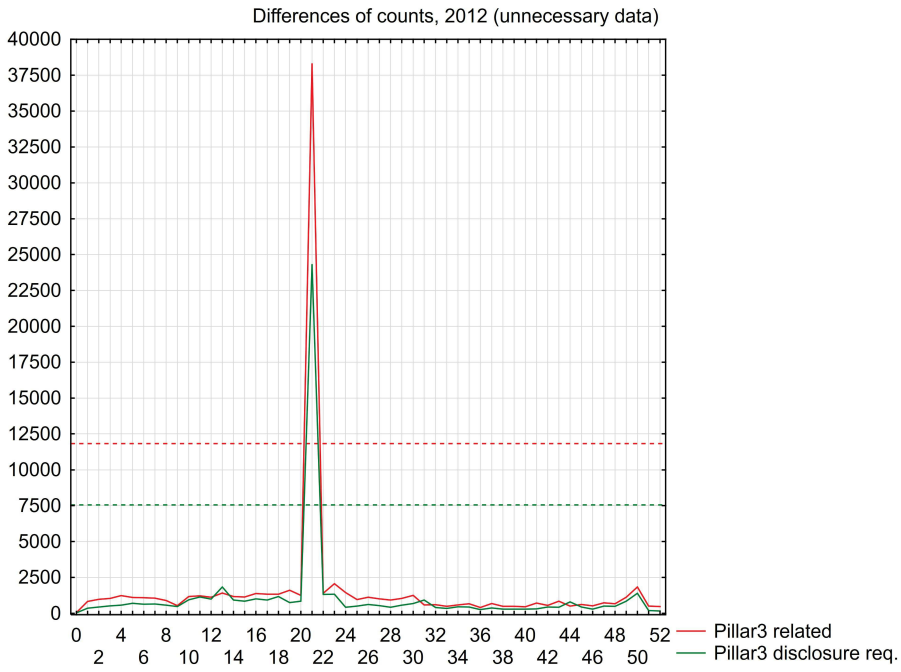



**Figure 9.1.** Counts differences of the model with error data

After cleaning the unnecessary data from the log file, the evaluation was repeated, and the difference is shown in Figure 9.2. All of the estimated parameters already mentioned were done using the corrected log file, but it was meaningful to mention that sometimes it is possible to discover issues with the data preparation phase at the end of model evaluation.

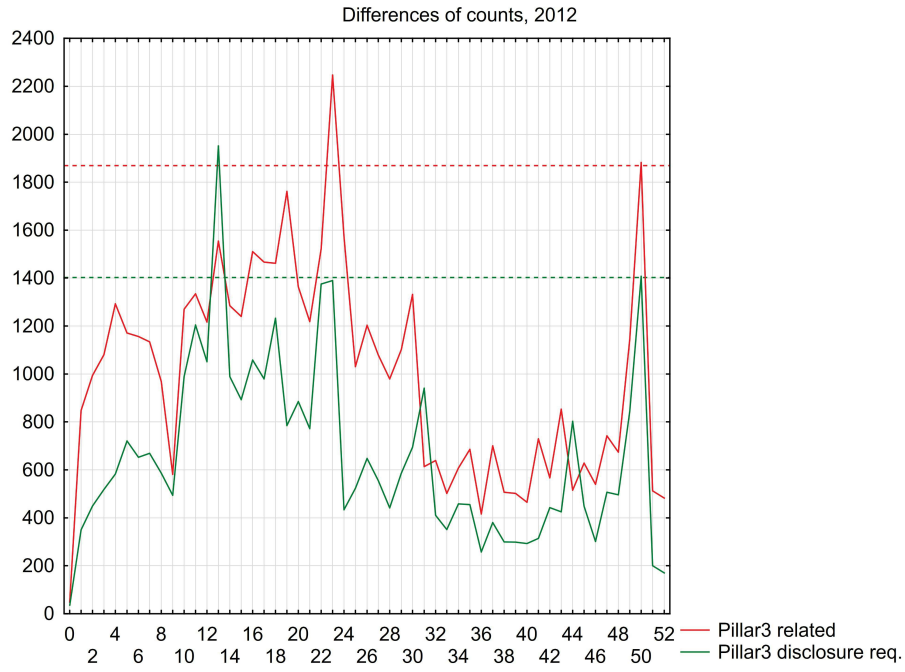



**Figure 9.2.** Differences of counts of the model with corrected data

A way to show the suitability of the model is also to evaluate theoretical and empirical logits. The idea is whether the estimated theoretical logits fit (model) the empirical logits calculated from the empirical relative counts of accesses $h_{ij} = ln\left(\frac{p_{ij}}{p_{iJ}}\right), j = 1, 2, ..., J-1$, where $p_{ij}$ is the empirical relative count of access to the web category $j$ in time $i$ and $p_{iJ}$ is the empirical relative count of access to the referential web category $J$ in time $i$. The visualization of observed and expected logits of each of the examined categories (except the referential web category) can show how the theoretical logits model the empirical logits. Based on the visualization for the critical year 2012 (Figure 9.3 and 9.4), it was seen that after the new data cleansing the theoretical logits fit the empirical logits better.

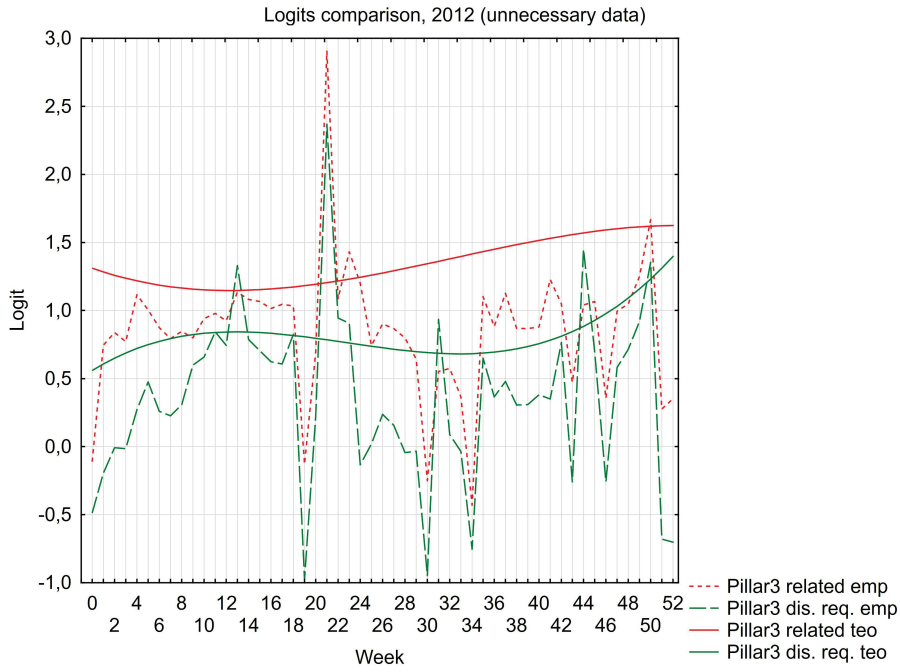



**Figure 9.3.** Logit visualisation of the model for the year 2012 with error data


**Figure 9.4.** Logit visualisation of the model for the year 2012 with corrected data

The plot (Figure 9.5) shows the visualization of probabilities of access to the market discipline related web categories (Pillar3) during the year 2009. This year is taken as the year of the financial crisis. It can be seen that the highest access during this year were to the web category *Pillar3 related* at the beginning of the year (the $0^{\text{th}}$ week has

the value 0.193, for the record this week is a week that contains days from the previous year and also from the actual year – at the turn of years). The lowest estimated values were identified later in the year (the 38$^{th}$ week has the value 0.160). The most interest for the category *Pillar3 disclosure requirements* was again at the turn of years but this time it was at the end of the year 2009 (the 52$^{nd}$ week has the value 0.100). The lowest access was in case of this category the same week as for the other category (the 38$^{th}$ week has the value 0.050). By studying both categories, it can be observed that both are interesting for stakeholders at the beginning of the year and then the interest lowers, whereas at the end of the year it starts to rise again. This can be analysed in more detail also using the other method by extracting the frequent itemsets of the web categories. The asterisks contained in the plot (Figure 9.5) represent the homogenous groups for occurence of frequent itemsets of the web categories for the year 2009. The zero hypothesis is rejected at the 5% significance level ($df = 3,\ Q = 8.258,\ p < 0.05$) for the quarters of the year 2009. Most frequent itemsets were identified in the first quarter (63.64%) and the lowest in the third quarter (38.64%). In the year 2009, two homogenous groups (2009 Q3, 2009 Q4, 2009 Q2) and (2009 Q4, 2009 Q2, 2009 Q1) were identified (Figure 9.5) based on the average occurence of extracted frequent itemsets of the web parts. These results verify the week analysis for this year. Where the most interest of the web users in Pillar3 categories was at the beginning of the year and the lowest in the third quarter (the 38$^{th}$ week is at the end of the third quarter where between 2009 Q1 and 2009 Q3 a statistically significant difference at the 5% significance level was identified).

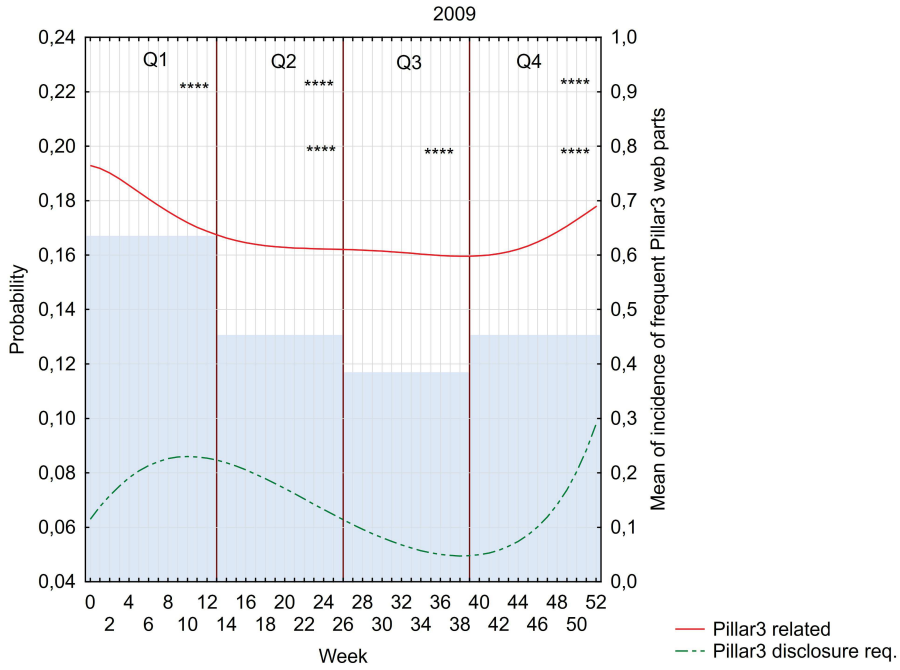



**Figure 9.5.** Probability visualization of market discipline related categories during the year 2009

The plot (Figure 9.6) shows the visualization of probabilities of access to the market discipline related web categories (Pillar3) during the year 2010. This year is taken as the year after the financial crisis. The highest access during this year was to the web category *Pillar3 related* at the beginning of the year (the 3$^{rd}$ week has the value

0.211). The lowest estimated values were identified at the end of the year in the last quarter (the 43$^{rd}$ week has the value 0.161). The most interest for the category *Pillar3 disclosure requirements* was at the beginning of the year at the end of the first quarter (the 11$^{th}$ week has the value 0.116). The lowest access was later in the year in the case of this category (the 40$^{th}$ week has the value 0.058). By studying both categories, it can be observed that both are interesting for the stakeholders at the beginning of the year and then the interest lowers, whereas at the end of the year it starts to rise again. Now the weekly results will be compared with the frequent itemsets. The asterisks contained in the plot (Figure 9.6) represent the homogenous groups for occurence of frequent itemsets of the web categories for the year 2010. The zero hypothesis is rejected at the 1% significance level ($df = 3,\ Q = 12.581,\ p < 0.01$) for the quarters of the year 2010. Most frequent itemsets were identified in the first quarter (40.91%) and the lowest in the third and fourth quarters (20.45% - 22.73%). In the year 2010 three homogenous groups (2010 Q4, 2010 Q3), (2010 Q3, 2010 Q2) and (2010 Q2, 2010 Q1) were identified (Figure 9.6) based on the average occurence of extracted frequent itemsets of the web parts. The most interest of the web users in Pillar3 categories was at the beginning of the year and the lowest in the last quarters, where between the 2010 Q1 and 2010 Q3/2010 Q4 and between 2010 Q2 and 2010 Q4 a statistically significant difference at the 5% significance level was identified.

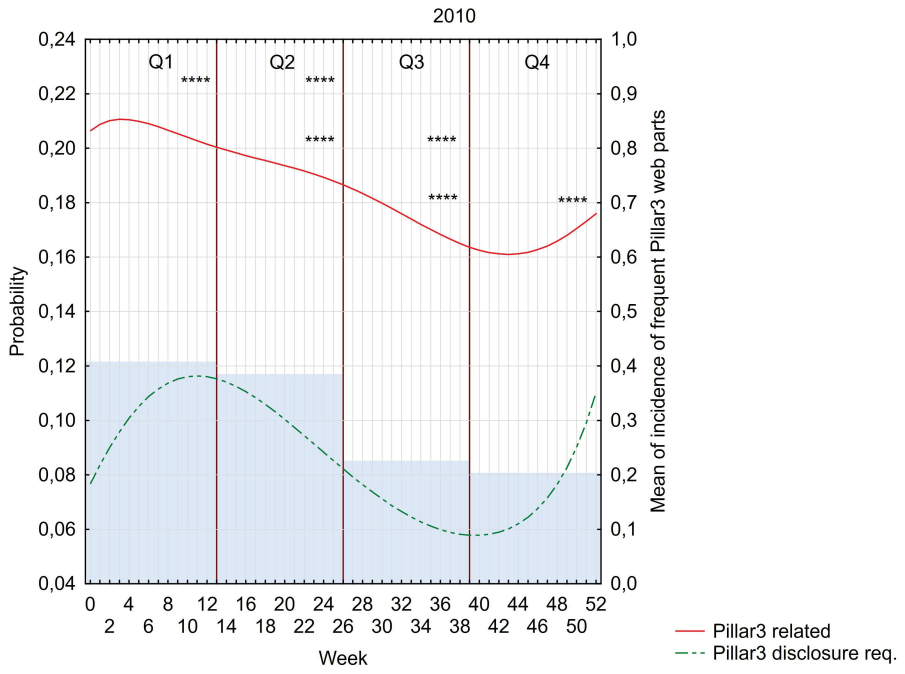



**Figure 9.6.** Probability visualization of market discipline related categories during the year 2010

The plot (Figure 9.7) shows the visualization of probabilities of access to the market discipline related web categories (Pillar3) during the year 2011. This year can be taken as the second year after the financial crisis. The highest access during this year was to the web category *Pillar3 related* again at the beginning of the year (the 5$^{th}$ week has the value 0.211). The lowest estimated values were identified at the same week as in the previous year with an even lower value (the 43$^{rd}$ week has the value 0.151). The most interest for the category *Pillar3 disclosure requirements* was also the same

as the previous year but with a little higher value (the 11$^{th}$ week has the value 0.132). The lowest access was also the same week as the previous year with an almost similar value in the case of this category (the 40$^{th}$ week has the value 0.059). By studying both categories, it can be observed that the behaviour is like the previous year with only a little deviation. Now the weekly results will be compared with the frequent itemsets. The asterisks contained in the plot (Figure 9.7) represent the homogenous groups for occurence of frequent itemsets of the web categories for the year 2011. The zero hypothesis is rejected at the 1% significance level ($df = 3$, $Q = 11.539$, $df = 3$, $p < 0.01$) for the quarters of the year 2011. The second quarter contained the most frequent itemsets (38.64%) and the lowest in the first and third quarters (18.18% - 20.45%). In the year 2011, two homogenous groups (2011 Q1, 2011 Q3, 2011 Q4) and (2011 Q4, 2011 Q2) were identified (Figure 9.7) based on the average occurence of extracted frequent itemsets of the web parts. There is a little difference to the previous years. Based on the quarters the highest interest is now in the second quarter, but the week analysis shows us that it is on the period interface.

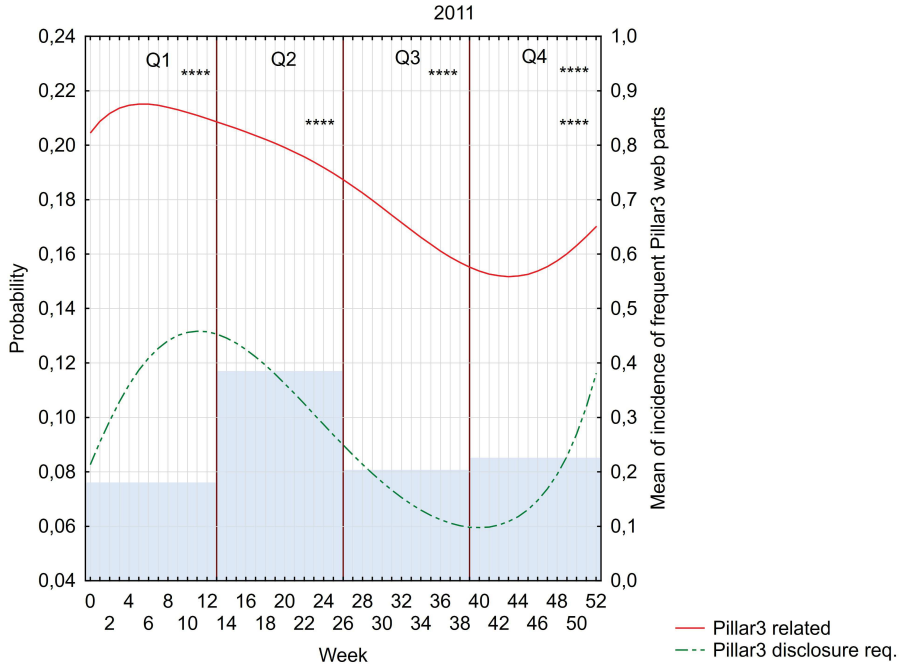



**Figure 9.7.** Probability visualization of market discipline related categories during the year 2011

The plot (Figure 9.8) shows the visualization of probabilities of access to the market discipline related web categories (Pillar3) during the year 2012. This year can be taken as one of the years after the financial crisis. The highest access during this year was to the web category *Pillar3 related* and has shifted more towards the end of the first quarter of the year (the 11$^{th}$ week has the value 0.135). The lowest estimated values were identified in the same period as in the previous two years (the 44$^{th}$ week has the value 0.067). The most interest for the category *Pillar3 disclosure requirements* was almost the same as the previous year (the 12$^{th}$ week has the value 0.107). The lowest access was in the case of this category stabilized at the same period (the 41$^{st}$ week has the value 0.034). By studying both categories, it can be observed that the behaviour has stabilized so that interest rises at the beginning of the year with the highest interest

in the Pillar3 information at the end of the first quarter. Consequently, the interest decreases with the lowest at the beginning of the fourth quarter and then starts to rise again. It can be observed that this year has lower interest for both categories in comparison to the previous years. In the case of frequent itemsets statistically significant differences for the year 2012 were not found ($df = 3$, $Q = 4.154$, $p = 0.2453$). Statistically significant differences for all of the next years were also not found (*2013:* $df = 3$, $Q = 3.255$, $p = 0.3539$; *2014:* $df = 3$, $Q = 4.565$, $p = 0.2066$; *2015:* $df = 3$, $Q = 3.001$, $p = 0.3916$) and the weekly analysis for these years was also not done. It can be said that the trend for the years after the crisis is similar to the years 2010, 2011 and 2012.

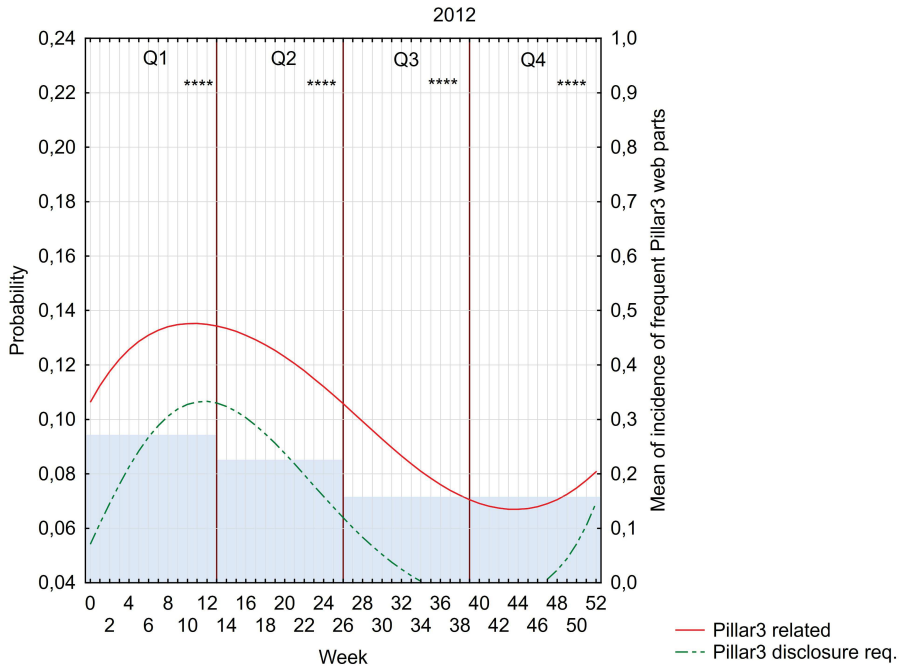



**Figure 9.8.** Probability visualization of market discipline related categories during the year 2012

The first quarters of the years 2009-2010 during the event- global financial crisis have a significant impact on the quantity of identified frequent itemsets of the parts. This can be seen also in the weekly analysis where the first quarters contained the highest interest from the examined Pillar3 web categories. On this basis, it can be concluded that the required quarterly frequency publication of the results is not necessary for market discipline. It would be enough to publish this information annually, ideally in the early weeks of the year. To obtain these results two different approaches with various time variables were used. Both approaches evaluate the behaviour of the users in time (mainly seasonality), (1) modelling the probabilities of access to the portal in time, (2) quantitative evaluation of frequent itemsets incidence in time. The results match and based on that can be regarded as robust. The combination of these methods improved the results of the data and helped to better understand the behaviour of the stakeholders with the Pillar3 information on the web categories.

## 5. Discussion and Conclusion

Pillar 3 of Basel II and Basel III regulation, called "market discipline" or according to some authors it should be named "information disclosure", is a very important component of banking supervision and regulation. Its importance has even increased after the crisis 2007 - 2009. Basel supervisors have designed Pillar 3 to be an effective market discipline tool. However, its current complexity, standardization and the whole architecture also puts impediments on the market discipline effectiveness. One such impediment is that regulator does not study stakeholders' interest and usage of disclosed information for some categories of commercial banks that are specific as far as their ownership structure, business model and funding. The impact of this situation is usually on one side no, or very low interest of stakeholders in information which is not in line with their interests and on the other side, the high costs of banks related to the preparation of enormous amounts of disclosed and unused information. One such market that might suffer due to current Pillar 3 architecture is the depositors' market, specifically in CEE countries where single foreign banks/group ownership are very frequent. Therefore, in this study the interests of depositors in the disclosed two groups of information were assessed: requirement of the Pillar 3 disclosure and Pillar 3 related information during the period 2009-2012 (year of crisis and subsequent years). The analysis was based on visits of the stakeholders to the web portal of commercial banks and analysing their interest in relation to time spent on the web page (time variables) and in relation to the events of the financial crisis in 2009. The analysis of time spent on categories of the web portal was based on weekly accesses and frequent itemsets in terms of quantity (based on quarters). The findings are as follows:

(1) the results of the analysis during and after the crisis suggest that stakeholders have expressed higher interest in *Pillar 3 related* information (such as annual reports, financial reports, annual reports, rating, group, general shareholder meeting, emitent prospects) rather than *Pillar 3 disclosure requirements*.
(2) the highest interest of stakeholders in disclosed information was in the year of the crisis and subsequently steadily decreased.
(3) the results of the analysis on the year of the financial crisis (weekly and quarterly analysis) have shown that in 2009 the highest interest was in the first quarter, at the beginning of the year, (exceptionally $52^{\text{nd}}$ week for *Pillar 3 disclosure requirements* in 2009) and was lowest in the third quarter for both categories.
(4) in the analysis of the years 2010-2012, the years after the financial crisis, similar results have been identified, with the highest interest being at the end of the first quarter, on a period interface (exceptionally $3^{\text{rd}}$ week for *Pillar 3 related* in 2010 and $5^{\text{th}}$ week for *Pillar 3 related* in 2011) and the lowest interest being identified in the fourth quarter. It is important to note, that interest decreased generally during 2010-2012 in comparison to 2009.
(5) the results are in line with Munk et al. (Munk et al., 2017), whose results show that studied CEE commercial banks stakeholders are particularly interested in Pillar 3 disclosures in the first quarter and that interest in disclosures decreased after the turbulence of 2009.
(6) the results also suggest that due to the significant impact of the first quarter with the highest interest of stakeholders, which was also validated by weekly analysis, quarterly disclosures seem less important for market discipline effects in comparison to annual disclosures. Annual disclosures imply higher interest than Pillar 3 disclosures and ideally should be disclosed at the beginning of the

year.

The above presented results suggest that changes in information disclosures' design in commercial banks operating according to the analysed model are inevitable to enhance the efficiency of market discipline mechanisms and to add value to key stakeholders (depositors).

This paper's conclusion fits with the study (Miklaszewska & Pawłowska, 2014, p. 264) that deeply analysed CEE banks' perspectives. Their conclusion is that in spite of fact that in the EU is currently applied regulatory and supervisory model that is complex but it may not have produced the required more efficient and stable banking system, particularly in CEE countries that have very competitive banking environments.

Moreover, it can be agreed with (Kuranchie-Pong, Bokpin & Andoh, 2016) that stakeholders in the banking industry are supposed to use market discipline to make risk management more effective but to assess the bank's risk profile they need sufficient relevant information disclosures.

Finally, according to the European Banking Authority (EBA), to disclose to markets a sufficient risk profile of financial institutions is the most important to ensure their correct functioning, creating trust between market participants and the efficiency of market discipline. The principles for adequate disclosures are clarity, meaningfulness, consistency over time and comparability across institutions, also in times of stress. There still exist open issues about the nature and potential impediments to disclosures in order to fulfil these principles and the authors hope that these findings and conclusions may also contribute to resolve these issues.

Due to some limitations of this research, and according to the findings, an analysis of the interest of stakeholders in the content of Pillar 3 disclosures has been identified as a future research topic.


## Acknowledgements

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contracts No. VEGA-1/0776/18 and by the scientific research project of the Czech Sciences Foundation Grant No:19-15498S.


## Disclosure statement

The authors declare that they have no conflict of interest.

## Notes on contributors


**Anna Pilková** is currently Professor of management at Faculty of Management at Comenius University in Bratislava, Slovakia. Before she worked at top managerial positions in commercial bank in Slovakia. Her research interests are focused on the banking regulation, risk management and strategic management at commercial banking in developing countries. In addition to that she has conducted research on entrepreneurial activities and entrepreneurship inclusivity as a national team leader for Global Entrepreneurship Monitor. She is a recipient of a few awards among them the Green

Group Award of Computational Finance and Business Intelligence (Best paper) of the International Conference on Computational Science 2013, the Workshop on Computational Finance and Business Intelligence (Barcelona, 2013).

**Michal Munk** received the M.S. degree in Mathematics and Informatics from Constantine the Philosopher University, Nitra, Slovakia, in 2003 and the Ph.D. degree in Mathematics from Constantine the Philosopher University, Nitra, Slovakia, in 2007. In 2018, he was appointed as a professor in System Engineering and Informatics at the Faculty of Informatics and Management, University of Hradec Kralove, Czechia. He is currently a professor with the Computer Science Department, Constantine the Philosopher University, Nitra, Slovakia. His research interests include data analysis, web mining, and natural language processing.

**Petra Blažeková** is currently external PhD student at Comenius University in Bratislava, Faculty of Management. She works at commercial bank in Slovakia as a risk management specialist. Her research interests are in the field of risk management, reporting and regulation.

**Ľubomír Benko** received the M.S. degree in Applied Informatics from Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Slovakia, in 2014. In 2018, he received the Ph.D. degree in Applied Informatics at University of Pardubice, Czechia. From 2014 to 2018, he was a researcher with the Computer Science Department, Constantine the Philosopher University, Nitra, Slovakia. He is currently an assistant professor with the Computer Science Department, Constantine the Philosopher University, Nitra, Slovakia. His research interests are in the field of web usage mining and natural language processing.

## References


Accornero, M., & Moscatelli, M. (2018). Listening to the Buzz: Social Media Sentiment and Retail Depositors' Trust. *SSRN Electronic Journal*. doi:10.2139/ssrn.3160570

Arsov, S., & Bucevska, V. (2017). Determinants of transparency and disclosure – evidence from post-transition economies. *Economic Research-Ekonomska Istraživanja*, *30*(1), 745–760. doi:10.1080/1331677X.2017.1314818

Bartlett, R. (2012). Making Banks Transparent. *Vand. L. Rev*, *65*, 293–386. Retrieved from http://scholarship.law.berkeley.edu/facpubs/1824

Bartulovic, M., & Pervan, I. (2012). Comparative analysis of voluntary internet financial reporting for selected CEE countries. *Recent Researches in Applied Economics and Management*, *1*(1), 296–301.

Benli, V. F. (2015). Basel's Forgotten Pillar: The Myth of Market Discipline on the Forefront of Basel III. *Financial Internet Quarterly*, *11*(3), 70–91.

Berger, A. N., & Bouwman, C. H. S. (2013). How does capital affect bank performance during financial crises? *Journal of Financial Economics*, *109*(1), 146–176. doi:10.1016/j.jfineco.2013.02.008

Berger, A. N., & Davies, S. M. (1998). The Information Content of Bank Examinations. *Journal of Financial Services Research*, *14*(2), 117–144. doi:10.1023/A:1008011312729

Bischof, J., Daske, H., Elfers, F., & Hail, L. (2016). A Tale of Two Regulators: Risk Disclosures, Liquidity, and Enforcement in the Banking Sector. *SSRN Electronic Journal*. doi:10.2139/ssrn.2580569

Bouaiss, K., Refait-Alexandre, C., & Alexandre, H. (2017). Will Bank Transparency really Help Financial Markets and Regulators? Retrieved from https://hal.archives-ouvertes.fr/hal-01637917

Bouvard, M., Chaigneau, P., & Motta, A. de. (2015). Transparency in the Financial System: Rollover Risk and Crises. *The Journal of Finance*, *70*(4), 1805–1837. doi:10.1111/jofi.12270

Calomiris, C. W. (2009). Bank regulatory reform in the wake of the financial crisis.
Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, *1*(1), 5–32.
Cubillas, E., Fonseca, A. R., & González, F. (2012). Banking crises and market discipline: International evidence. *Journal of Banking & Finance*, *36*(8), 2285–2298. doi:10.1016/J.JBANKFIN.2012.04.011
de Araujo, P., & Leyshon, K. I. (2016). The impact of international information disclosure requirements on market discipline. *Applied Economics*, *49*(10), 954–971. doi:10.1080/00036846.2016.1208361
Del Gaudio, B. L., Megaravalli, A. V., Sampagnaro, G., & Verdoliva, V. (2020). Mandatory disclosure tone and bank risk-taking: Evidence from Europe. *Economics Letters*, *186*, 108531. doi:10.1016/j.econlet.2019.108531
Distinguin, I. (2008). Market Discipline and Banking Supervision: The Role of Subordinated Debt. *SSRN Electronic Journal*. doi:10.2139/ssrn.1098252
Distinguin, I., Kouassi, T., & Tarazi, A. (2012). Interbank Deposits and Market Discipline: Evidence from Central and Eastern Europe. *SSRN Electronic Journal*, *41*(2), 544–560. doi:10.2139/ssrn.2119956
Distinguin, I., Rous, P., & Tarazi, A. (2006). Market Discipline and the Use of Stock Market Data to Predict Bank Financial Distress. *Journal of Financial Services Research*, *30*(2), 151–176. doi:10.1007/s10693-0016-6
Drlik, M., & Munk, M. (2019). Understanding time-based trends in stakeholders' choice of learning activity type using predictive models. *IEEE Access*, *7*, 3106–3121. doi:10.1109/ACCESS.2018.2887057
Ellahie, A. (2012). Capital Market Consequences of EU Bank Stress Tests. *SSRN Electronic Journal*. doi:10.2139/ssrn.2157715
Evanoff, D. D., & Wall, L. D. (2000). Subordinated debt as bank capital: a proposal for regulatory reform. *Economic Perspectives*, (Q II), 40–53. Retrieved from https://ideas.repec.org/a/fip/fedhep/y2000iqiip40-53nv.25no.2.html
Faria-e-Castro, M., Martinez, J., & Philippon, T. (2017). *Runs versus Lemons: Information Disclosure and Fiscal Capacity*. Cambridge, MA. doi:10.3386/w21201
Fijalkowska, J., Zyznarska-Dworczak, B., & Garsztka, P. (2017). The Relation between the CSR and the Accounting Information System Data in Central and Eastern European (CEE) Countries – The Evidence of the Polish Financial Institutions. *Journal of Accounting and Management Information Systems*, *16*(4), 490–521.
Flannery, M. J., & Bliss, R. R. (2019). Market discipline in regulation: Pre-and post-crisis. *Forthcoming, Oxford Handbook of Banking 3e.*
Fonseca, A. R., & González, F. (2010). How bank capital buffers vary across countries: The influence of cost of deposits, market power and bank regulation. *Journal of Banking & Finance*, *34*(4), 892–902. doi:10.1016/J.JBANKFIN.2009.09.020
Freixas, X., & Laux, C. (2011). Disclosure, transparency and market discipline. *CFS Working Paper*, *11*, 1–39. Retrieved from https://www.ifk-cfs.de/fileadmin/downloads/publications/wp/2011/11_11.pdf
Frolov, M. (2007). Why do we need mandated rules of public disclosure for banks? *Journal of Banking Regulation*, *8*(2), 177–191. doi:10.1057/palgrave.jbr.2350045
Gandrud, C., & Hallerberg, M. (2014). Supervisory transparency in the European banking union. *Bruegel Policy Contribution*, (2014/01). Retrieved from https://www.econstor.eu/handle/10419/106314
Giner, B., Allini, A., & Zampella, A. (2020). The Value Relevance of Risk Disclosure: An Analysis of the Banking Sector. *Accounting in Europe*. doi:10.1080/17449480.2020.1730921
Goldstein, I., & Leitner, Y. (2015). *Stress tests and information disclosure. No 15-10, Working Papers*. Federal Reserve Bank of Philadelphia. Retrieved from https://econpapers.repec.org/paper/fipfedpwp/15-10.htm
Goldstein, I., & Sapra, H. (2014). Should Banks' Stress Test Results be Disclosed? An Analysis of the Costs and Benefits. *Foundations and Trends® in Finance*, *8*(1), 1–54.

doi:10.1561/0500000038

Guillemin, F., & Semenova, M. (2018). Transparency and Market Discipline: Evidence from the Russian Interbank Market. *Higher School of Economics Research Paper No. WP BRP 67/FE/2018*, 32. doi:10.2139/ssrn.3225061

Habek, P. (2017). CSR Reporting Practices in Visegrad Group Countries and the Quality of Disclosure. *Sustainability*, *9*(12), 1–18.

Hadad, M. D., Agusman, A., Monroe, G. S., Gasbarro, D., & Zumwalt, J. K. (2011). Market discipline, financial crisis and regulatory changes: Evidence from Indonesian banks. *Journal of Banking & Finance*, *35*(6), 1552–1562. doi:10.1016/j.jbankfin.2010.11.003

Hasan, I., Jackowicz, K., Kowalewski, O., & Kozłowski, Ł. (2013). Market discipline during crisis: Evidence from bank depositors in transition countries. *Journal of Banking & Finance*, *37*(12), 5436–5451. doi:10.1016/j.jbankfin.2013.06.007

Hill, T., & Lewicki, P. (2013). *Electronic Statistics Textbook*. StatSoft Inc. Retrieved from http://www.statsoft.com/textbook/

Iren, P., Reichert, A. K., & Gramlich, D. (2014). Information Disclosure, Bank Performance and Bank Stability. *Int. Journal Banking, Accounting and Finance*, *5*(4), 39. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2874144

Jagtiani, J. A., Kaufman, G., & Lemieux, C. (1999). Do markets discipline banks and bank holding companies? evidence from debt pricing. *Emerging Issues*, (Jun). Retrieved from http://econpapers.repec.org/article/fipfedhei/y_3a1999_3ai_3ajun_3an_3asr-99-3r.htm

Jagtiani, J., & Lemieux, C. (2001). Market discipline prior to bank failure. *Journal of Economics and Business*, *53*(2), 313–324. Retrieved from http://econ.tu.ac.th/archan/Chalotorn/on%mkt%failure/jagtiani.pdf

Jordan, J. S., Peek, J., & Rosengren, E. S. (2000). The Market Reaction to the Disclosure of Supervisory Actions: Implications for Bank Transparency. *Journal of Financial Intermediation*, *9*(3), 298–319. doi:10.1006/jfin.2000.0292

Kapusta, J., Munk, M., & Drlik, M. (2012a). Cut-off time calculation for user session identification by reference length. In *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*. doi:10.1109/ICAICT.2012.6398500

Kapusta, J., Munk, M., & Drlik, M. (2012). User Session Identification Using Reference Length. In Capay, M and Mesarosova, M and Palmarova, V (Ed.), *DIVAI 2012: 9TH INTERNATIONAL SCIENTIFIC CONFERENCE ON DISTANCE LEARNING IN APPLIED INFORMATICS: CONFERENCE PROCEEDINGS* (pp. 175–184).

Karas, A., Pyle, W., & Schoors, K. (2013). Deposit Insurance, Banking Crises, and Market Discipline: Evidence from a Natural Experiment on Deposit Flows and Rates. *Journal of Money, Credit and Banking*, *45*(1), 179–200. doi:10.1111/j.1538-4616.2012.00566.x

Kozłowski, Ł. (2016). Cooperative banks, the internet and market discipline. *Journal of Co-Operative Organization and Management*, *4*(2), 76–84. doi:10.13140/RG.2.1.3768.6809

Kuranchie-Pong, L., Bokpin, G. A., & Andoh, C. (2016). Empirical evidence on disclosure and risk-taking of banks in Ghana. *Journal of Financial Regulation and Compliance*, *24*(2), 197–212. doi:10.1108/JFRC-05-2015-0025

Lapteacru, I. (2019). Do bank activities and funding strategies of foreign and state-owned banks have a differential effect on risk-taking in Central and Eastern Europe? *Economics of Transition and Institutional Change*, *27*(2), 541–576. doi:10.1111/ecot.12185

Li, Y., Li, C., & Gao, Y. (2020). Voluntary disclosures and peer-to-peer lending decisions: Evidence from the repeated game. *Frontiers of Business Research in China*, *14*(1), 1–26. doi:10.1186/s11782-020-00075-5

Matuszak, L., & Rozanska, E. (2017). An Examination of the Relationship between CSR Disclosure and Financial Performance: The Case of Polish Banks. *Journal of Accounting and Management Information Systems*, *16*(4), 522–533. Retrieved from https://econpapers.repec.org/RePEc:ami:journl:v:16:y:2017:i:4:p:522-533

Miklaszewska, E., & Pawłowska, M. (2014). Do safe banks create safe systems? Central and eastern European banks' perspective. *Revue de l'OFCE*, *132*(1), 243–267.

doi:10.3917/reof.132.0243
Moreno, D., & Takalo, T. (2016). Optimal Bank Transparency. *Journal of Money, Credit and Banking*, *48*(1), 203–231. doi:10.1111/jmcb.12295
Munk, M., Kapusta, J., Švec, P., & Turčáni, M. (2010). Data advance preparation factors affecting results of sequence rule analysis in web log mining. *E+M Ekonomie a Management*, *13*(4), 143–160.
Munk, M., Pilková, A., Drlik, M., Kapusta, J., & Švec, P. (2012). Verification of the fulfilment of the purposes of basel ii, pillar 3 through application of the web log mining methods. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *60*(2).
Munk, M, Drlik, M., & Vrabelova, M. (2011). Probability Modelling of Accesses to the Course Activities in the Web-Based Educational System. In *Computational Science And Its Applications - Iccsa 2011, Pt V* (Vol. 6786, pp. 485–499).
Munk, M., Benko, Ľ., Gangur, M., & Turčáni, M. (2015). Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management*, *18*(3), 144–159. doi:10.15240/tul/001/2015-3-013
Munk, M., Kapusta, J., & Švec, P. (2010). Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. In *Procedia Computer Science* (Vol. 1, pp. 2273–2280). doi:10.1016/j.procs.2010.04.255
Munk, M., Pilkova, A., Benko, L., & Blaželeková, P. (2017). Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management*, *18*(5), 954–973. doi:10.3846/16111699.2017.1360388
Munk, Michal, Vrábelová, M., & Kapusta, J. (2011). Probability modeling of accesses to the web parts of portal. *Procedia Computer Science*, *3*, 677–683. doi:10.1016/j.procs.2010.12.113
Naz, M., & Ayub, H. (2017). Impact of Risk-Related Disclosure on the Risk-Taking Behavior of Commercial Banks in Pakistan. *Journal of Independent Studies and Research-Management, Social Sciences and Economics*, *15*. doi:10.31384/jisrmsse/2017.15.2.9
Nier, E. W. (2005). Bank stability and transparency. *Journal of Financial Stability*, *1*(3), 342–354. doi:10.1016/J.JFS.2005.02.007
Nier, E. W., & Baumann, U. (2006). Market discipline, disclosure and moral hazard in banking. *Journal of Financial Intermediation*, *15*(3), 332–361. doi:10.1016/j.jfi.2006.03.001
Niessen-Ruenzi, A., Parwada, J. T., & Ruenzi, S. (2015). Information Effects of the Basel Bank Capital and Risk Pillar 3 Disclosures on Equity Analyst Research An Exploratory Examination. *SSRN Electronic Journal*. doi:10.2139/ssrn.2670418
Parwada, J. T., Lau, K., & Ruenzi, S. (2015). The Impact of Pillar 3 Disclosures on Asymmetric Information and Liquidity in Bank Stocks: Multi-Country Evidence. *CIFR Paper No. 82/2015*, 27. doi:10.2139/ssrn.2670403
Parwada, J. T., Ruenzi, S., & Sahgal, S. (2013). Market Discipline and Basel Pillar 3 Reporting. *SSRN Electronic Journal*. doi:10.2139/ssrn.2443189
Peristiani, S., Morgan, D. P., & Savino, V. (2010). *The Information Value of the Stress Test and Bank Opacity*.
Pilkova, A., Munk, M., Švec, P., & Medo, M. (2015). Assessment of the Pillar 3 Financial and Risk Information Disclosures Usefulness to the Commercial Banks Users. *Lecture Notes in Artificial Intelligence*, *9227*, 429–440.
Poshakwale, S., Aghanya, D., & Agarwal, V. (2020). The impact of regulations on compliance costs, risk-taking, and reporting quality of the EU banks. *International Review of Financial Analysis*, *68*, 101431. doi:10.1016/j.irfa.2019.101431
Sarker, N., & Sharif, J. (2020). SIMULTANEITY AMONG MARKET RISK TAKING, BANK DISCLOSURES AND CORPORATE GOVERNANCE: EMPIRICAL EVIDENCE FROM THE BANKING SECTOR OF BANGLADESH. *Academy of Accounting and Financial Studies Journal*, *24*(1), 1–21.
Scannella, E. (2018). *Market Risk Disclosure in Banks' Balance Sheet and Pillar 3 Report: The Case of Italian Banks*.
Scannella, E., & Polizzi, S. (2019). Do Large European Banks Differ in their Derivative Disclosure Practices? A Cross-Country Empirical Study. *Journal of Corporate Accounting &*

*Finance*, *30*(1), 14–35. doi:10.1002/jcaf.22373

Sironi, A. (2003). Testing for Market Discipline in the European Banking Industry: Evidence from Subordinated Debt Issues. *Journal of Money, Credit and Banking*, *35*(3), 443–472. Retrieved from http://econpapers.repec.org/article/mcbjmoncb/v_3a35_3ay_3a2003_3ai_3a3_3ap_3a443-72.htm

Song, V., Cavusoglu, H., Lee, G. M., & Ma, L. (2020). IT Risk Factor Disclosure and Stock Price Crashes. doi:10.24251/HICSS.2020.738

Sowerbutts, R., Zer, I., & Zimmerman, P. (2013). Bank disclosure and financial stability, 326–335.

Stephanou, C. (2010). Rethinking market discipline in banking: Lessons learned from the Financial Crisis. *Policy Research Working Paper, The World Bank*, *5227*, 1–37.

Vauhkonen, J. (2012). The Impact of Pillar 3 Disclosure Requirements on Bank Safety. *Journal of Financial Services Research*, *41*(1–2), 37–49. doi:10.1007/s10693-011-0107-x

Yang, W., & Koshiyama, A. S. (2019). Assessing qualitative similarities between financial reporting frameworks using visualization and rules: COREP vs. pillar 3. *Intelligent Systems in Accounting, Finance and Management*, *26*(1), 16–31. doi:10.1002/isaf.1441

Zer, I. (2015). Information Disclosures, Default Risk, and Bank Value. *Finance and Economics Discussion Series*, *2015*(104), 1–43. doi:10.17016/FEDS.2015.104

PRÍLOHA F: BLAŽEKOVÁ, PETRA, ĽUBOMÍR BENKO, ANNA PILKOVÁ A MICHAL MUNK, 2021. IS PILLAR 3 A GOOD TOOL FOR STAKEHOLDERS IN CEE COMMERCIAL BANKS? IN: *STUDIES IN SYSTEMS, DECISION AND CONTROL*. SPRINGER, S. 421–440. DOI:10.1007/978-3-030-76632-0_15 (**SCOPUS**) [SCOPUS: 0]

# Is Pillar 3 a Good Tool for Stakeholders in CEE Commercial Banks?


**Petra Blažeková, Ľubomír Benko, Anna Pilková, and Michal Munk**

**Abstract** The Pillar 3 is a supervisory tool for enhancement of market discipline, which was introduced as a response to the financial crisis in 2008. The aim of this paper is an assessment of Pillar 3 disclosures as a tool for market discipline enhancement, which is supported by the analysis of the website data dedicated to Pillar 3 disclosures of the banking institution operating in CEE country. The analysis of the web portal is based on the time spent on the web page by the web users and analysing their interest in relation to time spent and content of the Pillar 3 disclosures. The data consist of pre-processed data from a log file of a web portal of banking institution. The web portal pages were joined into logical web parts that were joined into specific categories. The results show statistically significant differences in the average time spent on the web parts, the most average time was spent on the web part Annual Reports and Emitent Prospects. The Pillar 3 Q-terly Info web part had the second least average time spent. This part was analysed in more detail and its categories with the most time spent by the web users were identified. They are Individual Financial Statements, Shareholders and Risk Management, which can indicate different importance of these web categories for Stakeholders based on its content or its high volume.

P. Blažeková · A. Pilková (✉)
Department of Strategy and Entreprenuership, Comenius University, Bratislava, Slovakia
e-mail: anna.pilkova@fm.uniba.sk

P. Blažeková
e-mail: petra.blazekova@fm.uniba.sk

Ľ. Benko · M. Munk
Department of Computer Science, Constantine the Philosopher University in Nitra, Nitra, Slovakia
e-mail: lbenko@ukf.sk

M. Munk
e-mail: mmunk@ukf.sk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

N. Kryvinska and A. Poniszewska-Marańda (eds.), *Developments in Information & Knowledge Management for Business Applications*, Studies in Systems, Decision and Control 376, https://doi.org/10.1007/978-3-030-76632-0_15

# 1 Introduction

Financial crisis in 2008 has been an important lesson for all market participants and market discipline, too. Market discipline failure due to the inadequate public disclosure was followed by the inability of investors to judge the risk and withdrawal of lending in times of systemic stress [1]. However, the market discipline's mechanism should avoid or remediate excessive risk taking by banks. When market discipline leads institutions to avoid excessive risk taking it is termed as an ex ante. From the perspective of regulators, ex ante market discipline is preferred [2]. Turbulences on the financial market has shown that integration of market discipline with regulatory discipline is crucial [3].

However, the process of integration of market discipline into regulatory framework reveals the problem of conflicts of interest between sustaining market discipline and safeguarding financial stability in the context of bank resolution. Whereas the former demands full enforcement of applicable legal rules when crises occur, the latter requires a more pragmatic and flexible approach. In order to achieve both objectives of these concepts, it is necessary to address this tension between them when designing a bank recovery and resolution framework [4]. Market discipline has been recognized as a key objective of the Basel II and III regulations. The higher capital requirements of the Basel III international banking regulations are used to monitor financial stability by market-solvency measures by central banks, because accounting measures did not readily indicate market risks [5]. Market discipline mechanism is operationalized through Pillar 3 information disclosure. The Pillar 3 disclosure requirements have been subject of the extensive regulatory review for the decade and the standards were revised few times aiming also to include comments of stakeholders. The last revision of the Pillar 3 standards has been published in 2016. Moreover, the Pillar 3 standards have been significantly changed and extended in all aspects, and also revised by EBA publishing Guidelines on disclosure requirements under Part Eight of Regulation (EU), which last revision was in 2017. These Guidelines are based on an update of the Pillar 3 requirements by the Basel Committee in January 2015 to enhance the consistency and comparability of institutions' disclosures to ensure market discipline.

The benefits of disclosure implementation are undeniable, which is also supported by empirical research. In many research studies the disclosures are broadly regarded as one of the most effective tool for market discipline enhancement [6]. Additionally, the quality of banks' financial reporting is central to its efficacy and enhancement of the stability of the financial system in downturns [7]. Moreover, banks, which disclose better content of risk management information, provide to investors information to evaluate their financial assets, reduce uncertainty concerning their risk environment and enhance the ability to monitor management practices. As a result, this reflects in stock prices, improving total investment return, which reduces the volatility in stock returns [8]. Disclosures, in general, offers lower costs of capital and help investors to maintain information about the bank's risk and its management. The lower risk disclosure generates ambiguity for potential stakeholders [9]

and the market response to risk disclosure is small when the expected level of risk is high [10]. However, acording to Goldstein and Leitner, even the disclosures are necessary in order to prevent market breakdown, but there is also the potential threat of disclosing too much information, which destroys risk-sharing opportunities [11]. Moreover, increase in disclosures is not sufficient to prevent crisis and authors stress the necessity of disclosing of useful information to investors [1].

Additionally, to achieve both objectives of market discipline and financial stability, in terms of implementation of Pillar 3 disclosures by relevant authorities, this process should also address interests of stakeholders and assess objectively its relevant cost-benefit information [4]. Stakeholders (shareholders, depositors, and loan customers) are an important source of market discipline, because of strong evidence that they discipline banks that restate financial statements. The soundness of market discipline varies by stakeholder as well as bank size and type [12]. Besides that, stakeholders need sufficient disclosure of risk related information to assess the risk profile of the banks [13] and disclosing information in a "clear and easily understandable way" [14]. It is important to include practical perception of stakeholders in the design of the useful model for disclosures, because regulatory reporting is mostly based on the theoretical point of view [15]. Based on that it is important to know, understand and critically evaluate stakeholders´ interests in disclosed information. In our research we assessed stakeholders' interest in information disclosures applying relevant methodology that was based on previous research [16–20].

The aim of this chapter is to propose the design of Pillar 3 content related to stakeholders´ interests and specific conditions under results obtained by applying relevant methodology.

The chapter is structured as follows: Sect. 2 is focused on overview of the research dealing with Pillar 3. Section 3 is devoted to the introduction of the used methodology and approaches used to obtain the results. Section 4 describes the obtained results and compares them on multiple layers. Subsequently, Sect. 5 provides a conclusion.

## 2 Related Work

Pillar 3 disclosures represent implementation of effective market discipline from regulators' point of view. Therefore, we have reviewed studies in which authors point out benefits connected to the implementation of Pillar 3 disclosures as an effective supervisory market discipline tool. Additionally, Pillar 3 decreases information asymmetry [21, 22], improve safety of the banking system [23], offers banks to raise cheaper capital [24] and quarterly reporting is useful to investors [22]. Chen and Du's study [25] concludes that more informative disclosure improves the ex-ante risk sharing provided by financial intermediation, which means positive impact of bank disclosure on optimal risk sharing. According to Naz and Ayub [26], market discipline plays a vital part of risk taking behavior in commercial banks and evidence valid increase in disclosures with reduction of risk-taking. On the other hand, De Araujo and Leyshon's [27] results suggest that, a majority of information disclosure

requirements do not impact market discipline practices, only several of the disclosure requirements have a weak positive impact. Moreover, Pillar 3 benefits are shadowed by excessive superstructure and monitoring costs [28] and it can bring range of issues [29].

Substantially, important factors in assessment of efficiency of Pillar 3 disclosures as a market disciplining tool, is also its content relevancy in relation to interest of key parties. Studies which analyzed Pillar 3 disclosures from its content relevancy show that investors value the information in the financial risk disclosure and the most relevant categories of Pillar 3 disclosures are credit risk and liquidity risk [30]. Scannella [31] highlights the market risk category of Pillar 3 disclosures and concludes that only with full disclosure of a bank's risk strategies and the effectiveness of its risk management policies is able to evaluate the bank's potential risks. Bischof et al. [32] conclude that the improved content of Pillar 3 disclosures translates into higher market liquidity. These results indicate that the success of regulation depends on the institutional fit among regulators, regulated institutions and stakeholders.

On the contrary, Faria-E-Castro et al. [33] points out the importance of disclosures content due to sensitivity of stakeholders to negative disclosures, which can also trigger inefficient bank runs. This is in conjunction with Del Gaudio et al. [34] whose analysis on a sample of European commercial banks (in the EU 15 area) over the time period from 2012 to 2017 finds that qualitative information disclosed in a negative tone contributes to explain the bank risk insolvency that lowers the distance to default. Furthermore, it is important to note, De Araujo and Leyshon's study [27] reveals a window in content relevancy of disclosures in relation to banks' risk profile because depositors and creditors are most responsive to information such as of the bank's asset, off-balance sheet items, and ratings for other banking activities. Moreover, authors highlight also important factors, which influence Pillar 3 content efficiency. These factors are overall quality of the risk disclosures, and valuation of the quantitative information more than the qualitative information.

Accordingly, we have reviewed studies, which analyse disclosures in CEE countries in which authors investigate the impact of the disclosures in CEE countries on the financial performance. Matuszak and Rózańska's study [35] revealed a positive relationship between banks' disclosures and their profitability measured by ROA and ROE. Fijalkowska et al.'s research [36] highlights the slack resources that are strongly related to disclosures. Additionally, the studies below name factors important in assessment of disclosures in CEE countries. Firstly, Bartulovič and Pervan [37] point out that disclosures in CEE countries have their weaknesses and Arsov and Bucevska [38] conclude that they lag transparency in comparison to their peers worldwide as measured one decade ago. Secondly, Habek's results [39] confirm existence of a positive relationship between external verification of a report and the level of quality of the disclosed reports. The reports in V4 countries lack independent verification of the data in these reports.

In addition, there is a lack of the studies assessing Pillar 3 information disclosures on the basis of its content relevancy in commercial banks in CEE countries. Research in this field is low and mostly in relation to Pillar 3 disclosures. Nevertheless, there

are arising issues about the nature of parts, categories, information of Pillar 3 disclosures, which are important to stakeholders and what type of information is efficient from their point of view. The factors, which are crucial in order to implement effective supervisory market discipline are relation to timing, content of the disclosures, importance of specific categories. Concretely, few studies focus on the interest of the stakeholders to disclosures, which is positively correlated with sufficient information disclosures [19]. The results show that the most visited part of the disclosures was part Group and stakeholders' interest in Pillar 3 information is only together with Annual reports or Information on group.

This study is focused on the research gap in the area of the content relevancy of Pillar 3 disclosures to web users of the commercials banks in CEE. We used methodology based on the time spent on the web page by the web users. Time can be an important part of the analysis. The log file as the source of the user activity on the web portal contains a timestamp information. This could be transformed into a time variable that could offer various possibilities to analyze. Munk et al. [19] proposed a methodology of the evaluation of frequent itemsets of web parts over a dedicated time. The time variable was used to distinguish time periods. Drlik and Munk [20] introduced a novel approach to learning analytics. The authors offer an insight into a virtual course schedule as it shows the peak times for different types of course activities. Moh and Saxena [40] introduced a novel system that improved the web recommendations utilizing the time spent on the web page. The authors used the time in combination with the frequent itemsets. Masseglia et al. [41] changes the perspective how to deal with the log file based on time periods. The authors deal with the log file based on periods and not as whole. The main feature is extraction of various behaviors that would be otherwise hidden in the log file.

Obtained results were analyzed and outcomes of this analysis present the stakeholders´ interests in the content of the disclosures.

## 3 Methodology

The examined dataset consists of pre-processed standard web server log files. The dataset contains data and variables that deal with the Pillar 3 regulations. The web usage analysis was done based on a sample of 2 071 235 logged accesses that were obtained after data preparation. The research methodology (Fig. 1) was inspired by [16–19]:

1. Data acquisition: obtaining log files from multiple bank webservers. The data is stored using a load balancer so the log files have to be joined into a one file.
2. Data cleaning of unnecessary files: removing of redundant data such as requests to images, styles, fonts, javascript, etc. These requests are not important for this analysis and need to be removed from the log file. The removal will be done based on the examination of the requested URL. As it contains any of the not needed filename extension.

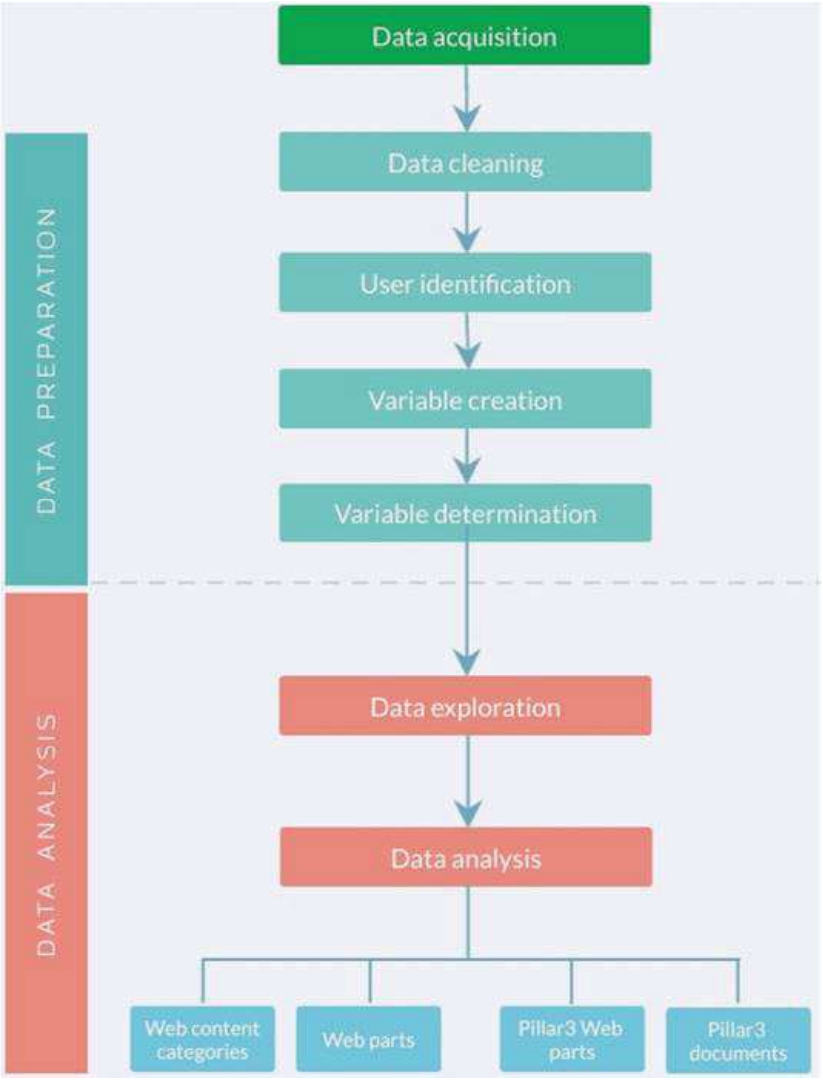



**Fig. 1** Flowchart of the research methodology

3. Data cleaning of robots: the log file is iterated and identified are the accesses of robots of search engines. These accesses are necessary to be removed for this part. The robots are there to browse through the whole web portal and create this way useless log entries [42].
4. User identification: to analyze the time spent on the web pages it is necessary to identify the users as the log file did not contain the information about the users login. The website was a site with an anonymous access. The standard user identification methods were used [42] where the log file fields IP address and the User Agent (contains the information of used web browser and computer operation system) were used to distinguish the users.
5. Variable creation: the next phase contains the working with already cleaned log file. The time variables *UnixTime* and *Length* are created. *UnixTime* represents the standard Unix time in seconds starting on January 1, 1970 and is calculated

using the difference between the actual date and the January 1, 1970 in seconds. *Length* variable represents the time spent on a web page by the user and is calculated from *UnixTime* in seconds.

6. Variables determination: a bank expert was asked to categorize every webpage according to the terminology used in the bank environment. The result of this process is the 9 different parts (grouped web pages) of the web represented by the *WebPart* variable (Rating, Group, Information for Banks, Annual Reports, General Shareholder Meeting, Financial Reports, Emitent Prospects, Pillar 3 Q-terly Info, Pillar 3 Semiannualy Info) each attached to 3 different categories—the variable *Category* (Pillar 3 related (the first seven web parts), Pillar 3 disclosure requirements (quarterly/semiannually information) and Other). The Other category contains web parts like History, Awards, Anti Money Laundering, Ethical Codex, We support, Organizational Structure, Pricing List, Business Conditions, Contacts, Branches and ATM.
7. Data exploration: based on the results of descriptive statistics were determined the hypotheses that claim that there is no statistically significant difference in the time spent on the web pages (*Length*) in the examined web content levels (*Category*/*WebPart*).
8. Data analysis: hypothesis testing using one-way analysis of variance and multiple comparisons. In order not to reduce the power of statistical tests, the assumptions of univariate significance tests was verified. The homogeneity of variance was verified by a Levene's non-parametric test with respect to the identification of deviations from normality. The analysis will be done on multiple layers. First will be analyzed all of the web content categories. As these categories were created from web parts, the next layer will be the analysis of web parts contained in the web content categories. As the interest is mainly in the Pillar 3 information then the next layer analysis will be focused only on the Pillar 3 web parts. The last layer will be the Pillar 3 Q-terly info categories created from the Pillar 3 Q-telry info web part based on the results of the research done by [19]. This division of the web parts and joining into categories can offer a detailed analysis of the interested Pillar 3 information.

## 4 Results

The analysis of web user time spent on specific web part and category is dealt with in the following experiment. The hypothesis is that the web users spend more time on web parts and categories related to the Pillar 3 information (e.g. bank annual reports, minutes from general assembly meetings, the prospect of eminent, information about group and information for banks, financial information, information on risk management). Firstly, the attention was given to the web content categories to observe the interest of the web users.

Based on the descriptive statistics can be seen (Table 1) that variability is almost similar for all the examined web content categories. Also, the standard error is close

**Table 1** Descriptive characteristics of time spent on the examined web content categories

| | Level of factor | *N* | Length mean | Length Std. Dev | Length Std. Err | Length −95% | Length +95% |
|---|---|---|---|---|---|---|---|
| Total | | 3,502,282 | 63.71 | 53.30 | 0.03 | 63.66 | 63.77 |
| Category | Other | 3,203,668 | 64.51 | 53.33 | 0.03 | 64.46 | 64.57 |
| Category | Pillar 3 related | 191,527 | 62.43 | 52.41 | 0.12 | 62.20 | 62.67 |
| Category | Pillar 3 disclosure requirements | 107,087 | 42.02 | 49.07 | 0.15 | 41.73 | 42.32 |

to zero. The variable *Length* is a dependent variable that represents time spent on the web page in seconds, and *Category* is an independent variable represents levels of the web content. From the point and interval estimates of the average (Table 1) can be seen the differences in the time spent on the pages of the examined web categories. The web users spent most of their time on sites in the Other and Pillar 3 related categories. On the contrary, less time spent the web users on the pages in the category Pillar 3 disclosure requirements. The results (Table 1) assume that time spent on the web page (*Length*) depends on the web content categories (*Category*). Based on these results, a null statistical hypothesis will be tested that claims that there is no statistically significant difference in the time spent on pages in the examined web categories.

Based on the ANOVA univariate results ($F = 9331.8$, $p < 0.001$), the zero hypotheses rejected at the 0.1% significance level, i.e. a statistically significant difference in the time spent on the examined website categories was identified. Homogeneous groups in terms of time spent on the examined web content categories (Table 2) were not identified.

While the time spent by stakeholders on the web pages from the *Other* category was above average, in the case of the *Pillar 3* category the time spent on the pages was below average (Fig. 2). In addition, in the case of the *Pillar 3 disclosure information* was the time even significantly below the overall average of the time spent on the web pages.

**Table 2** Unequal N HSD for web categories

| Category | Other | Pillar 3 related | Pillar 3 disclosure requirements |
|---|---|---|---|
| Other | | 0.000022 | 0.000022 |
| Pillar 3 related | 0.000022 | | 0.000022 |
| Pillar 3 disclosure requirements | 0.000022 | 0.000022 | |

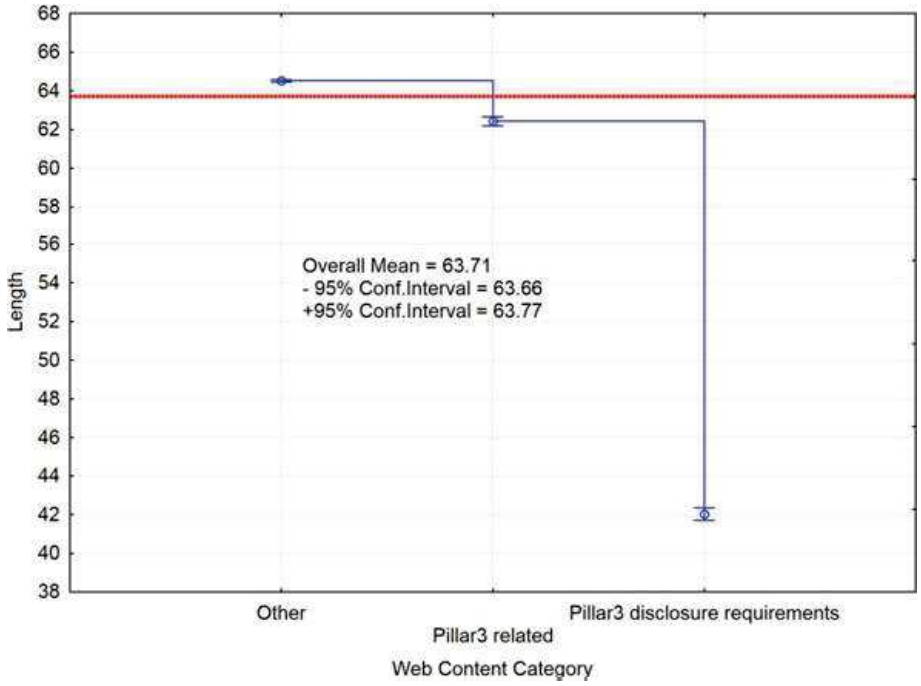



**Fig. 2** Visualization of the average time spent on the web categories in comparison to the total time

The next part of the analysis is focused on specific Pillar 3 web parts. The web categories have shown that there are differences in the interest of web users. First is the focus on all of the categories dealing with Pillar 3 related web parts.

The descriptive statistics (Table 3) show that the variability is almost similar for all the examined web parts. As well as the standard error is close to zero, however higher (>1) is only in the case of less frequent web parts *Values* and *Financial Reports* where the number of accesses was less than 2000.

Differences in the time spent on the web parts can be seen from the point and interval intervals of the average (Table 3). The users spent most of their time on web parts *Values* and *Branches and ATM*. On the other hand, the least time was spent by the users on the web part *Information of Banks*. Based on these results, a null statistical hypothesis will be tested that claims that there is no statistically significant difference in the time spent on the web parts.

The ANOVA univariate results ($F = 11191.8$, $p < 0.001$), the zero hypotheses is rejected at the 0.1% significance level, i.e. a statistically significant difference in the times spent on the examined web parts was identified.

From a multiple comparison (Fig. 3), were identified ten homogeneous groups ($p > 0.05$): (*Rating, Organizational Structure, Business Conditions*), (*Business Conditions, Financial Reports, ID*), (*Financial Reports, ID, Awards*), (*ID, Awards, Pillar 3 Semiannualy Info, Ethical Codex of.*), (*Awards, Pillar 3 Semiannualy Info, Ethical Codex of., Contacts*), (*General Shareholder Meeting, Group*), (*Mision,Vision., Pricing List*), (*Emitent Prospects, History*), (*History, Annual Reports*) and (*Values, Branches and ATM*).

**Table 3** Descriptive characteristics of the time spent on the examined web parts

| | Level of factor | *N* | Length mean | Length Std.Dev | Length Std.Err | Length −95.00% | Length + 95.00% |
|---|---|---|---|---|---|---|---|
| Total | | 3,502,282 | 63.71 | 53.30 | 0.03 | 63.66 | 63.77 |
| WebPart | History | 40,585 | 71.73 | 49.99 | 0.25 | 71.25 | 72.22 |
| WebPart | Awards | 20,038 | 55.77 | 50.64 | 0.36 | 55.07 | 56.47 |
| WebPart | Rating | 22,785 | 51.35 | 50.63 | 0.34 | 50.69 | 52.00 |
| WebPart | Mision, vision. | 18,219 | 63.28 | 49.42 | 0.37 | 62.56 | 64.00 |
| WebPart | Anti money laundering | 12,209 | 46.58 | 48.12 | 0.44 | 45.73 | 47.43 |
| WebPart | We support. | 59,536 | 37.07 | 49.19 | 0.20 | 36.67 | 37.46 |
| WebPart | Ethical codex of. | 4887 | 56.86 | 49.96 | 0.71 | 55.46 | 58.26 |
| WebPart | Values | 788 | 89.65 | 42.72 | 1.52 | 86.66 | 92.64 |
| WebPart | Organizational structure | 48,627 | 51.89 | 50.56 | 0.23 | 51.44 | 52.34 |
| WebPart | Group | 69,580 | 61.15 | 52.51 | 0.20 | 60.76 | 61.54 |
| WebPart | Information for banks | 12,538 | 30.79 | 41.15 | 0.37 | 30.07 | 31.51 |
| WebPart | Business conditions | 242,308 | 53.25 | 54.56 | 0.11 | 53.03 | 53.47 |
| WebPart | Pricing list | 568,218 | 63.59 | 53.49 | 0.07 | 63.45 | 63.73 |
| WebPart | Claim order | 4009 | 68.28 | 51.18 | 0.81 | 66.70 | 69.87 |
| WebPart | ID | 6452 | 55.15 | 49.08 | 0.61 | 53.95 | 56.35 |
| WebPart | Contacts | 1,536,016 | 57.69 | 53.56 | 0.04 | 57.61 | 57.78 |
| WebPart | Branches and ATM | 641,776 | 89.71 | 44.20 | 0.06 | 89.60 | 89.82 |
| WebPart | Pillar 3 Q-terly Info | 86,846 | 38.58 | 47.87 | 0.16 | 38.26 | 38.90 |
| WebPart | Pillar 3 semiannualy info | 20,241 | 56.79 | 51.36 | 0.36 | 56.08 | 57.50 |
| WebPart | Annual reports | 53,886 | 73.28 | 52.35 | 0.23 | 72.84 | 73.72 |
| WebPart | General shareholder meeting | 9298 | 60.25 | 51.08 | 0.53 | 59.21 | 61.29 |
| WebPart | Financial reports | 1748 | 54.05 | 54.39 | 1.30 | 51.50 | 56.60 |
| WebPart | Emitent prospects | 21,692 | 71.14 | 49.85 | 0.34 | 70.48 | 71.81 |

| WebPart | His. | Awa. | Rat. | Mis. Vis. | Anti Mon. Laun. | We sup. | Eth. Cod. | Val. | Org. Struc. | Group | Infor. for Banks | Bus. Cond. | Pric. List | Claim Order | ID | Cont. | Bran. ATM | Pil. Q-terly Info | Pil. Sem. Info | Ann. Rep. | Gen. Shar. Meet. | Fin. Rep. | Em. Pros. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| History | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.066 | 0.000 | 0.000 | 0.485 |
| Awards | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.401 | 0.000 | 0.000 | 0.000 | 0.000 | 0.015 | 0.000 | 0.000 | 0.460 | 0.104 | 0.000 | 0.000 | 0.227 | 0.000 | 0.000 | 0.103 | 0.000 |
| Rating | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.520 | 0.000 | 0.000 | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.007 | 0.000 |
| Mision.Vis. | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.012 | 0.000 | 0.000 | 0.714 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| Anti Money Laundering | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| We support.. | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.072 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Ethical Codex of .. | 0.000 | 0.401 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.177 | 0.324 | 0.000 | 0.000 | 0.936 | 0.000 | 0.000 | 0.008 | 0.000 |
| Values | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.942 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Organizational Structure | 0.000 | 0.000 | 0.520 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.107 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.028 | 0.000 |
| Group | 0.000 | 0.000 | 0.000 | 0.012 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.011 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.283 | 0.000 | 0.000 |
| Information for Banks | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Business Conditions | 0.000 | 0.015 | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.107 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.340 | 0.000 |
| Pricing List | 0.000 | 0.000 | 0.000 | 0.714 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.011 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Claim Order | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 |
| ID | 0.000 | 0.460 | 0.000 | 0.000 | 0.000 | 0.000 | 0.177 | 0.000 | 0.001 | 0.000 | 0.000 | 0.062 | 0.000 | 0.000 | | 0.022 | 0.000 | 0.000 | 0.125 | 0.000 | 0.000 | 0.193 | 0.000 |
| Contacts | 0.000 | 0.104 | 0.000 | 0.000 | 0.000 | 0.000 | 0.324 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.022 | | 0.000 | 0.000 | 0.535 | 0.000 | 0.002 | 0.000 | 0.000 |
| Branches and ATM | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.942 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Pillar3 Q-terly Info | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.072 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Pillar3 Semiannualy Info | 0.000 | 0.227 | 0.000 | 0.000 | 0.000 | 0.000 | 0.936 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.125 | 0.535 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.006 | 0.000 |
| Annual Reports | 0.066 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 | 0.030 |
| General Shareholder Meeting | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.283 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | | 0.000 | 0.000 |
| Financial Reports | 0.000 | 0.103 | 0.007 | 0.000 | 0.000 | 0.000 | 0.008 | 0.000 | 0.028 | 0.000 | 0.000 | 0.340 | 0.000 | 0.000 | 0.193 | 0.000 | 0.000 | 0.000 | 0.006 | 0.000 | 0.000 | | 0.000 |
| Emitent Prospects | 0.485 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.030 | 0.000 | 0.000 | |

**Fig. 3** Newman-Keuls test for the web parts

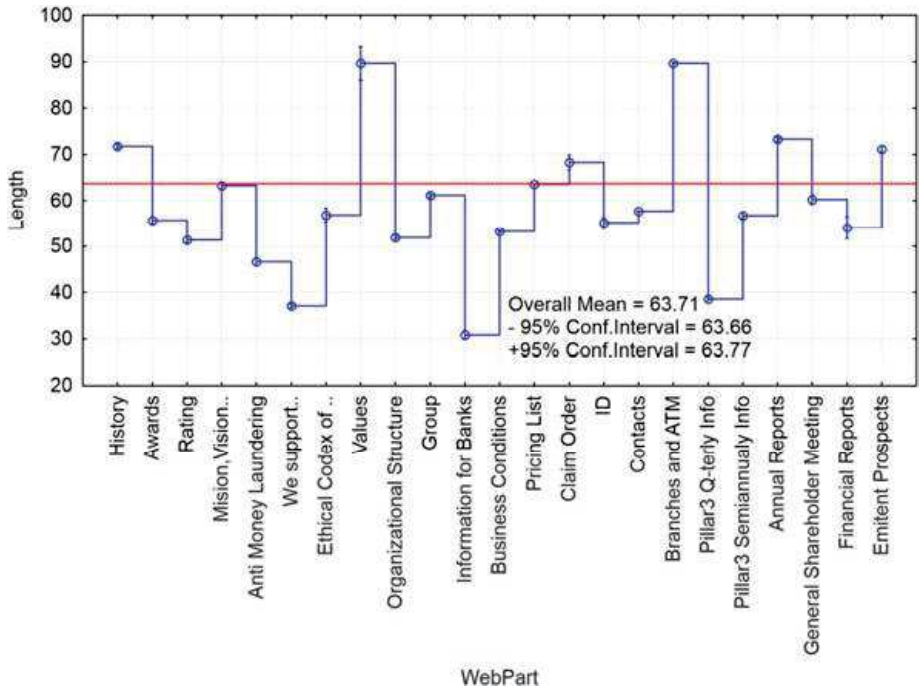



**Fig. 4** Visualization of the average time spent on the web parts in comparison to the total time

The time spent by the stakeholders (Fig. 4) was above average for the web parts: *Claim Order, Emitent Prospects, History, Annual Reports, Values* and *Branches and ATM*. The time spent was average for the web parts: *Mision,Vision.* and *Pricing List*. In the case of the remaining web parts, the time spent was below average.

The users spent the least time on the *We support.* web part. This web part offers general information of the bank category, as well as does the *Information for Banks* web part related to Pillar 3, and the *Pillar 3 Q-terly Info*, which belongs to the mandatory disclosure category (<40).

On the other hand, the web users spent most of their time on the *Mision, Vision., Pricing List, Claim Order, History, Values, Branches and ATM* web parts, offering general information on the bank and the *General Shareholder Meeting, Group, Issuer Prospects, Annual Reports* associated with Pillar 3 (>60).

The descriptive statistics (Table 4) show that the variability is almost similar for all the examined Pillar 3 web parts. Also, the standard error is close to zero, however higher ($>$ 1) is only in the case of less frequent web part *Financial Reports*, where the number of accesses was less than 2000.

A more detailed view on the results (Table 4) show that the users spent most of their time on Pillar 3 web parts *Annual Reports* and *Emitent Prospects*. On the other hand, the least time was spent by the users on the Pillar 3 web part *Information of Banks*. Based on these results, a null statistical hypothesis will be tested that claims that there is no statistically significant difference in the time spent on the Pillar 3 web parts.

**Table 4** Descriptive characteristics of the time spent on the examined Pillar 3 web parts

| | Level of factor | *N* | Length mean | Length Std.Dev | Length Std.Err | Length − 95.00% | Length + 95.00% |
|---|---|---|---|---|---|---|---|
| Total | | 298,614 | 55.11 | 52.16 | 0.10 | 54.93 | 55.30 |
| WebPart | Rating | 22,785 | 51.35 | 50.63 | 0.34 | 50.69 | 52.00 |
| WebPart | Group | 69,580 | 61.15 | 52.51 | 0.20 | 60.76 | 61.54 |
| WebPart | Information for banks | 12,538 | 30.79 | 41.15 | 0.37 | 30.07 | 31.51 |
| WebPart | Pillar 3 Q-terly info | 86,846 | 38.58 | 47.87 | 0.16 | 38.26 | 38.90 |
| WebPart | Pillar 3 semiannualy info | 20,241 | 56.79 | 51.36 | 0.36 | 56.08 | 57.50 |
| WebPart | Annual reports | 53,886 | 73.28 | 52.35 | 0.23 | 72.84 | 73.72 |
| WebPart | General shareholder meeting | 9298 | 60.25 | 51.08 | 0.53 | 59.21 | 61.29 |
| WebPart | Financial reports | 1748 | 54.05 | 54.39 | 1.30 | 51.50 | 56.60 |
| WebPart | Emitent prospects | 21,692 | 71.14 | 49.85 | 0.34 | 70.48 | 71.81 |

Based on the ANOVA univariate results ($F = 2851.6$, $p < 0.001$), the zero hypotheses rejected at the 0.1% significance level, i.e. a statistically significant difference in the time spent on the examined website parts was identified.

After rejecting the global zero hypotheses, the interest is in which pairs have a statistically significant difference. Three homogeneous groups (Table 5) were identified (*Rating, Financial Reports*), (*Financial Reports, Pillar 3 Semiannualy Info*) and (*General Shareholder Meeting, Group*), statistically significant differences were identified between the other pairs at the 1% significance level.

The time spent by the stakeholders (Fig. 5) was above average in the case of Pillar 3 web parts: *Pillar 3 Semiannualy Info, General Shareholder Meeting, Group, Emitent Prospects* and *Annual Reports* and average in the case of Pillar 3 web part *Financial Reports*. In the case of the remaining Pillar 3 web parts was the time by the stakeholders spent below average.

Based on the results of the previous research [19] it was decided to analyse the *Pillar 3 Q-terly Info* web part. The web part was divided into another seven categories (Table 6).

The descriptive statistics (Table 6) show that the variability is almost similar for all the examined Pillar 3 Q-terly info categories. Also, the standard error is close to zero, however higher ($> 2$) is only in the case of the least frequent Pillar 3 Q-terly info categories *Consolidated Statements* and *Financial Indicators,* where the number of accesses was less than 400. Also, higher (>1) is in the case of categories

**Table 5** Unequal N HSD for the Pillar 3 Q-terly categories

| WebPart | Rating | Group | Inform. for banks | Pillar 3 Q-terly info | Pillar 3 semi. info | Annual reports | General Shar. meeting | Finan. reports | Emit. prosp. |
|---|---|---|---|---|---|---|---|---|---|
| Rating | | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.80984 | 0.00001 |
| Group | 0.00001 | | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.95043 | 0.00101 | 0.00001 |
| Inform. for banks | 0.00001 | 0.00001 | | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 |
| Pillar 3 Q-terly info | 0.00001 | 0.00001 | 0.00001 | | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 |
| Pillar 3 Semiann. info | 0.00001 | 0.00001 | 0.00001 | 0.00001 | | 0.00001 | 0.00013 | 0.79927 | 0.00001 |
| Annual reports | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | | 0.00001 | 0.00001 | 0.00036 |
| General Sharehol. meeting | 0.00001 | 0.95043 | 0.00001 | 0.00001 | 0.00013 | 0.00001 | | 0.00828 | 0.00001 |
| Financial reports | 0.80984 | 0.00101 | 0.00001 | 0.00001 | 0.79927 | 0.00001 | 0.00828 | | 0.00001 |
| Emitent prospects | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00036 | 0.00001 | 0.00001 | |

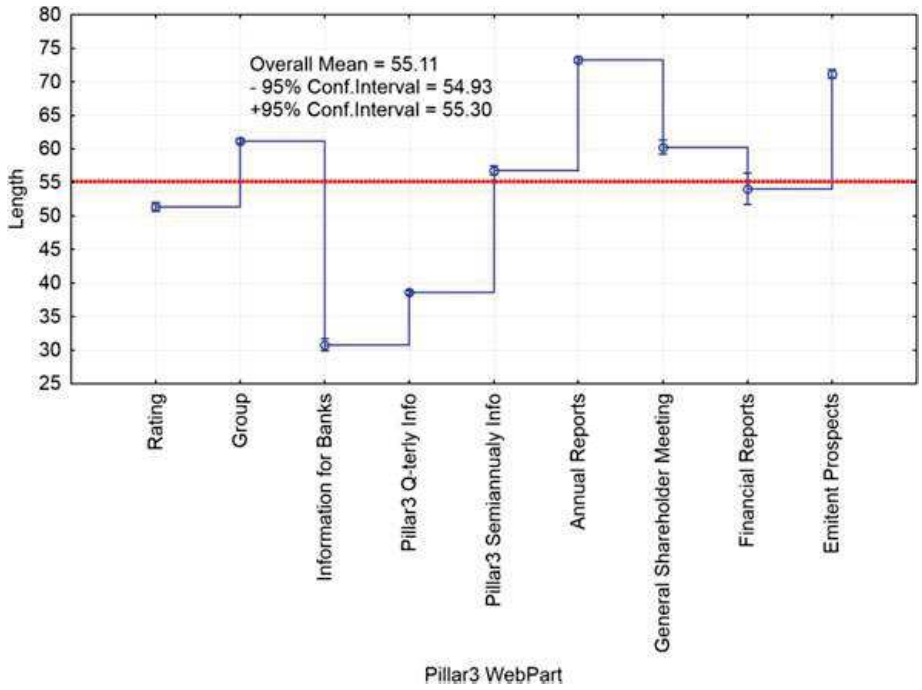



**Fig. 5** Visualization of the average time spent on the Pillar 3 web parts in comparison to the total time

**Table 6** Descriptive characteristics of the time spent on the examined Pillar 3 Q-terly info categories

| | Level of factor | *N* | Length mean | Length Std.Dev | Length Std.Err | Length 95,00% | Length + 95,00% |
|---|---|---|---|---|---|---|---|
| Total | | 86,846 | 38.58 | 47.87 | 0.16 | 38.26 | 38.90 |
| Pillar 3 | Other information | 64,721 | 28.38 | 41.74 | 0.16 | 28.06 | 28.70 |
| Pillar 3 | Individual financial statements | 8612 | 92.22 | 43.62 | 0.47 | 91.30 | 93.14 |
| Pillar 3 | Information on bank | 9976 | 43.76 | 49.73 | 0.50 | 42.79 | 44.74 |
| Pillar 3 | Financial indicators | 325 | 57.63 | 42.95 | 2.38 | 52.94 | 62.32 |
| Pillar 3 | Shareholders | 1657 | 85.44 | 43.39 | 1.07 | 83.35 | 87.54 |
| Pillar 3 | Consolidated statements | 394 | 55.67 | 47.39 | 2.39 | 50.98 | 60.36 |
| Pillar 3 | Risk management | 1161 | 87.05 | 43.95 | 1.29 | 84.52 | 89.58 |

*Risk management* and *Shareholders,* where the number of accesses was less than 2000.

The results (Table 6) show there are differences between the visit times of the examined Pillar 3 Q-terly info categories. The web users spent the most time on the categories *Individual Financial Statements, Shareholders* and *Risk Management*. Least time was spent on the categories *Other information* and *Information on Bank*. Based on these results, a null statistical hypothesis will be tested that claims that there is no statistically significant difference in the time spent on the examined Pillar 3 Q-terly info categories. Based on the ANOVA univariate results ($F = 3458.1$, $p < 0.001$), the zero hypotheses rejected at the 0.1% significance level, i.e. a statistically significant difference in the time spent on the examined Pillar 3 Q-terly info website categories was identified.

After rejecting the global zero hypotheses, the interest is in which pairs have a statistically significant difference. Three homogeneous groups (Table 7) have been identified (*Consolidated Statements, Financial Indicators*), (*Shareholders, Risk Management*) and (*Risk Management, Individual Financial Statements*), statistically significant differences were identified between the remaining pairs at the 1% significance level.

The time spent by the stakeholders (Fig. 6) was above average in the case of all Pillar 3 Q-terly info categories except the *Other information* category. In the case of the *Other information* category was the time spent by the stakeholders below average.

**Table 7** Unequal N HSD for the Pillar 3 Q-terly info categories

| Pillar 3 | Other infor | Individ. financ. statem | Informat. on bank | Financ. indicat | Shareh | Consol. statem | Risk manag |
|---|---|---|---|---|---|---|---|
| Other information | | 0.000026 | 0.000026 | 0.000026 | 0.000026 | 0.000026 | 0.000026 |
| Individual financial statements | 0.000026 | | 0.000026 | 0.000026 | 0.000138 | 0.000026 | 0.058041 |
| Information on bank | 0.000026 | 0.000026 | | 0.000792 | 0.000026 | 0.001974 | 0.000026 |
| Financial indicators | 0.000026 | 0.000026 | 0.000792 | | 0.000026 | 0.997349 | 0.000026 |
| Shareholders | 0.000026 | 0.000138 | 0.000026 | 0.000026 | | 0.000026 | 0.972818 |
| Consolidated statements | 0.000026 | 0.000026 | 0.001974 | 0.997349 | 0.000026 | | 0.000026 |
| Risk management | 0.000026 | 0.058041 | 0.000026 | 0.000026 | 0.972818 | 0.000026 | |

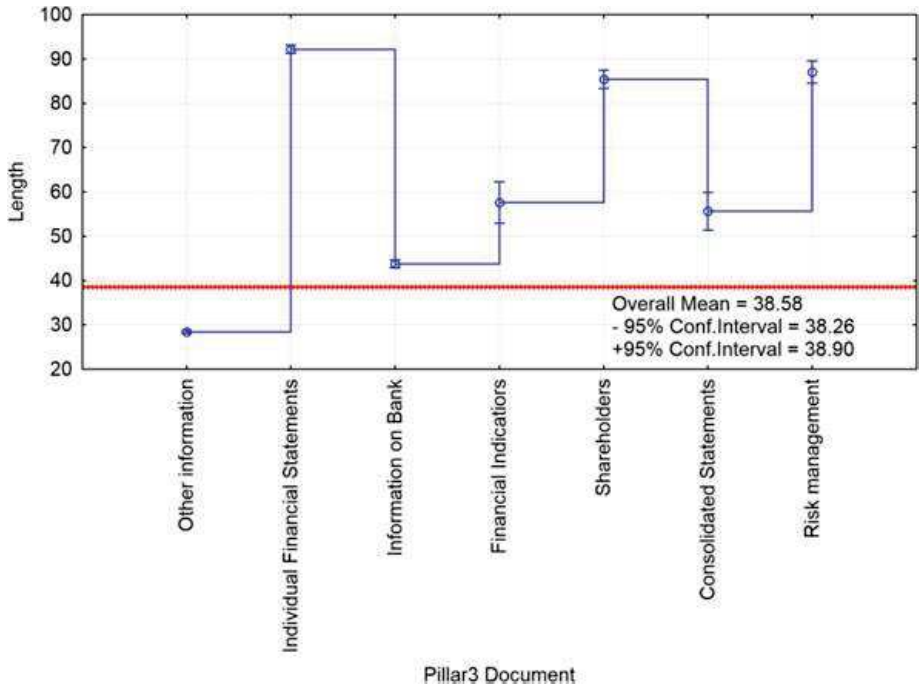



**Fig. 6** Visualization of the average time spent on the Pillar 3 Q-terly info categories in comparison to the total time

## 5 Conclusion

This paper analyzed Pillar 3 disclosures in relation to market discipline by web portal analysis. The analysis was based on the time spent on the web portal of banking institution by the web users and analyzing their interest in relation to time spent and content of the Pillar 3 disclosures. The results of the analysis of time spent on a web portal of banking institution have shown that the most average time was spent on the web part Annual Reports and Emitent Prospects. Pillar 3 Q-terly web part was with the second least average time spent and was analyzed in more detail. Its categories Individual Financial Statements, Shareholders and Risk Management have been identified as the web pages with the most time spent by the web users. Information of a long time spent on the pages can indicate that there is important content for web users. On the other hand, it can also imply that there is too much content on the pages. These results are in conjunction with Giner et al. [30], higher value of financial risk disclosure than Pillar 3 disclosures) and De Araujo and Leyshon [27], higher value of other type of information than Pillar 3 disclosures. Moreover, these results are also in line with Munk et al. [19], whose results show that the most visited part of the disclosures was part Group and web users' highest interest is in not in Pillar 3 disclosures as solo information, but together with Annual reports or Information on Group. Our findings could be used by regulators in the discussion process of designing new regulation, for institutions in the process of designing risk disclosure and web design. Moreover, information relevancy of the disclosures is

important for key interested parties in the financial market analysis. Lastly, we have identified analysis of categories of Pillar 3 disclosures in more detail as a topic for future research.

**Acknowledgements** This paper was created and its results supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contracts No. VEGA-1/0776/18.

## References

1. Sowerbutts, R., Zimmerman, P., Zer, I.: Banks' Disclosure and Financial Stability. (2013)
2. Flannery, M.J., Bliss, R.R.: Market discipline in regulation: pre-and post-crisis. Forthcoming, Oxford Handb. Bank. 3e. (2019)
3. Hadad, M.D., Agusman, A., Monroe, G.S., Gasbarro, D., Zumwalt, J.K.: Market discipline, financial crisis and regulatory changes: Evidence from Indonesian banks. J. Bank. Financ. **35**, 1552–1562 (2011). https://doi.org/10.1016/j.jbankfin.2010.11.003
4. Biljanovska, B.: Aligning market discipline and financial stability: a more gradual shift from contingent convertible capital to bail-in measures. Eur. Bus. Organ. Law Rev. **17**, 105–135 (2016). https://doi.org/10.1007/s40804-016-0028-0
5. González, L.O., Rodríguez Gil, L.I., Martorell Cunill, O., Merigó Lindahl, J.M.: The effect of financial innovation on European banks' risk. J. Bus. Res. **69**, 4781–4786 (2016). https://doi.org/10.1016/j.jbusres.2016.04.030
6. Fonseca, A.R., González, F.: How bank capital buffers vary across countries: the influence of cost of deposits, market power and bank regulation. J. Bank. Financ. **34**, 892–902 (2010). https://doi.org/10.1016/J.JBANKFIN.2009.09.020
7. Acharya, V.V., Ryan, S.G.: Banks' financial reporting and financial system stability. J. Account. Res. **54** (2016). https://doi.org/10.1111/1475-679X.12114
8. Jizi, M.I., Dixon, R.: Are risk management disclosures informative or tautological? Evidence from the U.S. Banking Sector. Account. Perspect. **16** (2017). https://doi.org/10.1111/1911-3838.12134
9. Nahar, S., Azim, M., Anne Jubb, C.: Risk disclosure, cost of capital and bank performance. Int. J. Account. Inf. Manag. **24**, 476–494 (2016). https://doi.org/10.1108/IJAIM-02-2016-0016
10. Heinle, M.S., Smith, K.: A theory of risk disclosure. SSRN Electron. J. (2015). https://doi.org/10.2139/ssrn.2635074
11. Goldstein, I., Leitner, Y.: Stress tests and information disclosure. Federal Reserve Bank of Philadelphia (2015).
12. Marsh, W.B., Roman, R.: Bank financial restatements and market discipline. Fed. Reserv. Bank Kansas City Econ. Rev. **103**, 1–29 (2018). https://doi.org/10.18651/ER/2q18MarshRoman
13. Kuranchie-Pong, L., Bokpin, G.A., Andoh, C.: Empirical evidence on disclosure and risk-taking of banks in Ghana. J. Financ. Regul. Compliance. **24**, 197–212 (2016). https://doi.org/10.1108/JFRC-05-2015-0025
14. Ojo, M.: The impact of capital and disclosure requirements on risks and risk taking incentives. MPRA Pap. (2010)
15. Fontes, A., Rodrigues, L.L., Craig, R.: A response to commentaries on a theoretical model of stakeholder perceptions of a new financial reporting system. Account. Forum. **41** (2017). https://doi.org/10.1016/j.accfor.2017.04.001
16. Munk, M., Kapusta, J., Švec, P.: Data preprocessing evaluation for web log mining: Reconstruction of activities of a web visitor. In: Procedia Computer Science. pp. 2273–2280 (2010). https://doi.org/10.1016/j.procs.2010.04.255

17. Munk, M., Kapusta, J., Švec, P., Turčáni, M.: Data advance preparation factors affecting results of sequence rule analysis in web log mining. E+M Ekon. Manage. **13**, 143–160 (2010)
18. Munk, M., Benko, Ľ., Gangur, M., Turčáni, M.: Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. E+M Ekon. Manage. **18**, 144–159 (2015). https://doi.org/10.15240/tul/001/2015-3-013
19. Munk, M., Pilkova, A., Benko, L., Blaženková, P.: Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. J. Bus. Econ. Manage. **18**, 954–973 (2017). https://doi.org/10.3846/16111699.2017.1360388
20. Drlik, M., Munk, M.: Understanding time-based trends in stakeholders' choice of learning activity type using predictive models. IEEE Access. **7**, 3106–3121 (2019). https://doi.org/10.1109/ACCESS.2018.2887057
21. Niessen-Ruenzi, A., Parwada, J.T., Ruenzi, S.: Information effects of the Basel bank capital and risk Pillar 3 disclosures on equity analyst research an exploratory examination. SSRN Electron. J. (2015). https://doi.org/10.2139/ssrn.2670418
22. Parwada, J.T., Ruenzi, S., Sahgal, S.: Market discipline and basel Pillar 3 reporting. SSRN Electron. J. (2013). https://doi.org/10.2139/ssrn.2443189
23. Vauhkonen, J.: The impact of Pillar 3 disclosure requirements on bank safety. J. Financ. Serv. Res. **41**, 37–49 (2012). https://doi.org/10.1007/s10693-011-0107-x
24. Frolov, M.: Why do we need mandated rules of public disclosure for banks? J. Bank. Regul. **8**, 177–191 (2007). https://doi.org/10.1057/palgrave.jbr.2350045
25. Chen, Y., Du, K.: The role of information disclosure in financial intermediation with investment risk. J. Financ. Stab. **46**, 100720 (2020). https://doi.org/10.1016/j.jfs.2019.100720
26. Naz, M., Ayub, H.: Impact of risk-related disclosure on the risk-taking behavior of commercial banks in Pakistan. J. Indep. Stud. Res. Soc. Sci. Econ. **15** (2017). https://doi.org/10.31384/jisrmsse/2017.15.2.9
27. de Araujo, P., Leyshon, K.I.: The impact of international information disclosure requirements on market discipline. Appl. Econ. **49**, 954–971 (2016). https://doi.org/10.1080/00036846.2016.1208361
28. Benli, V.F.: Basel's Forgotten pillar: the myth of market discipline on the forefront of Basel III. Financ. Internet Q. **11**, 70–91 (2015)
29. Wilms, W.: The dark side of the Basel committee's Pillar 3 framework. http://www.eurofiling.info/201411/presentations/20141125TheDarkSideOfTheBasselCommitteesWilfriedWilms.pdf
30. Giner, B., Allini, A., Zampella, A.: The value relevance of risk disclosure: an analysis of the banking sector. Account. Eur. (2020). https://doi.org/10.1080/17449480.2020.1730921
31. Scannella, E.: Market risk disclosure in banks' balance sheet and Pillar 3 report: The Case of Italian Banks (2018)
32. Bischof, J., Daske, H., Elfers, F., Hail, L.: A tale of two regulators: risk disclosures, liquidity, and enforcement in the banking sector. SSRN Electron. J. (2016). https://doi.org/10.2139/ssrn.2580569
33. Faria-e-Castro, M., Martinez, J., Philippon, T.: Runs versus lemons: information disclosure and fiscal capacity. Cambridge, MA (2017). https://doi.org/10.3386/w21201
34. Del Gaudio, B.L., Megaravalli, A.V., Sampagnaro, G., Verdoliva, V.: Mandatory disclosure tone and bank risk-taking: evidence from Europe. Econ. Lett. **186**, 108531 (2020). https://doi.org/10.1016/j.econlet.2019.108531
35. Matuszak, L., Rozanska, E.: An examination of the relationship between CSR disclosure and financial performance: the case of Polish Banks. J. Account. Manage. Inf. Syst. **16**, 522–533 (2017)
36. Fijalkowska, J., Zyznarska-Dworczak, B., Garsztka, P.: The Relation between the CSR and the accounting information system data in central and Eastern European (CEE) countries—the evidence of the polish financial institutions. J. Account. Manag. Inf. Syst. **16**, 490–521 (2017)
37. Bartulovic, M., Pervan, I.: Comparative analysis of voluntary internet financial reporting for selected CEE countries. Recent Res. Appl. Econ. Manage. **1**, 296–301 (2012)

38. Arsov, S., Bucevska, V.: Determinants of transparency and disclosure—evidence from post-transition economies. Econ. Res. Istraživanja. **30**, 745–760 (2017). https://doi.org/10.1080/1331677X.2017.1314818
39. Hąbek, P.: CSR reporting practices in visegrad group countries and the quality of disclosure. Sustainability. **9**, 1–18 (2017)
40. Moh, T.S., Saxena, N.S.: Personalizing web recommendations using web usage mining and web semantics with time attribute. Commun. Comput. Inf. Sci. **54**, 244–254 (2010). https://doi.org/10.1007/978-3-642-12035-0_24
41. Masseglia, F., Poncelet, P., Teisseire, M., Marascu, A.: Web usage mining: Extracting unexpected periods from web logs. Data Min. Knowl. Discov. **16**, 39–65 (2008). https://doi.org/10.1007/s10618-007-0080-z
42. Cooley, R., Mobasher, B., Srivastava, J.: others: data preparation for mining world wide web browsing patterns. Knowl. Inf. Syst. **1**, 5–32 (1999)

PRÍLOHA G: SVEC, PETER, LUBOMIR BENKO, MIROSLAV KADLECIK, JAN KRATOCHVIL A MICHAL MUNK, 2020. WEB USAGE MINING: DATA PRE-PROCESSING IMPACT ON FOUND KNOWLEDGE IN PREDICTIVE MODELLING. *PROCEDIA COMPUTER SCIENCE*. 171, 168–178. DOI:10.1016/J.PROCS.2020.04.018 (**SCOPUS**) [SCOPUS: 10]

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 168–178

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

# Web Usage Mining: Data Pre-processing Impact on Found Knowledge in Predictive Modelling

Peter Svec[a*], Lubomir Benko[a], Miroslav Kadlecik[a], Jan Kratochvil[b], Michal Munk[a]

[a]*Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia*
[b]*University of Pardubice, Studentska 95, 532 10 Pardubice, Czech republic*

**Abstract**

We describe the importance of pre-processing data in the web usage mining and the impact of mistakes to the analysis of the data during this phase. We analyse data of commercial bank and accesses of stakeholders to the selected part of the website according to the rules of Basel 2 Pillar 3. These rules are focused on the Market discipline and were created after the financial crisis in 2008. We model the time dependent behaviour of the web user. Modelling the probabilities of the accesses to web categories depending on time was done using the multinomial logit model, which is a part of generalised linear models. We found non-human access to the web portal during the modelling phase, which significantly influenced the obtained knowledge. We had to repeat the pre-processing phase and repeat the modelling. After the identification and removing of the problem accesses to web portal pages, we identified a new model parameter from which we calculated the logit estimates and subsequently estimated the probabilities of accesses to the web parts of the web portal. Our estimates of theoretical logits fit (model) empirical logits. It is essential to not underestimate the data pre-processing phase in the process of web usage mining, as the pre-processing phase directly affects the quality of the acquired knowledge.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).

*Keywords:* Web Log Mining; Multinomial Logit Model; Data pre-processing;

* Corresponding author. Tel.: +421 37 6408 675
*E-mail address:* psvec@ukf.sk

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19).
10.1016/j.procs.2020.04.018

## 1. Introduction

One of the essential parts of web usage mining is data pre-processing. We used data cleaning, integration, reduction and data conversion methods in the pre-processing level of data analysis. The quality of the analysis results depends on the quality of data pre-processing. Data cleaning, as the simplest step of data pre-processing, is non–trivial as the analysed content is highly specific. Even if we make the best during pre-processing, methods used for analysis can show that there is a hidden problem in the data pre-processing.

Moreover, after excluding problematic data, the results of the analysis differ. We focused on the data-preprocessing methods of the portal with frequent changes in the content and the structure too. We used the multinomial logit model to model the distribution of the category variable. Specially we modelled the time dependent probabilities of access to web parts of a web portal. We applied the model on academic (university/faculty web portal, virtual learning environment) as well as on commercial environment (e-shop systems, a web portal of commercial bank).

This article is focused on the results of the analysis of the web portal of the commercial bank. We show the influence of the data pre-processing to the obtained knowledge. Inconsistent data cleaning has a negative influence on the correct type of dependence or the creation of dependence itself. The article aims to introduce the pre-processing techniques that led to obtaining reliable data of web usage as well as alternative results evaluation techniques through which we can identify problem parts of the file.

## 2. Background

The source of data for further analysis was the web server log file. It doesn't matter what web server was installed on the portal; the structure of log files standardised. Some system administrators change the standard structure of the log file but always keep the mandatory information. A web server log is in its default form known as Common Log File and keeps information about IP address; date and time of visit; accessed and referenced resource. If we use the extended version of the log file, we can collect more information, e.g. type of browser (User-Agent Field). Webserver logs every information generated by the user's browser, including requests for cascading style-sheets, images, fonts etc. As the log file is saved as the text file on the webserver, the size of the log file is several hundred megabytes each day. We are able to reduce the size of the log by removing all client requests on multimedia files, cascade styles sheet, javascript and other non-content file types.

A lot of traffic on portals is generated in an artificial way (robots). Usually, we see robots of search engines like Google, Bing, etc. but there are many not well-knows robots of various small indexing services [1]. If the robot is polite, it accesses the *robots.txt* file and shows us the robot identification string. But there are many internet services, which crawl the internet and collect information for illegal activities, like collecting e-mail addresses for SPAM delivery, phone numbers, photos etc. The robot visits each site of the page and access of a robot is not valuable information for the analysis. We need to distinguish regular user and the robot. The difference between the robot and the regular user can be determined by various standard methods (click rate, HTML-to-Image ratio, percentage of file requests, percentage of 4xx error responses, percentage of HEAD requests, or access to robot.txt file) or we can employ methods like standard deviation of requested page's depth [1], navigational patters analysis [2], cluster analysis [3], classification techniques [4]. By removing unnecessary data and robots accesses; the log file reduces its size to just 10 per cent of the original one [5].

The analysis we used depends on the identification of the session [6–8]. The session identification method needs the IP address of the user. We can distinguish various users who share the same IP address based on the different browser they use. There is also a new phenomenon, access from the private, anonymous network.

The most popular anonymous network is Tor, also known as Onion Network. The principle of the Onion Network was designed in 1995 David Goldschlag, Mike Reed a Paul Syverson from the US Naval Research Lab (NRL) to keep the identity of the source and destination secret [9]. Tor anonymize the identity of users by connecting through a series of tunnels and nodes which are changing during the web session.

A user does not connect to the destination web server directly. A user sends a request anonymously to one Tor router from the directory servers. Once a connection is established, traffic is relayed to the first router (Entry Guard)

and generates a session key using the Deffie-Hellman key exchange. This process repeats among other routers (hops). The last hop (exit router) communicates with the destination web server as a socks proxy. [10]

Another popular service to hide the identity of the user is JonDonym, which implements web mixes [11] and mix networks [12] for communication. The JonDonym selects a static cascade of (usually) three mixes over which he intends to relay his web traffic [13]. We can find just last IP of the mix in the webserver log file.

The third popular service which hides user identity is the I2P. Each client of the I2P network creates a series of tunnels for incoming and outgoing traffic – direct p2p network. The number of tunnels is the key factor in achieving the anonymity, latency and throughput [14]. There are no official exit points from the network – each node can be the exit point.

## 3. Methods

While visiting the website, the browser sends a lot of data to the webserver. Each request for data of the webpage is logged to a log file. Logs provide primary data and cannot make relevant distinctions such as distinguishing between the time a user spends reading the webpage and the time a user left the screen [15]. To analyse the user's behaviour, we tagged log lines to form a session. The session represents a complete sequence of views. We used log file from the commercial bank web server, which covers a time period from 2009 to 2012. This web server provides bank public information. The selected period was selected by design as it covers website accesses before and during the Recent Financial Crisis [16,17].

We focused on modelling the bank visitors' behaviour using a multinomial logit model [18]. The log file kept access records to one part of the bank portal over an extended period of time (2009 – 2012). The size of the log files was several gigabytes, but most records in the log file were useless for further analysis and were removed using. We can use traditional approach [19], using big data approach [20] or use neural networks [21] to remove unnecessary data. The remaining log file was just 10 % of its original size.

The methodology of pre-processing the data consist of several steps. Firstly, we removed all client requests for multimedia files, cascade styles sheets, javascripts and other file types which are not valuable for the content analysis. We also removed all request with return code other than 2xx and 3xx. During this phase of data pre-processing, we removed search engines robots based on the *useragent* field, based on the access of client browser to *robots.txt* file [22] or based on the well-know IPs of search engines robots. We also used the number of different user agents per IP, where a high number of user agents identify robot [23]. We were not able to use the Image-to-Html ratio as we removed the request for images.

### *3.1. User/Session Identification*

The second step in data preparation was a user/session identification. A regular user uses the back button of the browser as the natural process of web navigation. The use of the back button is popular on mobile devices. If the user hits the back button, there is no request to the web server and the browser shows the content from its cache. This behaviour of the browser breaks the chain of recorded lines in the log file. We had to add those missing lines employing the Path completion [24]. We also distinguished employees of the bank and regular users outside the bank. Bank employees are accessing from the internal network, and their IP addresses are from the private IP pool according to RFC 1918.

While various users can share the same IP address or computer, we used the Reference Length method [8,25–27] for user/session identification. Sharing the IP is typical for home routers, proxies or VPN services. We also had to identify users who used anonymization tools – TOR, JonDonym or I2P. We were not able to detect users who use I2P.

As the IP of the users changes during the visit of the website, we had to remove accesses from the TOR network in our analysis. To detect TOR, we cannot rely on the user agent, as the TOR browser is using Mozilla user agent string. The browser is build based on Firefox. There are some approaches to identify TOR access [28–32]. We use the official list of exit points published on https://check.torproject.org, unofficial one published on https://www.dan.me.uk/tornodes and API available on https://www.ipqualityscore.com/tor-ip-address-check.

We can detect access from JonDonym, similar to TOR. We cannot rely on the user agent, as it is also a fake one. The list of IPs used as the last IP of the mix is available at http://78.129.207.59:8080/exitaddresses. In comparison to TOR, the user browses the website just from one endpoint, and we can use the user/session identification. We don't need to know the origin IP of the user as we focus only on his/her behaviour.

### *3.2. Variables determination*

The third step of data pre-processing is the determination of the variable for the user behaviour analysis. We asked a bank expert to tag each webpage of the portal with the terminology used in the financial environment according to Basel 2 Pillar 3. The expert created 19 different part of the web (*webpart*) and created six categories (*Pricing list, Reputation, Business Conditions, Pillar3 related, Pillar3 disclosure requirements, We support..*) [33]. An example of the tagged record is in Table 1.

Table 1. Tagged URL example

| URL | Web part | Category |
|---|---|---|
| /about/bank/sadzobnik.html | Pricing List | Pricing List |
| /att/sadzobnikFO010109.pdf | Pricing List | Pricing List |
| /about/bank/ocenenia.html | Awards | Reputation |
| /about/bank/obchodnepodm.html | Business Conditions | Business Conditions |
| /att/76823/VOP010109.pdf | Business Conditions | Business Conditions |
| /att/76823/VOP010109.pdf | Business Conditions | Business Conditions |
| /about/bank/ocenenia.html | Awards | Reputation |

Pillar 3 related information can be found in bank annual reports or minutes from general assembly meetings. The Pillar 3 Disclosure Requirements consists of general information about the bank, e.g. organizational chart, different activities of the bank, information about employees), financial information or information on risk management.

To use the multinomial logit model, we had to identify independent variables-predictors which consisted of the week number, the quartal of the year and the year. We also distinguished the period before the financial crisis and after the financial crisis. We modelled the time dependent access probabilities to content categories of the web portal and the period of the financial crisis where time was represented by the variable *t* and its square and the period of the financial crisis by a variable *crisis*.

## 4. Results

The input for the analytical procedures is pre-processed and transformed data. The output of the analytical procedures is knowledge. The first step of the methodology was the parameter estimation using the maximizing the logarithm of the multinomial likelihood function. The parameter estimation for individual data was done using the linear/non-linear models of the STATISTICA system. We test the significance of the parameters using the Wald test.

Based on the test results of all effects (Table 2), the parameters of the model were statistically significant. Significant parameters are discoloured.

Table 2 Test of all effects of the model

| | df | Wald Stat. | p |
|---|---|---|---|
| Intercept | 5 | 83200.9 | 0.0001 |
| t | 5 | 130029.6 | 0.0000 |
| t2 | 5 | 74714.7 | 0.0001 |
| *Crisis* | 5 | 155158.7 | 0.0000 |

We see in Table 3 that logits for all of the categories are significantly dependent on time and the square of time. The values of these logits are significantly influenced by the variable that identifies the period of crisis also.

Table 3 Parameter estimation of the model

| | Web Part | Estimate | Std. Error | Wald Stat. | P |
|---|---|---|---|---|---|
| Intercept 1 | Pricing List | -2.0144 | 0.0092 | 48084.9 | 0.0001 |
| t | Pricing List | 0.5504 | 0.0017 | 109626.5 | 0.0000 |
| $t^2$ | Pricing List | -0.0167 | 0.0001 | 60342.3 | 0.0001 |
| Crisis | Pricing List | -1.2745 | 0.0050 | 66001.2 | 0.0001 |
| Intercept 2 | Reputation | -1.5187 | 0.0098 | 23953 | 0.0001 |
| t | Reputation | 0.2145 | 0.0018 | 13534.7 | 0.0017 |
| $t^2$ | Reputation | -0.0054 | 0.0001 | 4891.5 | 0.0047 |
| Crisis | Reputation | -0.4721 | 0.0061 | 6055.5 | 0.0023 |
| Intercept 3 | Business Conditions | -2.3758 | 0.0116 | 41901.4 | 0.0001 |
| t | Business Conditions | 0.5298 | 0.0020 | 69708.4 | 0.0001 |
| $t^2$ | Business Conditions | -0.0155 | 0.0001 | 38394.3 | 0.0001 |
| Crisis | Business Conditions | -2.1470 | 0.0059 | 131618 | 0.0000 |
| Intercept 4 | Pillar3 related | -0.9698 | 0.0085 | 13045.2 | 0.0016 |
| t | Pillar3 related | 0.2541 | 0.0017 | 23695 | 0.0001 |
| $t^2$ | Pillar3 related | -0.0077 | 0.0001 | 11850.7 | 0.0012 |
| Crisis | Pillar3 related | -0.8076 | 0.0054 | 22394.1 | 0.0001 |
| Intercept 5 | Pillar3 disclosure requirements | -2.4173 | 0.0123 | 38715.5 | 0.0001 |
| t | Pillar3 disclosure requirements | 0.4117 | 0.0022 | 34467.8 | 0.0001 |
| $t^2$ | Pillar3 disclosure requirements | -0.0133 | 0.0001 | 21289.7 | 0.0001 |
| Crisis | Pillar3 disclosure requirements | -1.0369 | 0.0065 | 25375.9 | 0.0001 |

We can calculate the logits estimates using the estimated parameters (Table 3) and subsequently estimate the probabilities of each category in each hour of the day. The results of Table 3 correlate with the calculated probabilities (Figure 1a).

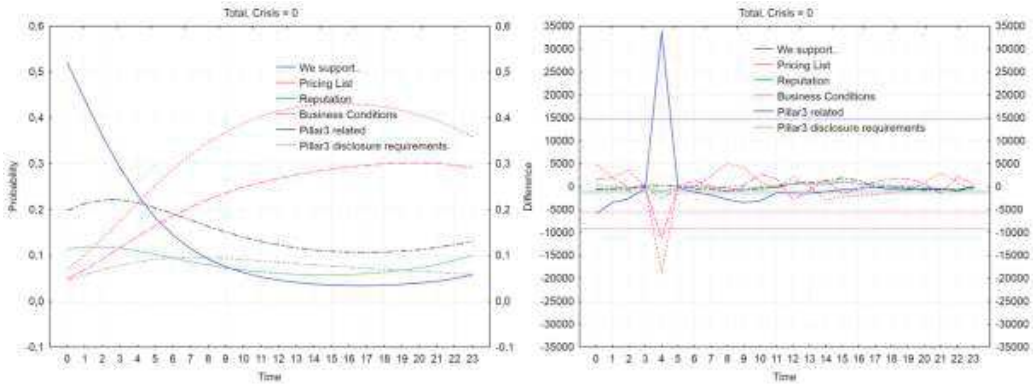



Figure 1 (a) Probability visualization for the examined categories (b) Differences of counts of the model

The plot shows the access probabilities for the examined web parts by the stakeholders in the period after the financial crisis. We can see from Figure 1a that the access probabilities for the web part *We support..*, *Pricing List* and *Business Conditions* are dependent on time quadratic.

We can see in Figure 1a that the probability of access for the category *We support..* is in early morning hours relatively high and the evening it gradually decreases. On the other hand, the probability of access for *Pricing List* or *Business Conditions* is small in the morning and increases during the day. The access probabilities for other web parts are relatively small for the period after the financial crisis.

### *4.1. Results evaluation*

Assuming the expected counts are large enough, i.e. they are non-zero and no more than 20 % of the expected counts are less than 5, to compare the current model with the saturated model that estimates the probabilities independently for $i = 0, 1, \dots, 23$, we can use the LR test

$$LR(\hat{\pi}) = 2\sum_{i=0}^{23}\sum_{j=1}^{J} y_{ij}\, ln\frac{y_{ij}}{\hat{y}_{ij}}. \quad (1)$$

We can compare our estimates of the theoretical count with empirical counts using the LR test (1). In our case, the value of the LR test was small ($df = 1034055, LR\ stat. = 6442301.28, LR\ stat./df = 0.62301$) so we regard the model as good.

In the given application domain, the condition to use of the LR test is often violated. The examined variable usually has many levels that interpret the web portal (system) or web parts (pages, content categories, activities, etc.). This results in the violation of the LR test condition, i.e. the expected counts are not large enough. From this reason, we used alternative techniques to evaluate the model. The techniques were based on a comparison of empirical and theoretical values on the level of counts (Figure 1b), probabilities and logits (Figure 2). Also, these techniques allowed us to identify the problem parts of the pre-processed log file based on time and web parts.

The evaluation on the level of counts was realised by visualization of the differences of empirical and theoretical counts and identification of extreme values (Figure 1b).

Figure 1b visualizes the differences between empirical and theoretical counts of accesses of stakeholders during the examined period. We can see that greater differences were found at the fourth hour. After the application of the rule "2*standard deviation," we identified five extreme cases. The prediction was overvalued for categories *Pricing List*, *Reputation*, *Business Conditions* and *Pillar3 disclosure requirements* at the fourth hour and on the contrary, understated for the category *We support...*

Another evaluation technique is the comparison of the distribution of the probabilities of empirical relative access count and the estimated probabilities of the selected web part *j* in time *i*. To test the zero hypothesis that distribution of the pair's differences is symmetrical around zero, we can use the Wilcoxon pair test. In the case of web part *We support.*. we rejected the zero hypothesis ($N = 24, T = 24.00, Z = 3.60, p = 0.0003$), the distribution of the pair's differences is symmetrical around zero, i.e. we identified statistically significant difference between the empirical relative access counts and the estimated probabilities for the selected web part *We support.*. in time *i*. Subsequently, we can use another option- visualization of empirical and theoretical logits for each web part expect the reference web part (Figure 2).

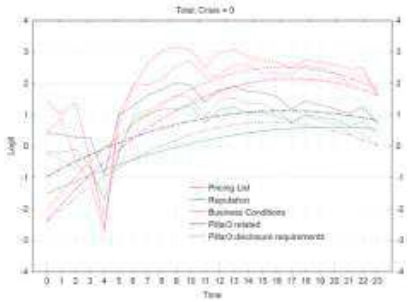



Figure 2 Logit visualization of the examined categories

Figure 2 shows the empirical and theoretical logits for the examined categories. We can see that our estimates of theoretical logits do not fit (do not model) the empirical logits. The biggest variations are at the fourth hour of the day.

Based on the evaluation of the results, we identified a systematic error that occurred during the fourth hour. After we reviewed data, we regularly identified direct accesses to web pages of the examined content (Figure 3). It is a script- automated process (it is related to maintenance- control, backup …).

| | 4 IP | 5 Time | 6 UnixTime | 7 Length | 8 URL | 9 Referrer | 10 UserAgent |
|---|---|---|---|---|---|---|---|
| 2014945 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/rating.html | - | Java/1.5.030 |
| 2014946 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/Misia-Vizia-Hodnoty.html | - | Java/1.5.030 |
| 2014947 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/Ochrana-proti-pranu-spinavych-penazi.h | - | Java/1.5.030 |
| 2014948 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora.html | - | Java/1.5.030 |
| 2014949 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie.html | - | Java/1.5.030 |
| 2014950 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/vzdelanieproj | - | Java/1.5.030 |
| 2014951 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/vzdelanieproj | - | Java/1.5.030 |
| 2014952 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/vzdelanieproj | - | Java/1.5.030 |
| 2014953 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/vzdelanieproj | - | Java/1.5.030 |
| 2014954 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/Projekty-200 | - | Java/1.5.030 |
| 2014955 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporavzdelanie/projekty2009 | - | Java/1.5.030 |
| 2014956 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura.html | - | Java/1.5.030 |
| 2014957 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/kulturaprojekty( | - | Java/1.5.030 |
| 2014958 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/podporakulturaç | - | Java/1.5.030 |
| 2014959 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/kulturaprojekt0( | - | Java/1.5.030 |
| 2014960 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/kulturaprojekt0E | - | Java/1.5.030 |
| 2014961 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/kulturaprojekt07 | - | Java/1.5.030 |
| 2014962 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/Projekty-2008.h | - | Java/1.5.030 |
| 2014963 | 10.238.10.12 | 30.11.2012 4:05:0 | 1354244702 | 0 | /cms/page/sk/about/podpora/podporakultura/projekty2005un | - | Java/1.5.030 |

Figure 3 Selection of identified problem accesses in the fourth hour of the day

After the identification and removing of the problem access to the portal, we identified a new model parameter. We calculated the logit estimates and subsequently estimated the probabilities of accesses to the web parts of the web portal again. Figure 4a shows the new estimates of probabilities of stakeholder's access to selected web parts in the period after the financial crisis.

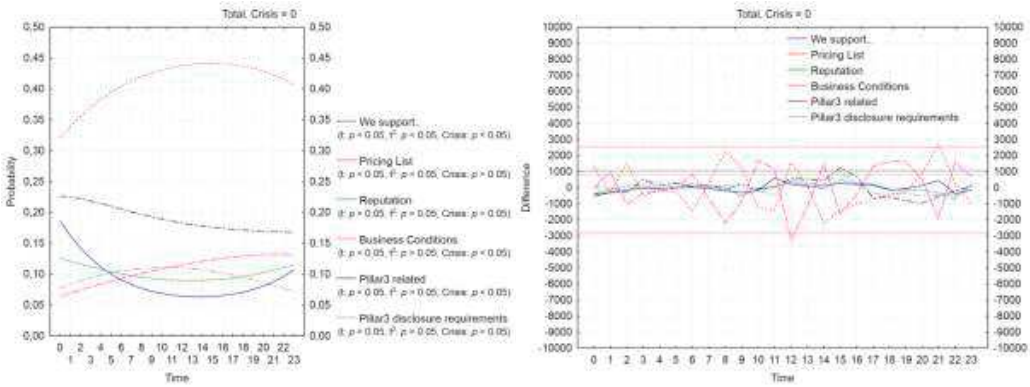



Figure 4 (a) Probability visualization of examined categories during the years after the financial crisis (b) Differences of counts of the model after removing the problematic accesses

Logits for web parts *We support..* and *Pricing List* are dependent from time quadratic, as well as logits of web parts *Business Conditions* and *Pillar3 related* are dependent from time linear, other web parts logits are not dependent from time (Figure 4a) which corresponds with the estimated probabilities (Figure 4a).

Figure 4b visualizes the differences between empirical and theoretical stakeholder's access counts during the examined period. As we can see the biggest differences were only sporadically, especially for the web part *Pricing List* and web parts *Pillar3* was the prediction slightly understated and on the other hand in case of category *Business Conditions* slightly overvalued. The suitability of the model is evidenced by the average of the differences that was approximately equal to zero.

We can see in Figure 5 that theoretical logits for the examined categories fit (model) empirical logits. We can also see that the logits are a quadratic function of time. We did not identify statistically significant differences between the empirical relative access counts and the estimated probabilities of selected web parts in time *i.* For this purpose, we used the Wilcoxon pair test.

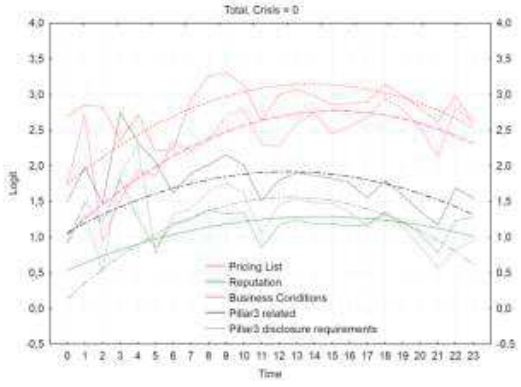



Figure 5 Logit visualization of the examined categories after removing the problematic accesses

## 5. Discussion

We showed the importance of data pre-processing as the initial phase of data analysis and modelling. We focused on deleting records from the log file to have only human accesses to the website. We deleted visits of search engines

robots, other robots and users from anonymizing networks like Tor or JonDonym. Even if we tried to delete all records of non-human access, there are still records we didn't catch using the automated process. Some problem accesses in our case caused by the script-automated process, significantly influenced the obtained knowledge, whether in terms of the type of dependency or in creating a dependency where none has been. After the analysis phase, the created model showed that there is a problem in the source of data. We were able to see differences in counts of the model. We took a closer look at the exact time of access to the website base on the model, and we found non-human access to the website. We had to repeat the pre-processing, repeat the analysis and create a new model. We were surprised that access of automatic script which access website regularly each day changed the model.

It is essential to not underestimate the data pre-processing phase in the process of web usage mining, where the pre-processing phase directly affects the quality of the acquired knowledge.

**Acknowledgements**

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and of Slovak Academy of Sciences under the contract No. VEGA-1/0776/18 and Cultural and Educational Grant Agency of the Ministry of Education of the Slovak Republic under contract No. KEGA-041UK-4/2017.

This publication was supported by the Operational Program: Research and Innovation project "Fake news on the Internet - identification, content analysis, emotions", co-funded by the European Regional Development Fund.

**References**

1. Stevanovic D, An A, Vlajic N. Detecting Web Crawlers from Web Server Access Logs with Data Mining Classifiers. In: Kryszkiewicz M, Rybinski H, Skowron A, Raś ZW, editors. Foundations of Intelligent Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. page 483–9.
2. Tan P-N, Kumar V. Discovery of Web Robot Sessions Based on Their Navigational Patterns [Internet]. In: Intelligent Technologies for Information Analysis. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. page 193–222.Available from: https://doi.org/10.1007/978-3-662-07952-2_9
3. Suchacka G, Sobków M. Detection of Internet robots using a Bayesian approach. In: 2015 IEEE 2nd International Conference on Cybernetics (CYBCONF). 2015. page 365–70.
4. Stevanovic D, An A, Vlajic N. Feature evaluation for web crawler detection with data mining techniques. Expert Syst. Appl. [Internet] 2012;39:8707–17. Available from: http://www.sciencedirect.com/science/article/pii/S0957417412002382
5. Munk M, Kapusta J, Švec P. Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor. Procedia Comput. Sci. [Internet] 2010 [cited 2015 Oct 5];1:2273–80. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-78649584518&partnerID=tZOtx3y1
6. Kapusta J, Svec P, Munk M, Skalka J. User Session Identification Using Extended Href Method. DIVAI 2014 10TH Int. Sci. Conf. DISTANCE Learn. Appl. INFORMATICS [Internet] [cited 2016 May 13];581–8. Available from: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=1&SID=P2LHOY1PuFok2vuTFiF&page=1&doc=3
7. He D, Harper DJ. Detecting session boundaries from Web user logs. In: Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research. 2000. page 57–66.
8. Kapusta J, Munk M, Drlík M. Cut-off time calculation for user session identification by reference length [Internet]. In: 2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings. Department of Informatics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Nitra, Slovakia: 2012. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84872843897&partnerID=40&md5=be72201dd81f886ecbe140efab4d6412
9. Reed MG, Syverson PF, Goldschlag DM. Anonymous connections and onion routing. IEEE J. Sel. Areas Commun. 1998;16:482–94.

10. Dingledine R, Mathewson N, Syverson P. Tor: The Second-generation Onion Router [Internet]. In: Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13. Berkeley, CA, USA: USENIX Association; 2004. page 21.Available from: http://dl.acm.org/citation.cfm?id=1251375.1251396
11. Berthold O, Federrath H, Köpsell S. Web MIXes: A System for Anonymous and Unobservable Internet Access. In: Workshop on Design Issues in Anonymity and Unobservability. 2000.
12. Chaum DL. Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms. Commun. ACM [Internet] 1981;24:84–90. Available from: http://doi.acm.org/10.1145/358549.358563
13. Herrmann D, Wendolsky R, Federrath H. Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. In: CCSW. 2009.
14. Astolfi F, Kroese J, van Oorschot J. I2P-The Invisible Internet Project.
15. Thomas P. Explaining Difficulty Navigating a Website Using Page View Data [Internet]. In: Proceedings of the Seventeenth Australasian Document Computing Symposium. New York, NY, USA: ACM; 2012. page 31–8.Available from: http://doi.acm.org/10.1145/2407085.2407090
16. Munk M, Pilkova A, Kapusta J, Svec P, Drlik M. Pillar 3 and Modelling of Stakeholders' Behaviour at the Commercial Bank Website during the Recent Financial Crisis. Procedia Comput. Sci. [Internet] 2013 [cited 2015 Oct 8];18:1747–56. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84896923179&partnerID=tZOtx3y1
17. Drlik M, Pilkova A, Munk M, Švec P. Modelling of domestic and foreign visitors' behaviour at commercial bank website during the recent financial crisis. Acta Univ. Agric. Silvic. Mendelianae Brun. [Internet] 2013;61:2065–70. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84893250276&doi=10.11118%2Factaun201361072065&partnerID=40&md5=1dbaaa91b541687588edbe97506c3da2
18. Munk M, Drlik M, Vrabelova M. Probability Modelling of Accesses to the Course Activities in the Web-Based Educational System. In: Computational Science And Its Applications - Iccsa 2011, Pt V. 2011. page 485–99.
19. Munk M, Drlík M. Analysis of stakeholders' behaviour depending on time in virtual learning environment. Appl. Math. Inf. Sci. [Internet] 2014;8:773–85. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84893155565&partnerID=40&md5=f48e614bb8250bdc3b2a1cb82ad5b301
20. Svec P, Chylo L, Filipik J. Web Log Data Preprocessing using Raspberry Pi Cluster and hadoop cluster. In: Turcani, M and Balogh, Z and Munk, M and Kapusta, J and Benko, L, editor. DIVAI 2018: 12TH INTERNATIONAL SCIENTIFIC CONFERENCE ON DISTANCE LEARNING IN APPLIED INFORMATICS. U NAKLADOVEHO NADRAZI 6, PRAHA 3, PRAGUE 130 00, CZECH REPUBLIC: WOLTERS KLUWER CR A S; 2018. page 513–21.
21. Stencl M, St'astny J. Artificial Neural Networks Numerical Forecasting of Economic Time Series. In: Hui, CL, editor. ARTIFICIAL NEURAL NETWORKS - APPLICATION. JANEZA TRDINE9, RIJEKA, 51000, CROATIA: INTECH EUROPE; 2011. page 13–28.
22. Sun Y, Zhuang Z, Councill IG, Giles CL. Determining Bias to Search Engines from Robots.Txt [Internet]. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington, DC, USA: IEEE Computer Society; 2007. page 149–55.Available from: http://dx.doi.org/10.1109/WI.2007.45
23. AlNoamany YA, Weigle MC, Nelson ML. Access Patterns for Robots and Humans in Web Archives [Internet]. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries. New York, NY, USA: ACM; 2013. page 339–48.Available from: http://doi.acm.org/10.1145/2467696.2467722
24. Kapusta J, Munk M, Svec P, Pilkova A. Determining the Time Window Threshold to Identify User Sessions of Stakeholders of a Commercial Bank Portal. Procedia Comput. Sci. [Internet] 2014 [cited 2015 Oct 29];29:1779–90. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84902797377&partnerID=tZOtx3y1
25. Cooley R, Mobasher B, Srivastava J, others. Data preparation for mining world wide web browsing patterns. Knowl. Inf. Syst. 1999;1:5–32.
26. Cooley R, Mobasher B, Srivastava J. Grouping Web page references into transactions for mining WorldWide Web browsing patterns. Proc. 1997 IEEE Knowl. Data Eng. Exch. Work. 1997;
27. Kapusta J, Munk M, Drlik M. User Session Identification Using Reference Length. In: DIVAI 2012: 9TH INTERNATIONAL SCIENTIFIC CONFERENCE ON DISTANCE LEARNING IN APPLIED INFORMATICS: CONFERENCE PROCEEDINGS. 2012. page 175–84.
28. Winter P, Lindskog S. How the Great Firewall of China is Blocking Tor. Free Open Commun. Internet 2012;
29. Dingledine R, Mathewson N. Design of a blocking-resistant anonymity system. 2006.
30. Gilad Y, Herzberg A. Spying in the dark: TCP and Tor traffic analysis. In: Lecture Notes in Computer Science

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012.
31. Chakravarty S, Portokalidis G, Polychronakis M, Keromytis AD. Detecting traffic snooping in tor using decoys. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2011.
32. McCoy D, Bauer K, Grunwald D, Kohno T, Sicker D. Shining light in dark places: Understanding the tor network. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2008.
33. Pilkova A, Munk M, Svec P, Medo M. Assessment of the Pillar 3 Financial and Risk Information Disclosures Usefulness to the Commercial Banks Users. In: Advanced Intelligent Computing Theories and Applications. Springer International Publishing; 2015. page 429–40.

PRÍLOHA H: MUNKOVA, DASA, MICHAL MUNK, ĽUBOMÍR BENKO A JIRI STASTNY, 2021B. MT EVALUATION IN THE CONTEXT OF LANGUAGE COMPLEXITY. *COMPLEXITY*. 2021, 1–15. DOI:10.1155/2021/2806108 (**WEB OF SCIENCE, 2021IF: 2.121, Q2**) [WOS: 2, SCOPUS: 0]

Hindawi
Complexity
Volume 2021, Article ID 2806108, 15 pages
https://doi.org/10.1155/2021/2806108

WILEY | Hindawi

*Research Article*

# MT Evaluation in the Context of Language Complexity

**Dasa Munkova**,[1] **Michal Munk**,[1] **Ľubomír Benko**,[1] **and Jiri Stastny**[2,3]

[1]*Department of Computer Science, Constantine the Philosopher University, Nitra, SK-949 01, Slovakia*
[2]*Institute of Automation and Computer Science, Brno University of Technology, Brno, CZ-619 69, Czech Republic*
[3]*Department of Informatics, Mendel University in Brno, Brno, CZ-613 00, Czech Republic*

Correspondence should be addressed to Ľubomír Benko; lbenko@ukf.sk

Received 18 May 2021; Revised 3 November 2021; Accepted 1 December 2021; Published 17 December 2021

Academic Editor: Wen-Long Shang

Copyright © 2021 Dasa Munkova et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The paper focuses on investigating the impact of artificial agent (machine translator) on human agent (posteditor) using a proposed methodology, which is based on language complexity measures, POS tags, frequent tagsets, association rules, and their summarization. We examine this impact from the point of view of language complexity in terms of word and sentence structure. By the proposed methodology, we analyzed 24 733 tags of English to Slovak translations of technical texts, corresponding to the output of two MT systems (Google Translate and the European Commission's MT tool). We used both manual (adequacy and fluency) and semiautomatic (HTER metric) MT evaluation measures as the criteria for validity. We show that the proposed methodology is valid based on the evaluation of frequent tagsets and rules of MT outputs produced by Google Translate or of the European Commission's MT tool, and both postedited MT (PEMT) outputs using baseline methods. Our results have also shown that PEMT output produced by Google Translate is characterized by more frequent tagsets such as verbs in the infinitive with modal verbs compared to its MT output, which is characterized by masculine, inanimate nouns in locative of singular. In the MT output, produced by the European Commission's MT tool, the most frequent tagset was verbs in the infinitive compared to its postedited MT output, where verbs in imperative and the second person of plural occurred. These findings are also obtained from the use of the proposed methodology for MT evaluation. The contribution of the proposed methodology is an identification of systematic not random errors. Additionally, the study can also serve as information for optimizing the translation process using postediting.

## 1. Introduction

Tasks play a crucial role in human behaviour or performance. Liu and Li ([1], p. 553) stated that human behaviour and/or performance depend on the interaction among task characteristics (such as complexity, which has a significant influence on behaviour and predicting human performance), task performer characteristics (such as performer's competencies), and environment characteristics. When the task is translation, especially machine translation (MT) or postediting of MT output (PE), one of the essential task characteristics is a complexity of MT output or postedited MT output (PEMT). Complexity is an intrinsic property (qualitative) of a translation task, which is given by internal textual structure and represents the objective characteristic of a task ([2], p. 2). Liu and Li ([1], p. 559) define task complexity as the aggregation of any intrinsic task characteristic that influences the performance of a task (translation). Complexity ([3], p. 40) can be understood as (a) primarily a psychological experience (purely subjective psychological approach), (b) an interaction between task and person characteristics (tasks are more or less complex relative to the capabilities of the individuals who perform the task), and (c) a function of objective task characteristics (in terms of objective task qualities).

Complex tasks are characterized [3, 4] (a) by unknown or uncertain alternatives of action (there is not only one alternative of translation; multiple outcomes), (b) by inexact or unknown means-ends connections (there are many ways to express the same meaning in translation; multiple potential ways), and (c) by the existence of a number of subtasks which may or may not be easily factored into nearly

independent parts (analysis of source text and analysis of target text in terms of intratextual and extratextual factors; conflicting interdependence among ways to outcomes, and also uncertain or probabilistic links among ways and outcomes). These objective task qualities contribute to complexity which placed high demands on the translator and also on the posteditor. Linguistic complexity contributes not only to the difficulty in solving word problems but also to the difficulty in translation tasks ([5], p. 1). There are many factors hidden by word complexity (e.g., patterns, grammatical constructions, lexicon, or physical circumstances in the given language); therefore, there is no general definition of linguistic complexity. Ma and Wang ([6], p. 3) state qualitative characteristics of language complexity such as uncertainty, incompleteness, sensitivity to initial conditions, dynamicity, nonlinearity, instability, path dependency, openness, and adaptivity, while the core of language complexity is nonlinearity. Nonlinearity consists of imbalance, emergence, and interactivity features ([6], p. 5), which means that expression of language patterns or their combinations deviates from linearity ([7], p. 53–54). Besides the qualitative characteristics of language complexity, there are also quantitative characteristics such as high organizational depth features, since quality and quantity are a pair of interdependent contradictions ([6], p. 13). In terms of language systems, high organizational depth refers to multilevel, which is an essential way to organize complex systems [8, 9]. Levels of language are used to understand the language complexity in oral or written form [10]. Ma and Wang ([6], p. 26) state that the higher the language complexity, the longer the minimum description length of the information, and the greater the resource/cost consumption.

Only the use of computer programming technology, e.g., oral or written corpora, to analyze and study language complexity will help improve the ability of language processing [11]. Complexity of text can be measured at word- or sentence-level using corpus-based methods—readability or lexico-grammatical features: word/sentence length and frequency of part of speech [12].

Complexity of text features always implies independent variables, i.e., textual elements (word-level, sentence-level, and discourse-level variables) that can be examined and analyzed ([13], p. 236). Based on the complexity of textual elements (words, syntax, and discourse structure), we can examine performance, such as translation or PE tasks.

*1.1. Evaluation of MT System.* The invention of neural machine translation (NMT) brought several fundamental changes to the translation industry, both from the point of view of the translation process or task [14] and of the business model [15]. Current NMT systems provide fluent translations of fairly good quality [16], but often, this fluency is at the expense of accuracy or intelligibility [17]. NMT and its predecessor, statistical MT (SMT), were commonly used not only for personal use but also to reduce the cost of translation in the translation industry for many years. NMT and SMT operate on a statistical basis using a corpus-based approach to MT. NMT took a huge leap forward when sequence-to-sequence models were first introduced [18]. So far, it has already achieved excellent performance on a great volume of translations from English into French [12, 19] as well as from English into German [20]. MT systems for translating dialectal sentences into their standard language form can be of great benefit as Farhan et al. [18] have shown. Translation technologies have become an integral part of the translator's work, so it is very important to know what machines can and, conversely, what they cannot adequately translate. They also serve to alert users to errors that occur in MT [21].

MT system is a complex natural language processing system composed of a large number of heterogeneous modules [22]. MT systems, like language, can be considered a complex adaptive system, which involves multiple agents (both natural and artificial) interacting with one another to achieve a common goal—a translation task ([23], p. 261). Since a natural language is not static but dynamic, i.e., previous behaviour influences the current and future behaviour of natural language, it can be considered as a complex adaptive system [23]. Complex systems have shifted from reductionist analyses of component parts (agents) and simple linear change to the study of interconnected elements ([24], p. 2).

The behaviour of agents in the language system is influenced by several elements at different levels, whether internal or external (locutionary, illocutionary, and perlocutionary acts). If we add a translation task as an act of communication, it brings a new level of complexity—a metalevel. The complexity of the system together with translation errors increases, since at least two languages are considered in the translation process—both source and target languages. Systems may differ from each other not because of differences in their features, but because of differences in how these features depend on and affect one another ([25], p. 2). Siegenfeld and Bar-Yam [25] liken it to steam and ice, both of which are made of the same water molecules but have different properties due to differences in the interactions between molecules, like MT output and its postedited MT output (PEMT). Both translations come from the same original, but due to differences in interactions, they have different properties, i.e., different translation qualities.

In MT systems, at the locutionary level, a language model of the source text is created using a neural network, i.e., the neural network is trained on a large amount of source text data. It creates patterns and identifies grammatical constructions and a lexicon to be able to assign corresponding patterns to them in the target language, and the same is done in the target language, in which another neural network is also trained. The key concept behind MT is to capture the linguistic knowledge of the locutionary level of the languages involved by means of translation pairs linking constructions across languages ([23], p. 269). The illocutionary level lies in the transfer itself, from source to target language. At the perlocutionary level, it is the quality of MT output from the MT system that plays a key role in the given communication.

Progress in MT depends on the results of the evaluation of MT quality. NMT, as a metric to evaluate the development of artificial intelligence, plays a crucial role in the current

natural language processing (NLP) community [26]. Many experts [27, 28] have sought various ways to assess MT quality, whether in the form of manual evaluation, automatic evaluation or both, in the form of a framework (e.g., dynamic quality framework or multidimensional quality metrics).

MT output can be evaluated manually or automatically with intrinsic and extrinsic methods applied [29]. Castilho et al. [30] distinguish manual methods according to six criteria: (1) adequacy and fluency; (2) readability and comprehensibility; (3) acceptability; (4) ranking; (5) usability and performance; and (6) evaluators. Adequacy and fluency are the most commonly used measures in translation assessment [31]. Alongside the standard criteria of fluency and adequacy, some researchers have focused on examining the linguistic features of a text, specifically on the identification of differences among original text and different translation outputs, i.e., human translation (HT), MT output, or PEMT output [32–34]. Methods are typically based on linguistic (e.g., word frequency) and extralinguistic features (formatting) [30]. Vanmassenhove et al. [34] have shown (for translation directions—English into French or Spanish) that MT texts contain lesser lexical variety compared to their source English texts or compared to their human translations in French or Spanish. Loock [12] has shown that the linguistic characteristics of MT texts from English into French differ from the original French texts.

Intrinsic methods involve comparisons of translation quality between MT output and reference (high-quality HT) or a fixed set of references. Manual intrinsic measures determine MT quality through human subjective judgments such as fluency and adequacy. The biggest issue that manual intrinsic methods face is their subjectivity and nonreproducibility, apart from their price and timeliness. Automatic intrinsic measures, such as ranking, compute sentence similarity among MT outputs and a fixed set of references to produce rankings among MT systems [35]. Unlike intrinsic measures which are focused on accuracy and text coherence, extrinsic methods focus on the effectiveness or usability of MT output in terms of the specific task such as PE [36–38]. PE as a specific task directly assesses the MT output in terms of the time and effort needed to correct the MT. It provides information about difficulty, but it does not provide sufficient information about task characteristics such as linguistic complexity. PE is a result of the linguistic complexity of MT output, which is related to posteditor-task interaction. A closer measure than time, which is related to linguistic complexity, is edit distance (error rate). It represents the number of changes within the sentence including insertion, deletion, substitution, or shifts, which required some correction of the MT output.

Availability of reference translations allows us to use not only manual evaluation methods but also measures of automatic evaluation to evaluate the translation quality [28]. Automatic MT evaluation measures provide quick feedback on translation quality, but this feedback is only a score. According to the criterion of lexical concordance, we divided them into automatic metrics of accuracy and metrics of error rate [39]. Metrics of accuracy are based on the closeness of the MT output/hypothesis ($h$) with the reference ($r$) in terms of $n$-grams. They calculate their lexical overlap in ($A$) the number of common words ($h \cap r$), ($B$) the length (number of words) of MT output, and ($C$) the length (number of words) of the reference. The higher the values of these metrics, the higher the translation quality [40]. Metrics of error rate are based on edit distance. They calculate the Levenshtein distance between an MT output/hypothesis ($h$) and a reference/human translation ($r$). The higher the values of these metrics, the lower the translation quality [39].

Automatic measures are a good objective indicator of how to improve system performance and are cheap and achieve more consistent results compared to manual. However, their main drawback is that they are not able to sufficiently assess the syntactic and semantic equivalence of translation (linguistic complexity). We are not able to perform a deeper linguistic analysis. Besides the overall scores, it is helpful to have additional information, i.e., the strengths and weaknesses of the system or types of MT errors [28]. Another problem with automatic measures is that its metrics operate mainly at the sentence/segment level and not at the document level, and they do not take context into account when assessing translation quality [30].

*1.2. Error Analysis in the Context of NLP and MT Evaluation.* According to Popović [28], error analysis and classification provide the basis for determining what type of errors are produced by the system and whether and how they can be eliminated. It can be carried out not only by classification and annotation of erroneous words but also by analyzing words or parts of speech (POS). In the translation industry, the evaluation usually relies on error analysis [30, 41]. Error analysis offers a number of answers to improve the system, better understanding of human or artificial agent behaviour or performance such as translation or PE tasks. However, it is time-consuming and requires extensive knowledge of annotator(s). Feng et al. [26] showed that the performance of an NMT system benefits from POS tag information of target language (Chinese-English and German-English translation datasets). POS tag is more informative and concise than combinatory categorial grammar (CCG) supertag [42]. Loock [12] showed how a linguistic analysis of a corpus of MT texts can also be used in translator education. Hládek et al. [43] aimed at the present alternative view of the task of morphological tagging and focused on Slovak. They proposed a rule-based system using expert knowledge. The system generates an outcome based on the rule that a certain tag was chosen from the match set. They summarized the whole decision process into three phases (matching, maximization, and minimization). The rules were created using the learning process and then were pruned for more specific rules offering better accuracy. They compared their proposed algorithm with the morphological tagger HunPos [44]. Laki et al. [45] presented a novel universal morphological feature schema as a set of features expressed by inflectional morphology across languages. They examined the variability of inflectional morphology by comparing multiple translations of the same source (the Bible). The results

showed that the schema offers potential benefits for NLP and MT by facilitating direct meaning-to-meaning translation between the language pairs, regardless of form-related differences.

It motivated us to apply POS tagging to determine the error rate and linguistic complexity. Just as word and sentence, POS tagging is implemented in text analysis, and it can also be used to compare two texts (MT output and PEMT output) or to determine the linguistic complexity through the quality of MT output, PEMT output, or HT.

*1.3. Research Objectives.* The study of multiagent behaviour motivated us in our research, in which we focus on the influence of the agent-machine translator on the behaviour of the human agent-posteditor within one complex adaptive system. In other words, we identify the behaviour of the agent-machine translator using POS tagging and association rules found and then identify its influence on the behaviour of the agent-human posteditor, whose task is to achieve the perlocutionary level of natural language, i.e., to postedit MT output to be both fluent and adequate. We focus on examining the behaviour and/or performance of an artificial agent and a human agent, which depends on the interaction between tasks characteristics and task performer characteristics from the point of view of language complexity. We investigate the translation task through language complexity, which is defined by frequent tagsets and rules.

The aim of the study is to present a new approach to the evaluation of MT quality and subsequently to validate the proposed MT evaluation methodology. The proposed methodology is based on the evaluation of frequent MT and PEMT tagsets and also on frequent POS tagsets and rules summarization. The aim consists of two consecutive objectives.

The first objective comprises three tasks. The first is to analyze MT outputs from two MT systems: Google Translate and the European Commission's MT tool as well as their postedited MT outputs based on POS tags in terms of task characteristics and language complexity at word- and sentence-level. The second is to examine the relationships between individual tags and tagsets within the four examined translations (as described in Section 3.1). The last task focuses on the comparison of translation quality based on the summarization of the incidence of frequent tagsets and rules (as described in Section 3.2).

We examine the extent to which the MT quality in terms of language complexity is identical to its PEMT version based on the frequency of tagsets and rules.

For this study, we have set as null hypotheses:

H01: the incidence of frequent tagsets does not depend on the method of translation (machine translation vs. postediting)

H02: the incidence of extracted rules does not depend on the method of translation (machine translation vs. postediting)

The second research objective is to validate the proposed methodology of MT evaluation using baseline methods. We used both manual and semiautomatic MT evaluation measures as criteria for validity (as described in Section 3.3).

*1.4. Implications and Limitations.* The study offers new insight into the evaluation of MT quality. The results and findings of the research offer one key theoretical contribution and two practical contributions to the field of complex adaptive systems, including MT evaluation.

The theoretical contribution consists of the design and verification of a novel methodology for evaluating MT quality in the context of inflectional languages. The proposed and verified methodology is unique, combining the advantages of using both intrinsic and extrinsic methods, focusing on translation into the inflectional language, which is characterized by a rich morphology and free word order. It analyzes and subsequently compares the translation quality, based on text complexity, i.e., based on the frequent tags and rules and their quantitative evaluation—summarizing the frequent tags and rules incidence. The proposed methodology allows us to identify the complexity of MT outputs, especially errors that are systematic and not random. The principle of the proposed methodology is applicable to any language pair as well as translation directions, but it is necessary to take into account the character of the target language when determining tags and/or part-of-speech tagging. For instance, declension is typical for inflectional languages such as Slovak but not for analytical languages like English, i.e., Slovak uses suffixes for grammatical cases, in contrast to English, in which cases are expressed by prepositions.

Nominative: *auto* (SK)—*a car* (EN)

Genitive: *auta* (SK)—*from a car* (EN)

Dative: *autu* (SK)—*to a car* (EN)

Accusative: *auto* (SK)—*a car* (EN)

Locative: *aute* (SK)—*about a car* (EN)

Instrumental: *autom* (SK)—*with a car* (EN)

We were inspired by the research of Conforti et al. [46] focusing on machine translation to morphologically rich language using POS tagging. We adopted a similar approach to assessing the quality of MT output but using text complexity from the perspective of word and sentence structure. Callison-Burch et al. [47] or Popović [28, 48, 49] have shown that metrics based on POS analysis correlate very well with human evaluation. Popović [50] provided a useful approach for quality estimation based on morphemes and POS tags. The proposed methodology can also be used to evaluate students' translation performance within their translation education or in language learning.

From a practical point of view, the findings offer a closer understanding of the text complexity of MT outputs, i.e., they allow us to reveal the linguistic features of MT texts from an analytical into a synthetic language. The second practical contribution, which follows on from the first, consists in the identification of "machine translationese" [12], what kind of translation task the machine can and cannot do correctly for the given direction of translation and the genre of the text.

The research also has certain limitations in the aspect that (a) the examined texts are not extensive and come from one genre (technical documentation), as well as the posteditor himself/herself, who has subjective sensitivity to errors within the text. However, in the evaluation, specifically, when assessing the adequacy and fluency of MT and subsequently when postediting the MT outputs, a large amount of manual work is required. In our case, it was done by students and translators during one day. Human evaluation is time- and labour-consuming, but it is considered highly reliable. For this reason, it is sometimes better to have a smaller dataset, but with more reliable data. We are working on expanding the dataset, but we are faced with the problem of evaluators' consistencies, as not all participants wanted to continue in the research (to repeat the same procedure with different genres and translation direction or different source language). (b) We focused only on the influence of an artificial agent's behaviour (MT system) on a human agent's behaviour (PE) using a word and sentence complexity. We did not consider the influence of a human agent's behaviour (preediting) on an artificial agent's behaviour (MT system) and subsequently its influence on the human agent' behaviour (PE). For this reason, we want to focus our future work on text volume as well as the diversity of genres, consistency of posteditors, and also on preediting.

The structure of the paper is as follows. Section 2 describes the research methodology, and the subsequent section focuses on the research results based on the association rules analysis and aims at the validation of the proposed methodology for MT evaluation. The penultimate section offers a discussion of the results. The last section comprises research conclusions.

## 2. Materials and Method

We examined unstructured textual data, namely technical texts—consisting of 606 sentences (more than 6 000 tags). The source texts (ST) written in English were translated into Slovak by two MT systems/engines—Google Translate (GT) and MT@EC.

For our research, the most important step was tokenization, which was done after sentence alignment, since we analyzed two MT engines (MT systems). We also used the TreeTagger tool for tokenization, developed by Schmid [51–53]. It supports morphological annotation of the Slovak language and automatically annotates Slovak texts with POS tagging and lemma information [54].

*2.1. Proposed Approach.* The applied methodology includes the following stages (in Figure 1):

(1) Acquisition of unstructured textual data: source texts (technical texts)

(2) Data preparation: it consists of multiple tasks:

  (a) Machine translation: translation of the source texts using both MT engines

  (b) Sentence alignment: the generated MT output is aligned with the source text based on the 1-to-1 principle

  (c) Postediting: the MT output is postedited by professional translators and students in M.A. degree

  (d) Evaluation: each MT sentence of both MT outputs is assessed by participants using the scale of fluency and adequacy (scale range is from 1 to 5)

  (e) POS tagging: the MT output and PEMT output are tokenized separately, which generates the tags and lemmas for annotated aligned words (see Supplementary Table 1 for more details on Slovak POS tags)

(3) Data analysis consists of searching the frequent POS tags (tagsets) of MT output (MT@EC_MT or GT_MT) and PEMT output in the examined text. The results were processed by association rule analysis using STATISTICA Sequence, Association, & Link Analysis, which is an implementation of the algorithm using apriori algorithm together with a tree-structured procedure that requires only one pass through data. The support for a tagset is given by a proportion of records in the transactions data set that have the tagset, i.e., for a tagset ($A$), the support can be calculated as follows:

$$\text{support}(A) = \frac{\text{frequency of } (A)}{\text{number of transactions in the dataset}} * 100. \tag{1}$$

Lift of rules can be similarly calculated. Based on support and confidence, a lift for a rule can be defined and computed ($A$–tagset, $C$–tagset)

$$\text{lift}(\textit{if } A \text{ then } C) = \frac{\text{confidence}(\text{if } A \text{ then } C)}{\text{support}(C)}, \tag{2}$$

where

$$\text{confidence}(\text{if } A \text{ then } C) = \frac{\text{support}(\text{if } A \text{ then } C)}{\text{support}(A)} * 100. \tag{3}$$

We focused on frequent tagsets extracted with the minimum support of 10%.

(4) Data understanding based on the results of association rule analysis.

(5) Comparison of found rules and frequent tagsets in examined translations.

We will validate the proposed methodology of MT evaluation, which is based on the evaluation of frequent MT and PEMT tagsets, by manual and semiautomatic MT evaluation.

*2.2. Manual and Semiautomatic MT Measures.* Adequacy, manual MT measure, represents the extent to which the translation transfers the meaning of the source text into the target language. Fluency, manual MT measure, represents

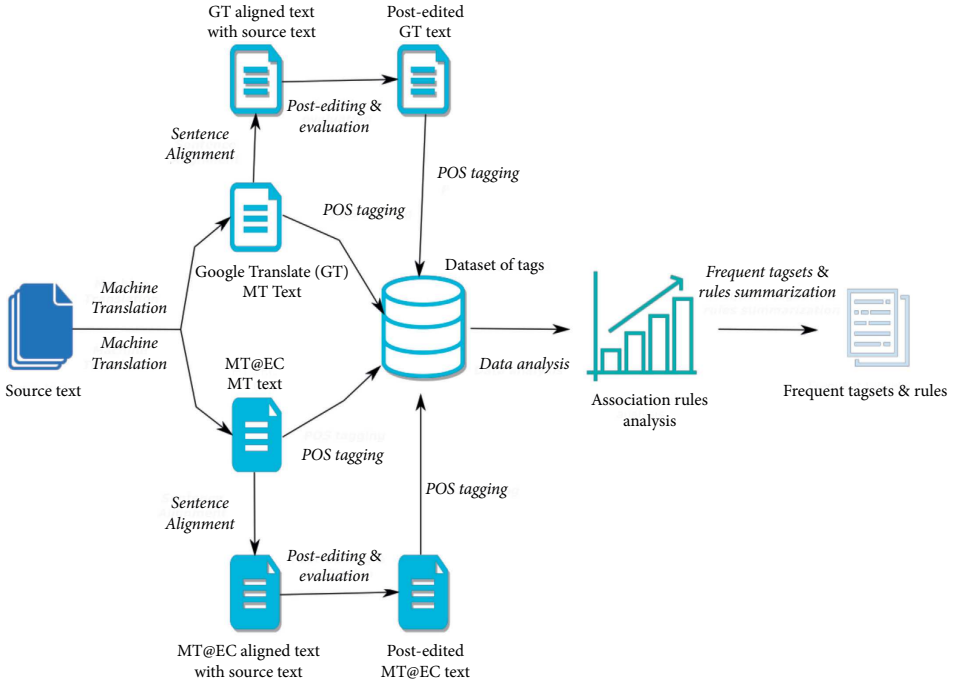



Figure 1: The proposed methodology.

the extent to which the translation follows the norms of the target language. Both measures assess the translation using a five-point Likert scale for each segment, where 1 means "none (adequacy)/incomprehensible (fluency)," 2 means "little meaning/disfluent Slovak," 3 means "much meaning/nonnative Slovak," 4 means "most meaning/good Slovak," and 5 is "all meaning (adequacy)/flawless Slovak (fluency)."

HTER (human-targeted translation error rate) [55] is a more complex semiautomatic MT measure; humans do not score translations directly but rather generate a new reference translation (PEMT) that is closer to the MT output but retains the fluency and meaning of the original reference ([56], p. 259). Specifically, HTER = # of edits (substitutions + insertions + deletions + shifts)/# of reference words.

*2.3. Dataset.* PE task and evaluation (fluency and adequacy) were conducted in the OSTPERE system [57]. MT outputs were postedited by professional translators and students of translation studies at the master's degree level (extrinsic method). The human translators also assessed each MT sentence using the scale of adequacy and fluency (intrinsic method). Due to the time and laborious complexity of the manual evaluation of translation quality, our dataset is not extensive but specialized on one text type. The data provide the possibility to perform more precise analyzes (e.g., linguistic analysis) for one specific domain. Data were obtained during a one-day workshop in order to keep the consistency of posteditors and evaluators (the average translator translates a maximum of 10 standard pages per day).

The final dataset consists of 24 733 tags: MT outputs (translated by GT and MT@EC) and their corresponding PEMT outputs. Subsequently, we created a program, using C#, to calculate the HTER scores for each MT sentence. Based on the source sentence ID, the corresponding HTER score and scores of adequacy and fluency were merged into a single data matrix. The data matrix was used to create a baseline for analysis.

Each sentence was annotated by the TreeTagger tool. Tokenization produced four files (GT_MT, GT_PEMT, MT@EC_MT, and MT@EC_PEMT) containing annotated tags. The composition of each file (GT_MT, GT_PEMT, MT@EC_MT, and MT@EC_PEMT) is depicted based on two feature types (Table 1) where each file consisted of approximately 6000 tags (including interpunction). It was necessary to adjust the files before merging them into a single data matrix because we wanted to keep a record of each sentence (ID), and the tool works only with text files. For this task, a simple JAVA program was created. This allows us to create a single data matrix incorporating all four files with the corresponding tags and also to compare both translations. A transaction/sequence model [58] was used for text representation. The results were processed by the association rule analysis. Found rules and frequent tagsets were summarized by a Cochran $Q$ test and using multiple comparisons.

Table 1: Dataset composition.

| Feature type | Feature name | GT_MT | GT_PEMT | MT@EC_MT | MT@EC_PEMT |
|---|---|---|---|---|---|
| Readability | Average sentence length (words) | 8.45 | 8.81 | 7.82 | 8.75 |
| | Average word length (characters) | 5.51 | 5.89 | 5.70 | 5.89 |
| | Number of short sentences ($n < 10$) | 63.37% | 56.44% | 67.49% | 58.35% |
| | Number of long sentences ($n \geq 10$) | 36.63% | 43.56% | 32.51% | 41.65% |
| Lexico-grammatical | Frequency of nouns | 32.81% | 36.82% | 31.25% | 36.23% |
| | Frequency of adjectives | 8.22% | 9.00% | 11.81% | 9.05% |
| | Frequency of adverbs | 3.16% | 2.70% | 3.11% | 2.73% |
| | Frequency of verbs | 16.57% | 16.11% | 15.00% | 15.87% |
| | Frequency of pronominals | 3.13% | 3.03% | 3.02% | 3.27% |
| | Frequency of participles | 1.63% | 1.89% | 1.62% | 1.97% |
| | Frequency of morphemes | 1.45% | 1.32% | 1.34% | 1.36% |
| | Frequency of abbreviation | 3.01% | 2.40% | 3.94% | 2.44% |
| | Frequency of numbers | 3.87% | 3.65% | 4.32% | 3.53% |
| | Frequency of undefinable POSs | 0.29% | 0.23% | 1.06% | 0.34% |
| | Frequency of foreign words | 6.98% | 4.84% | 6.02% | 5.14% |
| | Frequency of interjections | 0.02% | 0.02% | 0.02% | 0.02% |
| | Frequency of numerals | 0.75% | 0.49% | 0.70% | 0.42% |
| | Frequency of prepositions & conjunctions | 18.10% | 17.49% | 16.80% | 17.63% |

## 3. Results

The section Results is divided into two subsections: the first describes the identified relations between tagsets, and the second represents their quantitative summarization.

*3.1. Identification of Relations between Tagsets.* The association rule analysis represents a nonsequential approach to the data being analyzed. We will not analyze the sequences but transactions, so we will not include the tag order in the analysis. In our case, a transaction represents a set of tags observed in the MT sentence.

The web graphs (in Figures 2 and 3) depict the discovered association rules for the sentences, namely, the size of a node represents a support of the tag, the thickness of the line represents the support of rule—a pair of tags, and darkness of the line colour represents a lift of the rule.

In the GT_MT output (in Figure 2(a), see also Supplementary Table 2(a) for detailed analysis), the tag (*O*), conjunctions, belongs to the tags with the highest incidence within the text with almost 50% of the support and the tag (%), foreign language citation, with a probability of more than 35%. Other very frequent tags, after conjunctions and foreign language citation (untranslated or domesticated), with less probability of incidence (around 20%) were (*VMdpb+*), i.e., verb in imperative, perfective aspect, second person of plural in the affirmative (stlačte/press, pripojte/connect, vyberte/select, použite/use), and (*Eu4*), i.e., nonvocalized preposition in accusative (na/on, to/k, pre/for), which were tied with substantive in accusative whether in the masculine, inanimate gender (*SSns4*), or in the neuter (SSis4). Furthermore, the verbs in infinitive (*VId+*) were observed (spojiť/connect). The other identified tags (not depicted in Figure 2(a), see Supplementary Table 2(a)) do not meet the minimum support, i.e., the likelihood of occurrence in the identified sentences (transactions) is less than 10% (see Supplementary Table 2). Among the most found pairs (in Figure 2(a), see also Supplementary Table 2(a)), a pair of the tags in the sentence belong (*O*, *VMdpb+*), (*O*, *VKepb+*), and (%, *O*) with more than 17% of the *support*, i.e., conjunctions with verbs in imperative or present, in plural, and in affirmative. Subsequently, conjunctions with foreign language citations (použite kábel HDMI alebo ultra HD) use the cord HDMI or ultra HD.

Another large group of pairs, with the probability of around 15%, were (*SSis4*, *O*)—a noun in the inanimate masculine gender or in the neuter, singular, in accusative with conjunctions and also pair (*O*, *VId+*)—conjunctions with verbs in infinitive or infinitives with verbs in present, in the second person of plural (môžete poškodiť modul CAM a televízor/you can damage module CAM and TV, zvoľte a stlačte tlačidlo/select and press the button). Tags not presented in the analysis do not meet the minimum support and confidence of 10%; that is, these tags are identified in sentences with a probability of less than 10% (Figure 2).

The greatest degree of positive correlation (*lift* = 5.11) was identified by the (*SSis6, Eu6*) pair (in Figure 2(a), see also Supplementary Table 2(a)). Lift, in case (SSis6, Eu6), indicates a certain rule, i.e., substantives in the inanimate masculine gender in singular, locative case are tied with nonvocalized prepositions in locative (v prípade/in case, na televízore/on TV), less in case (SSis4, Eu4), where substantives in inanimate masculine or neuter gender in singular, accusative case are tied with nonvocalized prepositions in accusative (na nastavenie/for setting, pre vstup/to enter). Similarly, a greater degree of positive correlation (lift = 3.5) was reached for the pairs (*VId+*, *VKepb+*), i.e., verbs in imperative are tied with verbs in the present, in the second person of plural (môžete pripojiť/ you can connect). Tag pairs (*SSns4, Eu4*), (*SSis4, Eu4*), and (*VKepb+, O*) reached also a positive correlation (*lift* = 2). The remaining pairs, apart from the pair (%, O), achieved the lift degree higher than 1 (see Supplementary Table 2).

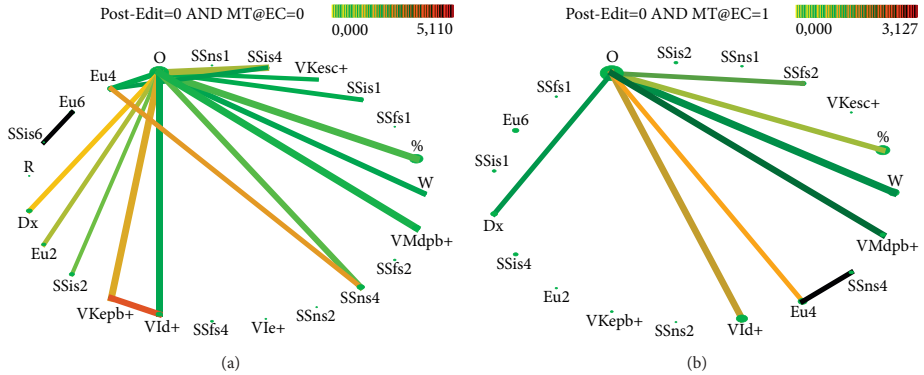



Figure 2: Visualization of tags identified in MT output translated by GT (a) and MT@EC (b).

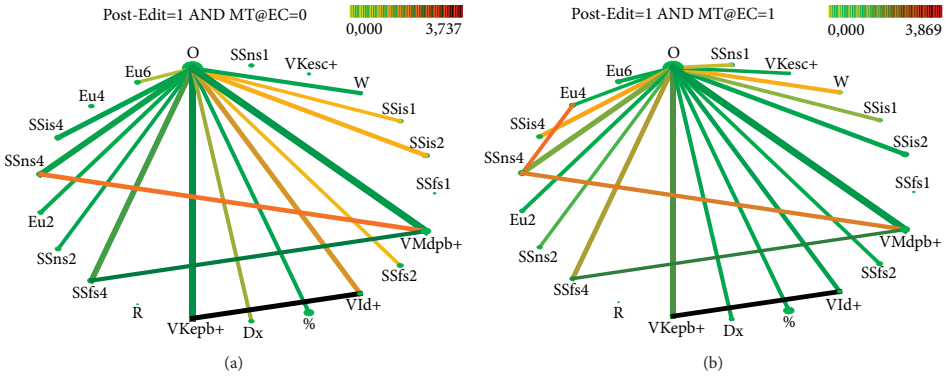



Figure 3: Visualization of tags identified in PEMT output translated by GT (a) and MT@EC (b).

For MT@EC_MT output (in Figure 2(b), see also Supplementary Table 2(b)), tag (*O*), conjunctions, was identified as the tag with the highest incidence in the sentence with the *support* of 45%. Similar to GT_MT output, the tag (*%*), foreign language citation has occurred in a sentence with a probability of more than 30%. The tags, with the probability of more than 10%, are (*VId+*) verbs in the affirmative and infinitive (*pozrieť/to see, poslať/to send*), (*W*) abbreviations (*DVD, HD*), (*Eu4*) prepositions in accusative and locative (*na/on, v/in*), and what is quite interesting, (*SSfs2*) nouns in feminine and in singular genitive cases (*potreby/the need*). The other tags did not again meet the minimum *support* of 10% (see Supplementary Table 2). MT@EC_MT output (in Figure 2(b), see also Supplementary Table 2(b)), with the probability of around 15%, contained mostly the following combinations—(*W, O*), (*O, VId+*), and (*O, VMdpb+*), i.e., abbreviations with conjunctions (DVD a/DVD and), and conjunctions with imperative or with imperative in the second person of plural (zaznamenať/to record, použite/use). It is most unexpected that nouns are not represented here. The rest of tag pairs (*%, O*), (*O, Dx*), (*Eu4, SSns4*), (*O, Eu4*), and (*SSfs2, O*) were identified with a probability of more than 10%. Tags with support of less than 10% are not depicted (in Figure 2(b), see also Supplementary Table 2(b)). For the pair (*SSns4, Eu4*), substantive in neuter gender, in singular, in accusative with the preposition in accusative, the greatest degree of positive correlation ($lift = 3.12$) was found (in Figure 2(b), see also Supplementary Table 2(b)). Slightly less positive correlation ($lift = 1.8$) was reached for (*VMdpb+, O*), imperative in the second person in plural and preposition (vyberte a/select and). Remaining tag pairs, apart from the pair (%, O), achieved the lift degree higher than 1 (see Supplementary Table 2).

In the case of PEMT outputs, both are very similar, because they are translations of the same source texts regardless of the MT system used. The only factor that plays a key role is the posteditor, i.e., the extent of his/her intervention and his/her lexical and stylistic preferences in postediting. The evidence lies in the very similar rules found in the PEMT outputs (in Figures 3(a) and 3(b), see also Supplementary Table 3).

The tag (*O*), conjunctions, also belonged to the tags with the highest occurrence with a probability of 50%. Tags (%), (VMdpb+), and (SSns4) were imperatives in the second person of plural, occurred in the sentences with the support of around 25%. Pairs of tags (*O, VMdpb+*), (*O, VKepb+*), and (*O, SSns4*) were identified with the *support* of around 20% in GT_PEMT output (in Figure 3(a), see also Supplementary Table 3(a)). In the case of MT@EC_PEMT output (in Figure 3(b), see also Supplementary Table 3(b)), (O, VMdpb+) with more than 20% of the support and (O, VKepb+) or (O, SSns4) with a probability of around 15% were found.

Based on the lift, we can claim that PEMT outputs are characterized by more frequent pairs (VId+, VKepb+)—verb in infinitive with modal verb (môžete použiť/can use) or a verb itself (použite/use). The highest interestingness of rule was found for the pair of tags (VKepb+, VId+) with lift = 3.74 in GT_PEMT and in MT@EC_PEMT for the same pair with lift = 3.87 (see Supplementary Table 2).

*3.2. Frequent Tagsets and Rules Summarization.* Based on the *Q* test results (Tables 2 and 3), the zero hypothesis, which reasons that the incidence of frequent tagsets does not depend on a way of translation (task), is rejected at the 0.001 significance level. The most frequent tagsets (almost 85%) were identified in MT@EC_PEMT, the lowest (almost 53%) in MT@EC_MT (Table 4).

From multiple comparisons (Table 4), two homogenous groups (MT@EC_MT) and (GT_PEMT, GT_MT, MT@EC_PEMT) were identified in terms of the average incidence of found frequent tagsets. Statistically significant differences were proved at the 0.05 significance level in the average incidence of frequent tagsets found between MT@EC_MT output and others.

Based on the *Q* test results (Table 3), the zero hypothesis, which reasons that the incidence of extracted rules does not depend on a way of translation (task), is rejected at the 0.001 significance level. The most extracted rules were found in translation MT@EC_PEMT output (almost 92%), the lowest in MT@EC_MT output (almost 34%) (Table 5).

From multiple comparisons (Table 5), three homogenous groups (MT@EC_MT), (GT_PEMT, GT_MT), and (MT@EC_PEMT) were identified in terms of the average incidence of extracted rules. Statistically significant differences were proved at the 0.05 significance level in the average incidence of found rules between MT@EC_MT output and others as well as between translation MT@EC_PEMT output and others. On the other hand, in both cases (Tables 4 and 5), a statistically significant difference between GT_MT output and GT_PEMT output was not found.

*3.3. Validation of the Proposed Methodology.* We have validated the proposed MT evaluation methodology based on the evaluation of frequent POS tags (tagsets) of MT outputs (MT@EC_MT or GT_MT) and PEMT outputs using the baseline methods. We used both, manual and semiautomatic MT evaluation measures as criteria for validity. In the case of manual evaluation, the criteria for validity are the scores of fluency (*F*) and adequacy (*A*). In the case of semiautomatic evaluation, we apply the HTER metric. Due to deviations from normality for testing differences between dependent variables, we used (Table 6) the Wilcoxon matched-pairs test.

Statistically significant differences were proved in the case of manual MT evaluation. The null hypotheses are rejected at the 0.001 significance level. We can see (in Figure 4(a)) differences in adequacy of MT output in favour of GT_MT output. Differences can be seen in the quartile range where 50% of the central values, for GT_MT output, were from the range [2, 5], contrary to MT@EC_MT output, where 50% of the central values were from the range [2, 4]. Similarly, in the case of the fluency of MT output (in Figure 4(b)), there are differences in favour of the GT_MT output. The differences can be seen in the median, where the estimation of the central value was 3 for the GT_MT output and 2.5 for MT@EC_MT output. In the case of both MT outputs, human translators used a range of the whole scale [from 1 to 5] to assess individual sentences, which indicates the heterogeneous quality of the examined MT sentences in case of adequacy and fluency.

Statistically significant differences were also shown in the case of the semiautomatic evaluation of MT, where H0 is rejected at the 0.001 significance level. We can see (in Figure 5(a)) differences in the HTER score in favour of GT_MT output. Based on the comparison of MTs with their corresponding PEMTs, a statistically significant lower error rate of MT output translated by GT compared to MT@EC_MT output was achieved.

Similar to manual evaluation, in the semiautomatic evaluation of individual sentences, the implemented HTER metric achieved the values of the whole range [0, 1], which refers to the heterogeneous quality of the examined MT segments for both MT outputs.

After rejecting the zero hypothesis, we are interested in MT segments with the highest differences in error rate (HTER) given to the used MT engine (MT@EC or GT). To identify segments, we use a method drawn from the residual analysis [59, 60]. We used this method to compare the results of semiautomatic MT evaluation of error rate between MT@EC_MT and GT_MT output (segment by segment). The aim of the analysis is to identify the segments (sentences) in which significant differences were found in the score of HTER of MT output (MT@EC and GT) from English into Slovak

$$\begin{aligned}(\text{residual value})_i = {} & (\text{value of } MT@EC_{MT})_i \\ & - (value\ of\ GT_{MT})_i,\ i = 1, 2, \ldots, I,\end{aligned} \tag{4}$$

where $I$ is a number of examined segments (sentences) in the dataset.

To identify extreme values (in Figure 5(b)), we use a rule $\pm 2\sigma$, i.e., residual values outside the interval we consider as extreme values

$$\begin{aligned}&\text{mean of residuals } (MT@EC_{MT} - GT_{MT}) \\ &\pm 2st.\text{dev.of residuals } (MT@EC_{MT} - GT_{MT}).\end{aligned} \tag{5}$$

Table 2: Cochran $Q$ test for incidence of frequent tagsets in examined translations.

| | GT_MT | | MT@EC_MT | | GT_PEMT | | MT@EC_PEMT | |
|---|---|---|---|---|---|---|---|---|
| Frequent tagset | Sup | Inc | Sup | Inc | Sup | Inc | Sup | Inc |
| (R) | 11.28 | 1 | 10.80 | 1 | 10.80 | 1 | | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| (*VMdpb+*) | 24.71 | 1 | 26.25 | 1 | 27.24 | 1 | 17.08 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| (*VIe+*) | 11.77 | 1 | | 0 | | 0 | | 0 |
| Cochran $Q$ test | $Q = 18.38298$; df = 3; $p < 0.001$ | | | | | | | |

Table 3: Cochran $Q$ test for incidence of extracted rules in examined translations.

| | GT_MT | | | MT@EC_MT | | | GT_PEMT | | | MT@EC_PEMT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rule | Sup | Lift | Inc | Sup | Lift | Inc | Sup | Lift | Inc | Sup | Lift | Inc |
| O = => *SSis1* | 11.77 | 1.21 | 1 | | | 0 | 10.47 | 1.08 | 1 | 10.80 | 1.17 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| *SSfs4* = => *VMdpb+* | | | 0 | | | 0 | 12.29 | 1.97 | 1 | 9.97 | 2.01 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| *SSns4* = => *Eu4* | 12.11 | 2.08 | 1 | 11.11 | 3.13 | 1 | | | 0 | 11.30 | 2.25 | 1 |
| Cochran $Q$ test | $Q = 39.90476$; df = 3; $p < 0.001$ | | | | | | | | | | | |

Table 4: Homogeneous groups for incidence of frequent tagsets in examined translations.

| Translation | Mean | 1 | 2 |
|---|---|---|---|
| MT@EC_MT | 0.529 | | **** |
| GT_PEMT | 0.765 | **** | |
| GT_MT | 0.765 | **** | |
| MT@EC_PEMT | 0.843 | **** | |

Table 5: Homogeneous groups for incidence of extracted rules in examined translations.

| Translation | Mean | 1 | 2 | 3 |
|---|---|---|---|---|
| MT@EC_MT | 0.333 | | **** | |
| GT_MT | 0.708 | **** | | |
| GT_PEMT | 0.750 | **** | | |
| MT@EC_PEMT | 0.916 | | | **** |

Table 6: Comparison of dependent samples MT@EC output and GT output.

| | $T$ | $Z$ | $p$ value |
|---|---|---|---|
| adequacy (MT@EC_MT) & adequacy (GT_MT) | 7457.0000 | 6.0682 | <0.001 |
| fluency (MT@EC_MT) & fluency (GT_MT) | 10870.5000 | 4.9426 | <0.001 |
| HTER (MT@EC_MT) & HTER (GT_MT) | 18269.5000 | 9.3839 | <0.001 |

Figure 5(b) visualizes the residuals for MT outputs (MT@EC_MT and GT_MT). Residual values above the average of the residuals indicate an above-average error rate of MT output produced by MT@EC against MT output produced by GT; residual values below the average of the residuals indicate an above-average error rate of GT_MT output against MT@EC_MT output. It identifies segments where the significant differences in the evaluation of error rate between MT@EC_MT output and GT_MT output exist. In the case of MT@EC_MT output (in Figure 5(b)), we identified 28 segments that showed a significant error rate against GT_MT output. In contrast to GT_MT output (in Figure 5(b)), only 15 segments showed a significant error rate against MT@EC_MT output. The identified segments were subsequently manually analyzed, which resulted in the determination of the main issue of the entire MT process from English into Slovak. The difficulty consists of an incorrect determination of predication (subject, verb, and object) leading to mistranslation or incorrect translation, either grammatically or semantically (different parts of

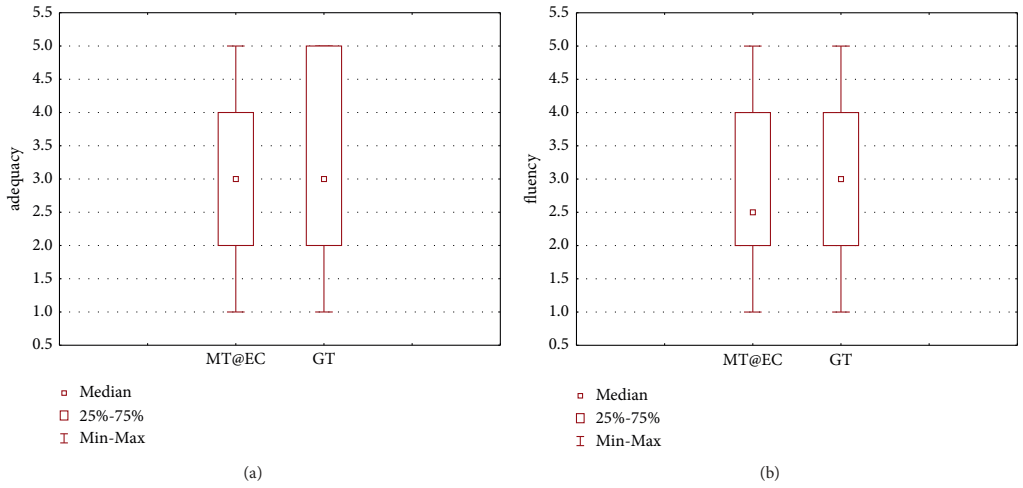



Figure 4: Visualization of descriptive statistics for manual MT evaluation: adequacy (a) and fluency (b).

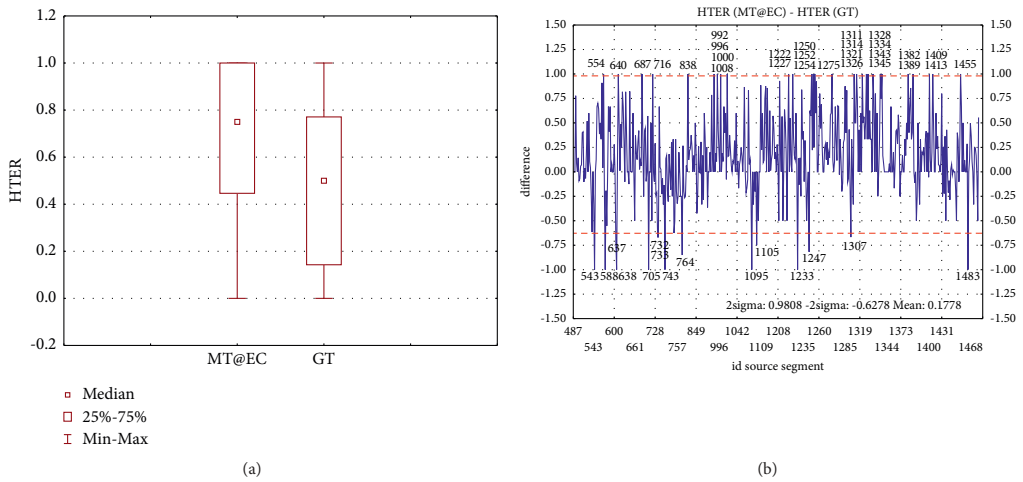



Figure 5: Visualization of descriptive statistics for semiautomatic MT evaluation: HTER metric (a) and residuals for MT output (b).

speech related to incorrect declension). These findings also confirm the results of the proposed MT evaluation methodology (MT@EC and GT) which is based on the evaluation of frequent tagsets.

## 4. Discussion

We agree with Mesmer et al. [13] that the text complexity must be understood to increase the knowledge about the interaction among the text characteristic, translator, or posteditor, and tasks such as translation or PE. The best approach to understand text complexity is through the analysis of words, sentences, and discourse [10]. For this reason, in the discussion, we will focus on the analysis of words and relationships among them.

In the case of GT, the highest value of the symmetric measure lift (*lift* = 5.11) was reached for a pair substantive in the inanimate masculine gender in the singular locative case with a preposition in locative. When analyzing this pair using the asymmetric measure confidence, we came across a discrepancy in the values (see Supplementary Table 4(a)). For the rule, SSis6 = => Eu6 the confidence value is 1.00, but for the rule, Eu6 = => SSis6 the confidence is only 0.58, which means that more often (100%) a preposition in

locative appears in segments (transactions) that contain substantive in the inanimate masculine gender in the singular locative only, as substantive in the inanimate masculine gender in the singular locative case in segments (transactions) that contain preposition in locative only (58.47%). The confidence measure points us to the fact that in the texts, prepositions are tied only to nouns, but nouns can also be found in a sentence as a subject or object, which do not require the presence of prepositions.

By MT@EC, it is substantive in the neuter gender, in the singular accusative case with the preposition in the accusative. As in the case of GT, a high lift value does not guarantee the same conditional probabilities of both directions of the rule (see Supplementary Table 4(b)). Even in the case of MT@EC, if the sentence contains a noun in the neuter gender, in the singular accusative case, then with 70% of confidence, it also contains the preposition in the accusative. However, the incidence of the noun in the neuter gender and in the singular accusative case is only 49.26% if the sentence contains a preposition in the accusative in the sentence only. A slightly less positive correlation occurred between imperative in the second person in the plural and conjunction (vyberte a/select and), which means if the sentence contains imperative in the second person in the plural, then with 81.55% of confidence, the sentence contains conjunction, but the incidence of imperative in the second person in the plural in the sentence is only 31.46% if there is conjunction only. Again, confidence values point to the fact that conjunctions are tied to either nouns or verbs in technical texts.

We can claim that GT_MT output was of a higher quality with respect to the rules and principles of the examined language. MT output has reached a greater congruence in gender, number, and case of a given language. The MT@EC engine rather translated word-by-word and did not focus on phrases and relations of the language. It literally translated from English (EN) to Slovak (SK) following grammar and the rules of the source language (EN) rather than the target language (SK). It was also confirmed by semiautomated evaluation, namely, by identifying specific segments (#554, #640, #1455), where the above-average error rate of MT@EC_MT output against GT_MT output was identified. These identified segments point to the specific errors that correspond to general (systematic) errors identified just by the analysis of frequent tagsets.

Each MT output either translated by GT or MT@EC was postedited by human translators. The aim of PE was to find out to which extent it was necessary to do PE to obtain a translation of publishing quality and to find out whether PE is more effective than a translation from scratch. We compared the quality of MT output with the quality of PEMT output based on language complexity, i.e., to what extent the tagsets and rules summarization for MT output and PEMT output are similar as well as the relation among tagsets that characterize language complexity.

The MT engine (GT) was relatively accurate (in Figure 3(a), see also Supplementary Table 3(a)) with only 5% of conjunctions inserted into the MT output (text). After a deeper analysis, we discovered that there are many conjunctions such as *and*, *or*, *when*, *after*, *then*, *to*, *before*, *if*. The conjunctions are tied to either the compound sentence or multiple sentence elements (verbs with objects). This is also shown by the high values of confidence measure (see Supplementary Table S5(a)), and despite the fact that the rules VKepb+ = => O and VId+ = => O do not reach the highest values of the lift (1.69; 1.59), they reach the highest confidence (84.38%; 79.46%). Other significant differences between the GT_MT output and GT_PEMT output were that the posteditors had to mainly correct or complete the masculine inanimate nouns in the singular genitive case together with prepositions in the genitive (*pomocou kábla/using cable*) as well as modal verbs (*môžete/you can*). This correction is closely linked with the flexion of the target language. Slovak consists of 4 paradigms of nouns (*S, A, F, U*), 4 genders (*m, f, n*), 2 numbers (*s, p*), and 7 cases (*1-7*). Compared to MT output, the posteditors had to mainly finish the translation by translating words that were not translated yet. It mainly referred to nouns in the neuter gender, in the singular accusative case, i.e., the adequate object was missing.

Based on the lift, we can claim that GT_PEMT output is characterized by more frequent phrase—verb in infinitive with modal verb (*môžete použiť/can use*) or a verb itself (*použite/use*), which is however in contrast with the values of lift in GT_MT output, where the most common phrase was a masculine, inanimate noun in locative of singular. It implies that the posteditors had to mostly correct the object, i.e., inflections and gender of nouns (*SSis6* to *SSns4* or *SSfs4*) corresponding to the agreement in the case of the preposition (*Eu6* to *Eu4*).

In the case of MT@EC, we received different results. In the MT@EC_MT output, the most frequent POSs were verbs in the infinitive (*VId+*), but in the MT@EC_PEMT output, there were imperatives in the second person of plural (*VMdpb+*). We deduce that the posteditors had not only to translate the nontranslated words (especially verbs and nouns in the subject) but also to a large extent modify verbs (the MT engine did not accept the rules of Slovak grammar, it kept the source language with its grammar, and it took only into account the basic form). Compared to MT@EC_MT output, differences have occurred in combination—conjunction and verb, where the verb has changed from infinitive to imperative (*VId* + to *VMdpb+*), as well as a noun which has changed from inanimate masculine to feminine, while the case and number were preserved (*SSis4* to *SSfs4*). The same rule was shown also in the MT@EC_PEMT output, i.e., if the sentence is not simple, then it is a copulative or conditional sentence, without agent expression, expressing only verb and object (*VMdpb+, SSfs4*) or multiple sentence element (*VMdpb+, O*), also one-syllable prepositions, which are tied to nouns in the accusative (*Eu4, SSns4*). The results of our error analysis also confirm the confidence values for MT@EC_MT output and its postedited version, i.e., the highest confidence (81.55%) for machine translation was achieved for the rule VMdpb + ==> O and for its postedited version, and the highest confidence (82.93%) was achieved for the rule VKepb+ = => O (see Supplementary Table 4(b) and Table 5(b)).

## 5. Conclusions

We focused on investigating the impact of an artificial agent (MT) on a human agent (posteditor) using the proposed methodology, which is based on POS tagging, frequent tagsets, association rules, and their summarization. We examined this impact from the point of view of perlocutionary acts, which includes the evaluation of machine translation. We have shown that the feature of adaptivity of a complex system requires human agents. Through human intervention, in our case by PE, MT systems can also contain the feature of adaptivity.

We proposed a new methodology for automatic MT evaluation using POS tags and association rules (in Figure 1). We compared two different MT engines—Google Translate and MT@EC (European Commission MT engine). We examined technical texts because they are the most frequently machine-translated texts. Based on the results of the analysis and found rules, we are able to characterize not only the text quality but also the text in terms of microstructure (morpho-syntactic relations).

Moreover, we validated the proposed methodology of MT evaluation using both manual and semiautomatic methods of MT evaluation (in Table 6). The results can be considered valid. The contribution of the proposed methodology is an identification of systematic, not random errors. In addition, the proposed methodology takes into account morpho-syntactic relations, which are important in evaluating translations between analytical and inflectional languages. Given that, we examined 4 translations conducted in four different ways—GT_MT, MT@EC_MT, GT_PEMT, and MT@EC_PEMT, and we investigated whether there are differences in the occurrence of frequent tagsets. Based on the $Q$ test result ($Q = 18.38298$; $df = 3$; $p < 0.001$), we discovered that there is a difference in method way of translation (translation process) with respect to tags' occurrence. Using multiple comparisons, we identified two homogenous groups (MT@EC_MT) and (GT_PEMT, GT_MT, MT@EC_PEMT), i.e., there is a statistically significant difference between MT output translated by MT@EC and others. In other words, the GT_MT output was very similar to PEMT whether the translation engine was GT or MT@EC. In the terms of found rules, a statistically significant difference was also proven (based on the result of $Q$ test ($Q = 39.90476$; $df = 3$; $p < 0.001$) and from multiple comparisons) between MT@EC_MT and MT@EC_PEMT, as well as between MT@EC_MT and others and also between MT@EC_PEMT and GT_MT or GT_PEMT output. The GT_MT output was very similar to GT_PEMT output. The posteditors similarly postedited both MT outputs, but to a large extent (statistically significant), the corrections were made in case of MT@EC_MT output.

To sum up our findings, we can state that for technical texts such as manuals, MT systems produce an output with an acceptable level of quality. A statistically significant difference between the GT_MT output and the GT_PEMT output in terms of the meaning or grammar was not proven. Last but not least, to answer the question concerning how to evaluate the MT quality or which methodology to use, we have shown an original and previously unused unique approach using text complexity measures. In our view, it is an objective evaluation of MT output by statistical, NLP, and machine learning methods. It can also be used for the automatic identification of MT errors into the inflectional language (e.g., Slovak).

The proposed methodology can serve as an alternative to the current, which use manual evaluation metrics, and which are not only time but also labour-consuming, but also use standard automatic evaluation metrics such as BLEU. We see the use of the methodology itself not only in evaluating MT quality but also in teaching MT and PE in the study programs of translation studies.

Another interdisciplinary contribution or future work lies in providing information to focus on during the PE process, which can finally improve translators' performance as expected by today's market.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the Slovak Research and Development Agency under the contract no. APVV-18-0473. This research was funded by European Commission under the ERASMUS + Programme, KA2, grant number: 2021-1-SK01-KA220-HED-000032095 "Future IT professionals Education in Artificial Intelligence," Ministry of Education of Slovakia, grant no. 004UKF-2-1/2021 "Preparation and development of teaching courses in English with a focus on artificial intelligence in the form of blended-learning," and Ministry of Education of Slovakia, grant number: 2020/8148: 34-A1101 "Support for the development of practical skills of UKF students in Nitra."

## Supplementary Materials

Supplementary Table 1: POS tagging for Slovak morphological annotation. Supplementary Table 2: tabulation of tags identified in MT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 3: tabulation of tags identified in PEMT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 4: tabulation of rules confidence in MT output translated by Google Translate (a) and MT@EC (b). Supplementary Table 5: tabulation of rules confidence in PEMT output translated by Google Translate (a) and MT@EC (b). (*Supplementary Materials*)

## References

[1] P. Liu and Z. Li, "Task complexity: a review and conceptualization framework," *International Journal of Industrial Ergonomics*, vol. 42, 2012.

[2] R. Pelánek, T. Effenberger, and J. Čechák, "Complexity and difficulty of items in learning systems," *International Journal of Artificial Intelligence in Education*, 2021.
[3] D. J. Campbell, "Task complexity: a review and analysis," *Academy of Management Review*, vol. 13, 1988.
[4] J. G. March and H. A. Simon, *Organizations*, Wiley, Oxford, England, 1958.
[5] G. Daroczy, M. Wolska, W. D. Meurers, and H.-C. Nuerk, "Word problems: a review of linguistic and numerical factors contributing to their difficulty," *Frontiers in Psychology*, vol. 06, 2015.
[6] Q. Ma and X. Wang, "What is language complexity?" *Macrolinguistics*, vol. 7, 2019.
[7] Ö. Dahl, *The Growth and Maintenance of Linguistic Complexity*, John Benjamins Publishing Company, Amsterdam, 2004.
[8] A. Andrason, "language complexity: an insight from complex-system theory," *International Journal of Language and Linguistics*, vol. 2, pp. 74–89, 2014.
[9] X. Guo-zhi, *Systems Science*, Shanghai Technology and Education Press, Shanghai, 2000.
[10] G. Berninger and C. Garvey, "Tag constructions: structure and function in child discourse," *Journal of Child Language*, vol. 9, 1982.
[11] S. Zhou and W. Liu, "English grammar error correction algorithm based on classification model," *Complexity*, vol. 2021, Article ID 6687337, 11 pages, 2021.
[12] R. Loock, "No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment," *J. Spec. Transl.*vol. 34, pp. 150–170, 2020.
[13] H. A. Mesmer, J. W. Cunningham, and E. H. Hiebert, "Toward a theoretical model of text complexity for the early grades: learning from the past, anticipating the future," *Reading Research Quarterly*, vol. 47, no. 3, pp. 235–258, 2012.
[14] M. L. Forcada, "Making sense of neural machine translation," *Transl. Spaces*, vol. 6, 2017.
[15] O. De Clercq, G. De Sutter, R. Loock, B. Cappelle, and K. Plevoets, "Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French," *Transl. Q*, vol. 101, pp. 1–21, 2021.
[16] H. Hassan, A. Aue, C. Chen et al., *Achieving Human Parity on Automatic Chinese to English News Translation*, http://arxiv.org/abs/1803.05567 ArXiv. accessed, 2018.
[17] L. Macken, L. Van Brussel, and J. Daems, "NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output," *Comput. Linguist. Netherlands J.*vol. 9, pp. 67–80, 2019.
[18] W. Farhan, B. Talafha, A. Abuammar et al., "Unsupervised dialectal neural machine translation," *Information Processing & Management*, vol. 57, 2020.
[19] R. Loock, "Traduction automatique et usage linguistique: une analyse de traductions anglais-français réunies en corpus," *Meta Le J. Traducteurs/Meta Transl. J*, vol. 63, pp. 786–806, 2018.
[20] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process*, vol. 1, pp. 1–10, Association for Computational Linguistics, Beijing, China, 2015.
[21] M. Yamada, "The impact of Google neural machine translation on post-editing by student translators," *J. Spec. Transl.*vol. 31, 2019.
[22] P. Kumar, R. Ahmad, B. D. Chaudhary, and R. Sangal, "Machine translation system as virtual appliance: for scalable service deployment on cloud," in *Proceedings of the 2013 IEEE Seventh Int. Symp. Serv. Syst. Eng.*, pp. 304–308, IEEE, San Francisco, CA, USA, Mar 2013.
[23] F. Gobbo, "Machine translation as a complex system: the role of Esperanto," *Interdisciplinary Description of Complex Systems*, vol. 13, no. 2, pp. 264–274, 2015.
[24] A. M. Hayes and L. A. Andrews, "A complex systems approach to the study of change in psychotherapy," *BMC Medicine*, vol. 18, p. 197, 2020.
[25] A. F. Siegenfeld and Y. Bar-Yam, "An introduction to complex systems science and its applications," *Complexity*, vol. 2020, Article ID 6105872, 16 pages, 2020.
[26] X. Feng, Z. Feng, W. Zhao, B. Qin, and T. Liu, "Enhanced neural machine translation by joint decoding with word and POS-tagging sequences," *Mobile Networks and Applications*, vol. 25, 2020.
[27] J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., *Translation Quality Assessment*, Springer International Publishing, Cham, 2018.
[28] M. Popović, "Error classification and analysis for machine translation quality assessment," in *Mach. Transl. Technol. Appl.*, J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., Springer, Cham, 2018.
[29] D. Shterionov, F. d. Carmo, J. Moorkens et al., "A roadmap to neural automatic post-editing: an empirical approach," *Machine Translation*, vol. 34, no. 2-3, pp. 67–96, 2020.
[30] S. Castilho, S. Doherty, F. Gaspari, and J. Moorkens, "Approaches to human and machine translation quality assessment," in *Transl. Qual. Assessment. Mach. Transl. Technol. Adn Appl.*Springer, Cham, 2018.
[31] K. Hu, S. O'Brien, and D. Kenny, "A reception study of machine translated subtitles for MOOCs," *Perspectives*, vol. 28, no. 4, pp. 521–538, 2020.
[32] G. De Sutter, M.-A. Lefer, and I. Delaere, Eds., *Empirical Translation Studies*, De Gruyter, Berlin, Boston, 2017.
[33] P. Isabelle, C. Cherry, and G. Foster, "A challenge set approach to evaluating machine translation," in *Proc. 2017 Conf. Empir. Methods Nat. Lang. Process.*Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.
[34] E. Vanmassenhove, D. Shterionov, and A. Way, "Lost in translation: loss and decay of linguistic richness in machine translation," in *Proc. Mach. Transl. Summit XVII*, vol. 1, pp. 222–232, Res. Track, European Association for Machine Translation, Dublin, Ireland, 2019.
[35] B. Dorr, M. Snover, and N. Madnani, "Part 5: machine translation evaluation," in *Handb. Nat. Lang. Process. Mach. Transl. DARPA Glob. Auton. Lang. Exploit.*, J. M. Joseph Olive and C. Christianson, Eds., , p. 936, Springer, 2011.
[36] J. Daems, S. Vandepitte, R. J. Hartsuiker, and L. Macken, "Identifying the machine translation error types with the greatest impact on post-editing effort," *Frontiers in Psychology*, vol. 8, p. 1282, 2017.
[37] C. Lo and D. Wu, "MEANT: an inexpensive, high-accuracy, semiautomatic metric for evaluating translation utility via semantic frames," *ACLPPinforma*, vol. 11, pp. 220–229, 2011.
[38] A. Toral, "Post-editese: an exacerbated translationese," in *Proc. Mach. Transl. Summit XVII*, vol. 1, pp. 273–281, Res. Track, European Association for Machine Translation, Dublin, Ireland, 2019.
[39] M. Munk, D. Munkova, and L. Benko, "Towards the use of entropy as a measure for the reliability of automatic MT

evaluation metrics," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 5, pp. 3225–3233, 2018.

[40] L. Benkova, D. Munkova, Ľ. Benko, and M. Munk, "Evaluation of English–Slovak neural and statistical machine translation," *Applied Sciences*, vol. 11, 2021.

[41] A. Lommel, "Metrics for translation quality assessment: a case for standardising error typologies," in *Transl. Qual. Assess.*, J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty, Eds., Springer International Publishing, Cham, 2018.

[42] M. Nadejde, S. Reddy, R. Sennrich et al., "Predicting target language CCG supertags improves neural machine translation," in *Proc. Second Conf. Mach. Transl.*Association for Computational Linguistics, Stroudsburg, PA, USA, 2017.

[43] D. Hládek, J. Staš, and J. Juhár, "Rule-based morphological tagger for an inflectional language," in *Cogn. Behav. Syst. Lect. Notes Comput. Sci.*, pp. 208–215, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[44] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: an open source trigram tagger," in *Proc. 45th Annu. Meet. Assoc. Comput. Linguist. Companion Vol. Proc. Demo Poster Sess.*, pp. 209–212, Association for Computational Linguistics, Prague, Czech Republic, 2007.

[45] L. J. Laki, G. Orosz, and A. Novák, "HuLaPos 2.0 - decoding morphology," in *Advances in Artificial Intelligence and Its Applications*, pp. 294–305, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[46] C. Conforti, M. Huck, and A. Fraser, "Neural morphological tagging of lemma sequences for machine translation costanza Conforti," in *Proc. 13th Conf. Assoc. Mach. Transl. Am. (Volume 1 Res. Track), Association for Machine Translation in the Americas*pp. 39–53, Boston, MA, 2018.

[47] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 workshop on statistical machine translation," *Proc. Fourth Work. Stat. Mach. Transl*, pp. 1–28, 2009.

[48] M. Popović, "Morphemes and POS tags for n-gram based evaluation metrics," *Proc. Sixth Work. Stat. Mach. Transl.*, 2011.

[49] M. Popović, *Machine Translation: Statistical Approach with Additional Linguistic Knowledge*, RWTH Aachen University, Aachen,Germany, 2009.

[50] M. Popović, "rgbF: an open source tool for n-gram based automatic evaluation of machine translation output," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 99–108, 2012.

[51] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proc. Int. Conf. New Methods Lang. Process*pp. 44–49, Manchester, UK, 1994.

[52] H. Schmid, "Improvements in part-of-speech tagging with an application to German," in *Text, Speech and Language Technology*, S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, Eds., Kluwer Academic Publishers, Dordrecht, pp. 13–25, 1999.

[53] H. Schmid, M. Baroni, E. Zanchetta, and A. Stein, "The enriched TreeTagger system," in *Proc. EVALITA 2007 Work.*, 2007.

[54] V. Benko, "Compatible sketch grammar experiment," in *Proc. Int. Conf. «Corpus Linguist*pp. 21–29, St. Petersburg, 2013.

[55] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. Assoc. Mach. Transl*, pp. 223–231, Am., 2006.

[56] M. Snover, N. Madnani, B. J. Dorr, and R. Schwartz, "Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric," in *Proc. Fourth Work. Stat. Mach. Transl.*, pp. 259–268, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, http://dl.acm.org/citation.cfm?id=1626431.1626480.

[57] D. Munková, J. Kapusta, and M. Drlík, "System for post-editing and automatic error classification of machine translation," in *DIVAI 2016 11th Int. Sci. Conf. Distance Learn. Appl. Informatics, Sturovo, May 2 – 4, 2016*, pp. 571–579, Wolters Kluwer, Sturovo, 2016.

[58] D. Munková, M. Munk, and M. Vozár, "Data pre-processing evaluation for text mining: transaction/sequence model," *Procedia Comput. Sci.*vol. 18, pp. 1198–1207, 2013.

[59] M. Munk and D. Munkova, "Detecting errors in machine translation using residuals and metrics of automatic evaluation," *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 5, pp. 3211–3223, 2018.

[60] D. Munková and M. Munk, "Automatic evaluation of machine translation through the residual analysis," in *Lecture Notes in Computer Science*, D. S. Huang and K. Han, Eds., Springer, Nitra,Slovakia, pp. 481–490, 2015.

PRÍLOHA I: BENKO, Ľ.UBOMÍR, DASA MUNKOVÁ A MICHAL MUNK, 2023. RELATIONSHIP BETWEEN LINGUISTIC COMPLEXITY AND MT ERRORS IN THE CONTEXT OF INFLECTIONAL LANGUAGES. IN: *RECENT CHALLENGES IN INTELLIGENT INFORMATION AND DATABASE SYSTEMS. ACIIDS 2023*. SPRINGER, CHAM, S. 546–557. DOI:10.1007/978-3-031-42430-4_45 (**SCOPUS**) [SCOPUS: 0]

# Relationship Between Linguistic Complexity and MT Errors in the Context of Inflectional Languages

Ľubomír Benko(✉), Dasa Munková, and Michal Munk

Department of Informatics, Constantine the Philosopher University in Nitra, 949 01 Nitra, Slovakia
{lbenko,dmunkova,mmunk}@ukf.sk

**Abstract.** The quality produced by the MT tool varies, from very high to very low, depending on the intrinsic linguistic features of the source and target languages. The more lexical words the text content, the more information contains. Lexical richness is a multidimensional construct for assessing language development. In our study, we apply this concept for assessing the quality of machine translation. We attempt to link lexical richness with types of errors occurring in machine translation. We focus on the identification of the relationship between five types of MT errors and examined measures of language complexity. We used Goodman and Kruskal's gamma to determine which measures of language complexity are associated with the MT error types. The results revealed that not all measures of lexical richness are associated with the MT errors in each examined error category. We showed that readability does not associate with an error rate, namely with error categories used in the present study.

**Keywords:** Machine Translation · Lexical Complexity · Text Readability · Error Classes · Natural Language Processing

## 1 Introduction

The last decade has brought many advances in the field of language that are mainly related to artificial intelligence and machine translation. With artificial intelligence, translation is faster not only for professional translators but also for people like tourists or businessmen. Artificial intelligence (AI) is developing extremely fast and is bringing a revolution in the language and translation industries. Machine translation (MT), as one of the applications of natural language processing, as well as natural language processing itself, has become an integral part of the global technological competition for primacy in the field of artificial intelligence [1]. This rapidly developing technology field changes not only the way people work but also how they understand textual data. Understanding a language in its written form is related to internal text properties such as text readability or text diversity (granularity). MT depends on computer language, i.e. on binary language, which distinguishes only 0 and 1, but by applying AI to translation, MT becomes more

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
N. T. Nguyen et al. (Eds.): ACIIDS 2023, CCIS 1863, pp. 546–557, 2023.
https://doi.org/10.1007/978-3-031-42430-4_45

similar to human translation in terms of fluency and accuracy [2]. Under the assumption of a sufficient training corpus, the performance of an MT model with integrated AI was higher than a statistical MT model, therefore more researchers began to conduct more experiments with AI in machine translation [3].

MT is the automatic conversion of text from one natural language to another natural language. The quality produced by MT tools varies, ranging from very high to very low, depending on the intrinsic properties of the given text and language pair. Its accuracy depends on the source and target languages, industry (domain), source text quality, training corpus, and other factors [2]. Nowadays, the most widely used freely available MT tool is Google Translate [4], which we also used in our study. Google Neural MT is an AI technology that uses deep learning to produce a translation.

MT tools, which are currently used, are based on neural networks. Neural MT offers better results in product quality, especially in terms of text readability, than its predecessor statistical MT. For example, Pangeanic's neural MT tools guarantee very high parity with a human translator (90 - 95%) and this is one of the reasons why many companies and government institutions in the EU have decided to use neural MT tools for translation [5]. MT is also increasingly being applied outside the translation industry. According to Lihua [2: 7] especially in healthcare, the automotive industry, and the IT industry.

Quality assessment plays a key role in optimizing MT tool or engine performance. Based on the results of the manual or automatic evaluation, MT tools are optimized. The biggest disadvantage of manual evaluation is its subjectivity and is labour- and time-consuming. On the other hand, manual evaluation is highly desirable and is considered very reliable. The advantage of automatic evaluation is speed, objectivity, and reusability, but it does not provide detailed information about the translation error rate or its accuracy. The result of the automatic evaluation is a score between 0 and 1, which is calculated based on a lexical comparison of the MT output with a human translation (reference). Despite the fact that results of manual or automatic evaluation provide very valuable information and help to improve MT systems, developers and researchers often look for additional information which could affect the performance of a given system or engine and help them to answer questions such as: What are the biggest problems of MT systems? What are the strengths and weaknesses of the given system? Finding the relationship between the strengths or weaknesses of the system and the result of manual or automatic evaluation is not easy [6]. It motivated us to implement error analysis to create an error profile of MT output. However, during error identification and classification, we encountered the same problems as during manual evaluation, which is mainly associated with time and labour. So besides error annotation, we decided to apply measures of language complexity, characterizing text complexity and examine the relationship between these measures and individual MT error types (categories).

The aim of the study is to determine the relationship between individual types of MT errors and selected measures of text complexity. We attempt to characterize MT errors using the internal properties of the text, i.e., to determine which properties of the text are associated with the error rate of the given category.

The remaining part of this paper is organized as follows: Sect. 2 briefly reviews the research background, Sect. 3 provides information about our data, Sect. 4 demonstrates our results, as well as the analysis and finally Sect. 5 concludes our work.

## 2 Research Background

The complexity of natural language is one of the current topics that is discussed not only in natural language processing and AI but also in linguistic research. Language complexity is considered a suitable measurable approach that describes performance, indicates proficiency, and measures an individual's development in language [7]. This approach is based on a theoretical framework considering language as a complex adaptive system [8], recognizing the relationship between the target language and its use as a double system, in which linear and non-linear links between several factors are formed by predicative, but also non-predicative tasks of contextual and linguistic elements [9]. Linguistic complexity represents an inherent property of a system by which languages can be compared according to several formal properties, such as the number of rules to obtain a certain output, the number of exceptions to the rules, or the size of the lexicon differing at different levels of representation [10]. It is a property of language that can be measured in several language subsystems [11].

Text Complexity is an internal characteristic of a written text affecting the performance of computer applications which process the text [12]. Text complexity is independent either from the reader of the text or the environmental conditions, i.e. it is an independent typographic representation of the text, which can manifest itself at all three levels - lexical, syntactic, and discourse. At the lexical level, the factors affecting lexical complexity (the meaning of words, or the lexical choice used in the text) mainly consist of rich vocabulary, long words, infrequent and technical terms, ambiguous words, vague quantifiers, inconsistent terminology, and figurative language. Several formulas are applied to calculate the frequencies of individual linguistic units. This approach is based on the premise that the more lexically rich and diverse the text is, the more complex it is. Lexical complexity deals with lexical density, diversity, and rareness [13]. Kalantari and Gholami [13: 3] define lexical density as the proportion of lexical (content) words to the total number of words in a text while lexical diversity measures the number of different words and specific word types used in a text. According to Mat Daud et al. [14], lexical density is inversely proportional to readability, and therefore readability formulas could be applied to lexical density. Modern approaches in natural language processing such as machine learning techniques are applied to automate the process of the assessment of readability [15].

At the syntactic level, the main factors [16] that affect the length and syntactic structure of the sentence are long sentences, complicated syntax, passive voice, and negative constructions. Formulas for measuring syntactic complexity are generally based on quantifying linguistic units [17]. Syntactic complexity is a quality of sentence construction and its underlying principles [11: 456].

In the context of translation quality evaluation, readability measures are often used to assess or determine the complexity of the source and target text [18]. Currently, there are more than 200 readability formulas [19] that are used for different types of documents or for various industries. Even military and governmental agencies develop formulas themselves, as they consider them a very good quality indicator when writing technical manuals. Belonging to the most used formulas are the Automated Readability Index (for technical documents and manuals), Flesch Reading Ease (any kind of text), and Flesch-Kincaid (manuals, forms, and other technical documents).

For the publishing industry, which deals with newspapers, journals, online media etc., it is suitable to use Flesch Reading Ease (any kind of text), SMOG (text aimed at secondary-age readers) or Gunning Fog (business publications and journals).

## 3 Research Method

We were inspired by the previous research [20–24], who investigated the relationship between automatic MT evaluation metrics and the types of errors that occurred in neural MT. In this research, we focus on the internal characteristics of a complex adaptive system, in our case an MT system, and attempt to determine the relationship between measures of text complexity and distribution of errors over the defined error classes. We attempt to determine which of the examined measures of lexical and syntactic complexity (30 measures) associate the best with individual error classes of a categorical framework for error analysis [25: 100]. The examined texts (1903 sentences/66 texts) were taken from the British online newspaper The Guardian. In 2021, the texts were translated by the freely available Neural Google Translate (NGT) engine. The dataset (Table 1) consists of neural MT texts translated from English into Slovak (NMTs).

**Table 1.** Dataset composition

| Feature type | Feature name | NMTs_SK |
|---|---|---|
| Readability | Average sentence length | 17.12034 |
| | Average word length | 5.696361 |
| | #short sentences ($n < 10$) | 469 |
| | #long sentences ($n \geq 10$) | 1434 |
| Lexico-grammatical | Frequency of proper nouns | 1501 |
| | Frequency of nouns | 10070 |
| | Frequency of adjectives | 3324 |
| | Frequency of adverbs | 933 |
| | Frequency of verbs | 5198 |
| | Frequency of pronominals | 2371 |
| | Frequency of particles | 592 |
| | Frequency of foreign words | 841 |
| | Frequency of numerals | 617 |
| | Frequency of prepositions & conjunctions | 6028 |
| | Frequency of interpunction | 5958 |

Subsequently, the NMTs were manually annotated according to the framework for error analysis [25: 100] by three Slovak linguists. The categorical framework consists of five error classes (categories):

1. Predication,
2. Modal and communication sentence framework,
3. Syntactic-semantic correlativeness,
4. Compound/complex sentences,
5. Lexical semantics.

After manual error classification, we identified 3081 error segments in our corpus. Specifically, we identified 686 errors in the category of Prediction, 52 errors in the category of Modal and communication sentence framework, 1486 errors in Syntactic-semantic correlativeness, 671 errors in Compound/complex sentences, and 2778 errors in the category of Lexical semantics (Table 2).

### 3.1 Hypothesis

Based on the descriptive statistics we assume that there is a statistically significant difference between the error categories of the framework. We state the following hypothesis:

*H0: Between the error categories, there are no differences in the frequency of errors obtained from the examined neural MT texts.*

### 3.2 Methods

For the metrics of lexical and syntactic complexity, the Python Natural Language Toolkit (NLTK) library was used. To calculate the score of lexical diversity of texts [26] we applied the Type-Token Ratio (TTR), Herdan's lexical diversity measure (Herdan's C), Guiraud's Root TTR (Guiraud's R), Carroll's Corrected TTR (CTTR), Summer's lexical diversity measure (Summer's index S), Dugast's lexical diversity measure (Dugast's Uber Index U), Maas's lexical diversity measure (Maas's indices (a, $\log V_0$ & $\log_e V_0$)), Mean Segmental Type-token Ratio (MSTTR), Moving Average Type-token Ratio (MATTR), Measure of Textual Lexical Diversity (MTLD), Hypergeometric distribution diversity measure (HD-D), and Hapax legomenon ratio (Hapax). To calculate the score of lexical density, we applied the following readability formulas: Flesch reading ease (FRE), Flesch-Kincaid Grade Level (FKG), Fog Scale (Gunning FOG Formula) (FOG), SMOG Index (SMOG), Automated Readability Index (ARI), Coleman-Liau Index (CLI), McAlpine EFLAW Readability Score (MAR), and Reading Time (RT). To calculate syntactic complexity, we applied Sentence Count (SenC) and Word count (WordC). We also applied traditional features such as Syllable Count (SC), Character Count (CharC), Letter Count (LetC), Polysyllable Count (PolyC), Monosyllable Count (MonoC), and Lexicon Count (LC), which may contribute to the complexity of a text.

## 4 Results

Based on the Friedman ANOVA test ($ChiSqr = 246.584, N = 66, df = 4, p < 0.001$) we proved statistically significant differences between the categories (Predication, Modal and communication sentence, framework, Syntactic-semantic correlativeness, Compound/complex sentences, and Lexical semantics) in the frequency of errors found in neural MT texts.

**Table 2.** Descriptive Statistics

| | N | Mean | Median | Sum | Min | Maxi | Lower Q | Upper Q | Quartile R |
|---|---|---|---|---|---|---|---|---|---|
| Error_segments | 66 | 46.68 | 41.00 | 3081.00 | 7.00 | 141.00 | 26.00 | 61.00 | 35.00 |
| Correct_segments | 66 | 2.88 | 2.00 | 190.00 | 0.00 | 13.00 | 1.00 | 4.00 | 3.00 |
| Predication | 66 | 10.39 | 9.00 | 686.00 | 1.00 | 36.00 | 6.00 | 14.00 | 8.00 |
| Modal and communication framework | 66 | 0.79 | 0.00 | 52.00 | 0.00 | 4.00 | 0.00 | 2.00 | 2.00 |
| Syntactic-semantic correlativeness | 66 | 22.52 | 19.50 | 1486.00 | 4.00 | 65.00 | 12.00 | 30.00 | 18.00 |
| Compound/complex sentences | 66 | 10.17 | 9.00 | 671.00 | 0.00 | 32.00 | 6.00 | 12.00 | 6.00 |
| Lexical semantics | 66 | 42.09 | 37.00 | 2778.00 | 6.00 | 125.00 | 23.00 | 54.00 | 31.00 |

After rejecting the global null hypothesis, we are interested in the relationship between error categories that have a statistically significant difference. We identified ($p > 0.05$) only one homogeneous group (Compound/complex sentences and Predication). For both categories, the frequency of errors was approximately equal. A statistically significant difference ($p < 0.05$) was identified between the remaining error categories, whereby the lowest frequency of errors was obtained for the Modal and communication framework and the highest for Syntactic-semantic correlativeness and Lexical semantics.

Based on the results of multiple comparisons (Table 3), we showed statistically significant differences between Syntactic-semantic correlativeness/Lexical semantics/Modal and communication framework and other categories, but there is no statistically significant difference between Compound/complex sentences and Predication. This is also confirmed by the median (9), which is the same for both error categories.

We were also interested in how the measures of lexical and syntactic complexity explain the frequency of errors for individual categories, i.e., which of the automatic measures was the best, or conversely, the worst associated with the frequency of found errors for each category. For this purpose, we applied 30 measures of text complexity (including traditional linguistic features) and through Goodman and Kruskal's gamma, we determine the rank associations between error categories and measures of text complexity.

**Table 3.** Multiple comparisons

| | Median | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Modal and communication framework | 0 | 0.79 | | **** | | |
| Compound/complex sentences | 9 | 10.17 | **** | | | |
| Predication | 9 | 10.39 | **** | | | |
| Syntactic-semantic correlativeness | 19.5 | 22.52 | | | **** | |
| Lexical semantics | 37 | 42.09 | | | | **** |

Note: *** * p > 0.05

Gamma represents the probability (Table 4) of whether two variables are in the same or opposite order. It represents the degree of association between two variables.

The errors in the category of Prediction (Table 4) are partially identified by measures UniWC, RTTR, CTTR, WordC LC, MonoC, LetC, CharC, SC, RT, PolyC, and SenC, where were achieved low to moderate measures of positive and statistically significant association ($Gamma > 0.27/0.302$, $p < 0.01/0.001$). Similarly, it achieved a low measure of association, but a statistically significant negative association between Prediction and Hapax and/or TTR ($Gamma < -0.190$, $p < 0.05/0.01$).

The error rate in the Prediction category is best explained by the metric of syntactic complexity (SenC) and traditional linguistic features that are based on the count, i.e., the higher the score of these metrics and linguistic features, the higher the error rate in the category of Prediction. And vice versa, measures of lexical diversity such as Hapax or TTR are negatively associated with error rate in the category of Prediction, i.e., the greater the score of the total number of unique words (types) divided by the total number of words, the lower the association with the given category.

We proceeded in the same way with the other categories. For the category of Modal and communication framework, a low positive measure of association with measures such as SenC, MonoC or MTLD ($Gamma > 0.244$, $p < 0.05/0.01$) was achieved. For this category, we also showed a small, but statistically significant association between traditional linguistic features and the occurrence of MT errors as well as between lexical diversity and MT errors in the given category.

For the category of Syntactic-semantic correlativeness (Table 5), we have achieved a low to a high positive measure of association between measures SenC, MonoC (*Gamma* $>$ 0.505, $p < 0.001$), WordC, RT, CharC, LC, SC, LetC, UniWC, RTTR, CTTR, PolyC (*Gamma* $>$ 0.400, $p < 0.001$), HD-D (*Gamma* $= 0.178$, $p < 0.05$) and error rate of neural MT within this category. Besides the positive measure of association, we also achieved a low to moderate negative, but also a statistically significant measure of association between Herdan, Hapax, TTR, and occurrence of MT errors in this category (*Gamma* $< -0.247/-0.319$, $p < 0.01/0.001$).

**Table 4.** Goodman and Kruskal's gamma – Prediction

| | Valid N | Gamma | Z | p-value |
|---|---|---|---|---|
| **predication & SenC_NMT** | 66 | **0.334*****| 3.832 | **0.0001** |
| **predication & PolyC_NMT** | 66 | **0.308***** | 3.555 | **0.0004** |
| **predication & RT_NMT** | 66 | **0.302***** | 3.502 | **0.0005** |
| **predication & SC_NMT** | 66 | **0.302***** | 3.501 | **0.0005** |
| **predication & CharC_NMT** | 66 | **0.302***** | 3.502 | **0.0005** |
| **predication & LetC_NMT** | 66 | **0.297***** | 3.450 | **0.0006** |
| **predication & MonoC_NMT** | 66 | **0.296***** | 3.432 | **0.0006** |
| **predication & LC_NMT** | 66 | **0.288***** | 3.350 | **0.0008** |
| **predication & WordC_NMT** | 66 | **0.286***** | 3.312 | **0.0009** |
| **predication & UniWC_NMT** | 66 | **0.285**** | 3.302 | **0.0010** |
| **predication & RTTR_NMT** | 66 | **0.270**** | 3.133 | **0.0017** |
| **predication & CTTR_NMT** | 66 | **0.270**** | 3.133 | **0.0017** |
| **predication & MTLD_NMT** | 66 | 0.132 | 1.538 | 0.1240 |
| **predication & HD-D_NMT** | 66 | 0.123 | 1.425 | 0.1541 |
| **predication & FRE_NMT** | 66 | 0.061 | 0.709 | 0.4782 |
| **predication & MSTTR_NMT** | 66 | 0.034 | 0.397 | 0.6914 |
| **predication & Maas_NMT** | 66 | 0.032 | 0.373 | 0.7090 |
| **predication & MATTR_NMT** | 66 | 0.016 | 0.181 | 0.8564 |
| **predication & Dugast_NMT** | 66 | -0.032 | -0.373 | 0.7090 |
| **predication & SMOG_NMT** | 66 | -0.056 | -0.645 | 0.5191 |
| **predication & MAR_NMT** | 66 | -0.076 | -0.879 | 0.3795 |
| **predication & FKG_NMT** | 66 | -0.077 | -0.885 | 0.3762 |
| **predication & CLI_NMT** | 66 | -0.078 | -0.902 | 0.3673 |
| **predication & Summer_NMT** | 66 | -0.078 | -0.905 | 0.3655 |
| **predication & ARI_NMT** | 66 | -0.103 | -1.194 | 0.2326 |
| **predication & FOG_NMT** | 66 | -0.108 | -1.257 | 0.2089 |
| **predication & Herdan_NMT** | 66 | -0.161 | -1.866 | 0.0620 |
| **predication & Hapax_NMT** | 66 | **-0.190*** | -2.206 | **0.0274** |
| **predication & TTR_NMT** | 66 | **-0.231**** | -2.681 | **0.0073** |

Note: 0.00 to 0.10 (0.00 to −0.10) – a trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – a low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – a moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – a high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – a very high positive (negative) measure of association; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

For the category compound/complex sentences, we obtained similar results as for syntactic-semantic correlativeness. We have achieved a low to moderate positive measure of association between measures UniWC, SenC, RT, CharC, WordC, SC, LC, MonoC, LetC, RTTR, CTTR, PolyC (*Gamma* $>$ 0.398, $p < 0.001$), MTLD, HD-D (*Gamma* $>$ 0.193, $p < 0.05$) and occurrence of MT errors within this category. On

**Table 5.** Goodman and Kruskal's gamma - Syntactic-semantic correlativeness

| | Valid N | Gamma | Z | p-value |
|---|---|---|---|---|
| **syn-sem corr & SenC_NMT** | 66 | **0.507*** ** | 5.876 | **0.0000** |
| **syn-sem corr & MonoC_NMT** | 66 | **0.505*** ** | 5.916 | **0.0000** |
| **syn-sem corr & WordC_NMT** | 66 | **0.496*** ** | 5.808 | **0.0000** |
| **syn-sem corr & RT_NMT** | 66 | **0.494*** ** | 5.787 | **0.0000** |
| **syn-sem corr & CharC_NMT** | 66 | **0.494*** ** | 5.787 | **0.0000** |
| **syn-sem corr & LC_NMT** | 66 | **0.490*** ** | 5.742 | **0.0000** |
| **syn-sem corr & SC_NMT** | 66 | **0.489*** ** | 5.735 | **0.0000** |
| **syn-sem corr & LetC_NMT** | 66 | **0.486*** ** | 5.706 | **0.0000** |
| **syn-sem corr & UniWC_NMT** | 66 | **0.483*** ** | 5.658 | **0.0000** |
| **syn-sem corr & RTTR_NMT** | 66 | **0.449*** ** | 5.269 | **0.0000** |
| **syn-sem corr & CTTR_NMT** | 66 | **0.449*** ** | 5.269 | **0.0000** |
| **syn-sem corr & PolyC_NMT** | 66 | **0.400*** ** | 4.669 | **0.0000** |
| **syn-sem corr & HD-D_NMT** | 66 | **0.178*** | 2.089 | **0.0367** |
| **syn-sem corr & FRE_NMT** | 66 | 0.142 | 1.657 | 0.0975 |
| **syn-sem corr & MTLD_NMT** | 66 | 0.124 | 1.450 | 0.1470 |
| **syn-sem corr & Maas_NMT** | 66 | 0.038 | 0.442 | 0.6582 |
| **syn-sem corr & MAR_NMT** | 66 | -0.022 | -0.258 | 0.7962 |
| **syn-sem corr & Dugast_NMT** | 66 | -0.038 | -0.442 | 0.6582 |
| **syn-sem corr & MSTTR_NMT** | 66 | -0.075 | -0.876 | 0.3810 |
| **syn-sem corr & MATTR_NMT** | 66 | -0.079 | -0.924 | 0.3555 |
| **syn-sem corr & FKG_NMT** | 66 | -0.117 | -1.362 | 0.1731 |
| **syn-sem corr & ARI_NMT** | 66 | -0.128 | -1.492 | 0.1358 |
| **syn-sem corr & Summer_NMT** | 66 | -0.141 | -1.652 | 0.0985 |
| **syn-sem corr & CLI_NMT** | 66 | -0.147 | -1.724 | 0.0848 |
| **syn-sem corr & FOG_NMT** | 66 | -0.154 | -1.799 | 0.0720 |
| **syn-sem corr & SMOG_NMT** | 66 | -0.160 | -1.865 | 0.0623 |
| **syn-sem corr & Herdan_NMT** | 66 | **-0.247**** | -2.895 | **0.0038** |
| **syn-sem corr & Hapax_NMT** | 66 | **-0.319***** | -3.746 | **0.0002** |
| **syn-sem corr & TTR_NMT** | 66 | **-0.389***** | -4.564 | **0.0000** |

Note: 0.00 to 0.10 (0.00 to −0.10) – a trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – a low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – a moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – a high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – a very high positive (negative) measure of association; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

the other hand, we also achieved a negative low to a moderate measure of association between measures Herdan, Hapax, TTR and error rate for this category ($Gamma < -0.184/-0.305$, $p < 0.05/0.01/0.001$).

For the category of Lexical semantics (Table 6), we have achieved a moderate to high positive measure of association between measures SenC, MonoC, SC, RT, CharC, WordC, LC, LetC, UniWC, RTTR, CTTR, PolyC ($Gamma > 0.430/0.529$, $p < 0.001$) and occurrence of the error rate of neural MT within this category. Besides the positive measure of association, we also achieved a low to moderate negative, but also a statistically significant measure of association between Summer, Herdan, Hapax, TTR, and occurrence of MT errors in this category ($Gamma < -0.188/-0.338$, $p < 0.05/0.001$).

**Table 6.** Goodman and Kruskal's gamma – Lexical semantics

| | Valid N | Gamma | Z | p-value |
|---|---|---|---|---|
| **lexical sem & SenC_NMT** | 66 | **0.529*** ** | 6.161 | **0.0000** |
| **lexical sem & MonoC_NMT** | 66 | **0.497*** ** | 5.852 | **0.0000** |
| **lexical sem & SC_NMT** | 66 | **0.496*** ** | 5.850 | **0.0000** |
| **lexical sem & RT_NMT** | 66 | **0.489*** ** | 5.768 | **0.0000** |
| **lexical sem & CharC_NMT** | 66 | **0.489*** ** | 5.768 | **0.0000** |
| **lexical sem & WordC_NMT** | 66 | **0.489*** ** | 5.756 | **0.0000** |
| **lexical sem & LC_NMT** | 66 | **0.487*** ** | 5.746 | **0.0000** |
| **lexical sem & LetC_NMT** | 66 | **0.482*** ** | 5.688 | **0.0000** |
| **lexical sem & UniWC_NMT** | 66 | **0.482*** ** | 5.673 | **0.0000** |
| **lexical sem & RTTR_NMT** | 66 | **0.440*** ** | 5.186 | **0.0000** |
| **lexical sem & CTTR_NMT** | 66 | **0.440*** ** | 5.186 | **0.0000** |
| **lexical sem & PolyC_NMT** | 66 | **0.430*** ** | 5.043 | **0.0000** |
| **lexical semas & FRE_NMT** | 66 | 0.155 | 1.828 | 0.0676 |
| **lexical sem & HD-D_NMT** | 66 | 0.124 | 1.465 | 0.1429 |
| **lexical sem & MTLD_NMT** | 66 | 0.103 | 1.209 | 0.2267 |
| **lexical sem & Maas_NMT** | 66 | 0.090 | 1.064 | 0.2873 |
| **lexical sem & MAR_NMT** | 66 | -0.043 | -0.503 | 0.6152 |
| **lexical sem & MSTTR_NMT** | 66 | -0.073 | -0.855 | 0.3927 |
| **lexical sem & Dugast_NMT** | 66 | -0.090 | -1.064 | 0.2873 |
| **lexical sem & MATTR_NMT** | 66 | -0.123 | -1.454 | 0.1460 |
| **lexical sem & FKG_NMT** | 66 | -0.130 | -1.513 | 0.1304 |
| **lexical sem & SMOG_NMT** | 66 | -0.131 | -1.529 | 0.1263 |
| **lexical sem & ARI_NMT** | 66 | -0.141 | -1.652 | 0.0986 |
| **lexical sem & FOG_NMT** | 66 | -0.154 | -1.812 | 0.0700 |
| **lexical sem & CLI_NMT** | 66 | -0.157 | -1.843 | 0.0653 |
| **lexical sem & Summer_NMT** | 66 | **-0.188*** | -2.223 | **0.0262** |
| **lexical sem & Herdan_NMT** | 66 | **-0.291*** ** | -3.437 | **0.0006** |
| **lexical sem & Hapax_NMT** | 66 | **-0.338*** ** | -3.983 | **0.0001** |
| **lexical sem & TTR_NMT** | 66 | **-0.421*** ** | -4.963 | **0.0000** |

Note: 0.00 to 0.10 (0.00 to −0.10) – a trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – a low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – a moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – a high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – a very high positive (negative) measure of association; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

## 5 Conclusion

Our study brings two insights into the evaluation of MT quality and into the complexity of language as a system. We find that errors that arise in neural MT of journalistic texts associate the best with linguistic properties which are based on countability. This is mainly the sentence count property. The identified association is logical in terms of a larger number of sentences so that the frequency of error categories will also increase. Other properties that showed strong associations with error categories were based on the count or frequency, such as monosyllabic or polysyllabic words.

A notable finding is that not all examined measures of lexical diversity are associated with the frequency of errors in each category. We found only two measures out of all twelve examined measures, RTTR and CTTR, which show a moderate positive statistically significant association with error frequency in the categories of Syntactic-semantic correlativity, Compound/complex sentences, and Lexical semantics.

Another notable finding is that readability scores do not associate with error categories. We explain this by the fact that our corpus consists of journalistic texts, which themselves are easily readable and therefore we found no correlations. The second explanation is that the neural MT tool is much more fluent than its predecessor and does not produce a large number of MT errors that affect readability. An interesting finding is related only to the RTTR and CTTR measures, despite the fact that both are derived from the TTR measure, but the TTR measure is negatively moderately associated with the three error categories mentioned above, but this association is statistically significant. Here, further research is called for to investigate the reasons for this paradox. Besides the TTR measure, the lexical diversity measures Hapax and Herdan are also negatively associated with the occurrence of MT errors within these categories.

**Acknowledgement.** This work was supported by the Slovak Research and Development Agency under contract No. APVV-18-0473 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0821/21. This research was funded by the European Commission under the ERASMUS+ Programme 2021, KA2, grant number: 2021-1-SK01-KA220-HED-000032095 "Future IT Professionals EDucation in Artificial Intelligence".

## References

1. Koehn, P.: Machine Translation Overview. https://omniscien.com/machine-translation/
2. Lihua, Z.: The relationship between machine translation and human translation under the influence of artificial intelligence machine translation. Mob. Inf. Syst. **2022**, 1–8 (2022). https://doi.org/10.1155/2022/9121636
3. Wang, Y., Wang, Y., Wang, Y.: A new approach to machine translation and human translation in the Era of Artificial Intelligence. Overseas Engl. **7**, 179–180 (2021)
4. GreatContent: The 11 Best Machine (AI) Translation Tools in 2022. https://greatcontent.com/machine-ai-translation-tools/#introduction
5. Virino, V.: Artificial Intelligence applied to Machine Translation at FITUR 2021. https://blog.pangeanic.com/ai-applied-to-mt-at-fitur-2021
6. Popović, M.: Error classification and analysis for machine translation quality assessment. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. MTTA, vol. 1, pp. 129–158. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_7
7. Bulté, B., Housen, A.: Evaluating short-term changes in L2 complexity development. Círculo de lingüística aplicada a la comunicación **63** (2015). https://doi.org/10.5209/rev_CLAC.2015.v63.50169
8. De Bot, K.: A History of Applied Linguistics. Routledge, New York (2015)
9. Kovačević, E.: The relationship between lexical complexity measures and language learning beliefs <b/>. Jezikoslovlje **20**, 555–582 (2019). https://doi.org/10.29162/jez.2019.20
10. Brunato, D.: A study on linguistic complexity from a computational linguistics perspective. A corpus-based investigation of Italian bureaucratic texts (2015)
11. Kovačević, E.: The relationship between language learning beliefs and syntactic complexity. In: Gudurić, S., Radić-Bojanić, B. (eds.) Jezici i Kulture u Vremenu i Prostoru 6, pp. 455–464. University of Novi Sad, Novi Sad (2017)
12. Temnikova, I.: Text Complexity and Text Simplification in the Crisis Management Domain (2012)

13. Kalantari, R., Gholami, J.: Lexical complexity development from dynamic systems theory perspective: lexical density, diversity, and sophistication. Int. J. Instr. **10**, 1–18 (2017). https://doi.org/10.12973/iji.2017.1041a
14. Mat Daud, N., Hassan, H., El-Tingari, S., Abdul Aziz, N.: Web-based Arabic text readability index. In: 8th International Technology, Education and Development Conference, Valencia, Spain, pp. 1574–1581. IATED Academy (2014)
15. Imperial, J.M., Ong, E.: Application of Lexical Features Towards Improvement of Filipino Readability Identification of Children's Literature. arXiv 7 (2021)
16. Jurafsky, D., Martin, J.: Speech and Language Processing (2020)
17. Lu, X.: Automatic analysis of syntactic complexity in second language writing. Int. J. Corpus Linguist. **15**, 474–496 (2010)
18. Doherty, S.: Investigating the Effects of Controlled Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach (2012)
19. ReadabilityFormulas: How Do I Decide Which Readability Formula Or Formulas To Use On My Document? https://readabilityformulas.com/search/pages/Readability_Formulas/
20. Benko, L., Benkova, L., Munkova, D., Munk, M., Shulzenko, D.: Error classification using automatic measures based on n-grams and edit distance. In: Guarda, T., Portela, F., Augusto, M.F. (eds.) ARTIIS 2022. CCIS, vol. 1675, pp. 345–356. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20319-0_26
21. Munkova, D., Munk, M., Benko, Ľ., Stastny, J.: MT evaluation in the context of language complexity. Complexity **2021**, 1–15 (2021).https://doi.org/10.1155/2021/2806108
22. Munkova, D., Munk, M., Welnitzova, K., Jakabovicova, J.: Product and process analysis of machine translation into the inflectional language. Sage Open **11**, 215824402110545 (2021). https://doi.org/10.1177/21582440211054501
23. Munkova, D., Munk, M., Benko, Ľ, Hajek, P.: The role of automated evaluation techniques in online professional translator training. PeerJ Comput Sci. **7**, e706 (2021). https://doi.org/10.7717/peerj-cs.706
24. Kapusta, J., Benko, Ľ, Munkova, D., Munk, M.: Analysis of edit operations for post-editing systems. Int. J. Comput. Intell. Syst. **14**, 197 (2021). https://doi.org/10.1007/s44196-021-00048-3
25. Vaňko, J.: Kategoriálny rámec pre analýzu chýb strojového prekladu. In: Munkova, D., Vaňko, J. (eds.) Mýliť sa je ľudské (ale aj strojové), pp. 83–100. UKF v Nitre, Nitra (2017)
26. Zhang, Y., Lin, N., Jiang, S.: A study on syntactic complexity and text readability of ASEAN English news. In: 2019 International Conference on Asian Language Processing (IALP), pp. 313–318. IEEE (2019). https://doi.org/10.1109/IALP48816.2019.9037695

PRÍLOHA J: BENKO, Ľ.UBOMÍR, LUCIA BENKOVA, DASA MUNKOVA, MICHAL MUNK A DANYLO SHULZENKO, 2022. ERROR CLASSIFICATION USING AUTOMATIC MEASURES BASED ON N-GRAMS AND EDIT DISTANCE. IN: *ADVANCED RESEARCH IN TECHNOLOGIES, INFORMATION, INNOVATION AND SUSTAINABILITY. ARTIIS 2022*. SPRINGER, CHAM, S. 345–356. DOI:10.1007/978-3-031-20319-0_26 (**WEB OF SCIENCE, SCOPUS**) [WOS: 0, SCOPUS: 0]

# Error Classification Using Automatic Measures Based on n-grams and Edit Distance

L'ubomír Benko(✉), Lucia Benkova, Dasa Munkova, Michal Munk, and Danylo Shulzenko

Constantine the Philosopher University in Nitra, 949 01 Nitra, Slovakia
lbenko@ukf.sk

**Abstract.** Machine translation (MT) evaluation plays an important task in the translation industry. The main issue in evaluating the MT quality is an unclear definition of translation quality. Several methods and techniques for measuring MT quality have been designed. Our study aims at interconnecting manual error classification with automatic metrics of MT evaluation. We attempt to determine the degrees of association between automatic MT metrics and error classes from English into inflectional Slovak. We created a corpus, which consists of English journalistic texts, taken from the British online newspaper The Guardian and their human and machine translations. The MT outputs, produced by Google translate, were manually annotated by three professionals using a categorical framework for error analysis and evaluated using reference proximity through the metrics of automated MT evaluation. The results showed that not all examined automatic metrics based on n-grams or edit distance should be implemented into a model for determining the MT quality. When determining the quality of machine translation in respect to syntactic-semantic correlativeness, it is sufficient to consider only the Recall, BLEU-4 or F-measure, ROUGE-L and NIST (based on n-grams) and the metric CharacTER, which is based on edit distance.

**Keywords:** Machine translation · Automatic metrics · Error classification

## 1 Introduction

Machine translation (MT) is one of the most popular natural language processing applications. It is the automatic translation of text from one natural language into another natural language. The quality of the translation, its accuracy or, on the other hand, its error rate, plays a key role in interpersonal communication. Evaluating the MT quality is essential for improving MT systems, as it presents a strong indicator of the correlation between an MT output and its corresponding human translation [1:2]. The biggest issue in evaluating MT quality is an unclear definition of translation quality together with its criteria and measures for translation quality. There are no explicit criteria for "good translation" [1:2]. For this reason, several methods and techniques for measuring translation quality have been designed. In general, they can be divided into a manual approach to MT quality assessment and an automatic approach to MT quality assessment [2]. Both

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
T. Guarda et al. (Eds.): ARTIIS 2022, CCIS 1675, pp. 345–356, 2022.
https://doi.org/10.1007/978-3-031-20319-0_26

approaches have their advantages, but also their disadvantages. The manual evaluation assesses the translation more likely as a whole, i.e. it assesses cohesiveness and coherence of the translation, but this evaluation is very subjective and time-consuming [3]. The standard criteria used within manual evaluation are fluency (grammatical correctness), adequacy (preservation of the meaning) or usability. In addition to standard criteria of MT quality, human evaluators also use task oriented methods for quality evaluation such as post-editing or error analysis, and/or error classification. Error classification (i.e. identification and classification of errors occurring in a machine translated text) is not only a time-consuming, but also a resource intensive task. It provides a distribution of errors over the defined error classes, but it suffers from low consistency of human evaluators [4].

On the other hand, automatic evaluation brings speed, objectivity, and reusability to the measurement. The objective of automatic MT evaluation is to calculate the numerical score (between 0–1), which represents the quality of MT output and/or the performance of the MT system. This evaluation is less reliable compared to manual evaluation, as the evaluation lies in a lexical comparison of two strings - MT output with reference/human translation - in a target language. Within automatic MT evaluation, there are two main approaches for evaluating quality (MT output) automatically - reference proximity and performance-based techniques [5]. In this study we focus on reference proximity techniques, which are based on statistical principles (lexical similarities) or linguistic features [6]. They compare translation to the human reference in that way, that the closer MT output is to the reference the better the quality is considered to be. Distance between MT output and reference translation is calculated automatically (e.g. WER, TER or CharacTER) or their overlap (e.g. BLEU, F-measure, METEOR or NIST).

Our study aims at interconnecting manual error classification with automatic metrics of MT evaluation. Through error analysis, we point out the degree of association between automatic MT metrics and error classes from English into inflectional Slovak.

The structure of the paper is as follows. The second section introduces automatic MT metrics based on reference proximity. The third section focuses on the methodology of experiment with assumptions, methods, and dataset. The fourth section describes the results of the experiment. Subsequently, the last two sections discuss the obtained results and draw conclusions.

## 2 Automatic MT Metrics Based on Reference Proximity

Automatic MT metrics provide quantified scores of overall translation quality. They do not require high human effort and they can be used quite easily to compare the performance of two or more MT systems. Therefore, they are not only popular, but also in great demand. Based on their results, MT systems are subsequently developed or optimized.

In this study, we focus on automatic MT metrics that compare MT output with reference based on exact lexical matches between MT words, and/or phrases and reference.

Lexical similarity is a measure of the degree to which the word or phrase of MT output is similar to the corresponding word or phrase in reference. A lexical similarity

of 1 means a total overlap between MT output and reference, whereas 0 means there is no match. *Precision* and *recall* belong to the basic MT metrics [7], where precision is the proportion of words in MT output/hypothesis ($Y$) that are present in the reference ($X$), and recall is the proportion of words in reference ($X$) that are present in the hypothesis ($Y$). F-measure is a harmonic mean of precision and recall:

$$P = precision(Y|X) = \frac{|X \cap Y|}{|Y|}, \tag{1}$$

$$R = recall(Y|X) = \frac{|X \cap Y|}{|X|}, \tag{2}$$

$$F1 = \frac{2PR}{P+R}. \tag{3}$$

*Bilingual Evaluation Understudy* (BLEU) is a standard automatic measure, which is a precision-oriented metric. *BLEU-n* [8] is a geometric mean of n-gram precisions with a *brevity penalty* (BP), i.e. penalty to prevent very short sentences:

$$BLEU(n) = exp \sum_{n=1}^{N} w_n \log p_n \times BP \tag{4}$$

where $w_n$ is weights for different $p_n$,

$$BP = \begin{cases} 1, if\ h > r \\ e^{1-\frac{r}{h}}, if\ h \leq r \end{cases} \tag{5}$$

where $r$ is a reference of a hypothesis $h$.

The *BLEU* represents two features of translation quality- *adequacy* and *fluency* by calculating words or lexical *precisions* [9]. The *BLEU* score has several variations, depending on the number of words in the reference used to compute the brevity penalty. The IBM version of *BLEU* uses the average value of the length of the reference. The *NIST* version of *BLEU* uses the shortest references to compute the brevity penalty. To not get confused, there exists the *NIST* metric which is not equal to the *NIST* version of *BLEU*, using the arithmetic mean of the n-grams counts instead of the geometric mean, which is used in the ordinary *BLEU-n* metric.

*Measure for Evaluation of Translation with Explicit Ordering* (METEOR) is a recall-oriented measure. It calculates not only *precision* (like *BLEU*), but also *recall*. Both are combined with a preference to *recall* when calculating the harmonic mean. It is based on a combination of unigram-precision and unigram-recall, and on direct capture of how well-ordered the matched words/phrases in MT outputs are in respect to the reference [10]:

$$METEOR = \frac{10PR}{R+9P}(1 - BP), \tag{6}$$

where the unigram-recall and unigram precision are given by $P$ and $R$, and

$$BP = 0.5\left(\frac{\#chunks}{\#unigrams\_matched}\right), \tag{7}$$

where chunk (a group of matched unigrams between MT output and reference) is a minimum number of words required to match unigrams in the MT output with corresponding references [11].

*NIST* [12] is a metric based on *BLEU*. It was designed to improve *BLEU* by rewarding the translation of infrequently used words, i.e. it uses heavier weights for rarer words [11]. The *BLEU* metric calculates n-gram precision with equal weight to each one, but the *NIST* metric calculates how much information is preserved in a particular n-gram.

*Character n-gram F-measure* (ChrF) is a language- and tokenization-independent metric, which correlates well with human judgments on the system- and segment-level [13]:

$$chrF\beta = (1 + \beta^2)(\frac{chrP \cdot chrR}{\beta^2 chrP + chrR}), \tag{8}$$

where the character n-gram *precision* and *recall* are given by *chrP* (percentage of n-grams in the hypothesis) and *chrR* (percentage of n-grams in the reference). β is a parameter which assigns β times more important to recall than to precision. For instance, if $\beta = 1$, both (precision and recall) have the same weight and if $\beta = 2$, recall is two times more important than precision and vice versa, if $\beta = ½$, precision is two times more important than recall [4, 14].

*Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) counts the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans [15]. It includes several automatic evaluation measures that determine the similarity between summaries. In this study, we used *ROUGE-N* and *ROUGE-L*. *ROUGE-N* is an n-gram recall between a hypothesis summary and a set of reference summaries. *ROUGE-L* is the longest common subsequence F-measure and counts only in sequence co-occurrences.

The second approach to measure the lexical similarity of two words, and/or phrases is to calculate the minimum edit distance to transform an MT output/hypothesis into a reference (to transform one string into another) through edit operations. Sets of string operations depend on the type of edit distance. One of the simplest sets of edit operations is defined by Levenshtein [16:107-111]:

- Insertion of a character. If $a = uv$, then insert the character $x$ produces $uxv$. This can also be denoted $\varepsilon \rightarrow x$, using ε to denote the empty string.
- Deletion of a character $x$ changes $uxv$ to $uv(x \rightarrow \varepsilon)$.
- Substitution of a character $x$ for a character $y \neq x$ changes $uxv$ to $uyv(x \rightarrow y)$.

*Word Error Rate* (WER) counts the Levenshtein distance between the hypothesis and reference, without allowing the words reordering [17]:

$$WER\,(h, r) = \frac{min\#(I+D+S)}{|r|}), \tag{9}$$

where $r$ is a reference of a hypothesis $h$, $I$- insertion, $D$ - deletion, and $S$ - substitution.

The minimum number of edit operations (insertions, substitutions, and deletions of the words necessary to transform the hypothesis/MT output into the reference) is divided by the number of words in the reference [7].

*Translation Edit Rate* (TER) is defined as the minimum number of edit operations required to change a hypothesis/machine translation to an exact match with the reference [18]:

$$TER\,(h, r) = \frac{min\#(I + D + S + shift)}{|r|}), \tag{10}$$

where $r$ is a reference of a hypothesis/machine translation $h$, $I$ - insertion, $D$ - deletion, $S$ - substitution and shift (number of changes in word order).

*CharacTER* [19] is an edit distance metric, which is based on character-level and calculates the character-level edit distance while performing the shift edit on word level. Like *TER*, *CharacTER* also calculates the minimum number of character edit operations required to change a hypothesis to the exact match of the reference, divided by the length of the hypothesis:

$$CharacTER\,(h, r) = \frac{min\#(shift + I + D + S)}{|h|} \tag{11}$$

where $r$ is a reference of a hypothesis $h$, $I$- insertion, $D$ - deletion, $S$ - substitution and shift (number of changes in word order).

## 3 Experiment

Our objective is to investigate the relationship between automatic MT metrics and a distribution of errors over the defined error classes. We attempt to determine which of the examined metrics (based on lexical similarity and edit distance) associate the best with individual error classes of a categorical framework for error analysis [20:100]. The examined texts (1903 sentences/3271 segments) were of the journalistic style, taken from the British online newspaper The Guardian. In 2021, the texts were translated by the freely available Neural Google Translate (NGT) engine and subsequently manually annotated by three professionals. The annotation was performed according to the categorical framework for error analysis for translation into Slovak [20:100]. The framework consists of five error classes (categories):

1. Predication,
2. Modal and communication sentence framework,
3. Syntactic-semantic correlativeness,
4. Compound/complex sentences,
5. Lexical semantics.

In this study, we focus only on one particular category - Syntactic-semantic correlativeness - characterizing inflectional languages like Slovak. This category corresponds to the category of *language*, and/or *fluency*, both belonging to the core of harmonized DQF-MQM Error typology [21].

The category of Syntactic-semantic correlativeness is more deeply divided into subcategories: Nominal morphosyntax, Pronominal morphosyntax, Numeral morphosyntax, Verbal morphosyntax, Word order, Other morphosyntactic phenomena, and Others.

### 3.1 Assumption

Given that the metrics of automatic evaluation are constantly developing, we have been encouraged to examine which of the MT metrics (based on lexical similarity or edit distance) used so far are appropriate and/or best capture the errors that occurred in machine translation into the inflectional language. Besides free word order, inflectional languages are also characterized by inflection and declension. Both linguistic features are particularly captured in the category of Syntactic-semantic correlativeness.

We assume that:

*Automatic MT metrics based on lexical similarity (precision, recall, F-measure, ChrF, NIST, ROUGE, METEOR, and BLEU) associate better with the occurrence of errors in a given category than automatic MT metrics based on edit distance (CharacTER, WER, and TER).*

To prove our assumption, we used Goodman and Kruskal's gamma. Gamma represents the degree of association between two variables, i.e. the probability of whether two variables are in the same or opposite order.

### 3.2 Dataset

The dataset consists of machine-translated journalistic texts from English (STs) to Slovak (NMTs). The readability and lexico-grammatical features of our corpus are as follows (Table 1):

**Table 1.** Dataset composition

| Feature type | Feature name | NMTs_SK | STs_EN |
|---|---|---|---|
| Readability | Average sentence length | 17.12034 | 19.26274 |
| | Average word length | 5.696361 | 4.996122 |
| | #short sentences ($n < 10$) | 469 | 395 |
| | #long sentences ($n \geq 10$) | 1434 | 1508 |
| Lexico-grammatical | Frequency of proper nouns | 1501 | 3078 |
| | Frequency of nouns | 10070 | 8627 |
| | Frequency of adjectives | 3324 | 2968 |
| | Frequency of adverbs | 933 | 1667 |
| | Frequency of verbs | 5198 | 6473 |
| | Frequency of pronominals | 2371 | 2124 |
| | Frequency of particles | 592 | 149 |
| | Frequency of foreign words | 841 | 0 |
| | Frequency of interjections | 3 | 3 |
| | Frequency of numerals | 617 | 777 |
| | Frequency of prepositions & conjunctions | 6028 | 6697 |
| | Frequency of interpunction | 5958 | 3547 |

### 3.3 Methods

For the metrics as *BLEU*, *NIST*, *METEOR*, and *ChrF* Python Natural Language Toolkit (NLTK) library was used.

```
from nltk.translate.bleu_score import sentence_bleu, sen-
tence_nist, meteor_score, chrf_score
bleu_scores_1.append(sentence_bleu([ref], hyp,
weights=(1,0,0,0)))
bleu_scores_2.append(sentence_bleu([ref], hyp,
weights=(0,1,0,0)))
bleu_scores_3.append(sentence_bleu([ref], hyp,
weights=(0,0,1,0)))
bleu_scores_4.append(sentence_bleu([ref], hyp,
weights=(0,0,0,1)))
nist_scores.append(sentence_nist([ref], hyp, n=1))
meteor_scores.append(meteor_score([ref], hyp))
chrf_scores.append(chrf_score.sentence_chrf(ref, hyp))
```

For *ROUGE*, *TER*, and *WER* open-source libraries were used.

```
import jiwer
import pyter
from rouge_metric import PyRouge
wer_scores.append(jiwer.wer (ref, hyp))
rogue_scores.append(rouge.evaluate_tokenized([hyp], [ref]))
ter_scores.append(pyter.ter(hyp,ref))
```

*Precision*, *recall*, *F-measure* were implemented separately from the others. The *CharacTER* was implemented as an edit distance function.

## 4 Results

After manual error classification, we identified 1851 errors in the category of syntactic-semantic correlativeness, of which 394 errors were identified in nominal morphosyntax, 88 errors in pronominal morphosyntax, 4 errors in numeral morphosyntax, 276 errors in verbal morphosyntax, 453 errors in word order, 617 errors in other morphosyntactic phenomena, and 19 errors in the subcategory others.

Based on a Cochran Q test ($N = 3271$, $Q = 1371.86$, $df = 6$, $p < 0.001$) we showed that there are statistically significant differences between the individual subcategories. These results were also proved by *Kendall's Coeff. of concordance* (0.07), where were identified a small agreement, and/or almost no agreement between the examined subcategories.

Based on the results of multiple comparisons, we showed statistically significant differences between Other morphosyntactic phenomena/Word order/Pronominal morphosyntax and other subcategories and, conversely, there were no statistically significant differences between Numeral morphosyntax and Others, or between Nominal morphosyntax and Word order (Table 2).

**Table 2.** Multiple comparisons: Homogenous groups, $p < 0.05$

| | Incidence | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Numeral morphosyntax | 0.12% | **** | | | | |
| Others | 0.58% | **** | | | | |
| Pronominal morphosyntax | 2.69% | | | **** | | |
| Verbal morphosyntax | 8.44% | | | | **** | |
| Nominal morphosyntax | 12.05% | | **** | | | |
| Word order | 13.85% | | **** | | | |
| Other morphosyntactic phenomena | 18.86% | | | | | **** |

Using Goodman and Kruskal's gamma, we determined the rank associations between the individual subcategories and the automatic MT metrics based on lexical similarity or edit distance (Tables 3 and 4).

**Table 3.** Nominal morphosyntax - rank association

| Error category & automatic metrics | Valid *N* | *Gamma* | *Z* | *p*-value |
|---|---|---|---|---|
| Nominal morphosyntax & BLEU-4 | 3271 | **0.08**** | 2.9868 | **0.0028** |
| Nominal morphosyntax & NIST | 3271 | **0.05*** | 2.0931 | **0.0363** |
| Nominal morphosyntax & BLEU-3 | 3271 | 0.05 | 1.9018 | 0.0572 |
| Nominal morphosyntax & BLEU-2 | 3271 | 0.04 | 1.7635 | 0.0778 |
| Nominal morphosyntax & precision | 3271 | 0.04 | 1.5416 | 0.1232 |
| Nominal morphosyntax & ChrF | 3271 | 0.04 | 1.4759 | 0.1400 |
| Nominal morphosyntax & F-measure | 3271 | 0.04 | 1.4083 | 0.1591 |
| Nominal morphosyntax & METEOR | 3271 | 0.03 | 1.2716 | 0.2035 |
| Nominal morphosyntax & recall | 3271 | 0.03 | 1.1908 | 0.2337 |
| Nominal morphosyntax & BLEU-1 | 3271 | 0.03 | 1.1790 | 0.2384 |
| Nominal morphosyntax & WER | 3271 | −0.03 | −1.1804 | 0.2379 |
| Nominal morphosyntax & TER | 3271 | −0.03 | −1.2338 | 0.2173 |
| Nominal morphosyntax & ROUGE1 | 3271 | −0.04 | −1.7095 | 0.0874 |

(*continued*)

**Table 3.** (*continued*)

| Error category & automatic metrics | Valid *N* | *Gamma* | *Z* | *p*-value |
|---|---|---|---|---|
| Nominal morphosyntax & ROUGE2 | 3271 | −0.04 | −1.7095 | 0.0874 |
| Nominal morphosyntax & ROUGE-L | 3271 | **−0.07**** | −2.8404 | **0.0045** |
| Nominal morphosyntax & CharacTER | 3271 | **−0.09***** | −3.6744 | **0.0002** |

Note: 0.00 to 0.10 (0.00 to −0.10) – trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – very high positive (negative) measure of association; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The subcategory Nominal morphosyntax (Table 3) is partially identified by the metrics *BLEU-4* and *NIST*, where a trivial, but statistically significant degree of positive association was achieved (*Gamma* $< 0.1$, $p < 0.01/0.05$), similarly, in the case of the metrics *ROUGE-L* and *CharacTER*, there were achieved statistically significant, but trivial degrees of a negative association (*Gamma* $< -0.1$, $p < 0.01/0.001$).

The automatic metrics *BLEU-4* and *NIST*, both based on precision, associated best with MT errors in the subcategory of nominal morphosyntax. On the other hand, in terms of edit distance, the metric *CharacTER* associated best with this subcategory.

In the case of the subcategories of pronominal morphosyntax and other morphosyntactic phenomena, there were achieved only trivial, statistically insignificant degrees of association between automatic MT metrics and the given subcategories (*Gamma* $\approx$ 0.00).

In the case of the subcategory of numeral morphosyntax, the degree of association oscillates between a low (0.10–0.30 and/or −0.10–−0.30) and a very high (0.70–1.00 and/or −0.70–−1.00) either positive or negative degrees of association (Table 4).

**Table 4.** Numeral morphosyntax - rank association

| Error category & automatic metrics | Valid *N* | *Gamma* | *Z* | *p*-value |
|---|---|---|---|---|
| Numeral morphosyntax & CharacTER | 3271 | 0.32 | 1.3681 | 0.1713 |
| Numeral morphosyntax & TER | 3271 | 0.30 | 1.2536 | 0.2100 |
| Numeral morphosyntax & WER | 3271 | 0.30 | 1.2519 | 0.2106 |
| Numeral morphosyntax & ROUGE-L | 3271 | 0.30 | 1.2566 | 0.2089 |
| Numeral morphosyntax & ROUGE1 | 3271 | 0.22 | 0.9428 | 0.3458 |
| Numeral morphosyntax & ROUGE2 | 3271 | 0.22 | 0.9428 | 0.3458 |
| Numeral morphosyntax & BLEU-3 | 3271 | −0.26 | −1.0836 | 0.2786 |
| Numeral morphosyntax & ChrF | 3271 | −0.40 | −1.7115 | 0.0870 |
| Numeral morphosyntax & BLEU-2 | 3271 | **−0.49*** | −2.0522 | **0.0401** |

(*continued*)

**Table 4.** (*continued*)

| Error category & automatic metrics | Valid *N* | *Gamma* | *Z* | *p*-value |
|---|---|---|---|---|
| Numeral morphosyntax & METEOR | 3271 | **−0.57*** | −2.4312 | **0.0151** |
| Numeral morphosyntax & BLEU-4 | 3271 | **−0.62*** | −2.2520 | **0.0243** |
| Numeral morphosyntax & NIST | 3271 | **−0.63**** | −2.6330 | **0.0085** |
| Numeral morphosyntax & BLEU-1 | 3271 | **−0.65**** | −2.7494 | **0.0060** |
| Numeral morphosyntax & precision | 3271 | **−0.70**** | −2.9358 | **0.0033** |
| Numeral morphosyntax & F-measure | 3271 | **−0.74**** | −3.1129 | **0.0019** |
| Numeral morphosyntax & recall | 3271 | **−0.77**** | −3.2238 | **0.0013** |

Note: 0.00–0.10 (0.00 to −0.10) – trivial positive (negative) measure of association; 0.10–0.30 (−0.10 to −0.30) – low positive (negative) measure of association; 0.30–0.50 (−0.30 to −0.50) – moderate positive (negative) measure of association; 0.50–0.70 (−0.50 to −0.70) – high positive (negative) measure of association; 0.70–1.00 (−0.70 to −1.00) – very high positive (negative) measure of association; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

In the case of verbal morphosyntax, we achieved similar results as for the subcategory of nominal morphosyntax, i.e. only for *ROUGE-L*, *ROUGE1*, *ROUGE2*, and *CharacTER* were achieved a statistically significant, but trivial degrees of negative association ($Gamma > -0.1$, $p < 0.01$).

We obtained slightly better results for the subcategories Word order and Others, but still with a low positive, and/or negative degree of association. Only for metrics *Recall*, *Precision*, *F-measure*, *BLEU-3*, and *BLEU-4* ($Gamma \geq -0.1$, $p < 0.001/0.01/0.05$) were achieved a statistically significant negative degree of association, in the case of the category of Word order. For the category Others, only the *ChrF* metric has achieved a low, but statistically significant positive degree of association ($Gamma = 0.23$, $p = 0.0345$).

## 5 Discussion

Metrics like *Precision*, *Recall*, *F-measure*, *BLEU-n*, *NIST*, *METEOR*, *WER*, *TER*, and *ROUGE* are more reliable and have a higher association with linguistic errors within these subcategories: word order, nominal morphosyntax, and numeral morphosyntax. Although they have high associations, the *CharacTER* metric (based on edit distance) has the highest statistical significance among them in nominal morphosyntax. The *ChrF* metric compared to other metrics, which are based on n-grams, showed a poor performance and is not suitable for this linguistic subcategory (error class).

In the case of numeral morphosyntax, the metrics based on n-gram outperform the metrics based on edit distance in all aspects, i.e. in terms of a degree of association with linguistic category, they achieved a higher level of statistical significance ($p < 0.01$). Linguistic categories like verbal morphosyntax, other morphosyntactic phenomena, pronominal morphosyntax, and others do not show the clear associations to automatic metrics (based on n-grams or edit distance) due to approximately the same low degree of association and a low level of statistical significance ($p < 0.05$).

## 6 Conclusions

The results of our study showed that not all automatic metrics based on n-grams or edit distance should be implemented into a model for determining the MT quality of journalistic texts translated from English into inflectional Slovak. When determining the quality of machine translation in respect to syntactic-semantic correlativeness, it is sufficient to consider only *Recall*, *BLEU-4* or the *F-measure*, *ROUGE-L* and *NIST* (based on n-grams) and the metric *CharacTER*, which is based on edit distance. The results can be also applicable to other inflectional languages.

The results of our study also showed certain pitfalls and limitations that open up space for further research. The first question that arises here is whether automatic MT metrics based on statistical principles (lexical similarity) are suitable for determining the quality of machine translation into the inflectional Slovak language? Or rather to accept into the model automatic MT metrics based on linguistic features? On the other hand, whether the categorical framework used for error analysis is suitable (for translation of journalistic texts from English into Slovak), as the strong associations between automatic MT metrics and the error category under study were not proved.

We consider the size of the corpus to be the main limitation of our study along with the limitation to only one style and genre. In future work, we want to focus on the expansion of our corpus in terms of size and style.

**Acknowledgements.** This work was supported by the Slovak Research and Development Agency under contract No. APVV-18-0473 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0821/21.

## References

1. Chow, J.: Lost in translation: fidelity-focused machine translation evaluation (2019). https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1819-ug-projects/ChowJ-Lost-in-translation-fidelity-focused-machine-translation-evaluation.pdf
2. Castilho, S., Doherty, S., Gaspari, F., Moorkens, J.: Approaches to human and machine translation quality assessment. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. MTTA, vol. 1, pp. 9–38. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_2
3. Sepesy Maučec, M., Donaj, G.: Machine translation and the evaluation of its quality. In: Recent Trends in Computational Intelligence. IntechOpen (2020). https://doi.org/10.5772/intechopen.89063
4. Popović, M.: Error classification and analysis for machine translation quality assessment. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Machine Translation: Technologies and Applications. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_7
5. Babych, B.: Automated MT evaluation metrics and their limitations. In: evista Tradumàtica: Tecnologies De La Traducció, 12 (2014). https://doi.org/10.5565/rev/tradumatica.70
6. Munk, M., Munková, D., Benko, Ľ: Identification of relevant and redundant automatic metrics for MT evaluation. In: Sombattheera, C., Stolzenburg, F., Lin, F., Nayak, A. (eds.) MIWAI 2016. LNCS (LNAI), vol. 10053, pp. 141–152. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49397-8_12

7. Munk, M., Munkova, D.: Detecting errors in machine translation using residuals and metrics of automatic evaluation. J. Intell. Fuzzy Syst. **34**, 3211–3223 (2018). https://doi.org/10.3233/JIFS-169504
8. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia (2002)
9. Munk, M., Munkova, D., Benko, L.: Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics. J. Intell. Fuzzy Syst. **34**, 3225–3233 (2018). https://doi.org/10.3233/JIFS-169505
10. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (ACL-05), pp. 65–72. Michigan (2005)
11. Wołk, K., Koržinek, D.: Comparison and Adaptation of Automatic Evaluation Metrics for Quality Assessment of Re-Speaking (2016)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, pp. 138–145 (2002)
13. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Stroudsburg, PA, USA (2015). https://doi.org/10.18653/v1/W15-3049
14. Popović, M.: chrF deconstructed: beta parameters and n-gram weights. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 499–504. Association for Computational Linguistics, Stroudsburg, PA, USA (2016). https://doi.org/10.18653/v1/W16-2341
15. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
16. Jurafsky, D., Martin, J.: Speech and Language Processing (2020)
17. Nießen, S., Och, F.J., Leusch, G., Ney, H.: An evaluation tool for machine translation: Fast evaluation for MT research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000), pp. 39–45 (2000)
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
19. Wang, W., Peter, J.-T., Rosendahl, H., Ney, H.: CharacTer: translation edit rate on character level. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp. 505–510. Association for Computational Linguistics, Stroudsburg, PA, USA (2016). https://doi.org/10.18653/v1/W16-2342
20. Vaňko, J.: Kategoriálny rámec pre analýzu chýb strojového prekladu. In: Munkova, D. and Vaňko, J. (eds.) Mýliť sa je ľudské (ale aj strojové), pp. 83–100. UKF v Nitre, Nitra (2017)
21. Lommel, A.: Metrics for translation quality assessment: a case for standardising error typologies. In: Moorkens, J., Castilho, S., Gaspari, F., Doherty, S. (eds.) Translation Quality Assessment. MTTA, vol. 1, pp. 109–127. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91241-7_6

PRÍLOHA K: BENKO, ĽUBOMÍR, DASA MUNKOVA, MICHAL MUNK, LUCIA BENKOVA A PETR HAJEK, 2024A. THE USE OF RESIDUAL ANALYSIS TO IMPROVE THE ERROR RATE ACCURACY OF MACHINE TRANSLATION. SCIENTIFIC REPORTS (*V RECENZNOM KONANÍ OD 2023, 3. KOLO*) (**WEB OF SCIENCE, 2022IF: 4.6, Q2**)

# The use of residual analysis to improve the error rate accuracy of machine translation

**Ľubomír Benko[1*], Dasa Munkova[1], Michal Munk[1,2], Lucia Benkova[1], Petr Hajek[2]**

[1] Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, SK 949 01, Nitra, Slovakia

[2] Science and Research Centre, University of Pardubice, Studentská 84, CZ 532 10, Pardubice, Czech Republic

[*] lbenko@ukf.sk

## ABSTRACT

The aim of the study is to compare two different approaches to machine translation – statistical and neural – using automatic MT metrics of error rate and residuals. We examined four available online MT systems (statistical Google Translate, neural Google Translate, and two European commission's MT tools – statistical mt@ec and neural eTranslation) through their products (MT outputs). We propose using residual analysis to improve the accuracy of machine translation error rate. Residuals represent a new approach to comparing the quality of statistical and neural MT outputs. The study provides new insights into evaluating machine translation quality from English and German into Slovak through automatic error rate metrics. In the category of prediction and syntactic-semantic correlativeness, statistical MT showed a significantly higher error rate than neural MT. Conversely, in the category of lexical semantics, neural MT showed a significantly higher error rate than statistical MT. The results indicate that relying solely on the reference when determining MT quality is insufficient. However, when combined with residuals, it offers a more objective view of MT quality and facilitates the comparison of statistical MT and neural MT.

## Introduction

Although relying on human translation offers more accuracy and fluency, human translation is of limited efficiency and it is challenging for it to meet the needs of long text translation[1]. This limitation stimulates the search for new approaches to translation. One such approach is the implementation of intelligent algorithms within machine translation (MT) system. Various algorithms address the issues of MT system such as RNN encoding–decoding in existing log-linear SMT, transfer learning method, self-attention mechanism, unsupervised training algorithm, the adversarial augmentation method, reinforcement learning, neural MT (LSTM and transformer), hybrid (neural MT + statistical MT), rule-based MT, phrase-based MT, and others[2]. Currently, machine translation employs deep neural network (NN) learning, which initially learns rules and then automatically produces translations. This approach has yielded very good results for tasks with sufficiently labelled data for learning. However, if there is little tagged data, machine translation produces poor performance[3]. The primary obstacle for market-oriented neural MT systems or applications lies in its weak translation quality that fails to meet users' needs[4]. MT evaluation is a fundamental step in improving the performance of MT systems. The continuous enhancement of the performance of current neural MT systems is closely tied to research on evaluating the quality of MT output based on sentence comparison[4]. This comparison involves two inseparable aspects – qualitative/human and quantitative/automatic evaluation. The first serves as the foundation and guiding principle for the second, while the latter represents the digital outcome of the former.

Two main approaches exist for evaluating MT systems – human/manual and automatic evaluation. Blur criteria and scales for manual translation quality, along with different human evaluator sensitivity to translation errors may result in the judge subjectivity, which can be reflected in the poor consistency and instability of the evaluation results[5]. Human evaluation is an effective way to assess translation quality, but is challenging to find reliable bilingual annotators[6]. In addition to poor consistency and subjectivity, manual evaluation is both financially and time-consuming; however, unlike automatic evaluation, it does not require a reference translation. The advantages of automatic evaluation lie in its objectivity, consistency, stability, speed, reusability, and language independence. It is cost-effective and easy to use for comparing multiply systems, but at the expense of quality [6]. Furthermore, automatic evaluation requires reference – human translation (gold standard) – since the evaluation is based on comparing MT output with reference translation. Automatic MT evaluation metrics compare overlapping words between MT output and reference (e.g. [7–9] and others). Automatic metrics of MT evaluation only capture lexical similarity and correctly measure neither semantic and grammatical diversity, nor syntactic structures[6,10].

The automatic evaluation metrics applied in MT evaluation can be divided into untrained - lexical metrics and trained - supervised and unsupervised metrics[6]. The untrained metrics measure lexical similarity/distance (overlap with a reference) using a mathematical formula or heuristic methods at the word-level (N-gram-based metrics such as BLEU, METEOR or NIST; or edit distance-based metrics such as TER or WER) or at the character-level (such as characTER). Within the trained metrics, we can distinguish embedding-based metrics from supervised metrics. The embedding-based metric measures lexical similarity using machine learning techniques or deep learning algorithms (automatic metrics such as MEANT or BEER). The supervised metric trains regression models using labelled data, annotated by humans such as the COMET metric[6].

The BLEU (Bilingual Evaluation Understudy) evaluation index is the standard method used in the automatic evaluation of the quality of the MT system through its product. The BLEU index is calculated through three components that require minimal human intervention: 1) n-gram-based precision of the MT output and the reference, 2) brevity penalty to prevent overfitting of sentence length, and 3) clipping for calibration of continuous word appearance[6]. The BLEU index ranges from 0 to 1, with a higher value indicating more accurate matching (overlap) with the reference translation. On the other hand, the closer to zero, the more machine translation deviates (differs) from reference translation[11]. However, this index does not reflect the degree of this deviation and its gravity. Its poor performance at sentence-level and inadequate handling of recall are its main limitations[12].

Moreover, many studies have shown that BLEU has a low correlation with human evaluation, especially in cases of inflectional languages, which has led to the development of various BLEU variants[13–15]. Additionally, BLEU is not always reported consistently due to divergences in the tokenization and normalization schemes used[16].

The BLEU index only determines accuracy with the reference (similarity), which is neither helpful in improving nor in optimizing the MT system. To know or understand what needs to be improved, we need to know the error

rate. For this reason, in our paper, we focused on the error rate (edit distance) and not on the lexical similarities (accuracy) of MT systems.

Word error rate (WER) measures Levenshtein edit distance, i.e., it computes the minimum edit distance to transform a MT output into a reference through edit operations such as insertions, substitutions, and deletions of words necessary to transform one string into another without allowing the words reordering[17]:

$$WER = \frac{\text{\# of insertions} + deletions + substitutions}{reference\ lenght}.$$

The limitation of the WER metric lies in penalizing word changes within a sentence. Several variants overcome this limitation, such as position-independent word error rate metric[18] or translation edit rate metric[8].

Position-independent word error rate (PER) is based on the WER metric but ignores the word order in both MT output and reference translation[18].
Translation edit rate (TER) is defined as the minimum number of edit operations, including shift (a moving action or block movement), required to change a MT output to an exact match with a reference[8]:

$$TER = \frac{\text{\# of insertions} + deletions + substitutions + shifts}{reference\ lenght}.$$

In comparison with WER, which focuses on word operations only, TER considers shifts as part of edit operations. The higher the score of the error rate metrics, the worse the translation quality, and vice versa. The main motivation for using character-based metrics is their improved performance in evaluating morphologically rich languages like Slovak or other Slavic languages[19,20].

CharacTER is a character-level metric inspired by the TER metric[21]. It is defined as the minimum number of character edit operations required to match a MT output with a reference, normalized by the length of the MT output:

$$CharacTER = \frac{\text{\# of insertions} + deletions + substitutions + shifts}{MT\ output\ lenght}.$$

CharacTER first performs shift edits at the word-level; then, the shifted MT output sequence and the reference are split into characters, and the Levenshtein distance between them is computed.

Cross-lingual optimized metric for the evaluation of translation (COMET) is a PyTorch-based framework for training highly multilingual and adaptable MT evaluation models that can function as metrics[22]. It supports both architectures: the estimator model (trained to regress directly on a quality score) and the translation ranking model (trained to minimize the distance between "good" MT output and its corresponding reference/original source).

The most commonly used approach to determine the ability of automatic metrics to substitute human evaluation metrics is to compare the correlations between human evaluation metrics and the scores of automatic metrics[6]. However, it is still only a score (a number from the <0, 1> interval) that does not indicate the level of translation error rate at the segment/sentence/text level within the corpus. Additionally, automated metrics provide varying results and varying degrees of correlation with human evaluations, which are often inconsistent themselves. The translation quality of a pair of MT systems often relies on the differences between automatic scores (BLEU index) to draw conclusions without performing any further assessment[23,24].

This motivated us to search for other techniques that would be suitable for comparing translation quality and help us to identify segments/sentences/texts within a corpus that vary extremely (significantly) in translation quality, but with minimal human intervention. The advantage of using residuals when comparing translations is the ability to detect specific segments/sentences/texts within the corpus that deviate significantly from the golden-standard translation.

Residual analysis and error analysis are closely related analyses; both measure a distance (deviation or error). Residual analysis evaluates a regression model's validity by examining the differences between observed values and predicted values by the model; in our case, the model is the MT model.

The deviation or error is the distance of the observed value from the predicted/expected value, i.e., residuals represent the distance of observed values from predicated values:

$$Residual\ value_i = Observed\ value_i - Predicted\ value_i, i = 1,2,\dots,N,$$

where, in our research, *N* represents the number of examined texts in the data set, observed value is represented by the neural MT error rate and the predicted/expected value by the statistical MT error rate of a given text.

Extreme distances between the examined models (MT systems in our research) are identified based on the $\pm 2sigma$ rule, similar to outliers in residual analysis:

$$Mean(Residual\ value_i) \pm StdDev(Residual\ value_i), i = 1,2,\dots,N,$$

where the residual values represent the differences in the error rate of the examined MT models, neural MT system and statistical MT system in our case.

Residuals allow us to identify patterns, better understand and interpret model errors, and subsequently eliminate, correct, or analyze them, as well as their influence on MT quality[25].The aim of the study is to compare two different approaches to machine translation – statistical and neural – using automatic MT metrics of error rate and residuals. We examined four available online MT systems (statistical Google Translate, neural Google Translate, and two European commission's MT tools - statistical mt@ec and neural eTranslation) through their products (MT outputs).

The statistical MT (SMT) systems are represented by Google Translate (GT_SMT)[26] and mt@ec (the European Commission's MT tool)[27] and their transformations into neural MT (NMT) systems, which are represented by Google Translate (GT_NMT)[28] and eTranslation (the European Commission MT tool)[29]. The shift from mt@ec to eTranslation improved the translation quality, speed, and security of the interface. Google team made the same transformation in September 2016; it switched to Google neural machine translation, focusing on an end-to-end learning framework that learns from millions of examples and provides significant improvements in translation quality[30].

The main objective consists of three partial objectives:

- The first objective lies in the comparison of statistical MT systems and neural MT systems based on the automatic MT metrics of error rate (WER, PER, and TER).
- The second objective aims to identify or detect machine-translated segments/sentences/texts that deviate significantly from human translations based on the score of error rate metrics and residuals. This includes identifying texts in which statistical MT was closer to human translation than neural MT or vice versa.
- The third objective involves verifying the validity of the obtained results through metrics such as BLEU and COMET, as well as the characTER metric, which correlates better with human evaluation in the case of morphologically richer languages[19,20].

The translation directions were from English and German into an inflectional and a low-resourced Slovak. Moreover, Slovak belongs to one of the official languages of the European Union.

The structure of the paper is as follows. The second section contains related work in the field of automated MT evaluation and a comparison of various MT systems. The third section describes the used data set and the applied research methodology. The subsequent section focuses on the research results based on the evaluation of error rate metrics and residuals. The fifth section offers a discussion of the results. The last section comprises research conclusions.

## Related Work

Statistical MT and neural MT are the most extensively used architectures within the MT systems[31].

Pinnis et al.[32] compared the NMT and phrase-based SMT systems for highly inflected and low-resourced languages. They compared large and small bilingual corpora, focusing on six language pairs: [Latvian, Estonian]-English, Estonian-Russian, and vice versa. MT evaluation was conducted using the automatic evaluation metrics (BLEU, NIST, and ChrF2) and manual error analysis. The error analysis was focused on identification morphological, syntactical, and lexical errors. The results showed that the NMT system produced twice as many errors in lexical choice (wrong or incorrect lexical choice) as the phrase-based SMT system. On

the other hand, the NMT system demonstrated much better grammatical accuracy (forms and structure of words, and word order) than the SMT system.

Yang et al.[33] examined translation quality from ancient Chinese to modern Chinese. They proposed a novel automatic evaluation model - dual-based translation evaluation, without multiple references. To compare the results, BLEU and the Levenshtein distance were used as baselines. They proved that dual-based translation evaluation achieved better agreement, and/or concordance with human evaluation (human judgements).

Fomicheva and Specia[34] conducted a broad meta-evaluation study of automatic evaluation metrics. They evaluated more than 20 automatic evaluation metrics on multiple data sets (WMT16 data set, MTSummit17 English–Latvian data set, Multiple-Translation Chinese data set, WMT17 Quality Estimation German–English data set, GALE Arabic–English data set, and EAMT11 French–English data set). Data sets also contained manual assessments based on different quality criteria (adequacy, fluency, or PE effort) collected using several different methods. The meta-evaluation was conducted based on three aspects: MT quality, MT system types, and manual evaluation type. They showed that the accuracy of automatic MT evaluation varies depending on the overall MT quality. They showed that the automatic metrics perform poorly when faced with low-quality translations, but additionally that evaluating low-quality translations is also more challenging for humans. They also showed that metrics are more reliable when evaluating neural MT than statistical MT systems. Metric performance can be affected by various factors, such as text-domain, language-pair, or type of MT system.

Moghe et al.[35] evaluated nine metrics consisting of string overlap metrics, embedding-based metrics, and metrics trained using scores from human MT evaluation on three extrinsic tasks (dialogue state tracking, question answering, and semantic parsing) covering 39 unique language pairs. They showed that interpreting the quality of the produced MT translation based on a number is unreliable and difficult. They also showed that scores provided by neural metrics (e.g. COMET) are not interpretable, in large part due to having undefined ranges, and also that it is unclear if automatic metrics can reliably distinguish good translations from bad at the sentence level.

Alvarez-Vidal & Olivier[23] found that automatic metrics such as BLEU were intended to be used as a development tool and cannot be blindly used to assess MT systems without taking into account the final use of the translated text. They recommend a two-step MT evaluation which can ensure the quality of the MT output. They compared two different NMT engines – the commercial online available DeepL NMT system and a system trained on news-domain by the authors for the English-Spanish language pair. They showed that automatic metrics used (BLEU, NIST, WER, TER, EdDist, and COMET) yield better results for the NMT system trained by the authors, except for COMET.

Almahasees[36] compared the MT outputs of Google Translate and Microsoft Bing Translator (both based on SMT). The used data contained political news in English and were translated into Arabic. The data were evaluated using the automatic evaluation metrics, and the results showed better results for MT outputs produced by Google Translate. Later, Almahasees[37] conducted similar research with journalistic texts for the language pair English-Arabic, but with MT systems operating on neural networks. He compared the MT outputs based on automatic MT evaluation metrics of error rate. The results showed similar results for both MT systems in orthography and grammatical accuracy. The difference was found in the case of lexis, where the neural MT (Google Translate) achieved better results than the statistical MT (Bing).

Marzouk & Hansen-Schirra[38] focused on controlled languages (CLs) to improve the quality of NMT output. They compare the impact of applying nine CL rules on the quality of MT output produced by five MT systems (Google, Bing, Lucy, SDL, and Systran, i.e., neural, rule-based, statistical, and two hybrid MT systems) by applying three methods: error annotation, human evaluation, and automatic evaluation (TERbase and hLEPOR). The data set consisted of 216 source sentences of technical-domain translated from German into English. They showed that the NMT does not require CL rules; i.e., before and after applying the CL rules, NMT system showed the lowest number of errors.

Li & Wang[39] focused on the optimization of automatic MT evaluation. They applied representative ‘list-wise learning to rank’ approach, ListMLE. The selection of features was motivated by the BLEU-n metrics, phrase-based SMT, and NMT. They used the data sets released for WMT'2014 and WMT'2015 metrics tasks. To evaluate the results of the experiment, they compared the list-wise approach with the most used metrics, such as

BLEU-n, METEOR, TER, etc. The results showed that the novel approach achieved better results than the above-mentioned metrics.

Singh & Singh[40] focused on MT quality for low-resource languages. They aimed at an NMT system that should improve the translation quality for the English-Manipuri language pair. They compared multiple approaches such as SMT, RNN, and transformer architecture. The results showed a higher quality translation in terms of statistically significant automatic scores and manual evaluation compared to the statistical and neural supervised baselines, as well as the pretrained mBART and existing semi-supervised models.

Shterionov et al.[41] compared phrase-based SMT and NMT systems based on lexical similarities. They applied automatic evaluation metrics (BLEU, TER, and F-measure) to assess the performance of MT systems. Based on the same data set, they built five NMT and phrase-based SMT engines for various language pairs. They showed that the quality evaluation scores indicated better performance for the PBSMT engines, contrary to human evaluation. They suggested that automatic evaluation metrics (BLEU, TER, and F-measure) are not always convenient for evaluation and do not correspond with NMT quality.

Tryhubyshyn et al.[42] examined the relationship between MT system quality and QE system performance. They showed that QE systems trained on lower-quality MT translations (a mix of translations from different MT models) tended to perform better than those trained on higher-quality MT outputs (translations from one MT system).

As mentioned in the introduction, automated metrics (such as BLEU) yield varying results depending on the reference translation, text domain and languages. They draw conclusions without performing further evaluation or analysis, such as error analysis. Moreover, when the results of automatic evaluation were compared with those of manual evaluation, their correlation reached different degrees of agreement. Additionally, evaluators in the manual evaluations were often inconsistent regarding the error rate of the machine translation. These findings are also supported by the studies focused on Slavic languages or low-resource languages.

The aforementioned studies (as well as ours in the first objective) found that, on average, NMT is better than SMT. However, our proposed approach through residual analysis (regardless of which automatic metric is used) identifies segments that, on the contrary, show higher SMT quality. We have shown that our approach is suitable not only for automatic metrics of accuracy, but also for automatic metrics of error rate, which distinguishes us from all previous studies focused on Slovak so far. Moreover, it turns out that it is more appropriate not to consider the raw score of automatic metrics within the MT quality evaluation and comparison of NMT and SMT, but their distance (difference). Analyzing differences will enable us to evaluate the MT quality of individual segments.

## Materials and Methods

This study focuses on comparing NMT systems (represented by Google Translate and eTranslation) and SMT systems (represented by Google Translate and mt@ec, the European Commission's MT tool, which has later transformed into neural MT- eTranslation).

The statistical machine translated articles were obtained in 2016 from both, Google Translate (GT_SMT) and the European Commission's DGT tool (mt@ec). Later, in 2021, the same articles were machine-translated by the NMT engines Google Translate (GT_NMT) and the European Commission's DGT tool (eTranslation).The translation directions were from English and German into Slovak, where Slovak is a synthetic language containing inflected morphology and with loose word order[43]. Human translation and post-editing of machine translation were conducted in interactive online system OSTEPERE[25,44–47].

The examined articles were tokenized and aligned using the Hunalign tool[48] in the following order: source sentence with one human translation (HT), four machine translations (MTs), and one post-edited machine translation (PEMT).

The evaluation of the two different MT systems was conducted through automatic metrics of error rate (WER, PER, and TER). We aimed to identify the errors produced by the examined MT systems and determine whether changing the architecture of the MT systems resulted in decreasing to produce the same errors or, on the contrary, whether they start to create new ones. To verify the validity of the obtained results of the error rate, we used the metrics of accuracy - BLEU and COMET, and also character-based metric of error rate – characTER.

## Dataset

The data set comprises articles published by the British online newspaper The Guardian and the German online newspaper Der Spiegel, along with their machine and human translations. The corpus consists of eight data sets, and/or two English-Slovak and German-Slovak corpora: 1) articles written in English and German as source texts, 2) articles machine-translated from English and German into Slovak by four different MT engines (by SMT in 2016 and by NMT 2021), 3) human-translated articles from English and German into Slovak by professional translators (both in 2016), and 4) post-edited machine-translated articles by professional translators (in 2016).

The lexico-grammatical structure of the dataset[49] was obtained using Stanza[50], an automatic morphological annotator tool (Table 1).

**Table 1 – Dataset composition of (a) English MT outputs /HT and (b) German MT outputs /HT**

| (a) Feature type | Feature name | GT_SMT | GT_NMT | mt@ec_SMT | eTransla-tion_NMT | Human translation |
|---|---|---|---|---|---|---|
| Readability | Average sentence length (words) | 19.39 | 19.12 | 17.91 | 19.09 | 19.84 |
| | Average word length (characters) | 5.43 | 5.59 | 5.75 | 5.55 | 5.69 |
| | Number of short sentences (n<10) | 18.13% | 18.75% | 21.88% | 18.13% | 15.63% |
| | Number of long sentences (n>=10) | 81.88% | 81.25% | 78.13% | 81.88% | 84.38% |
| Lexico-grammatical | Frequency of nouns | 22.65% | 22.06% | 23.80% | 21.90% | 21.99% |
| | Frequency of adjectives | 10.84% | 10.96% | 11.81% | 10.85% | 10.59% |
| | Frequency of verbs | 9.30% | 9.88% | 8.79% | 9.74% | 10.78% |
| | Frequency of determiners | 4.44% | 4.48% | 4.52% | 4.41% | 5.17% |
| | Frequency of adpositions | 9.63% | 9.68% | 9.42% | 9.46% | 9.35% |
| | Frequency of proper nouns | 4.77% | 4.45% | 4.67% | 4.88% | 4.23% |
| | Frequency of coordinating conjunctions | 3.31% | 3.14% | 3.56% | 3.26% | 3.40% |
| | Frequency of subordinating conjunctions | 3.34% | 3.53% | 2.69% | 3.65% | 2.75% |
| | Frequency of interjections | 0.17% | 0.17% | 0.03% | 0.11% | 0.13% |
| | Frequency of adverbs | 3.50% | 3.37% | 3.29% | 3.46% | 3.56% |
| | Frequency of pronouns | 2.48% | 3.26% | 2.09% | 3.21% | 3.99% |
| | Frequency of auxiliaries | 4.42% | 3.51% | 3.95% | 3.99% | 2.94% |
| | Frequency of numerals | 3.75% | 3.90% | 4.16% | 3.77% | 3.61% |
| | Frequency of particles | 1.46% | 1.70% | 1.53% | 1.59% | 1.97% |
| | Frequency of punctuations | 14.62% | 15.14% | 14.44% | 14.92% | 14.42% |
| | Frequency of others | 1.30% | 0.78% | 1.26% | 0.81% | 1.13% |

| (b) Feature type | Feature name | GT_SMT | GT_NMT | mt@ec_SMT | eTransla-tion_NMT | Human translation |
|---|---|---|---|---|---|---|
| Readability | Average sentence length (words) | 14.22 | 14.20 | 12.81 | 13.74 | 14.63 |
| | Average word length (characters) | 5.44 | 5.54 | 5.90 | 5.60 | 5.65 |
| | Number of short sentences (n<10) | 36.14% | 36.49% | 41.86% | 37.39% | 33.09% |
| | Number of long sentences | 63.86% | 63.51% | 58.14% | 62.61% | 66.91% |

| | (n>=10) | | | | | |
|---|---|---|---|---|---|---|
| **Lexico-grammatical** | Frequency of nouns | 22.00% | 22.88% | 25.22% | 22.79% | 23.02% |
| | Frequency of adjectives | 9.98% | 10.36% | 10.61% | 10.43% | 10.53% |
| | Frequency of verbs | 9.11% | 9.74% | 7.94% | 9.55% | 9.79% |
| | Frequency of determiners | 4.23% | 4.50% | 3.59% | 4.37% | 4.57% |
| | Frequency of adpositions | 9.38% | 9.74% | 9.33% | 9.80% | 9.57% |
| | Frequency of proper nouns | 5.50% | 5.40% | 4.76% | 5.59% | 5.24% |
| | Frequency of coordinating conjunctions | 3.14% | 2.90% | 2.95% | 2.84% | 3.00% |
| | Frequency of subordinating conjunctions | 2.90% | 2.63% | 2.57% | 2.78% | 2.50% |
| | Frequency of interjections | 0.06% | 0.04% | 0.01% | 0.03% | 0.01% |
| | Frequency of adverbs | 4.83% | 4.36% | 3.95% | 3.91% | 4.47% |
| | Frequency of pronouns | 2.28% | 2.95% | 2.03% | 2.84% | 3.32% |
| | Frequency of auxiliaries | 3.68% | 3.04% | 3.33% | 3.25% | 2.72% |
| | Frequency of numerals | 3.09% | 3.05% | 3.27% | 3.24% | 2.97% |
| | Frequency of particles | 2.68% | 2.70% | 2.92% | 2.61% | 3.69% |
| | Frequency of punctuations | 16.04% | 14.99% | 16.42% | 15.28% | 13.85% |
| | Frequency of others | 1.11% | 0.72% | 1.09% | 0.69% | 0.74% |

Due to the fact that the created corpora are composed of articles with the features of newspaper writing (own register), the examined corpora mainly consist of nouns, followed by verbs and adjectives. Regarding the readability of the examined translations (from EN to SK and also from DE to SK), there are unequal proportions of short ($n < 10$) and long ($n >= 10$) sentences among MTs. The reduction in words within the sentence occurs frequently in statistical MT (mt@ec), which indicates word omission and a shift in meaning, and/or a certain loss of meaning (e.g., short sentences ($n < 10$) for EN: GT_SMT = 18.13%; GT_NMT = 18.75%; mt@ec_SMT = 21.88%; eTranslation_NMT = 15.63%; and for DE: GT_SMT = 36.14%; GT_NMT = 36.49%; mt@ec_SMT = 41.86%; eTranslation_NMT = 37.39%).

The readability results are also confirmed by corpus statistics (Table 1), where only adjectives are approximately equally distributed in all four MT outputs in both language directions (e.g., adjectives for EN: GT_SMT = 10.84%; GT_NMT = 10.96%; mt@ec_SMT = 11.81%; eTranslation_NMT = 10.85%; and for DE: GT_SMT = 9.98%; GT_NMT = 10.36%; mt@ec_SMT = 10.61%; eTranslation_NMT = 10.43%), compared to verbs or nouns (Table 1). This motivated us to investigate the differences between individual MT outputs, whether these differences are statistically significant and whether these differences cause grammatical or lexical errors in translation.

## Applied Methodology

The applied methodology, inspired by other studies[51–53], consists of these stages (Figure 1):

(1) Acquisition of unstructured textual data – source text (journalistic texts). We focused on journalistic texts (newspaper writing) as they belong to the most read and translated texts by people. We chose the two most popular journals, from which we obtained all freely available texts from various fields (politics, sports, show business, and technology) published in the given year 2016.

(2) Data preparation – consisting of following tasks:

   (a) Text pre-processing – removing text formatting, which can influence the MT quality (images or tables can divide the text inappropriately and produce bad translation).

   (b) Human translation – the translation process was realized in the tailored system OSTEPERE, which offers user-friendly interface for human translators and post-editors. The system saved

the human translations and post-edited machine translations into a database for further processing.

(c) Machine translation – automatic translation of the source text by MT engines (Google Translate [SMT | NMT], mt@ec [SMT], and eTranslation [NMT].

(d) Sentence alignment – the generated MT outputs and human translations are aligned with the source texts using Hunalign tool (based on the 1-to-1 principle).

(3) Automatic MT evaluation using automatic metrics of error rate at the segment level.
We applied automatic MT metrics based on the Levenshtein distance, which computes the minimum edit distance to transform a MT output into a reference through edit operations (insertions, substitutions, deletions, and shift of words necessary to transform one string into another).
$WER(h,r) = \frac{min\#(I+D+S)}{|r|}$, where *r* is a reference of a hypothesis/MT output *h*, *I* - insertion, *D* - deletion, and *S* - substitution.
The minimum number of edit operations is divided by the number of words in the reference[54].
$PER(h,r) = 1 - \frac{n-\max(0,|h|-|r|)}{|r|}$, where *r* is a reference of a hypothesis/MT output *h*, *n* is the number of similar words[18].

$TER\ (h,r) = \frac{min\#(I+D+S+shift)}{|r|}$, where *r* is a reference of a hypothesis/MT output *h*, *I* - insertion, *D* - deletion, *S* – substitution, and *shift* (a number of changes in word order). Compared to WER, TER considers shifts as a part of edit operations. TER deals with more edit operations, allowing it to capture various differences in word order.
The higher the score of error rate metrics, the worse the translation quality, and vice versa.

(4) Comparison of MT quality based on (i) MT system used (Google Translate or European Commission's DGT system) and (ii) artificial intelligence approach to MT (statistical approach to MT or neural approach to MT).

(i) We test the differences in the score of automatic MT metrics between two MT systems (Google Translate (GT) and the European Commission's MT tool (EC)), separately for WER, PER, and TER.

(ii) We test the differences in the score of automatic MT metrics between artificial intelligence approached to MT (statistical vs neural), separately for WER, PER, and TER.
To test the differences between dependent samples (WER/PER/TER: EC_SMT, GT_SMT, EC_NMT, and GT_NMT), we use adjusted univariate tests for repeated measure due to the failure of the sphericity assumption (Mauchley sphericity test – WER: $W = 0.849$, *Chi-Square* = 25.886, $df = 5$, $p < 0.001$; PER: $W = 0.916$, *Chi-Square* = 13.795, $df = 5$, $p = 0.017$; TER: $W = 0.846$, *Chi-Square* = 26.336, $df = 5$, $p < 0.001$).

(5) Identification of extreme differences between statistical and neural MT. To identify extreme values, we apply the residual analysis, i.e.,
$(residual\ value)_i = (WER/PER/TER\ score\ of\ NMT\ text)_i - (WER/PER/TER\ score\ of\ SMT\ text)_i\ i = 1,2,\dots,N$, where *N* is the number of examined texts in the dataset.

(6) Validation of the obtained results –using automatic metrics BLEU, COMET, and characTER.
We used one of the main models of COMET: *wmt22-comet-da*. This model uses a reference-based regression approach and has been trained on direct assessments from WMT17 to WMT20. It provides scores ranging from 0 to 1, where 1 represents a perfect translation.

$CharacTER\ (h,r) = \frac{min\#(I+D+S+shift)}{|h(characters)|}$, where *h* is a hypothesis/MT output, *I* - insertion, *D* - deletion, *S* – substitution, and *shift*.
*BLEU-n*[7] is a geometric mean of *n-gram precisions* with a *brevity penalty* (BP), i.e. penalty to prevent very short sentences:

$BLEU\ (n) = exp \sum_{n=1}^{N} w_n \log p_n \times BP$, where $w_n$ is weights for different $p_n$,

$$BP = \begin{cases} 1, & if\ h > r \\ e^{1-\frac{r}{h}}, & if\ h \leq r \end{cases}$$, where $r$ is a reference of a hypothesis $h$.


**Fig. 1 – Methodology workflow diagram**

# Results

## Automatic MT evaluation based on metrics of error rate

For all automatic metrics (WER, PER, and TER), the Mauchley sphericity test is significant ($p < 0.05$), i.e., the assumption is violated. We adjusted the degrees of freedom using the Greenhouse-Geisser adjustment. Based on the results of adjusted univariate tests for repeated measure (Greenhouse-Geisser adjustment) among GT_SMT, GT_NMT, mt@ec_SMT, and eTranslation_NMT, there are significant differences in MT quality concerning the scores of metrics of error rate (WER, PER, and TER: *G-G Epsilon* < 0.944, *G-G Adj.* $p < 0.001$). NMTs were of statistically significantly better quality than SMTs regardless of which MT tool (GT or the European Commission's MT tool) was used. NMTs were lexically more similar to the references than SMTs.

**Table 2 – Bonferroni (adjustment) post-hoc test for multiple comparisons of the metric WER between different MT systems (GT tools or EC tools) and approaches (statistical or neural) in the English-Slovak language pair**

| English=1 | Mean | 1 | 2 | 3 |
|---|---|---|---|---|
| WER_GT_NMT | 0.679 | | **** | |
| WER_EC_NMT | 0.715 | | | **** |
| WER_GT_SMT | 0.778 | **** | | |
| WER_EC_SMT | 0.800 | **** | | |

Note: **** - homogenous groups $p > 0.05$

Based on multiple comparisons (Table 2), there are significant differences in the score of metric WER between NMT (GT) and the others, as well as between NMT (eTranslation_NMT) and the others, but there is no difference between SMT (GT) and SMT (EC). Were identified three homogeneous groups (**** - $p > 0.05$) in terms of the agreement/concordance of the examined texts. NMT produced by GT achieved the lowest error rate (0.679) compared to other MTs. On the other hand, SMT produced by mt@ec achieved the highest error rate (0.800), but is very close to SMT produced by GT (0.778).

**Table 3 – Bonferroni (adjustment) post-hoc test for multiple comparisons of the metric PER between different MT systems (GT tools or EC tools) and approaches (statistical or neural) in the English-Slovak language pair**

| English=1 | Mean | 1 | 2 | 3 |
|---|---|---|---|---|
| PER_GT_NMT | 0.548 | **** | | |
| PER_EC_NMT | 0.575 | **** | | |
| PER_GT_SMT | 0.642 | | **** | |
| PER_EC_SMT | 0.683 | | | **** |

Note: **** - homogenous groups $p > 0.05$

In terms of lexical similarity, regardless of word order (PER metric), there is a difference between SMT, produced by GT tool or EC tool and neural MT, but there is no difference between neural MT produced by GT tool and EC tool (Table 3). Based on multiple comparisons (Table 3), were identified three homogeneous groups (**** - $p > 0.05$) in terms of the agreement/text similarity of the examined texts. Moreover, three out of four MTs achieved lower PER scores of error rate ($PER \leq 0.642$) than all MTs evaluated by metric WER ($WER \geq 0.679$).

**Table 4 – Bonferroni (adjustment) post-hoc test for multiple comparisons of TER metrics between different MT systems (GT tools or EC tools) and approaches (statistical or neural) in the English-Slovak language pair**

| English=1 | Mean | 1 | 2 | 3 |
|---|---|---|---|---|
| TER_GT_NMT | 0.674 | | **** | |
| TER_EC_NMT | 0.711 | | | **** |
| TER_GT_SMT | 0.774 | **** | | |
| TER_EC_SMT | 0.796 | **** | | |

Note: **** - homogenous groups $p > 0.05$

The TER values copy the WER values (Table 2 and Table 4). Based on multiple comparisons (Table 4), were identified three homogeneous groups (**** - $p > 0.05$) in terms of the agreement/text similarity of the examined texts. There are significant differences in the score of the metric TER between GT_NMT (neural GT) and the others, as well as between EC_NMT (eTranslation_NMT) and the others (Table 4), but there is no difference between GT_SMT (statistical GT) and EC_SMT (mt@ec_SMT). Neural MT produced by GT achieved the lowest error rate (0.674) compared to other MTs. On the other hand, statistical MT produced by mt@ec achieved the highest error rate (0.796), but is very close to statistical MT produced by GT (0.774).

**Table 5 – Bonferroni (adjustment) post-hoc test for multiple comparisons of (a) the PER and (b) WER metrics between different MT systems (GT tools or EC tools) and approaches (statistical or neural) in the German-Slovak language pair**

| (a) English=0 | Mean | 1 | 2 | 3 | 4 | (b) English=0 | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PER_GT_NMT | 0.495 | **** | | | | WER_GT_NMT | 0.609 | **** | | | |
| PER_EC_NMT | 0.548 | | **** | | | WER_EC_NMT | 0.664 | | **** | | |
| PER_GT_SMT | 0.649 | | | **** | | WER_GT_SMT | 0.765 | | | **** | |
| PER_EC_SMT | 0.720 | | | | **** | WER_EC_SMT | 0.821 | | | | **** |

Note: **** - homogenous groups $p > 0.05$

**Table 6 – Bonferroni (adjustment) post-hoc test for multiple comparisons of TER metrics between different MT systems (GT tools or EC tools) and approaches (statistical or neural) in the German-Slovak language pair**

| English=0 | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| TER_GT_NMT | 0.607 | **** | | | |
| TER_EC_NMT | 0.662 | | **** | | |
| TER_GT_SMT | 0.763 | | | **** | |
| TER_EC_SMT | 0.820 | | | | **** |

Note: **** - homogenous groups $p > 0.05$

We applied the same analysis to machine-translated texts from German into Slovak. Due to the violation of the assumption of sphericity of the covariance matrix, we used modified tests for repeated measurements (Greenhouse-Geisser adjustment) to test the differences in MT quality among GT_SMT, GT_NMT, mt@ec_SMT, and eTranslation_NMT represented by the metrics of error rate (PER: $W = 0.868$, *Chi-sqr.* = 78.816, $df = 5$, $p < 0.001$; WER: $W = 0.873$, *Chi-sqr.* = 75.826, $df = 5$, $p < 0.001$; TER: $W = 0.889$, *Chi-sqr.* = 65.643, $df = 5$, $p < 0.001$). The highest rate of violation of the assumption was identified in the case of the metric WER (*G-G Epsilon* = 0.912), followed by PER (*G-G Epsilon* = 0.919), on the contrary, the lowest for the metric TER (*G-G Epsilon* = 0.923). Overall, the rate of violation of the assumption of sphericity of the covariance matrix was low for all applied metrics, we used adjusted significance tests (WER, PER, and TER: *G-G Epsilon* < 0.923, *G-G Adj.* $p < 0.001$) and subsequently, we compared them with unadjusted univariate tests for repeated measure ($F > 211.214$, $p < 0.001$).

Based on the results, we reject the global H0 at the 0.001 significance level in the case of all metrics, which claims that there is no statistically significant difference in the quality of MT when translating from German to Slovak, represented by the error rate metrics PER, WER, and TER, among GT_SMT, GT_NMT, mt@ec_SMT, and eTranslation_NMT. NMTs were of statistically significantly better quality than SMTs regardless of which MT tool was used (Tables 5 and Table 6). NMT produced by GT tool (Table 5 and Table 6) achieved statistically significant the lowest error rate (*PER* = 0.495, *WER* = 0.609, *TER* = 0.607). On the other hand, SMT produced by mt@ec, a EC tool (Table 5 and Table 6) achieved statistically significant the highest error rate (*PER* = 0.720, *WER* = 0.821, *TER* = 0.820).

We conclude that the assumption regarding better NMT quality compared to SMT has been confirmed, regardless of the language pair. We showed statistically significant differences between SMT and NMT in favor of NMT based on all metrics of error rate (WER, PER, and TER), regardless of MT tool used (Google Translate tool or the European Commission's MT tool).

These findings indicate that the error rate in the examined texts is probably related to recall (lexical accuracy). Considering the reference, the error rate of the examined MTs is more associated with lexical accuracy, i.e., vocabulary and word omission, than grammatical accuracy, i.e., forms and structure of words and word order. This motivated us to apply residual analysis to identify and specify in more detail MT errors that occurred in individual machine translations.

## Identification of extreme differences based on the score of error rate metrics between SMT and NMT - English-Slovak machine translations

We used residuals to identify texts with extreme values of error rate metrics (WER, PER, and TER) between SMT and NMT for each MT tool separately. We applied the rule ± *2sigma*, i.e., values outside the interval are considered extremes. The mean of NMT – SMT differences for all metric values (WER/PER/TER) is negative (Figures 2-7), which confirms our finding (previous subsection) that in terms of error rate, NMT achieved a statistically significantly lower error rate, i.e., better translation quality. The neural MT outputs were more similar to the references than the statistical MT outputs.

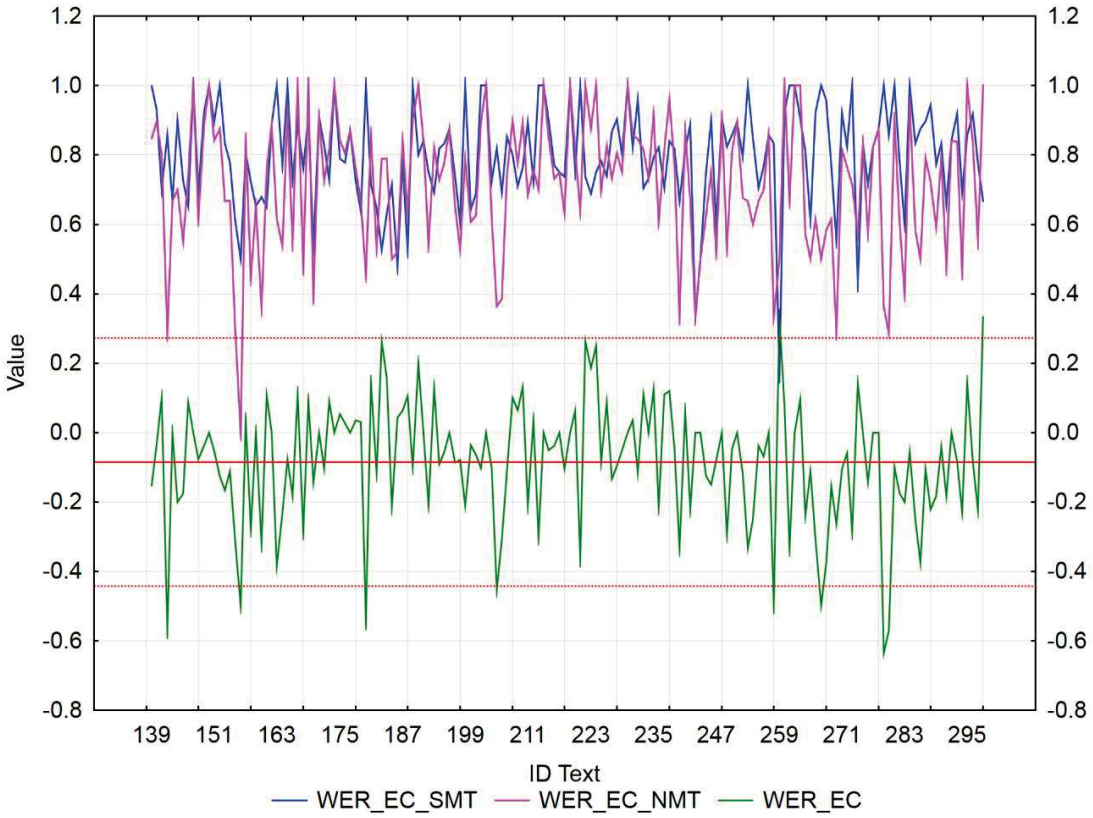



**Figure 2 – Visualization of NMT-SMT residuals for WER metric and the European Commission's MT tool**

In the case of the European Commission's MT tool (Figure 2), we identified 8 texts (ID_142, ID_156, ID_180, ID_205, ID_258, ID_267, ID_279, and ID_280), that showed a statistically significantly better WER score of NMT against SMT (*residuals* ≈ -0.5). Only 2 texts (ID_259 and ID_298) achieved a significantly better WER score of SMT against NMT (*residuals* ≈ 0.33), but both texts consist of short sentences (less than 7 words, including articles), which could have had an impact on the results.

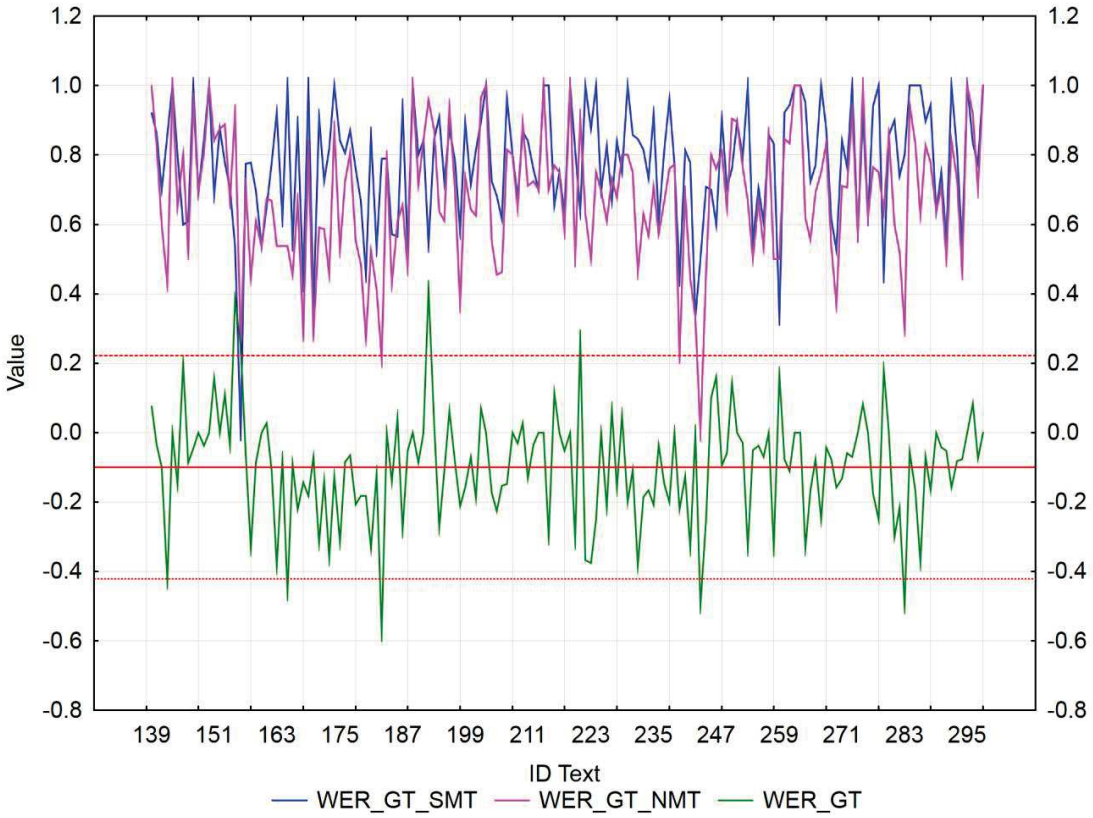



**Figure 3 – Visualization of NMT-SMT residuals for WER metric and Google translate**

In the case of Google translate (Figure 3), we identified 5 texts (ID_142, ID_156, ID_180, ID_205, ID_258, ID_267, ID_279, and ID_280), that showed a statistically significantly better WER score of NMT against SMT (*residuals* ≈ -0.5) and 4 texts (ID_155, ID_156, ID_192, and ID_221) with a significantly better WER score of SMT against NMT (*residuals* ≈ 0.35). These texts were more similar to the reference than NMT (NMT was correct, but used synonyms, which could have had an impact on the results).

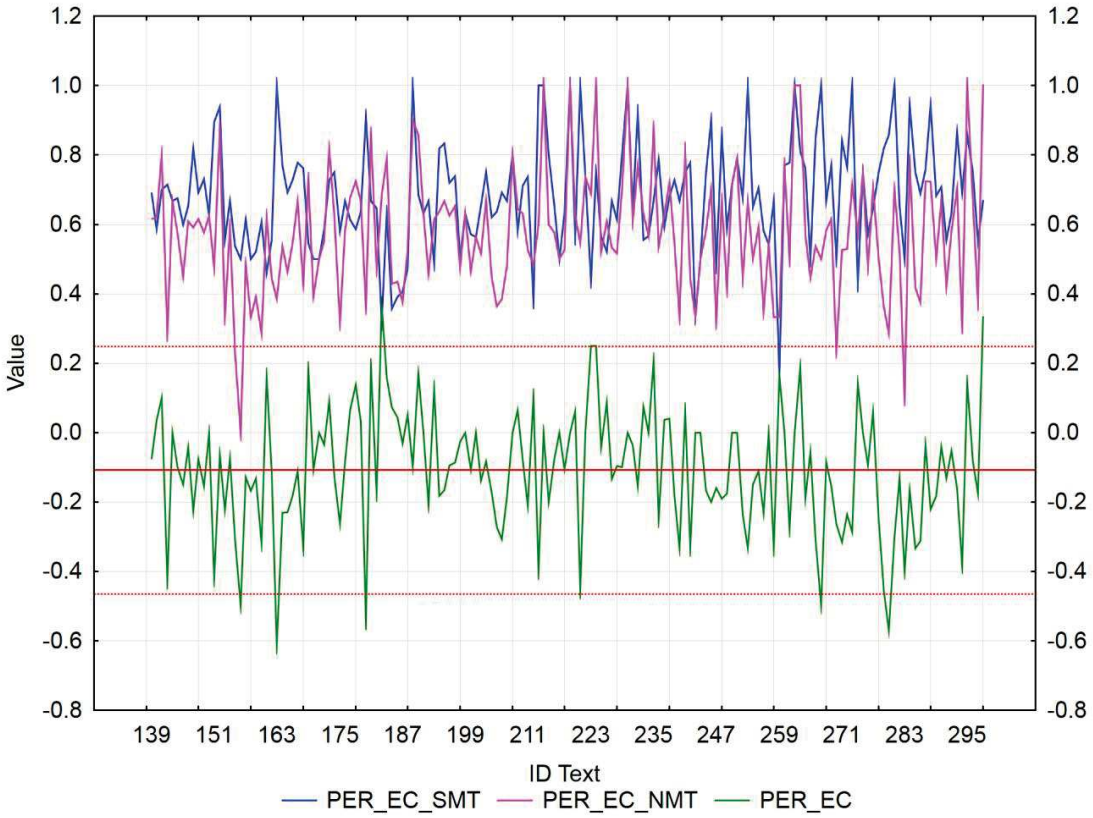



**Figure 4 – Visualization of NMT-SMT residuals for PER metric and the European Commission's MT tool**

In the case of the European Commission's MT tool (Figure 4), we identified 5 texts (ID_156, ID_163, ID_180, ID_267, and ID_280), that showed a statistically significantly better PER score of NMT against SMT (*residuals* ≈ -0.55). Only 4 texts (ID_183, ID_223, ID_224, and ID_298) achieved a significantly better PER score of SMT against NMT (*residuals* ≈ 0.3). Again, they were texts with short sentences, and NMT added extra words compared to the reference, which could have had an impact on the results.

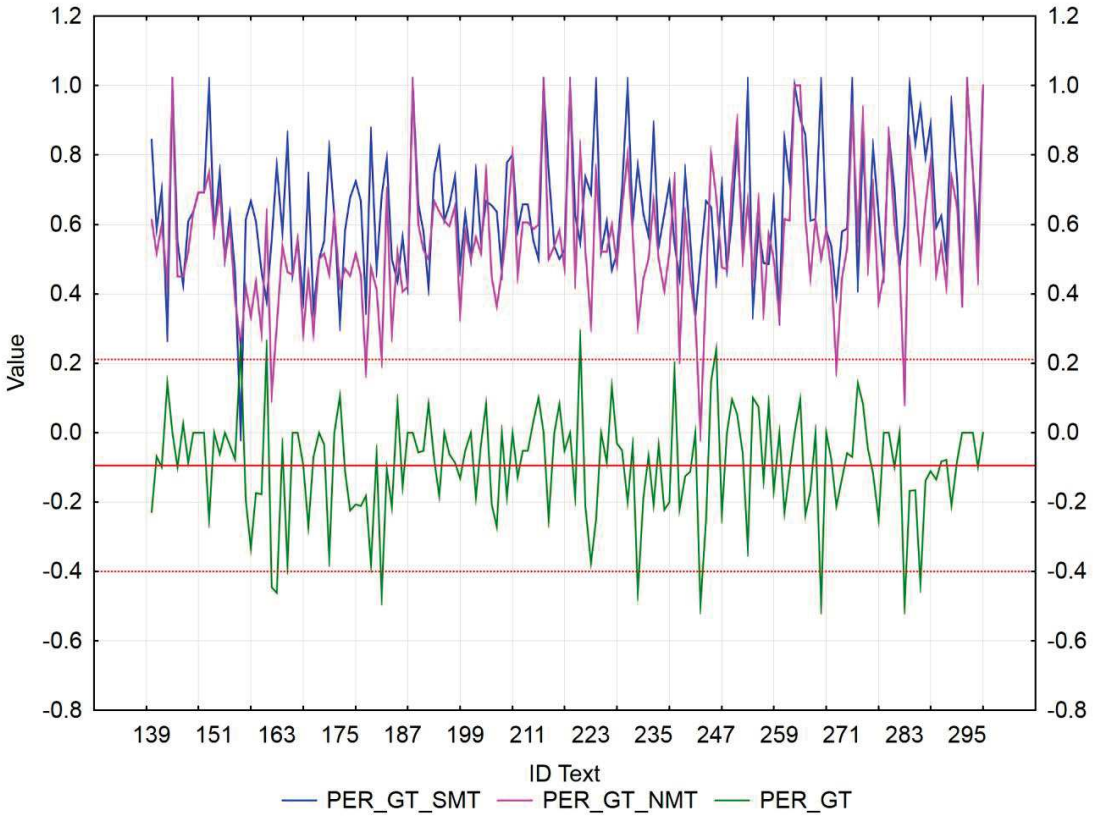



**Figure 5 – Visualization of NMT-SMT residuals for PER metric and Google translate**

In the case of Google translate (Figure 5), we identified 8 texts (ID_162, ID_163, ID_183, ID_232, ID_244, ID_267, ID_283, and ID_286), that showed a statistically significantly better PER score of NMT against SMT (*residuals* ≈ -0.45) and 4 texts (ID_156, ID_161, ID_221, and ID_247) with a significantly better PER score of SMT against NMT (*residuals* ≈ 0.25). These texts were more similar to the reference than NMT.

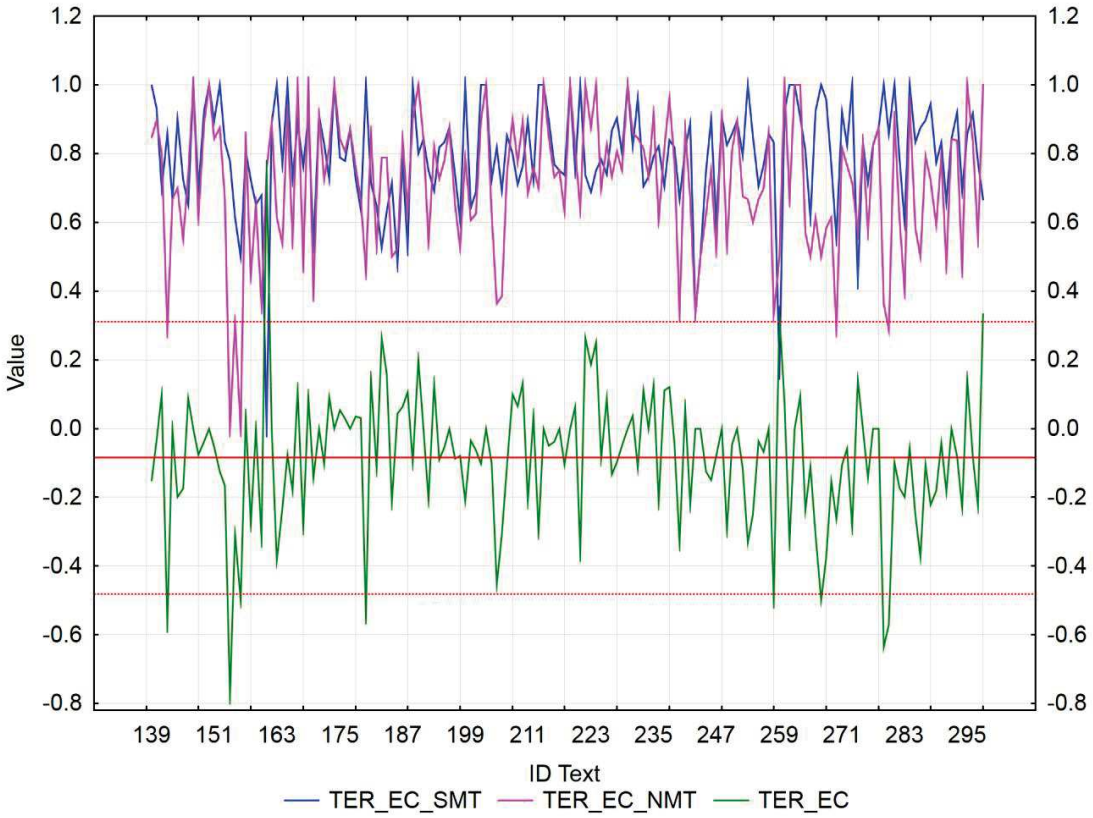



**Figure 6 – Visualization of NMT-SMT residuals for TER metric and the European Commission's MT tool**

In the case of the European Commission's MT tool (Figure 6), we identified 8 texts (ID_142, ID_154, ID_156, ID_180, ID_258, ID_267, ID_297, and ID_280), that showed a statistically significantly better TER score of NMT against SMT (*residuals* ≈ -0.6). Only 3 texts (ID_161, ID_259, and ID_298) achieved a significantly better TER score of SMT against NMT (*residuals* ≈ 0.33).

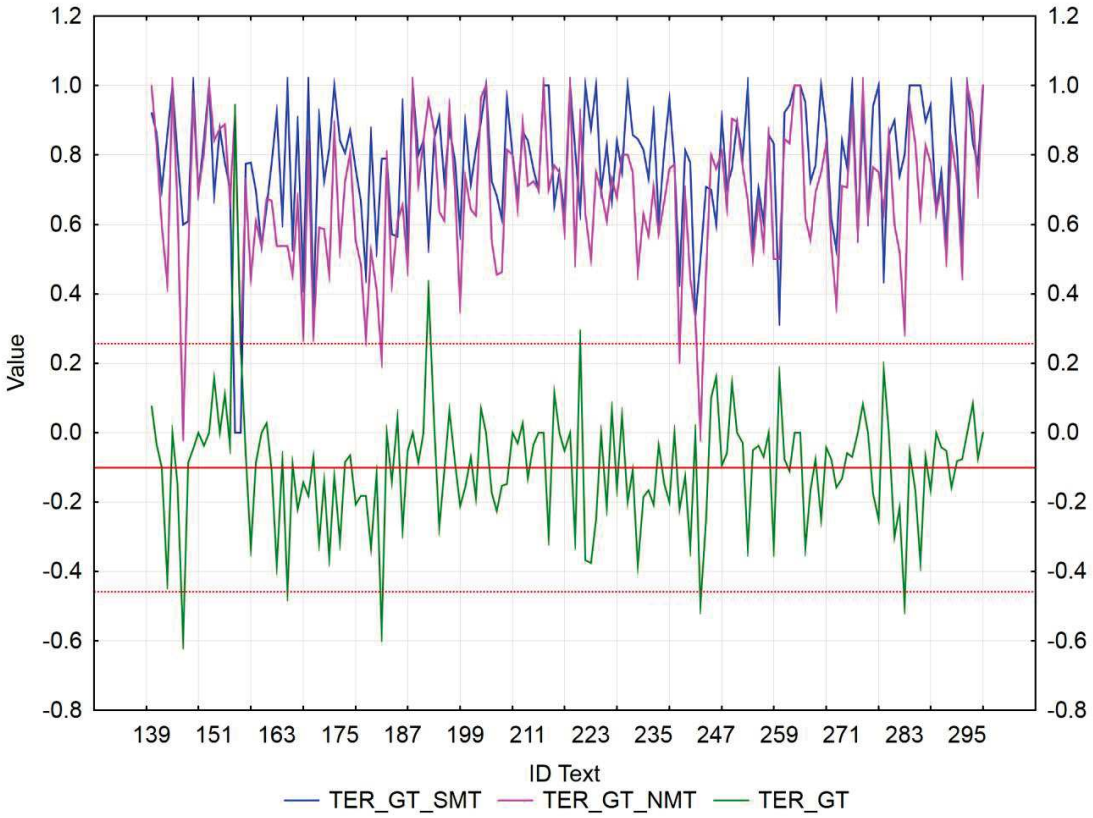



**Figure 7 – Visualization of NMT-SMT residuals for TER metric and Google translate**

In the case of Google translate (Figure 7), we identified 5 texts (ID_145, ID_165, ID_183, ID_244, and ID_283), that showed a statistically significantly better TER score of NMT against SMT (*residuals* ≈ -0.5) and 3 texts (ID_155, ID_192, and ID_221) with a significantly better score of SMT against NMT (*residuals* ≈ 0.35). These texts were more similar to the reference than NMT.

Based on our results, we can infer that the issue in MT systems lies in lexical semantics rather than in word order in the case of neural machine translation.

## Discussion

The applied automatic metrics are based on a comparison with a reference, which, in our case, was created independently (pure human translation, not affected by MT output). This could cause a distortion of MT quality, but it did not affect the comparison of SMT and NMT because we used the same reference in both cases.

Based on corpus statistics (Table 1), we assumed that NMT outperforms SMT with respect to the lexico-grammatical features of the examined texts (frequency of nouns, adjectives, and verbs).

Based on analysis results, we can conclude that NMT demonstrated higher quality than SMT in terms of error rate. All automatic metrics achieved lower scores for neural MT compared to statistical MT, i.e., NMT outperformed SMT. The most serious issues of SMT include a shift in part-of-speech, omission or addition of words, and inflection. The word order was not such a serious issue for neural MT, which we explain by the fact that it was a translation into Slovak, which has a loose word order, unlike English, which has a strict word order (SWO).

Regarding the accuracy of the strings (represented by metrics WER and TER), SMT produced approximately the same error rate, whether it was SMT produced by GT or produced by mt@ec, EC tool (Table 2 and Table 5b), which is noteworthy. Both MT tools performed at a very similar level. On the contrary, due to the similarity of

the strings (represented by metric PER), NMT produced approximately the same error rate, whether it was NMT produced by GT or produced by mt@ec, EC tool (Table 3 and Table 5a). We explain this fact by the character of the examined texts. They were of the journalistic style (newspaper writing) with no specific vocabulary or complex syntax, so MT tools did not require training on a specific text-domain. SMT showed similar error rates whether it was trained on a general text-domain (GT) or a specialized text-domain, such as administrative texts (EC). In the case of the metric PER, which only focuses on word similarity (independent of word position) and does not take into account word order and extra words, NMT achieved approximately the same error rate, whether it was NMT produced by GT or produced by mt@ec (Table 3 and Table 5b).

To validate the obtained results, we employed automatic metrics such as BLEU, COMET, and CharacTER to ensure the reliability of error rate metrics for both language directions (BLEU: *G-G Epsilon* < 0.940, *G-G Adj. p* < 0.001; COMET: *G-G Epsilon* = 0.812, *G-G Adj. p* < 0.001, and characTER: *G-G Epsilon* = 0.932, *G-G Adj. p* < 0.001). The results (Table 7) nearly fully correspond with the results for the metrics PER, WER, and TER for English-Slovak machine translation (Tables 2-4) as well as German-Slovak machine translation (Table 5 and Table 6). NMTs were of statistically significantly better quality than SMTs regardless of which MT tool and language direction were used (Table 7). NMT produced by GT (Table 7) achieved statistically significantly the lowest error rate (*CharacTER* = 0.481) and statistically significantly the highest accuracy (*COMET* = 0.887, *BLEU_1* = 0.514, *BLEU_2* = 0.227, *BLEU_3* = 0.164, *BLEU_4* = 0.097). On the other hand, SMT produced by mt@ec, EC tool (Table 7) achieved statistically significantly the highest error rate (*CharacTER* = 0.688) and statistically significantly the lowest accuracy (*COMET* = 0.662, *BLEU_1* = 0.303, *BLEU_2* = 0.115, *BLEU_3* = 0.044). According to the metric BLEU_4 (Table 7f), both SMT systems (mt@ec and GT) form one homogeneous group, i.e., they achieved the same lowest quality ($p > 0.05$).

**Table 7 – Bonferroni (adjustment) post-hoc test for multiple comparisons of (a) the CharacTER, (b) COMET, (c) BLEU_1, (d) BLEU_2, (e) BLEU_3, and (f) BLEU_4 metrics between different MT systems (GT tools or EC tools) and approaches (statistical and neural)**

| (a) All Groups | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| CharacTER_GT_NMT | 0.481 | **** | | | |
| CharacTER_EC_NMT | 0.544 | | **** | | |
| CharacTER_GT_SMT | 0.594 | | | **** | |
| CharacTER_EC_SMT | 0.688 | | | | **** |

| (b) All Groups | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| COMET_EC_SMT | 0.662 | **** | | | |
| COMET_GT_SMT | 0.739 | | **** | | |
| COMET_EC_NMT | 0.857 | | | **** | |
| COMET_GT_NMT | 0.887 | | | | **** |

| (c) All Groups | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| BLEU_1_EC_SMT | 0.303 | **** | | | |
| BLEU_1_GT_SMT | 0.382 | | **** | | |
| BLEU_1_EC_NMT | 0.463 | | | **** | |
| BLEU_1_GT_NMT | 0.514 | | | | **** |

| (d) All Groups | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| BLEU_2_EC_SMT | 0.115 | **** | | | |
| BLEU_2_GT_SMT | 0.153 | | **** | | |
| BLEU_2_EC_NMT | 0.220 | | | **** | |
| BLEU_2_GT_NMT | 0.277 | | | | **** |

| (e) All Groups | Mean | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| BLEU_3_EC_SMT | 0.044 | **** | | | |
| BLEU_3_GT_SMT | 0.072 | | **** | | |
| BLEU_3_EC_NMT | 0.119 | | | **** | |
| BLEU_3_GT_SMT | 0.164 | | | | **** |

| (f) All Groups | Mean | 1 | 2 | 3 |
|---|---|---|---|---|
| BLEU_4_EC_SMT | 0.017 | **** | | |
| BLEU_4_GT_SMT | 0.032 | **** | | |
| BLEU_4_EC_NMT | 0.062 | | **** | |
| BLEU_4_GT_NMT | 0.097 | | | **** |

Note: **** - homogenous groups $p > 0.05$

To analyze the relationships between the automatic metrics of error rate (PER, WER, and TER) and the metrics we chose as a baseline - valid criteria (BLEU_1-4, CharacTER, and COMET), we employed non-parametric correlations. Due to deviations from the normality of the automatic metrics (PER, WER, TER, BLEU_1-4, CharacTER, and COMET), we applied non-parametric Spearman rank order correlations to both language directions ($W < 0.993$, $p < 0.001$), but separately for statistical MT (Table 8) and for neural MT (Table 9).

**Table 8 – Non-parametric correlations for SMT: (a) PER x valid criterion (BLEU_1-4, CharacTER, COMET), (b) WER x valid criterion (BLEU_1-4, CharacTER, COMET), (c) TER x valid criterion (BLEU_1-4, CharacTER, COMET)**

| (a) PER_GT_SMT | | | | PER_EC_SMT | | | |
|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.92 | -65.107 | < 0.001 | BLEU_1 | -0.93 | -66.824 | < 0.001 |
| BLEU_2 | -0.75 | -30.742 | < 0.001 | BLEU_2 | -0.74 | -29.891 | < 0.001 |
| BLUE_3 | -0.58 | -19.152 | < 0.001 | BLUE_3 | -0.55 | -17.712 | < 0.001 |
| BLEU_4 | -0.39 | -11.369 | < 0.001 | BLEU_4 | -0.35 | -10.164 | < 0.001 |
| CharacTER | 0.49 | 15.083 | < 0.001 | CharacTER | 0.48 | 14.793 | < 0.001 |
| COMET | -0.37 | -10.659 | < 0.001 | COMET | -0.34 | -9.720 | < 0.001 |

| (b) WER_GT_SMT | | | | WER_EC_SMT | | | |
|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.73 | -28.521 | < 0.001 | BLEU_1 | -0.71 | -27.205 | < 0.001 |
| BLEU_2 | -0.67 | -24.155 | < 0.001 | BLEU_2 | -0.65 | -22.877 | < 0.001 |
| BLUE_3 | -0.54 | -17.097 | < 0.001 | BLUE_3 | -0.50 | -15.345 | < 0.001 |
| BLEU_4 | -0.36 | -10.441 | < 0.001 | BLEU_4 | -0.33 | -9.312 | < 0.001 |
| CharacTER | 0.60 | 20.005 | < 0.001 | CharacTER | 0.52 | 16.469 | < 0.001 |
| COMET | -0.35 | -10.124 | < 0.001 | COMET | -0.30 | -8.100 | < 0.001 |

| (c) TER_GT_SMT | | | | TER_EC_SMT | | | |
|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.73 | -28.720 | < 0.001 | BLEU_1 | -0.71 | -27.214 | < 0.001 |
| BLEU_2 | -0.67 | -24.288 | < 0.001 | BLEU_2 | -0.65 | -22.886 | < 0.001 |
| BLUE_3 | -0.54 | -17.116 | < 0.001 | BLUE_3 | -0.50 | -15.355 | < 0.001 |
| BLEU_4 | -0.36 | -10.402 | < 0.001 | BLEU_4 | -0.33 | -9.354 | < 0.001 |
| CharacTER | 0.59 | 19.801 | < 0.001 | CharacTER | 0.52 | 16.480 | < 0.001 |
| COMET | -0.35 | -10.157 | < 0.001 | COMET | -0.30 | -8.095 | < 0.001 |

In the case of SMT (Table 8), similar results were achieved for both MT systems (GT and EC). The examined metrics of error rate (PER, WER, and TER) positively correlate with the CharacTER metric (Table 8), indicating a moderate (> 0.3) to high (> 0.5) degree of statistically significant direct proportional dependency ($p < 0.001$). On the contrary, in the case of the metrics of accuracy (BLEU_1-4 and COMET), a negative correlation was identified (Table 8), revealing a moderate (< -0.3) degree of dependency between the automatic metrics (PER, WER, and TER) and the metric COMET/BLEU_4. A high (< -0.5) to very high (< -0.7) degree of statistically significant inverse-related dependency was observed between them and the metrics BLEU_1-3 ($p < 0.001$).

**Table 9 – Non-parametric correlations for NMT: (a) PER x valid criterion (BLEU_1-4, CharacTER, COMET), (b) WER x valid criterion (BLEU_1-4, CharacTER, COMET), (c) TER x valid criterion (BLEU_1-4, CharacTER, COMET)**

| PER_GT_NMT | | | | (a) | PER_EC_NMT | | | |
|---|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.96 | -91.080 | < 0.001 | | BLEU_1 | -0.95 | -81.806 | < 0.001 |
| BLEU_2 | -0.85 | -43.190 | < 0.001 | | BLEU_2 | -0.78 | -33.459 | < 0.001 |
| BLUE_3 | -0.72 | -27.440 | < 0.001 | | BLUE_3 | -0.67 | -24.399 | < 0.001 |
| BLEU_4 | -0.57 | -18.339 | < 0.001 | | BLEU_4 | -0.48 | -14.652 | < 0.001 |
| CharacTER | 0.68 | 24.687 | < 0.001 | | CharacTER | 0.62 | 20.922 | < 0.001 |
| COMET | -0.51 | -15.980 | < 0.001 | | COMET | -0.49 | -14.986 | < 0.001 |

| WER_GT_NMT | | | | (b) | WER_EC_NMT | | | |
|---|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.79 | -34.741 | < 0.001 | | BLEU_1 | -0.76 | -30.891 | < 0.001 |
| BLEU_2 | -0.80 | -35.190 | < 0.001 | | BLEU_2 | -0.70 | -26.229 | < 0.001 |
| BLUE_3 | -0.69 | -25.411 | < 0.001 | | BLUE_3 | -0.62 | -20.887 | < 0.001 |
| BLEU_4 | -0.56 | -18.204 | < 0.001 | | BLEU_4 | -0.47 | -14.101 | < 0.001 |
| CharacTER | 0.79 | 34.311 | < 0.001 | | CharacTER | 0.75 | 30.335 | < 0.001 |
| COMET | -0.48 | -14.494 | < 0.001 | | COMET | -0.44 | -12.941 | < 0.001 |

| TER_GT_NMT | | | | (c) | TER_EC_NMT | | | |
|---|---|---|---|---|---|---|---|---|
| **GT_SMT** | **R** | **t(N-2)** | **p-value** | | **EC_SMT** | **R** | **t(N-2)** | **p-value** |
| BLEU_1 | -0.79 | -34.716 | < 0.001 | | BLEU_1 | -0.75 | -30.828 | < 0.001 |
| BLEU_2 | -0.79 | -34.597 | < 0.001 | | BLEU_2 | -0.70 | -26.140 | < 0.001 |
| BLUE_3 | -0.68 | -24.863 | < 0.001 | | BLUE_3 | -0.61 | -20.624 | < 0.001 |
| BLEU_4 | -0.55 | -17.835 | < 0.001 | | BLEU_4 | -0.46 | -13.881 | < 0.001 |
| CharacTER | 0.78 | 33.488 | < 0.001 | | CharacTER | 0.75 | 30.019 | < 0.001 |
| COMET | -0.47 | -14.446 | < 0.001 | | COMET | -0.43 | -12.902 | < 0.001 |

Similar results were achieved in the case of NMT (Table 9). The automatic error rate metrics (PER, WER, and TER) positively correlate with the characTER error rate metric (Table 9), showing a high (> 0.5) to very high (> 0.7) degree of statistically significant direct proportional dependency ($p < 0.001$). On the contrary, in the case of the metrics of accuracy (BLEU_1-4 and COMET), a negative correlation was identified (Table 9). Between the automatic metrics (PER, WER, and TER) and the metric COMET/ BLEU_4 a moderate (< -0.3) to a high (< -0.5) degree of dependency was observed, and between the metrics BLEU_1-3 and automatic metrics (PER, WER, and TER), a high (< -0.5) to very high (< -0.7) degree of statistically significant inverse-related dependency was found ($p < 0.001$).

In the case of NMT (Table 9), higher dependencies were identified compared to SMT (Table 8), but in both cases, they reached at least a medium level of statistically significant dependency.

These results motivated us to conduct a manual error analysis for both SMT and NMT. We restricted the analysis to only 5 MT texts produced by GT tools (SMT_GT vs NMT_GT) due to its labour- and time-intensive nature. We divided the occurred errors into the following 4 categories that cover the text complexity of inflectional languages[55]: 1) predication, 2) syntactic-semantic correlativeness, 3) compound/complex sentences, and 4) lexical semantics.

SMT produced 184 errors in the category of predication, 279 errors in syntactic-semantic correlativeness, 76 errors in compound/complex sentences, and 370 errors in the category of lexical semantics. The results obtained for NMT were significantly different. In the sphere of predication 27 errors were identified, in syntactic-semantic correlativeness 106 errors, in compound/complex sentences 12 errors, and in the sphere of lexical semantics 442 errors were identified.

Our results correspond with the findings of similar studies[56,57] which showed that SMT is more accurate in meaning (lexical accuracy), but less fluent in grammar (grammatical accuracy), and vice versa, NMT is grammatically more fluent, but less accurate in meaning (lexical semantics).

Using residual analysis, we can reveal which errors persist and, conversely, which have been eliminated or have arisen.

In the case of the European Commission's DGT tools, when we compared SMT and NMT based on the WER metric, which takes into account not only lexical accuracy, but also grammatical correctness and word order, we found that errors most often occurred within the lexical semantics, either in (1) part of speech transformation, e.g. a noun becomes an adjective after translation with a shift in meaning, or in (2) a shift of gender, most often from masculine to feminine, or in (3) omission of commas.

Example,

SS: The other is *the opposite*: adaptable, empathetic, flexible. (noun)

SMT: Druhou je *opačná*. prispôsobiteľné empatický pružný. (adjective)

NMT: Druhý je *opak*: Prispôsobivý, empatický, flexibilný. (noun)

HT: Druhý je presným *opakom*: prispôsobivý, empatický, flexibilný. (noun)

Another (4) frequent issue was word omission and word order, e.g.

SS: Among their number were Belgian students, French schoolchildren and British lawyers.

SMT: Spomedzi nich boli belgické, francúzske a britské študentov, advokátov. (omission of word *students* or *schoolchildren*)

NMT: Medzi ich počet boli belgickí *študenti*, francúzski *žiaci* a britskí právnici.

HT: Nachádzajú sa medzi nimi belgickí *študenti*, francúzski školáci a britskí právnici.

We achieved similar results for Google Translate, but we also identified four texts in which SMT achieved a better WER score than NMT. However, after a deeper analysis we found that it was caused by using synonyms (different words with the same meaning) or by expanding the information with respect to the reference, e.g.

SS: That equates to 5am GMT back in the *United Kingdom*.

SMT: To sa rovná 5 hodín ráno GMT vzadu v *Spojenom kráľovstve*. (United Kingdom)

NMT: To predstavuje vo *Veľkej Británii* 5:00 GMT. (Great Britain)

HT: To je presne 5:00 ráno západoeurópskeho času v *Spojenom kráľovstve* (GMT).

Our findings are in line with the other studies[58–63] that focused on comparing SMT and NMT quality across various text genres. Benkova et al.[63] conducted similar research, using residual analysis and the automatic metric BLEU-n to compare quality between neural and statistical MT systems. They came to the conclusion that neural MT is more accurate or closer to reference than statistical MT in the translation of journalistic texts from English into Slovak. However, their focus was solely on the standard automatic metric of accuracy (BLEU), which does not always correlate with human evaluation in the case of machine translation into inflectional languages.

Our study provides new insights into the evaluation of MT quality from English and German into Slovak through automatic evaluation metrics of error rate and residuals. Residuals, combined with automatic metrics of error rate, represent, and/or indicate a new approach to quality evaluation and comparison between statistical and neural machine translation. To our knowledge, no study has applied residuals to identify extreme differences in the error rate of SMT and NMT. This approach is universal, independent of the languages, text-domains, or the MT tools used, which makes it original. Moreover, the issue related to reference translation is removed, and/or eliminated, as it is only a parameter when comparing two MT outputs.

The study has certain limitations, which are mainly related to the size of the dataset. We plan to expand our corpus size with more texts of the newspaper writing style, as well as of other styles.

## Conclusions

In our study, we demonstrated that through automatic evaluation metrics, neural machine translation achieved a lower error rate than statistical machine translation, regardless of the MT tool used. The manual error analysis of the selected smaller sub-corpus indicates that in the category of prediction (consisting of predicative categories, non-finite verb or other word class instead of finite verb functioning as a predicate, missing verb in predication, sentence with or without subject, sentence with or without agent, descriptive and reflexive passive verb forms, incorrectly identified subject in the sentence, incorrectly identified predicate in a sentence, incorrect form of a complex verb phrase, and others) and syntactic-semantic correlativeness (consisting of nominal morphosyntax, pronominal morphosyntax, numeral morphosyntax, verbal morphosyntax, word order, and other morphosyntactic phenomena), SMT showed a significantly higher error rate than NMT. Conversely, in the category of lexical semantics (adequate transfer of the words' meaning, polysemy, homonymy, semantic and stylistic compatibility, derivation, omission, literal translation, explication, and other), NMT showed a significantly higher error rate than SMT.

Remarkably, the research also revealed considerable diversity in translation quality. As mentioned in the methodology, the MT outputs were post-edited by professional translators. We assumed that human translations and post-edited MT outputs would be at least 80% similar, therefore, we included a calculation of the text similarity through the cosine similarity into our analysis. We found that in the case of SMT, PEMT_SMT and HT there is only about 50% similarity, and in the case of NMT, PEMT_NMT and HT, there is only about 54% similarity, as expected since these were two different translation techniques (post-editing of MT output and human translation). However, in the case of post-editing, a much higher agreement and/or text similarity was assumed, which was not confirmed. In the case of PEMT_SMT and PEMT_NMT, only about 61% text similarity was observed. Even the text similarity between two post-edited MT outputs was not high; it achieved only 75% in the case of post-editing of NMT produced by GT. However, the NMT error rate dropped from about 69% to about 58% when the post-edited SMT was used as a reference.

Relying solely on the reference when determining MT quality turns out to be insufficient, but in combination with residuals, it provides more reliable results, and/or a more objective view of MT quality and the comparison of SMT and NMT.

A significant contribution of residual analysis is the identification of specific segments, in our case short texts, in which neural MT achieved a significantly lower error rate, but mainly in the identification of segments in which, on the contrary, statistical MT achieved better results, regardless of MT systems and language directions, with a focus on machine translation into inflectional and low-resourced Slovak.

## References

1. Wu, Y. & Qin, Y. Machine translation of English speech: Comparison of multiple algorithms. *Journal of Intelligent Systems* **31**, 159–167 (2022).

2. Sharma, S. *et al.* Machine Translation Systems Based on Classical-Statistical-Deep-Learning Approaches. *Electronics (Basel)* **12**, 1716 (2023).

3. Zhou, M., Duan, N., Liu, S. & Shum, H. Y. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering* vol. 6 275–290 Preprint at https://doi.org/10.1016/j.eng.2019.12.014 (2020).

4. Liu, S. & Zhu, W. An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems. *Applied Artificial Intelligence* **37**, (2023).

5. Ghorbani, B. *et al.* Scaling Laws for Neural Machine Translation. Preprint at (2021).

6. Lee, S. *et al.* A Survey on Evaluation Metrics for Machine Translation. *Mathematics* **11**, 1006 (2023).

7. Papineni, K., Roukos, S., Ward, T. & Zhu, W. BLEU: a method for automatic evaluation of machine translation. in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (Philadelphia, 2002).

8. Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. A study of translation edit rate with targeted human annotation. in *Proceedings of Association for Machine Translation in the Americas* 223–231 (2006).

9. Lavie, A. Evaluating the Output of Machine Translation Systems. in *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts* (Xiamen, China, 2011).

10. Tatman, R. Evaluating Text Output in NLP: BLEU at your own risk. *Towards Data Science* https://towardsdatascience.com/evaluating-text-output-in-nlp-bleu-at-your-own-risk-e8609665a213 (2019).

11. Mathur, N., Baldwin, T. & Cohn, T. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 4984–4997 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.acl-main.448.

12. Callison-Burch, C., Koehn, P. & Osborne, M. Improved statistical machine translation using paraphrases. in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* - 17–24 (Association for Computational Linguistics, Morristown, NJ, USA, 2006). doi:10.3115/1220835.1220838.

13. Machacek, M. & Bojar, O. Results of the WMT14 Metrics Shared Task. in *Proceedings of the Ninth Workshop on Statistical Machine Translation* 293–301 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2014). doi:10.3115/v1/W14-3336.

14. Stanojević, M., Kamran, A., Koehn, P. & Bojar, O. Results of the WMT15 Metrics Shared Task. in *Proceedings of the Tenth Workshop on Statistical Machine Translation* 256–273 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2015). doi:10.18653/v1/W15-3031.

15. Bojar, O., Graham, Y. & Kamran, A. Results of the WMT17 Metrics Shared Task. in *Proceedings of the Second Conference on Machine Translation* 489–513 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2017). doi:10.18653/v1/W17-4755.

16. Post, M. A Call for Clarity in Reporting BLEU Scores. (2018).

17. Nießen, S., Och, F. J., Leusch, G. & Ney, H. An evaluation tool for machine translation: Fast evaluation for MT research. in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)* 39–45 (2000).

18. Popović, M. & Ney, H. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. in *Proceedings of the Second Workshop on Statistical Machine Translation* 48–55 (Association for Computational Linguistics, Prague, Czech Republic, 2007).

19. Sai, A. B., Mohankumar, A. K. & Khapra, M. M. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput Surv* **55**, 1–39 (2023).

20. Popović, M. chrF++: words helping character n-grams. in *Proceedings of the Second Conference on Machine Translation* 612–618 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2017). doi:10.18653/v1/W17-4770.

21. Wang, W., Peter, J.-T., Rosendahl, H. & Ney, H. CharacTer: Translation Edit Rate on Character Level. in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers* 505–510 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2016). doi:10.18653/v1/W16-2342.

22. Rei, R., Stewart, C., Farinha, A. C. & Lavie, A. COMET: A Neural Framework for MT Evaluation. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2685–2702 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.emnlp-main.213.

23. Alvarez-Vidal, S. & Oliver, A. Assessing MT with measures of PE effort. *Ampersand* **11**, 100125 (2023).

24. Marie, B., Fujita, A. & Rubino, R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 7297–7306 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2021). doi:10.18653/v1/2021.acl-long.566.

25. Munkova, D., Munk, M., Benko, Ľ. & Hajek, P. The role of automated evaluation techniques in online professional translator training. *PeerJ Comput Sci* **7**, e706 (2021).

26. Google. Google Translate API - Fast Dynamic Localization — Google Cloud Platform. https://cloud.google.com/translate/ (2016).

27. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. in *Proceedings of the MT Summit* vol. 5 79–86 (Phuket Island, 2005).

28. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. (2016).

29. eTranslation. https://webgate.ec.europa.eu/etranslation (2023).

30. Turovsky, B. Found in translation: More accurate, fluent sentences in Google Translate. *The Keyword Google Blog* https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ (2016).

31. Sheshadri, S. K., Gupta, D. & Costa-Jussà, M. R. A Voyage on Neural Machine Translation for Indic Languages. *Procedia Comput Sci* **218**, 2694–2712 (2023).

32. Pinnis, M., Krišlauks, R., Deksne, D. & Miks, T. Evaluation of neural machine translation for highly inflected and small languages. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 10762 LNCS 445–456 (Springer Verlag, 2018).

33. Yang, K., Liu, D., Qu, Q., Sang, Y. & Lv, J. An automatic evaluation metric for Ancient-Modern Chinese translation. *Neural Comput Appl* (2020) doi:10.1007/s00521-020-05216-8.

34. Fomicheva, M. & Specia, L. Taking MT Evaluation Metrics to Extremes: Beyond Correlation with Human Judgments. *Computational Linguistics* **45**, (2019).

35. Moghe, N., Sherborne, T., Steedman, M. & Birch, A. Extrinsic Evaluation of Machine Translation Metrics. in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 13060–13078 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2023). doi:10.18653/v1/2023.acl-long.730.

36. Almahasees, Z. M. Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English. *International Journal of Languages, Literature and Linguistics* **3**, 1–4 (2017).

37. Almahasees, Z. M. Assessment of Google and Microsoft Bing Translation of Journalistic Texts. *International Journal of Languages, Literature and Linguistics* **4**, 231–235 (2018).

38. Marzouk, S. & Hansen-Schirra, S. Evaluation of the impact of controlled language on neural machine translation compared to other MT architectures. *Machine Translation* **33**, (2019).

39. Li, M. & Wang, M. Optimizing Automatic Evaluation of Machine Translation with the ListMLE Approach. *ACM Transactions on Asian and Low-Resource Language Information Processing* **18**, (2019).

40. Singh, S. M. & Singh, T. D. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Syst Appl* **209**, 118187 (2022).

41. Shterionov, D. *et al.* Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* **32**, (2018).

42. Tryhubyshyn, I., Tamchyna, A. & Bojar, O. Bad MT Systems are Good for Quality Estimation. in *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track* 200–208 (Asia-Pacific Association for Machine Translation, Macau SAR, China, 2023).

43. Kosta, P. Targets, Theory and Methods of Slavic Generative Syntax: Minimalism, Negation and Clitics. In: Kempgen, Sebastian / Kosta, Peter / Berger, Tilman / Gutschmidt, Karl (eds.). Slavic Languages. Slavische Sprachen. An International Handbook of their Structure. in *Slavic Languages. Slavische Sprachen. An International Handbook of their Structure, their History and their Investigation. Ein internationales Handbuch ihrer Struktur, ihrer Geschichte und ihrer Erforschung.* (eds. Kempgen, S., Kosta, P., Berger, T. & Gutschmidt, K.) 282–316 (Berlin, New York: Mouton. de Gruyter, 2009).

44. Benko, Ľ. & Munková, D. Application of POS Tagging in Machine Translation Evaluation. in *DIVAI 2016 : 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, May 2 – 4, 2016* 471–489 (Wolters Kluwer, ISSN 2464-7489, Sturovo, 2016).

45. Munková, D., Kapusta, J. & Drlík, M. System for Post-Editing and Automatic Error Classification of Machine Translation. in *DIVAI 2016 : 11th International Scientific Conference on Distance Learning in Applied Informatics, Sturovo, May 2 – 4, 2016* 571–579 (Wolters Kluwer, ISSN 2464-7489, Sturovo, 2016).

46. Munková, D., Munk, M., Benko, Ľ. & Absolon, J. From Old Fashioned "One Size Fits All" to Tailor Made Online Training. in *Advances in Intelligent Systems and Computing* vol. 916 365–376 (Springer Verlag, 2020).

47. Kapusta, J., Benko, Ľ., Munkova, D. & Munk, M. Analysis of Edit Operations for Post-editing Systems. *International Journal of Computational Intelligence Systems* **14**, 197 (2021).

48. Varga, D. *et al.* Parallel corpora for medium density languages. *Proceedings of the RANLP 2005* 590–596 (2005).

49. Benko, Ľ., Munkova, D., Munk, M., Benková, L. & Hájek, P. Dataset of evaluation error-rate metrics for journalistic texts EN/SK and DE/SK. *Mendeley Data* **V1**, (2024).

50. Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 101–108 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.acl-demos.14.

51. Munk, M., Pilkova, A., Benko, L., Blazekova, P. & Svec, P. Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Syst Appl* **185**, 115503 (2021).

52. Munkova, D., Munk, M., Ľubomír Benko & Stastny, J. MT Evaluation in the Context of Language Complexity. *Complexity* **2021**, 1–15 (2021).

53. Munkova, D., Munk, M., Welnitzova, K. & Jakabovicova, J. Product and Process Analysis of Machine Translation into the Inflectional Language. *Sage Open* **11**, 215824402110545 (2021).

54. Munk, M., Munkova, D. & Benko, L. Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics. *Journal of Intelligent & Fuzzy Systems* **34**, 3225–3233 (2018).

55. Vaňko, J. Kategoriálny rámec pre analýzu chýb strojového prekladu. in *Mýliť sa je ľudské (ale aj strojové)* (eds. Munkova, D. & Vaňko, J.) 83–100 (UKF v Nitre, Nitra, 2017).

56. Welnitzova, K. POST-EDITING OF PUBLICISTIC TEXTS IN THE CONTEXT OF THINKING AND EDITING TIME. in *7th SWS International Scientific Conference on Arts and Humanities - ISCAH 2020, 25-27 August, 2020* (STEF92Technology, Sofia, 2020). doi:10.5593/sws.iscah.2020.7.1/s26.29.

57. Panisova, L. & Munkova, D. Peculiarities of Machine Translation of Newspaper Articles from English to Slovak. in *Forlang: cudzie jazyky v akademickom prostredí : periodický zborník vedeckých príspevkov a odborných článkov z medzinárodnej vedeckej konferencie konanej 23. - 24. júna 2021* 281–290 (Technická univerzita, Kosice, Kosice, Slovakia, 2021).

58. Skadiņš, R., Goba, K. & Šics, V. Improving SMT for Baltic Languages with Factored Models. *Frontiers in Artificial Intelligence and Applications* **219**, 125–132 (2010).

59. Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Comput Speech Lang* **49**, 52–70 (2018).

60. Volkart, L., Bouillon, P. & Girletti, S. Statistical vs. Neural Machine Translation: A Comparison of MTH and DeepL at Swiss Post's Language Service. in *Proceedings of the 40th Conference Translating and the Computer* 145–150 (London, UK, 2018).

61. Jassem, K. & Dwojak, T. Statistical versus neural machine translation - a case study for a medium size domain-specific bilingual corpus. *Poznan Studies in Contemporary Linguistics* **55**, 491–515 (2019).

62. Hasan, Md. A., Alam, F., Chowdhury, S. A. & Khan, N. Neural vs Statistical Machine Translation: Revisiting the Bangla-English Language Pair. in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)* 1–5 (IEEE, 2019). doi:10.1109/ICBSLP47725.2019.201502.

63. Benkova, L., Munkova, D., Benko, Ľ. & Munk, M. Evaluation of English–Slovak Neural and Statistical Machine Translation. *Applied Sciences* **11**, (2021).

## Acknowledgements

This work was supported by the Slovak Research and Development Agency under contract No. APVV-18-0473.

## Competing interests

The authors declare no competing interests.

## Data availability

The dataset analysed during the current study is available in the Mendely Data repository under doi: 10.17632/yrft7c64z6.1 .

## Contributions

All authors contributed to the study conception and design. Ľ.B. and L.B. performed data collection, metrics calculation and designed the methodology. M.M. conducted the data analysis and prepared Fig. 2-7 and all Tables. D.M. and P.H. performed the results interpretation and supervised the experiment. The first draft of the manuscript was written by D.M. and L.B. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

PRÍLOHA L: BENKO, ĽUBOMÍR, DASA MUNKOVA, MÁRIA PAPPOVÁ A MICHAL MUNK, 2024B. COMPARISON OF VARIOUS APPROACHES TO TAGGING FOR THE INFLECTIONAL SLOVAK LANGUAGE. *PEERJ COMPUTER SCIENCE* (*V RECENZNOM KONANÍ OD 2023, 2. KOLO*) (**WEB OF SCIENCE, 2022IF: 3.8, Q2**)

# Comparison of various approaches to tagging for the inflectional Slovak Language

Ľubomír Benko[1], Dasa Munkova[1], Mária Pappová[1], Michal Munk[1,2]

[1] Department of Computer Science, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, SK 949 01, Nitra, Slovakia
[2] Science and Research Centre, University of Pardubice, Studentská 84, CZ 532 10, Pardubice, Czech Republic

Corresponding Author:
Ľubomír Benko[1]
Tr. A. Hlinku 1, Nitra, SK 949 01, Slovakia
Email address: lbenko@ukf.sk

## Abstract

Morphological tagging provides essential insights into grammar, structure, and the mutual relationships of words within the sentence. Tagging text in a highly inflectional language presents a challenging task due to word ambiguity. This research aims to compare six different automatic taggers for the inflectional Slovak language, seeking for the most accurate tagger for literary and non-literary texts. Our results indicate that it is useful to differentiate texts into literary and non-literary and subsequently, based on the text style to deploy a tagger. For literary texts, UDPipe2 outperformed others in seven out of nine examined tagset positions. Conversely, for non-literary texts, the RNNTagger exhibited the highest performance in eight out of nine examined tagset positions. The RNNTagger is recommended for both types of the text, the best captures the inflection of the Slovak language, but UDPipe2 demonstrates a higher accuracy for literary texts. Despite dataset size limitations, this study emphasizes the suitability of various taggers for the inflectional languages like Slovak.

## Introduction

Part-of-speech (POS) tagging is one of the most essential tasks of natural language processing (NLP), aiming to assign the correct syntactic label to each word in the context of its appearance. It is an automatic text annotation process, in which assigned words or tokens correspond to the main word class categories (adjectives, nouns, verbs etc.), while they are mutually distinguished by morphosyntactic features (gender, tense, number, etc.). Together with lemmatization, both are fundamental tasks, and/or steps of linguistic pre-processing, which can be later used in NLP tasks such as machine translation, word sense disambiguation, question-answering analysis, etc. The genesis of POS tagging is based on the ambiguity of many words regarding their POS in context.

Morphology (with all its complexity) is ubiquitous among languages, which motivates researchers to design universal schemes with universal tags, such as UniMorph, or focus researchers on projects aiming at tagset universalization, such as the Universal Dependencies (UD) project or Interset for inflectional morphology,

including low-resource languages (Kirov et al., 2018; Karyukin et al., 2023) such as Slovak. The idea behind this is that a set of syntactic POS categories – universals, exists in similar form across languages, i.e. they represent their cross-lingual nature (Petrov, Das & McDonald, 2012). The UD project provides a token-level corpus complementary to the UniMorph type-level data (Kirov et al., 2018).

CLARIN (Common Language Resources and Technology Infrastructure) is a digital infrastructure governed by the European Research Infrastructure Consortium, established by the European Commission in 2009. Clarin provides access to a broad range of tools (and language data) to support research in the humanities and beyond (Branco et al., 2023). It offers 68 tools for part-of-speech tagging for a single language and also for multiple languages, including Slovak (Sparv, which is Språkbanken's corpus annotation pipeline infrastructure; GENIA or STEPP Tagger for annotating biomedical texts, and MorphoDiTa).

The Slovak language belongs to a family of a highly inflectional languages with complex rules for word formation and inflection. Due to the many possible word forms, classifying context for tasks like POS tagging, lemmatization, or semantic analysis is more challenging and requires larger search space and more complex classifier training (Hladek, Stas & Juhar, 2015). The Slovak National Corpus (SNC) (Horák et al., 2004) is a morphological annotated and lemmatized corpus, consisting of two sub-corpora. Both are annotated, but the smaller sub-corpus (r-mak) is annotated manually while the larger is annotated and lemmatized automatically. The tagset, designed within SNC is both positional and attributive; tags are of unequal length following inflectional paradigm, which describes the morphological (inflectional) behaviour of the word (Garabík & Šimková, 2012).

## Motivation

Manual assigning a POS tag to each word in text is very time and labour-consuming. This leads to the existence of various approaches and methods to automate the task, where the overall process takes a word or a sentence as input, assigns a POS tag to the word or each word in the sentence, and creates a tagged text as the output (Hladek, Stas & Juhar, 2015).

Only few POS tagging algorithms and tools exist which can be deployed for low-resource languages and inflectional Slavic languages. It motivated us to conduct our research, focusing on the efficiency of these algorithms and tools especially for the Slovak language, which does not only belong to above-mentioned language families, but it is also one of official European Union languages.

The aim of this paper is to compare six different automatic taggers for the inflectional Slovak language. We attempt to find the best performing tagger in terms of accuracy.

There are some studies focusing on evaluation of new proposed taggers. For example, Straka et al. (2019) evaluated contextualized embeddings (UDPipe2) on 54 languages in POS tagging or Qi et al. (2020) built on the highly accurate neural network components that enable efficient training and evaluation for more than 70 languages (Stanza). Even few of them focus on Slovak, they do not compare the taggers mutually and do not distinguish between literary and non-literary texts. Moreover, they evaluated tagger accuracy using the F1 measure for the entire tagset, which provides only an overall accuracy score with gold tokenization, but lacks detailed linguistic information. In our research, we evaluated the accuracy for each position within a 15-positional tagset, enabling us to capture various grammatical aspects of the language. By categorizing the texts into literary and non-literary style, we were able to compare taggers across different text styles, a comparison that none of the studies undertook. Therefore, this study attempts to fill this gap in literature and research focusing on POS taggers for the Slovak literary and non-literary texts.

## Contribution

The theoretical contribution of our research consists in designed methodology, how to compare automatic taggers with different output formats (POS tags).

The practical contribution lies in the verification of the effectiveness of six available automatic taggers for an inflectional Slovak language.

For automatic linguistic annotation of the Slovak text we recommend to use RNNTagger and UDPipe2 regardless of text type. Both tools best capture the inflection of the Slovak language. Moreover, our results indicate that for linguistic analysis of non-literary texts, the best approach is to combine RNNTagger with UDPipe2 (the latter mentioned mainly when determining tense and negation). However, for linguistic analysis of literary texts, the best approach is also to combine these two taggers, but in reverse order, i.e. only in the case of number and person we should prefer the RNNTagger over UDPipe2.

The structure of the paper is as follows. The POS tagging algorithms section briefly describes principles of rule-based taggers and stochastic taggers. The related work section summarizes POS tagging for low-resource languages and Slavic languages. The POS taggers section describes the most known and used taggers, suitable for POS tagging of Slovak texts. Materials and methods section covers used dataset, selected automatic annotation tools, and applied research methodology. The results and discussion sections focus on the research results and their interpretations based on the performance of the taggers in terms of accuracy. The conclusion section summarizes our findings.

## POS Tagging Algorithms

Currently, the process of morphological language analysis is often performed in two steps. The first step is the analysis itself, which involves assigning to each word a list of possible combinations of lemma and morphological tags. The second step is unification, where one (if possible, correct) combination of lemma-tag is selected. The analysis typically consists of selecting entries from a database of inflected word forms, followed by guessing the lemma and/or tags for words outside the dictionary. The second step is often executed using statistical methods, which require training on manually annotated corpora (Garabík & Šimková, 2012).

Algorithms providing POS tagging can be grouped into rule-based taggers and stochastic taggers. Rule-based taggers – require lexical knowledge – involve a large database of handwritten disambiguation rules based on the formal syntax of the given language; on the other hand, stochastic taggers – demand high computational resources – resolve tagging ambiguities using a training set to calculate the probability that a given word has a particular tag in a specific context (Izzi & Ferilli, 2020). Stochastic POS taggers do not rely on syntactic analysis of the input, but on the Hidden Markov model (HMM), which captures the lexical and contextual information.

HMM is a doubly stochastic process with a basic probabilistic process that is not observable (it is hidden), but it can only be observed through another set of stochastic processes that create a sequence of observed symbols (Rabiner & Juang, 1986; Jurafsky & Martin, 2020). Information about the model's state can be obtained from the probability distribution within possible output tokens, as each state of the model creates a different distribution. The sequence of output tokens provides an overview of the state sequence in the process known as pattern theory. However, algorithms associated with HMMs are efficient for performing tasks in many real-time systems; they are often applied in speech recognition, signal processing, and in some low-level NLP tasks such as morphological annotation, information extraction from documents, or speech-to-text conversion in speech recognition (Fink, 2008). In HMM words are treated as the observed events (e.g., words, that could be seen in the input) and hidden events (e.g. their parts of speech), which can

be considered as causal factors in the probabilistic model (Blunsom, 2004). We treat POS tags as hidden events and individual words as observed events.

HMM consists of triple parameters $\lambda = (A, B, \pi)$ defined on the set of states $Q$ and emissions $V$ (Blunsom, 2004). Let $Q$ be a set of states $Q = \{q_i\}_{i=1}^{N}$ and $V$ be a set of emissions $V = \{v_i\}_{i=1}^{N}$ where $\boldsymbol{\pi}$ is the probability distribution function denoting the probability $\pi(q_i) = P(S_1 = q_i)$, $A$ is an $NxN$ matrix (called the state transition matrix), where entry $a_{ij}$ is given by $a_{ij} = P(S_{t+1} = q_j \,|\, S_t = q_i)$ and $B$ is an $NxM$ matrix (called the emission matrix), where entry $b_{ij} = P(O_t = v_j \,|\, S_t = q_i)$. The parameters within HMM can be estimated from the tagged or untagged words or tokens.

## Related Work

Tagging text in a highly inflectional language is a complex task due to word ambiguity, resulting in many homographs and, due to segmentation, into a set of morphemes (Alosaimy & Atwell, 2018). Morphological tagging offers basic information about the grammar, and/or text structure and relationships among words within the sentence. Low-resourced morphological tagging is gaining increasing recognition (Afanasev, 2023). The current shift to a language-independent approach for morphological disambiguation is regarded as an extension of POS tagging, jointly predicting complex morphological tags (Toleu, Tolegen & Mussabayev, 2022).

There are different pre-trained monolingual and multilingual models that are used for the morphological tagging, but most of them are too universal or underprepared for low resource languages, expect for the Stanza tagger (Afanasev, 2023). That is why we also decided to employed Stanza in our research. Afanasev (2023) compared Stanza to UDPipe taggers for Belarusian-Khislavichi and also for Russian-Taiga languages, all belonging to the East Slavic family, and found that a modified Stanza tagger provides more effective tagging than UDPipe. Ljubešić and Dobrovoljc (2019) conducted an experiment with three Slavic – Slovenian, Croatian, and Serbian – morphosyntactic taggers and compared two state-of-the-art tools with different architecture, traditional Reldi-tagger with a modified neural Stanza (stanfordnlp+lex). They showed that the neural Stanza yields significant improvements in tagging compared to the traditional approach. Fehle et al. (2021) evaluated two POS-taggers for German - TreeTagger and Stanza. Concerning POS-tagging, they showed very few differences.

Spoustova et al. (2009) focused on evaluating the quality of morphological annotation generated by several different POS taggers. The quality assessment was conducted for the tools HMM tagger, Morče (the predecessor of the MorphoDiTa tagger), and Feature-Based Tagger. The results of the POS taggers were categorized into three cases: correct annotation, incorrect annotation, and vague annotation. The main contribution of the research was the methodology for identifying problematic tags without the need for a human-annotated baseline.

Rosen et al. (2014) dealt with the annotation scheme of texts produced by non-native speakers of Czech. The authors not only focused on manual annotation but also conducted experiments with automated linguistic annotation tools. The results were compared for a spell checker Korektor (Richter, 2010) and the POS tagger Morče, aiming to identify errors in automated tagging. Furthermore, the authors compared two taggers: Morče and TnT (Brants, 2000). The results demonstrated that TnT faced challenges in a context with many errors, but performed better than Morče on faulty forms. On the other hand, Morče exhibited a strong preference for verbs and demonstrated better overall performance.

Machura et al. (2019) compared the Czech morphological taggers MorphoDiTa and Majka Tagger. The experimental results indicated higher precision and recall for MorphoDiTa. During the experiment, the

authors enhanced the MorphoDiTa tool and significantly improved its accuracy. The authors examined the differences in the Czech language. However, the conclusion of the study emphasized that the input text has a significant impact on the quality of the output.
Straka and Strakova (2017) compared different versions of UDPipe and its subsequent enhancements. The taggers were evaluated using the TIRA platform for the CoNLL 2017 UD Shared Task, where all inputs were plain text, and the results were based on F1 scores. Overall, the system upgrades demonstrated improvements in POS tag annotation. The authors utilized the old version of UDPipe as a baseline. Straka (Straka, 2018) continued to improve the model by refining the model architecture, resulting in enhanced performance. The evaluation was not limited to POS tagging but also included lemmatization for the tool in the CoNLL 2018 UD Shared Task. The results of the competition motivated the author to further enhance the tool in the future.

## POS Taggers

### TreeTagger

TreeTagger is designed to analyze the morphological and syntactic structure of a sentence and assign parts of speech and lemmas to each word (21). It can be used for multiple languages, including Slovak, and is adaptable to other languages if a lexicon and manually annotated corpus are available. It is primarily based on decision trees guided by modified ID3 algorithms. The tree itself is modeled recursively on a training data sample, mainly consisting of trigrams. It combines rule-based and stochastic algorithms and uses a set of rules to identify possible parts of speech for each word in the text based on its morphological and contextual properties. These rules are then applied within a probabilistic framework to determine the most probable part of speech for each word.

### MorphoDita

Morphological Dictionary and Tagger (MorphoDiTa) is an open-source tool for morphological text analysis of natural language. It was developed within the LINDAT project by (Straka & Straková, 2014). It is one of the most widely used tools for morphological analysis focusing on English, Czech, and Slovak. It is based on a combination of rule-based algorithms and machine learning. Besides morphological analysis, it performs morphological generation, tagging, and tokenization. It is distributed as a standalone tool or a library, along with trained linguistic models (ibid). Its predecessors are the tagging library Morče and Featurama (Spoustová et al., 2009). The tagger is implemented as a supervised, averaged perceptron. It further utilizes two main machine learning algorithms:

- Morphological analyzer (Independent Feature Selection Classifier) for distinguishing various morphological features (e.g. number, case, gender, etc.). The classifier is trained on a large source of annotated data.
- Dependency parsing for deep syntactic analysis. It identifies relationships between words/tokens in the text and creates a tree structure of dependencies.

MorphoDiTa estimates regular patterns based on affixes, common morpheme endings, and automatically groups them into morphological "templates" without language-specific knowledge. MorphoDiTa Online operates on the same principle as the library itself, available for various operation systems (Linux/Windows/OS X) and various programming languages (C++, Python, Perl, Java, and C#). The trained model is available from 2017, and extensive changes and updates have only taken place within the Czech models.

**UDPipe2**

UDPipe2, similar to its predecessor UDPipe1, is a language-agnostic, trainable pipeline performing POS tagging, lemmatization, and dependency syntactic parsing using CoNLL-U format (Straka & Straková, 2017; Straka, 2018). UDPipe2 compare to UDPipe1 is a Python prototype. Trained models are available for almost all Universal Dependencies (UD) corpora. UDPipe2 utilizes multiple machine learning algorithms for morphological and syntactic analysis of texts. Specifically, the tool employs algorithms based on deep neural networks, allowing the tool to learn from data and patterns, creating more precise and efficient models for text analysis. It includes tokenizer, POS tagger, lemmatizer, and parser models for 99 treebanks of 63 languages of Universal Depenencies 2.6 Treebanks, created solely using UD 2.6 data (Straka, 2020).

**RNNTagger**

RNNTagger was developed in 2019, primarily aiming at morphological annotation of historical texts that preserve a certain dialect that many libraries struggled with (Schmid, 2019). It is a tool for annotating text with POS and lemma information. It is a type of sequence labeling model that utilizes recurrent neural networks, specifically Bi-LSTMs, to predict tags for each element in the sequence. It is a neural POS tagger implemented in Python using the PyTorch deep learning library. The model takes a sequence of words as input and processes them one by one while maintaining a hidden state that captures information about the context of previous words. The hidden state is updated at each time step using the current input and the previous hidden state, allowing the model to learn dependencies between words in the sequence. Compared to TreeTagger, RNNTagger lemmatizes all tokens, but requires Python and PyTorch. RNNTagger tries to lemmatize all words, including unknown words, compared to TreeTagger, which uses the word form, and RNNTagger suffers from attempting to lemmatize non-inflected tokens (Proisl et al., 2020).

**Stanza**

Stanza is an open-source Python NLP toolkit supporting many human languages (Qi et al., 2020). Stanza is built on top of the PyTorch library and utilizes deep learning models in the form of neural pipelines, where each phase is implemented using a deep neural network model trained on a large amount of annotated data to perform the corresponding task. The outputs of each phase are used as inputs for the next phase, allowing the pipeline to sequentially process the input text and produce a whole range of outputs, such as syntactic and semantic representations of the text. One of the key advantages of Stanza is its ease of use, contributing to its high popularity. It provides a simple and consistent interface for performing NLP tasks, making it accessible to users with varying levels of expertise. Additionally, pre-trained models are customizable, allowing users to fine-tune them based on their data. Stanza is used in various applications, including social media analysis, machine translation, and information extraction. Compared to UDPipe, Stanza supports 66 languages and is fully neural.

## Materials & Methods

### POS Categories

The part of speech category (POS) is fundamental, as each morphological interpretation of a word form is assigned a POS value (Petkevič et al., 2019). Tagsets for different languages are typically different. They can be entirely different for unrelated languages and very similar for related languages, but this is not always the rule. Tagsets can also vary in levels of granularity. Basic tagsets may contain tags for the most common parts of speech (N for noun, V for verb, A for adjective, etc.) (Universal Dependencies contributors, 2022). However, it is more common to go into detail and differentiate between singular and plural nouns, verb

phrases, tenses, aspects, voices, and more. Petrov et al. (2012) proposed a tagset consisting of twelve universal POS categories for 22 different languages, including Czech, but not Slovak: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV, (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (abbreviations or foreign words). Hajic developed the Czech Prague Dependency Treebank with its tagset format (PDT tagset), designed for the needs of Slavic languages (Hajič, 2006; Bejček & Straňák, 2010). The PDT tagset uses three layers of annotation – morphological, syntactical, and tectogrammatical. It is a string of 15 characters that can more precisely determine the meaning of the tagged word; one character symbol encodes one morphological category (Hajič et al., 2020). The PDT tagset consists of a fixed length, and each position encodes one grammatical category, while two positions (13th and 14th) are empty. The attributes in positional tags are as follows: the 1st position – part of speech, 2nd – a detailed part of speech, 3rd – gender, 4th – number, 5th – case, 6th – possessor´s gender, 7th – possessor´s number, 8th – person, 9th – tense, 10th – degree of comparison, 11th – negation, 12th – voice, 13th – empty, 14th – empty, and the last position, 15th – variant and style.

According to general principles, two main groups of tagsets are recognized - UPOS and XPOS. The Universal POS tag (UPOS) is represented by tags indicating the basic categories of parts of speech. Universal POS tags are categorized into open class words, closed class words, and others. Under open class words, words can be continually added and modified. Over the decades, words like "smartphone," "selfie," or "e-sport" have been added to nouns. On the other hand, closed class words represents a group of words that will be universally valid and immutable (Universal Dependencies contributors, 2022). Under the others category, various symbols, punctuation marks, or symbols for unrecognizable words that, for example, the model cannot identify, are recognized.

The abbreviation XPOS refers to a language-specific POS tag specific to a given language (e.g. English: Language-specific POS). Unique rules for encoding XPOS are defined by each library. A single character, which can be a letter of the Latin alphabet, a digit, or a mathematical symbol, represents the values of individual categories. Consequently, each letter corresponds to only one value, even across parts of speech. An exception is observed in the paradigm category, which reuses the part of speech code (Garábik & Bobeková, 2021). One tag (label) for one token and lemma is formed by a set of these characters.

## Dataset

Obtaining a dataset that both, best captures the diversity of the Slovak language and is sufficiently complex for evaluating tools specialized in the Slovak language tagging was challenging. Manually annotated data are crucial for training and evaluating statistical tools such as POS taggers and lemmatizers (Proisl et al., 2020). For this purpose, a manually annotated and lemmatized sub-corpus of the Slovak Dependency Corpus (SDC) was chosen (Gajdošová & Šimková, 2016). The dataset consists of 10,604 sentences and 106,043 tokens. The annotation follows the guidelines of the Prague Dependency Treebank (PDT) (Hajič, 2006) slightly modified to align with Slovak grammatical rules. Morphological tags, lemmas, and dependencies were manually assigned to each word. The sub-corpus includes only sentences in which two human annotators perfectly agreed on the tag. A drawback of the dataset is that it mainly contains short sentences (Benko & Benková, 2022). The dataset also only includes surface-dependent (analytical) trees and does not encompass a deep syntactic/semantic (tectogrammatical) layer (Majchráková et al., 2014). The primary starting point for the Slovak annotation is the functional-generative approach. Unlike the PDT, which exclusively contains journalistic texts, the stylistic-genre structure of SDC is more diverse (Šimková & Gajdošová, 2008). The texts were divided into two text types: non-literary and literary. Literary texts

consist of novels and fairy tales. Non-literary texts consist of historical texts, texts obtained from Wikipedia, and journalistic texts.

The data file is available in a specific dictionary-like format (CoNLL-X format) used for text processing and morphological annotation. In the CoNLL-X format, each word in a sentence is represented as a row with various columns of information, including the word form, part of speech, lemma, and syntactic head. Every dictionary created for machine learning and deep learning is stored in this format.

A more detailed description of individual columns can be found in (Gajdošová & Šimková, 2016). For research purposes, only "POSTAG" column was investigated, while it contains manually annotated words from the "FORM" column. This type of data cannot be analyzed using the automatic tools, so it was necessary to reconstruct these data to the original format. The individual words were extracted from the file and reconstructed into sentence structures, which were then implemented as input for the examined automatic taggers.

## Taggers

Open-source libraries or tools for morphological tagging were chosen as taggers for the Slovak language. However, upon closer examination, it was revealed that another library is already utilized by a given library or tool to process NLP for the Slovak language. Among such online tools, Sparv by Swedish developers (Hammarstedt et al., 2022) can be mentioned; it uses the Stanza tagging library for morphological annotation of Slovak. Their priority was primarily the analysis of the Swedish Språkbanken corpus. Another library is the GENIA Tagger by Tsuruoka et al. (2005), which annotates the Slovak language, but is only intended for biomedical texts, as is the STEPP Tagger (Piao, Tsuruoka & Ananiadou, 2009), which originated at the University of Tokyo and was further presented in 2012. Lastly, the Turku Neural Parse Pipeline by Finnish authors (Kanerva et al., 2018) is mentioned, which offers morphological annotation for over 50 languages, but the tagset for the Slovak language was not available. Finally, a library is mentioned that morphologically annotates the Slovak language, but its implementation was unsuccessful. The first such library was Dagger, which originated at the Technical University of Košice and was based on the principle of HMM, specifically the Viterbi algorithm and binary decision trees, particularly the ID3 algorithm (Hládek, Staš & Juhár, 2012). This library could not be implemented because it was no longer supported by newer versions of the Linux operating system. Another problematic tool was the RFTagger (Schmid & Laws, 2008), based on the HMM and decision trees method. It is suitable mainly for POS tagsets with many fine-grained tags, i.e., those that contain more details and distinguish between various subtypes of parts-of-speech and grammatical categories. Despite the excellent results achieved by this library on the German Tiger Treebank, it was not possible to annotate the entire text correctly, leading to the exclusion of RFTagger from further evaluation. For this reason, the following automated taggers were selected for comparison: TreeTagger, RNNTagger, Stanza, MorphoDiTa (application and online version will be separated), and UDPipe2.

## Applied Methodology

The applied methodology, inspired by other research (Hochreiter & Schmidhuber, 1997; Huang, Xu & Yu, 2015; Yao & Huang, 2016; Benkova et al., 2021; Munkova et al., 2021a,b; Kapusta et al., 2021), comprises the following steps (Fig. 1):

(1) Acquisition of dataset with manual morphological annotation – a gold tokenization (Gajdošová & Šimková, 2016). This gold tokenization is positional and attributive, and the tags are of unequal length following inflectional paradigm. It contains 85,929 tokens.

(2) Data preparation – Sentence reconstruction. Since the acquired dataset was already tokenized, we converted the tokenized texts into the original text forms, i.e., to sentences, giving us 10,604

sentences (85,929 tokens). Afterwards, we divided texts according to the text type into literary texts (45,819 tokens) and non-literary texts (40,110 tokens).

(3) Automatic POS tagging – converted texts (literary and non-literary texts) were annotated using the following tools:
   (a) RNNTagger (using the model, trained on the SNC),
   (b) TreeTagger (using the model Slovak parameter file, trained on the SNC),
   (c) MorphoDiTa (using the model slovak-morfflex-pdt-170914 (Hajič & Hric, 2017), trained on SNC),
   (d) MorphoDiTa online (using the model slovak-morfflex-pdt-170914 (Hajič & Hric, 2017)),
   (e) Stanza (we used the model from Universal Dependencies v2.12 (Zeman et al., 2023), trained on SNC),
   (f) UDPipe2 (we used the model slovak-snk-ud-2.12-230717, trained on SNC).

   The tools were run at two different personal computers because RNNTagger requires only Linux operating system. RNNTagger was implemented on ASUS Transformer Book Flip TP300LD (Intel Core i5 4210 Haswell, RAM 12GB DDR3L, NVIDIA GeForce GT 820 2GB, 500 GB SSD, Ubuntu Desktop 22.10). Other tagging tools where implemented on Apple M1 Max (Processor with 10 cores 2.06 – 3.22 GHz, 32-Core GPU, 16-Core Neural Engine, 32 GB RAM, 1 TB SSD, macOS Ventura 13.2.1).

(4) Tagset conversion – the automatic taggers (RNNTagger, TreeTagger, UDPipe2, and Stanza) do not employ the same PDT output tagset format, so their output tagsets need to be converted to the same tagset format. To compare the performance of the investigated taggers in terms of accuracy with the gold tokenization (reference), we unified the output tagset format. Since the tagset used for Slovak has tags of unequal length, i.e., a different number of characters (positions) for each part of speech, we decided to employ a universal, 15-positional tagset for all examined taggers.

(5) Dataset creation – the outputs of the six automated taggers were joint into one data matrix and dummy variables were created for each position (1st – 15th). These variables have a binary character, depending on the tag position match (agreement) with the reference (gold tokenization). Some positions did not contain any tags, so these positions were excluded from the experiment. We created two datasets depending on text type (literary and non-literary dataset).

(6) Data analysis – we applied non-parametric procedures based on both, frequency (Cochran Q test) and ranks, where the degree of concordance was expressed by the Kendall coefficient of concordance.


Figure 1: Methodology workflow diagram

## Results

As described in the methodology, some of the automatic POS taggers did not support the PDT tagset style. During the conversion into the PDT tagset style some empty positions were generated, which resulted in the elimination of the positions 6th, 7th, 13th, 14th and 15th from the experiment. Positions 13th and 14th are empty in general, and the position 6th and 7th within a tagset are associated with the possessor's gender and the possessor's number, and the position 15th with the style.

Since the tagsets, produced by the six examined automatic taggers, representing the indicators of accuracy (agreement with a reference tagset), have a binary character at the relevant tag position, we used non-parametric procedures. Based on frequencies (Cochran Q test) and ranks (Kendall coefficient of concordance), we tested the global null statistical hypotheses, which claim that there is no statistically significant difference in the performance of the investigated taggers in terms of accuracy with reference. We tested the hypothesis for each tag of the 15-positional tagset, except for the five excluded tags (6th, 7th, 13th, 14th, and 15th). We examined the performance of the automatic taggers separately for literary and non-literary texts, to determine whether the text type can affect the performance of given tagger.

**Literary and non-literary texts**

In the case of non-literary texts, based on the results of Cochran Q tests, the global null hypothesis (stating there are no statistically significant differences between individual tags within the tagset produced by examined POS taggers and a reference tags in tagset) is rejected at the 0.001 significance level for tags in the 1st, 2nd, 3rd, 4th, 5th, 8th, 9th, 10th, 11th, and 12th position in the tagset (1st: $N = 45819$, $Q = 18024.87$, $df = 5$, $p < 0.001$; 2nd: $N = 38967$, $Q = 34381.00$, $df = 5$, $p < 0.001$; 3rd: $N = 31285$, $Q = 24833.86$, $df = 5$, $p < 0.001$; 4th: $N = 34125$, $Q = 16198.12$, $df = 5$, $p < 0.001$; 5th: $N = 32531$, $Q = 12497.83$, $df = 5$, $p < 0.001$; 8th: $N = 6672$, $Q = 19756.04$, $df = 5$, $p < 0.001$; 9th: $N = 6169$, $Q = 535.99$, $df = 5$, $p < 0.001$; 10th: $N = 7223$, $Q = 4400.26$, $df = 5$, $p < 0.001$; 11th: $N = 7377$, $Q = 1078.99$, $df = 5$, $p < 0.001$; 12th: $N = 7210$, $Q = 15876.04$, $df = 5$, $p < 0.001$).

Similarly, for the literary texts, based on the results of Cochran Q tests, the global null hypotheses is rejected at the 0.001 significance level for tags in the 1st, 2nd, 3rd, 4th, 5th, 8th, 9th, 10th, 11th, and 12th position (1st: $N = 40110$, $Q = 20352.07$, $df = 5$, $p < 0.001$; 2nd: $N = 28956$, $Q = 40532.13$, $df = 5$, $p < 0.001$; 3rd: $N = 24871$, $Q = 38023.70$, $df = 5$, $p < 0.001$; 4th: $N = 28952$, $Q = 18599.09$, $df = 5$, $p < 0.001$; 5th: $N = 21590$, $Q = 11327.92$, $df = 5$, $p < 0.001$; 8th: $N = 11077$, $Q = 36999.44$, $df = 5$, $p < 0.001$; 9th: $N = 9714$, $Q = 303.93$, $df = 5$, $p < 0.001$; 10th: $N = 4515$, $Q = 7028.828$, $df = 5$, $p < 0.001$; 11th: $N = 11779$, $Q = 4927.626$, $df = 5$, $p < 0.001$; 12th: $N = 11643$, $Q = 30379.50$, $df = 5$, $p < 0.001$).

Our results indicate that there are differences in the accuracy of taggers' performance, whether the text is literary or non-literary. Moreover, these differences are statistically significant.

Non-parametric procedures, which we applied, work with absolute differences of sums of ranks, where critical values were obtained asymptotically. Based on the multiple comparisons, we identified homogeneous groups among which the statistically significant differences (Table 1-10) were proven.

**The 1st tagset position - part of speech**

In addition to the ten traditional parts of speech – noun (N), adjective (A), pronoun (P), verb (V), adverb (D), numeral (C), conjunction (J), preposition (R), interjection (I), and particle (T) – the 1st tagset position distinguishes also the abbreviation (B), foreign word (F), segment (S), isolated letter (Q), and punctuation (Z).

For both types of texts, MorphoDiTa Tagger, with less than 87% (86.65% for non-literary and 85.68% for literary texts), achieved the lowest performance in terms of matching with the reference (Table 1). On the other hand, the RNNTagger, with more than 99%, achieved the highest performance in terms of matching with the reference in the case of non-literary texts, and in the case of literary texts, UDPipe2 (99.26%) achieved the highest accuracy and/or performance with the reference (Table 1b).

For non-literary texts (Table 1a), six trivial single-element homogenous groups were identified among which statistically significant differences were observed ($p < 0.05$).

For literary texts (Table 1b), four trivial single-element homogenous groups and one two-element homogenous group were identified among which statistically significant differences were observed ($p <$

0.05). In terms of agreement with the reference, the RNNTagger and UDPipe2 form one two-element homogenous group ($p > 0.05$).

Table 1:
Ranking of taggers in the 1st tagset position a) non-literary texts, b) literary texts

| a) N = 45819 | 1's (%) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| **MorphoDiTa** | 86.65 | o | | | | | |
| **MorphoDiTa_Online** | 91.10 | | o | | | | |
| Stanza | 97.51 | | | o | | | |
| UDPipe2 | 98.10 | | | | o | | |
| TreeTagger | 98.71 | | | | | o | |
| RNNTagger | 99.37 | | | | | | o |

| b) N = 40110 | 1's (%) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **MorphoDiTa** | 85.68 | | o | | | |
| **MorphoDiTa_Online** | 89.08 | | | o | | |
| TreeTagger | 98.44 | | | | o | |
| Stanza | 98.80 | | | | | o |
| RNNTagger | 99.21 | o | | | | |
| UDPipe2 | 99.26 | o | | | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The taggers' performance for both text types is higher than 97% (except for MorphoDiTa and MorphoDiTa_online), which indicates that the part of speech identification (1st position in the PDT tagset format) is mostly accurate, even though the best performance was obtained from the RNNTagger for both text types (Table 1).

**The 2nd tagset position – a detailed part of speech**

The second tagset position contains values for fine-grained distinction of the major POS category (66 SUBPOS values) which serves as an indicator of applicability/non-applicability of other categories (Mikulová et al., 2020).

Ranking of taggers (Table 2) has again shown that the lowest performance in terms of matching with the reference in the 2nd position was achieved by the MorphoDiTa tagger (75.69% for non-literary texts and 66.30% for literary texts). Similar to the 1st position, the RNNTagger, with more than 99%, has achieved the highest performance in terms of matching with the reference for non-literary texts (Table 2a), and UDPipe2 (99.13%) achieved the highest performance and/or accuracy with the reference for literary texts (Table 2b).

For non-literary texts (Table 2a) five homogenous groups were identified among which statistically significant differences were observed ($p < 0.05$) - four single-element homogenous groups and one two-element homogenous group, which consists of UDPipe2 and TreeTagger.

For literary texts (Table 2b) four homogenous groups (two single-element and two two-element homogenous groups) were identified among which statistically significant differences were observed ($p < 0.05$). The TreeTagger and Stanza (with more than 97% of concordance) form a one two-element homogenous group ($p > 0.05$), and the RNNTagger and UDPipe2, similarly to the first position, form the second two-element homogenous group ($p > 0.05$) in terms of agreement with the reference.

Table 2:
Ranking of taggers in the 2nd position of the tag a) non-literary texts, b) literary texts

| a) N = 38967 | 1's (%) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **MorphoDiTa** | 75.69 | | o | | | |
| **MorphoDiTa_Online** | 80.23 | | | o | | |
| Stanza | 96.71 | | | | o | |
| UDPipe2 | 98.30 | o | | | | |
| TreeTagger | 98.52 | o | | | | |
| RNNTagger | 99.18 | | | | | o |

| b) N = 28956 | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **MorphoDiTa** | 66.30 | | | o | |
| **MorphoDiTa_Online** | 70.01 | | | | o |
| TreeTagger | 97.82 | o | | | |
| Stanza | 97.95 | o | | | |
| RNNTagger | 98.83 | | o | | |
| UDPipe2 | 99.13 | | o | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The performance of the four taggers (Stanza, UDPipe2, TreeTagger, and RNNTagger) was above 96% which indicates that the identification of the detailed part-of-speech (the 2nd tagset position) is very accurate. Moreover, we can observe similar performance for above-mentioned taggers as for the first tagset position, which confirms the link (relationship) between the first and second position.

**The 3rd tagset position – gender**

The third tagset position denotes grammatical gender for both, lexical gender of nouns and agreement gender of verbs, adjectives, pronouns, and numerals (Mikulová et al., 2020).

Similar results of the taggers ranking for non-literary texts in the 3rd tagset position (Table 3a) have been obtained as for the previous positions. The lowest performance in terms of matching with the reference was achieved by the MorphoDiTa Tagger, with less than 75%, and the highest performance was achieved by the RNNTagger, with more than 98%. The second-highest performance, with more than 98%, was achieved by TreeTagger, which is a predecessor of the RNNTagger.

Five homogenous groups (four single-element homogenous groups and one two-element homogenous group) were identified (Table 3a), and statistically significant differences among them were observed in terms of agreement with the reference in the 3rd position ($p < 0.05$).

In the case of determining the tag in the 3rd position within the literary texts (Table 3b), MorphoDiTa Tagger (with less than 64%) statistically significantly performed the worst. Four homogenous groups were identified, but between homogenous group consisting of TreeTagger and Stanza, and homogenous group containing Stanza, RNNTagger, and UDPipe2, a statistically significant difference was not proven, only between TreeTagger and RNNTagger, and/or TreeTagger and UDPipe2.

Table 3:
Ranking of taggers in the 3rd position of the tag a) non-literary texts, b) literary texts

| a) N = 31285 | 1's (%) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **MorphoDiTa** | 74.59 | | o | | | |
| **MorphoDiTa Online** | 79.59 | | | o | | |
| Stanza | 95.24 | o | | | | |
| UDPipe2 | 95.59 | o | | | | |
| TreeTagger | 98.19 | | | | o | |
| RNNTagger | 98.67 | | | | | o |

| b) N = 24871 | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **MorphoDiTa** | 63.17 | | | o | |
| **MorphoDiTa Online** | 67.64 | | | | o |
| TreeTagger | 97.78 | | o | | |
| Stanza | 98.21 | o | o | | |
| RNNTagger | 98.56 | o | | | |
| UDPipe2 | 98.65 | o | | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

Similarly, as for the 2nd position, the taggers TreeTagger, Stanza, RNNTagger, and UDPipe2 achieved the highest performance for both text styles (more than 95%). The results indicate that the RNNTagger is a suitable automatic tool for grammatical gender identification regardless of text type.

**The 4th tagset position – number**

The fourth tagset position has mostly two standard values – singular and plural, which are also applied to adjectives, pronouns, and numerals (Mikulová et al., 2020).

The ranking of taggers for both text types in the 4h position (Table 4) has shown the lowest performance in terms of matching with the reference for the MorphoDiTa Tagger, with less than 86% (85.13% for non-literary texts and 83.60% for literary texts). The highest performance in terms of matching with the reference was achieved by the RNNTagger with more than 99%.

Four homogenous groups were identified for non-literary texts (Table 4a) and three homogenous groups were identified for non-literary texts (Table 4b). Statistically significant differences were observed in terms of agreement with the reference for all the taggers determining the 4th position ($p < 0.05$). The highest

agreement was achieved for the homogenous group – Stanza, RNNTagger, TreeTagger, and UDPipe2 – with more than 99% in terms of matching with the reference ($p > 0.05$).

Table 4:
Ranking of taggers in the 4th position of the tag a) non-literary texts, b) literary texts

| a) N = 34125 | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **MorphoDiTa** | 85.13 | | | o | |
| **MorphoDiTa_Online** | 90.69 | | | | o |
| Stanza | 97.97 | o | | | |
| UDPipe2 | 98.24 | o | | | |
| TreeTagger | 99.24 | | o | | |
| **RNNTagger** | 99.51 | | o | | |

| b) N = 28952 | 1's (%) | 1 | 2 | 3 |
|---|---|---|---|---|
| **MorphoDiTa** | 83.60 | | o | |
| **MorphoDiTa_Online** | 88.10 | | | o |
| TreeTagger | 99.33 | o | | |
| Stanza | 99.47 | o | | |
| UDPipe2 | 99.55 | o | | |
| **RNNTagger** | 99.62 | o | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The results (Table 4) show that for both text types, the best performance was achieved by RNNTagger (>99.5%). MorphoDiTa and MorphoDiTa_Online achieved the lowest performance, but still with more than 83% matches with reference for both text types (a little higher performance was achieved for non-literary texts).

**The 5th tagset position – case**

Slovak usually distinguishes among six (seven) cases: nominative, genitive, dative, accusative, (vocative), locative, and instrumental.

For non-literary texts, the results of the taggers' ranking in the 5th position copies the previous positions, above all the first tagset position (Table 5a). The lowest performance in terms of matching with the reference was achieved by the MorphoDiTa Tagger (less than 84%). The highest performance, more than 98%, was achieved by the RNNTagger. Six trivial single-element homogenous groups were identified (Table 5a). A statistically significant differences were observed in terms of agreement with the reference for all examined taggers determining the 5th tagset position ($p < 0.05$).

For literary texts, we obtained different results compared to non-literary texts (Table 5b). The worst and statistically significant performance was achieved by the MorphoDiTa Tagger (less than 84%). The highest performance was achieved by the RNNTagger, which forms together with Stanza and UDPipe2 one homogenous group with more than 98% of matching with the reference (Table 5b). Overall, four homogenous groups were identified among which statistically significant differences were observed in terms of agreement with the reference ($p < 0.05$).

Table 5:
Ranking of taggers in the 5th position of the tag a) non-literary texts, b) literary texts

| a) N = 32531 | 1's (%) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| **MorphoDiTa** | 83.79 | o | | | | | |
| **MorphoDiTa_Online** | 88.95 | | o | | | | |
| Stanza | 95.58 | | | o | | | |
| UDPipe2 | 96.77 | | | | o | | |
| TreeTagger | 97.76 | | | | | o | |
| RNNTagger | 98.72 | | | | | | o |

| b) N = 21590 | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **MorphoDiTa** | 83.65 | | o | | |
| **MorphoDiTa_Online** | 88.75 | | | o | |
| TreeTagger | 97.72 | | | | o |
| Stanza | 98.81 | o | | | |
| RNNTagger | 98.86 | o | | | |
| UDPipe2 | 99.09 | o | | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The results (Table 5) for determining the 5th tagset position show a good performance from almost all taggers. Regardless of text type, the RNNTagger proves to be one of the most suitable automatic tools for POS tagging (achieving more than 98.7%).

**The 8th tagset position – person**

The eighth tagset position expresses the person of verb forms (if applicable) and person of personal pronouns; and it usually takes on three standard values – 1st person, 2nd person, and 3rd person (Mikulová et al., 2020).

For non-literary texts (Table 6a), taggers' ranking determining the 8th position has shown two homogenous groups ($p < 0.05$) – one containing MorphoDiTa and MorphoDiTa_Online, and the second homogenous group consisting of the remaining taggers – Stanza, RNNTagger, TreeTagger, and UDPipe2. The first taggers' homogenous group achieved the lowest performance in terms of matching with the reference (less than 40%). On the other hand, the second taggers' homogenous group achieved a performance of more than 99% in terms of matching with the reference.

For literary texts (Table 6b), the worst and statistically significant performance was again achieved by the MorphoDiTa and MorphoDiTa_Online Tagger (less than 32%) forming one homogenous group ($p > 0.05$). Similarly, to non-literary texts, the highest performance was achieved by UDPipe2, Stanza, and RNNTagger (more than 99%) which form one homogenous group ($p > 0.05$).

Table 6:

Ranking of taggers in the 8th position of the tag a) non-literary texts, b) literary texts

| a) N = 6672 | 1's (%) | 1 | 2 |
|---|---|---|---|
| **MorphoDiTa** | 39.40 | | o |
| **MorphoDiTa_Online** | 39.51 | | o |
| Stanza | 99.10 | o | |
| UDPipe2 | 99.21 | o | |
| TreeTagger | 99.43 | o | |
| **RNNTagger** | 99.70 | o | |

| b) N = 11077 | 1's (%) | 1 | 2 | 3 |
|---|---|---|---|---|
| **MorphoDiTa** | 31.23 | | o | |
| **MorphoDiTa_Online** | 31.26 | | o | |
| TreeTagger | 98.28 | | | o |
| UDPipe2 | 99.03 | o | | |
| Stanza | 99.22 | o | | |
| **RNNTagger** | 99.55 | o | | |

Note: o - Homogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The results (Table 6) indicate that MorphoDiTa and MorphoDiTa_Online lag behind in determining the person of verb forms or in person of personal pronouns compared to other taggers (less than 40% for non-literary texts and less than 32% for literary texts).

**The 9th tagset position – tense**

The ninth tagset position represents only verb forms, in the purely morphological sense – future, present, and past (Mikulová et al., 2020).

Taggers' ranking determining the 9th position for non-literary texts (Table 7a) and also for literary texts (Table 7b) has shown two homogenous groups, and a statistically significant difference was observed in terms of agreement with the reference in the 9th position ($p < 0.05$). There was a small difference in performance between the two identified homogenous groups. The lowest performance in terms of matching with the reference was achieved by the MorphoDiTa and MorphoDiTa_online Tagger (about 98% for non-literary texts and about 99% for literary texts). The highest performance was obtained by the second homogenous group consisting of TreeTagger, RNNTagger, UDPipe2, and Stanza (more than 99.8%).

Table 7:

Ranking of taggers in the 9th position of the tag a) non-literary texts, b) literary texts

| a) N = 6169 | 1's (%) | 1 | 2 |
|---|---|---|---|
| MorphoDiTa | 98.02 | | o |
| MorphoDiTa_Online | 98.02 | | o |
| TreeTagger | 99.82 | o | |
| RNNTagger | 99.82 | o | |

| b) N = 11077 | 1's (%) | 1 | 2 |
|---|---|---|---|
| MorphoDiTa_Online | 99.24 | | o |
| MorphoDiTa | 99.24 | | o |
| RNNTagger | 99.89 | o | |
| TreeTagger | 99.90 | o | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **UDPipe2** | 99.90 | o | | **UDPipe2** | 99.94 | o | |
| **Stanza** | 99.94 | o | | **Stanza** | 99.94 | o | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

When determining tense, but in the purely morphological sense (9th position), Stanza proves to be the most effective tool with respect to the reference and regardless of text type.

**The 10th tagset position – degree of comparison**

The tenth tagset position is used for adjective and adverbs – positive, comparative, and superlative, apart from possessive adjectives.

Taggers' ranking determining the 10th position for non-literary texts has shown five homogenous groups – four single-element and one two-element homogenous groups (Table 8a), and the statistically significant differences were observed among them in terms of agreement with the reference in the 10th position ($p < 0.05$). The lowest performance in terms of agreement with the reference was achieved by the MorphoDiTa Tagger (less than 81%). The highest performance was identified by the RNNTagger (more than 98%). Between UDPipe2 and TreeTagger was not identified a statistically significant difference in terms of agreement with the reference ($p > 0.05$).

For literary texts taggers' ranking determining the 10th position (Table 8b) has shown statistically significant differences among four homogenous groups ($p < 0.05$). A homogenous group formed by Stanza, TreeTagger, and RNNTagger achieved a high performance of more than 97%, but a statistically significant difference among them was not observed in terms of agreement with the reference ($p > 0.05$). The worst performance was achieved for the MorphoDiTa Tagger (less than 63%) followed by MorphoDiTa_online (less than 65%). Compared to the non-literary texts, in which the highest performance was achieved by the UDPipe2 Tagger with more than 99%.

Table 8:

Ranking of taggers in the 10th position of the tag a) non-literary, b) literary text

| a) N = 7223 | 1's (%) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **MorphoDiTa** | 80.42 | | o | | | |
| **MorphoDiTa_Online** | 83.65 | | | o | | |
| **Stanza** | 96.01 | | | | o | |
| UDPipe2 | 97.73 | o | | | | |
| TreeTagger | 98.10 | o | | | | |
| RNNTagger | 98.88 | | | | | o |

| b) N = 4515 | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **MorphoDiTa** | 62.97 | | o | | |
| **MorphoDiTa_Online** | 64.89 | | | o | |
| **Stanza** | 97.56 | o | | | |
| TreeTagger | 97.72 | o | | | |
| RNNTagger | 97.96 | o | | | |
| UDPipe2 | 99.56 | | | | o |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

When determining the tenth position (Table 8), the combination of UDPipe2 and RNNTagger is shown to be the most effective way of automatically determining the degree of comparison with respect to the reference, regardless of the type of text.

**The 11th tagset position – negation**

The eleventh tagset position is fully inflectional category, as the negation in Slovak is expressed by a prefix – affirmative or negated. Negation belongs to verbs, adverbs, adjectives, and nouns (Mikulová et al., 2020). Ranking of taggers for non-literary and also literary texts, determining the 11th position (Table 9), has shown a statistically significant difference in terms of agreement with the reference in the 11th position among three homogenous groups ($p < 0.05$). First two homogenous groups are single-element, and the third homogenous groups is formed by four taggers - TreeTagger, RNNTagger, Stanza, and UDPipe2 (Table 9). The lowest performance in terms of matching with the reference was achieved by the MorphoDiTa Tagger (less than 95% for non-literary texts and less than 91% for literary texts). The highest performance was

achieved by UDPipe2 (more than 99%), but there is no statistically significant difference between UDPipe2 and TreeTagger/RNNTagger/ Stanza ($p > 0.05$).

Table 9:
Ranking of taggers in the 11th position of the tag a) non-fiction, b) fiction text

| a) N = 7377 | 1's (%) | 1 | 2 | 3 |
|---|---|---|---|---|
| **MorphoDiTa** | 94.89 | | o | |
| **MorphoDiTa_Online** | 95.58 | | | o |
| **TreeTagger** | 98.98 | o | | |
| **RNNTagger** | 99.08 | o | | |
| **Stanza** | 99.27 | o | | |
| **UDPipe2** | 99.32 | o | | |

| b) N = 11779 | 1's (%) | 1 | 2 | 3 |
|---|---|---|---|---|
| **MorphoDiTa** | 90.14 | | o | |
| **MorphoDiTa_Online** | 90.98 | | | o |
| **TreeTagger** | 99.38 | o | | |
| **RNNTagger** | 99.58 | o | | |
| **Stanza** | 99.64 | o | | |
| **UDPipe2** | 99.69 | o | | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

When determining the eleventh position, UDPipe2 appears to be the most powerful tool, but there is no statistical difference between it and the other automatic tools (RNNTagger, TreeTagger, and Stanza) in determining negation with respect to the reference, regardless of text type.

**The 12th tagset position – voice**

The twelfth tagset position is only used for verb forms, mainly for verb participles (Mikulová et al., 2020). A similar situation, as obtained in determining tense or negation, can also be observed in the case of voice. Ranking of taggers determining the 12th position for non-literary and literary texts (Table 10) has identified two homogenous groups between which a statistically significant difference was observed in terms of agreement with the reference in the 12th position ($p < 0.05$). The lowest performance was again achieved by the MorphoDiTa and MorphoDiTa_online Tagger (less than 54% for non-literary texts and less than 47% for literary texts). The highest performance was by UDPipe2, Stanza, TreeTagger, and RNNTagger (more than 98% for non-literary texts and more than 99% for literary texts), which form one homogenous group ($p > 0.05$).

Table 10:
Ranking of taggers in the 12th position of the tag a) non-literary texts, b) literary texts

| a) N = 7210 | 1's (%) | 1 | 2 |
|---|---|---|---|
| **MorphoDiTa** | 53.37 | | o |
| **MorphoDiTa_Online** | 53.47 | | o |
| UDPipe2 | 98.53 | o | |
| **Stanza** | 98.59 | o | |
| TreeTagger | 99.24 | o | |
| RNNTagger | 99.29 | o | |

| b) N = 11643 | 1's (%) | 1 | 2 |
|---|---|---|---|
| **MorphoDiTa** | 46.58 | | o |
| **MorphoDiTa_Online** | 46.71 | | o |
| TreeTagger | 99.36 | o | |
| **Stanza** | 99.51 | o | |
| RNNTagger | 99.54 | o | |
| UDPipe2 | 99.54 | o | |

Note: o - H**o**mogenous Groups, p > 0.05, marked - similar tagger ranking for both text styles

The only difference which can be observed is the order of taggers (Table 10), i.e. in the case of literary texts, the most accurate tagger with a reference is UDPipe2, and in the case of non-literary texts, it is the RNNTagger.

## Discussion

In the case of non-literary texts, the highest degree of concordance among the examined taggers (*Kendall Coeff. of Concordance* > 0.4) was identified for the tags in 8th position (*Kendall Coeff. of Concordance* = 0.59), 9th position (*Kendall Coeff. Of Concordance* = 0.50), and 12th position in the tagset (*Kendall Coeff. of Concordance* = 0.44). We identified two homogeneous groups with similar performance in terms of accuracy (Table 6a, 7a, 10a).

One homogeneous group consisted of MorphoDiTa and MorphoDiTa_Online, which achieved the lowest accuracy in the automatic determination of person, tense, and voice. All three positions in the tagset focus on the verb (the person of verb forms, verb forms in the purely morphological sense, and verb participles) during linguistic annotation. Our results indicate that neither tool is suitable for linguistic analysis of Slovak non-literary texts. On the other hand, tools like RNNTagger, TreeTagger, Stanza, and UDPipe2, when tagging the non-literary texts, achieved a high accuracy with reference for determining person (8th position), tense (9th position), and voice (12th position). They were consistent when it came to analyzing verbs and their forms and persons.

When determining other tag positions within the 15 positional tagsets that represent - part of speech, a detailed part of speech, gender, number, case, degree of comparison, and negation - automatic taggers achieved different quality. In general, MorphoDiTa achieved the lowest accuracy, and a statistically significant difference between it and MorphoDiTa_Online was proven. RNNTagger appears to be the most effective automatic tool, especially when it comes to determining part of speech, a detailed part of speech, gender, number, case, person, degree of comparison, and voice, even though when determining some tag positions (12th, 8th, or 4th) this tool was comparable to other taggers (Stanza/UDPipe2/TreeTagger).

Similarly, in the case of literary texts; the highest degree of concordance among the examined taggers (*Kendall Coeff. of Concordance* > 0.4) was identified for the tags in 8th position (*Kendall Coeff. of Concordance* = 0.67), 9th position (*Kendall Coeff. of Concordance* = 0.86), and 12th (*Kendall Coeff. of Concordance* = 0.52). We also identified two homogeneous groups with similar performance in terms to accuracy (Table 6b, 7b, 10b), apart from the tag in 8th position (person) produced by the TreeTagger (Table 6b) and for the tag in 9th position (Table 7b) produced by the Stanza tagger (Table 10b).

When determining verbs, the same accuracy and performance of automatic taggers as for non-literary texts can be observed. However, when determining the other tags within the tagset, the most accurate determination is produced by UDPipe2, whether it is part of speech, a detailed part of speech, gender, case, degree of comparison, negation, and voice.

The results showed that the usage of automatic annotation tools could be proficient in the case of the Slovak language. Four of the six examined tools achieved a high performance for most of the tagset positions. TreeTagger, as the predecessor of RNNTagger, lacked in some tagset positions (1st, 2nd, 3rd, 5th, 8th). The difference in performance was not large, but the new tool RNNTagger offers a novel method using recurrent neural networks for annotating texts.

Similar results were achieved for both text types, and it can be concluded that usage of the RNNTagger should be preferred for both types. Stanza, as another representative of the neural network pipeline that is used for tagging, achieved high performance in almost all tagset positions (> 95%). UDPipe2 also achieved a high performance, but mostly in the case of literary texts. In seven out of nine examined tagset positions of literary texts, UDPipe2 achieved the highest performance. In the case of non-literary texts, the highest performance was achieved by the RNNTagger (eight out of nine examined tagset positions). Together with RNNTagger, UDPipe2 achieved the highest performance of the examined taggers.

On the other hand, MorphoDiTa and MorphoDiTa_online struggled in some tagset positions (3rd, 8th, 10th, 12th, 15th). The tool was designed by the same authors as UDPipe2 and was focused on the Czech language but supporting the Slovak language. The other issue that could have caused low performance could have been that they were the only tools that generated the output in the PST tagset format. As for the other tools a parser was used to convert the SNC tagset to PST tagset, it could have been that MorphoDiTa output was too detailed.

Our results indicate that it is useful to differentiate texts into literary and non-literary and subsequently, based on the text style to deploy a tagger. For literary text, UDPipe2 outperforms other taggers for morphological annotation of the inflectional Slovak language. However, for non-literary texts, RNNTagger is more effective (accurate) for morphological analysis compared to other taggers (Table 11). Moreover, our results show how linguistic aspects affect tagger performance in terms of accuracy with gold tokenization. For example, if we focus on number or person phenomena within the literary texts, it is more effective to deploy an RNNTagger; or in the case of morphological analysis focusing on tense, it is better to deploy Stanza, despite the fact that, in general, UDPipe2 performs the best.
Last but not least, our results reveal that the most effective approach to morphological annotation involves a combination of UDPipe2 and RNNTagger for general (non-specific) linguistic analysis.

Table 11:
Ranking of taggers accuracy according to aspect of linguistic analysis

| | **Literary texts** | | | | | | **Non-literary texts** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tag position | MorphoDiTa | MorphoDiTa_Online | UDPipe2 | Stanza | TreeTagger | RNNTagger | MorphoDiTa | MorphoDiTa_Online | UDPipe2 | Stanza | TreeTagger | RNNTagger |
| part of speech | | | **x** | | | x | | | | | | **x** |
| a detailed part of speech | | | **x** | | | x | | | | | | **x** |
| gender | | | **x** | x | | x | | | | | | **x** |
| number | | | x | x | x | **x** | | | | | x | **x** |
| case | | | **x** | x | | x | | | | | | **x** |
| person | | | x | x | | **x** | | | x | x | x | **x** |
| tense | | | x | **x** | x | x | | | x | **x** | x | x |
| degree of comparison | | | **x** | | | | | | | | | **x** |
| negation | | | **x** | x | x | x | | | x | x | x | **x** |
| voice | | | **x** | x | x | x | | | x | x | x | **x** |

Note: marked – the best performance in accuracy for individual text styles

## Conclusions

The research was focused on evaluation the tagging functionality of various automatic tools for the Slovak language. Morphological annotation is a time-consuming task that requires lot of manual work from experts. The six analyzed automatic tools were evaluated based on the performance of the taggers expressed in terms of accuracy with gold tokenization. The results showed that all tools offer a high performance in determining the part of speech. That is important and offers a good baseline to use the tools. A more accurate complex morphological annotation of the word POS tag offered mainly RNNTagger and UDPipe2. Non-literary texts offered various genres, and RNNTagger achieved the highest performance in terms of agreement with the gold tokenization. Literary texts comprised novels and fairy tales where UDPipe2 achieved the highest performance in terms of agreement with the gold tokenization. High performance results were achieved also for TreeTagger and Stanza taggers on both text types.
The study has certain limitations, which mainly consist of the size of the dataset. This is an issue that is hard to resolve as creating a corpus with manual annotation is very time-consuming and requires a great deal of manual work. That is also the reason so many automatic annotation tools have been developed. Despite that, the used dataset was sufficient to highlight that many tools already support an inflectional

language such as Slovak. This article also focused only on evaluating the tag generation performance within the 15 positional tagsets for the Slovak language. In future work, it would be appropriate to focus on lemmatization, as most of these tools also offer this functionality.

## References


Afanasev I. 2023. The Use of Khislavichi Lect Morphological Tagging to Determine its Position in the East Slavic Group. In: *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 174–186. DOI: 10.18653/v1/2023.vardial-1.18.

Alosaimy A, Atwell E. 2018. Web-based Annotation Tool for Inflectional Language Resources. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA),.

Bejček E, Straňák P. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation* 44:7–21. DOI: 10.1007/s10579-009-9093-0.

Benko Ľ, Benková L. 2022. Comparison of Novel Approach to Part-Of-Speech Tagging of Slovak Language. In: *DIVAI 2022 – The 14th international scientific conference on Distance Learning in Applied Informatics*. Štúrovo, Slovakia: Wolters Kluwer, 327–333.

Benkova L, Munkova D, Benko Ľ, Munk M. 2021. Evaluation of English–Slovak Neural and Statistical Machine Translation. *Applied Sciences* 11. DOI: 10.3390/app11072948.

Blunsom P. 2004.Hidden Markov Models

Branco A, Eskevich M, Frontini F, Hajič J, Hinrichs E, Jong F de, Kamocki P, König A, Lindén K, Navarretta C, Piasecki M, Piperidis S, Pitkänen O, Simov K, Skadiņa I, Trippel T, Witt A, Zinn C. 2023. The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. DOI: 10.1007/s10579-023-09658-z.

Brants T. 2000. TnT - a statistical part-of-speech tagger. In: *Proceedings of the sixth conference on Applied natural language processing* -. Morristown, NJ, USA: Association for Computational Linguistics, 224–231. DOI: 10.3115/974147.974178.

Fehle J, Schmidt T, Wolff C. 2021. Lexicon-based Sentiment Analysis in German: Systematic Evaluation of Resources and Preprocessing Techniques. In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*. Düsseldorf, Germany: KONVENS 2021 Organizers, 86–103.

Fink GA. 2008. *Markov Models for Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-71770-6.

Gajdošová K, Šimková M. 2016.Slovak Dependency Treebank

Garábik R, Bobeková K. 2021. Lematizácia, morfologická anotácia a dezambiguácia slovenského textu – webové rozhranie. *Slovenská reč* 86:104–109.

Garabík R, Šimková M. 2012. Slovak Morphosyntactic Tagset. *Journal of Language Modelling*. DOI: 10.15398/jlm.v0i1.35.

Hajič J. 2006. Complex Corpus Annotation: The Prague Dependency Treebank. *Insight into the Slovak and Czech Corpus Linguistics*:54–73.

Hajič J, Bejček E, Hlaváčová J, Mikulová M, Straka M, Štěpánek J, Štěpánková B. 2020. Prague Dependency Treebank - Consolidated 1.0. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: European Language Resources Association (ELRA), 5208–5218.

Hajič J, Hric J. 2017. MorfFlex SK 170914. *Available at http://hdl.handle.net/11234/1-3277* (accessed October 20, 2023).

Hammarstedt M, Schumacher A, Borin L, Forsberg M. 2022. *Sparv 5 User Manual*. Göteborg.

Hládek D, Staš J, Juhár J. 2012. Dagger: The Slovak morphological classifier. In: *Proceedings ELMAR-2012*. Zadar, Croatia : IEEE, 195–198.

Hladek D, Stas J, Juhar J. 2015. Morphological Analysis of the Slovak Language. *Advances in Electrical and Electronic Engineering* 13. DOI: 10.15598/aeee.v13i4.1491.

Hochreiter S, Schmidhuber J. 1997. Long Short-Term Memory. *Neural Computation* 9:1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Horák A, Gianitsová L, Šimková M, Šmotlák M, Garabík R. 2004. Slovak National Corpus. In: *Text, Speech and Dialogue, TSD 2004*. Springer, Berlin, Heidelberg, 89–93. DOI: 10.1007/978-3-540-30120-2_12.

Huang Z, Xu W, Yu K. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging.

Izzi GL, Ferilli S. 2020. UniBA @ KIPoS: A Hybrid Approach for Part-of-Speech Tagging. In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*. Accademia University Press, 501–506. DOI: 10.4000/books.aaccademia.7773.

Jurafsky D, Martin J. 2020. *Speech and Language Processing*.

Kanerva J, Ginter F, Miekka N, Leino A, Salakoski T. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium: Association for Computational Linguistics, 133–142. DOI: 10.18653/v1/K18-2013.

Kapusta J, Benko Ľ, Munkova D, Munk M. 2021. Analysis of Edit Operations for Post-editing Systems. *International Journal of Computational Intelligence Systems* 14:197. DOI: 10.1007/s44196-021-00048-3.

Karyukin V, Rakhimova D, Karibayeva A, Turganbayeva A, Turarbek A. 2023. The neural machine translation models for the low-resource Kazakh–English language pair. *PeerJ Computer Science* 9:e1224. DOI: 10.7717/peerj-cs.1224.

Kirov C, Cotterell R, Sylak-Glassman J, Walther G, Vylomova E, Xia P, Faruqui M, Mielke SJ, McCarty A, Kübler S, Yarowsky D, Eisner J, Hulden M. 2018. UniMorph 2.0: Universal Morphology. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA),.

Ljubešić N, Dobrovoljc K. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 29–34. DOI: 10.18653/v1/W19-3704.

Machura J, Geržová H, Masopustová M, Valíčková M. 2019. Comparing majka and MorphoDiTa for Automatic Grammar Checking. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019*. Brno, Czech Republic: Tribun EU, 3–14.

Majchráková D, Dušek O, Hajič J, Karčová A, Garábik R. 2014. Semi-automatic detection of Multiword Expressions in the Slovak Dependency Treebank. In: *Computational Linguistics in Bulgaria*. Sofia, Bulgaria: The Institute for Bulgarian Language Prof. Lyubomir Andreychin – Bulgarian Academy of Sciences, 32–38.

Mikulová M, Hajič J, Hana J, Hanová H, Hlaváčová J, Jeřábek E, Štěpánková B, Vidová Hladká B, Zeman D. 2020. *Manual for Morphological Annotation Revision for Prague Dependency Treebank - Consolidated 2020 release*. Prague, Czech Republic.

Munkova D, Munk M, Benko Ľ, Hajek P. 2021a. The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science* 7:e706. DOI: 10.7717/peerj-cs.706.

Munkova D, Munk M, Ľubomír Benko, Stastny J. 2021b. MT Evaluation in the Context of Language Complexity. *Complexity* 2021:1–15. DOI: 10.1155/2021/2806108.

Petkevič V, Hlaváčová J, Osolsobě K, Svášek M, Šimandl J. 2019. Parts of Speech in NovaMorf, A New Morphological Annotation of Czech. *Journal of Linguistics/Jazykovedný casopis* 70:358–369. DOI: 10.2478/jazcas-2019-0065.

Petrov S, Das D, McDonald R. 2012. A Universal Part-of-Speech Tagset. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2089–2096.

Piao S, Tsuruoka Y, Ananiadou S. 2009. Sentiment Analysis with Knowledge Resource and NLP Tools. *The International Journal of Interdisciplinary Social Sciences: Annual Review* 4:17–28. DOI: 10.18848/1833-1882/CGP/v04i05/52902.

Proisl T, Dykes N, Heinrich P, Kabashi B, Blombach A, Evert S. 2020. EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatizaion and Semantic Tagging to a German Web and CMC Corpus. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: European Language Resources Association (ELRA), 6142–6148.

Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 101–108. DOI: 10.18653/v1/2020.acl-demos.14.

Rabiner LR, Juang BH. 1986. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3:4–16. DOI: 10.1109/MASSP.1986.1165342.

Richter M. 2010. Pokročilý korektor češtiny. diploma thesis Thesis. Prague, Czech Republic: Charles University.

Rosen A, Hana J, Štindlová B, Feldman A. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation* 48:65–92. DOI: 10.1007/s10579-013-9226-3.

Schmid H. 2019. Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. New York, NY, USA: ACM, 133–137. DOI: 10.1145/3322905.3322915.

Schmid H, Laws F. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*. Morristown, NJ, USA: Association for Computational Linguistics, 777–784. DOI: 10.3115/1599081.1599179.

Šimková M, Gajdošová K. 2008. Slovenský závislostný korpus. *Grammar & Corpora*:135–141.

Spoustová D, Hajič J, Raab J, Spousta M. 2009. Semi-supervised training for the averaged perceptron POS tagger. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*. Morristown, NJ, USA: Association for Computational Linguistics, 763–771. DOI: 10.3115/1609067.1609152.

Straka M. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 197–207. DOI: 10.18653/v1/K18-2020.

Straka M. 2020.Universal Dependencies 2.6 models for UDPipe 2

Straka M, Straková J. 2014.MorphoDiTa: Morphological Dictionary and Tagger

Straka M, Straková J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal*

*Dependencies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 88–99. DOI: 10.18653/v1/K17-3009.

Toleu A, Tolegen G, Mussabayev R. 2022. Language-Independent Approach for Morphological Disambiguation. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 5288–5297.

Tsuruoka Y, Tateishi Y, Kim J-D, Ohta T, McNaught J, Ananiadou S, Tsujii J. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: *PCI 2005: Advances in Informatics*. Springer, Berlin, Heidelberg, 382–392. DOI: 10.1007/11573036_36.

Universal Dependencies contributors. 2022.Universal POS tags

Yao Y, Huang Z. 2016. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation. In: *ICONIP 2016: Neural Information Processing*. Springer, Cham, 345–353. DOI: 10.1007/978-3-319-46681-1_42.

Zeman D, Nivre J, Abrams M, Ackermann E, Agić Ž, Aepli N, Aghaei H, Ahrenberg L. 2023.Universal Dependencies 2.12. *Available at http://hdl.handle.net/11234/1-5150* (accessed November 22, 2023).

PRÍLOHA M: BENKO, ĽUBOMÍR, ANNA PILKOVA, MICHAL MUNK A SLAVKA ELEY, 2024C. PILLAR 3: THE IMPACT OF LANGUAGE COMPLEXITY ON THE PREFERENCES OF COMMERCIAL BANK WEBSITE USERS. *EXPERT SYSTEMS WITH APPLICATIONS* (V RECENZNOM KONANÍ OD 2024, 1. KOLO) (**WEB OF SCIENCE, 2022IF: 8.5, Q1**)

# Pillar 3: The impact of language complexity on the preferences of commercial bank website users

**Ľubomír Benko[a,*], Anna Pilkova[b], Michal Munk[a,c], Slavka Eley[b]**

**[a] Department of Computer Science, Constantine the Philosopher University in Nitra, Address: Tr. A. Hlinku 1, SK 949 01, Nitra, Slovakia;**
**email: {lbenko, mmunk}@ukf.sk**
**[b] Department of Strategy and Entrepreneurship, Comenius University in Bratislava, Address: Odbojárov 10, 820 05 Bratislava, Slovakia**
**email: anna.pilkova@fm.uniba.sk, eley1@uniba.sk**
**[c] Science and Research Centre, University of Pardubice, Address: Studentská 84, CZ 532 10, Pardubice, Czech Republic**
**email: michal.munk@upce.cz**

**Corresponding Author: [*]Ľubomír Benko, email: lbenko@ukf.sk**

**Abstract.**

This research explores the impact of document complexity and readability on user preferences for disclosure information on commercial bank web portals, with a focus on Pillar 3 disclosures. The study investigates the usage patterns of disclosed information by stakeholders. Through an analysis of web portal access variables and document complexity and readability measures, the study identifies correlations between text complexity/readability and user preferences. The findings reveal that while stakeholders show interest in financial documents despite their complexity, preferences vary based on document type and readability. Notably, documents with higher complexity tend to attract more attention, suggesting potential challenges in accessibility and comprehension. The paper highlights the importance of presenting disclosure information in a manner that enhances readability and accessibility to better engage stakeholders. Additionally, the research introduces user preference indicators aligned with web portal access metrics, providing insights into stakeholders' behavior. The research followed up on previous findings of very low interest in Pillar 3 information from commercial bank stakeholders. The results showed that stakeholders are more interested in less demanding and more readable texts such as annual reports, and as Pillar 3 and other Pillar documents are more complex and harder to read, they are less interested in them. The conclusion of the research is that increasing the interest in this information in commercial banks requires finding ways of presenting it that make it less complicated and more readable.

**Keywords:** Web Usage Mining; Web Content Mining; Text Complexity; Text Readability.

## 1. Introduction

The Basel Committee on Banking Supervision (BCBS), as the primary global standard setter for the prudential regulation of banks, aims to promote market discipline through meaningful disclosure of key risks borne by internationally active banks (BIS Connect, 2019). For this purpose, Pillar 3 of the Basel Framework sets out a comprehensive set of disclosure requirements, the aim of which is to provide market participants with sufficient information to assess the significant risks and capital adequacy of an internationally active bank. The Pillar 3 standard is now part of the Basel Consolidated Framework, which combines all BCBS regulatory requirements into one document. BCBS will continue to update the information on Pillar 3 as the Basel Committee issues or modifies its requirements. The Pillar 3 framework provides a comprehensive package of all existing disclosure requirements covering all main parts of the Basel Framework. It includes the disclosure on the composition of capital and exposures to the main risks (credit risk, market risk, counterparty credit risk, operational risk, securitization positions and others), and disclosures related to other parts of the Basel Framework, such as regulatory liquidity ratios, leverage ratio, asset encumbrance, remuneration or interest rate risk in the banking book. Unless otherwise stated, the framework applies to all internationally active banks at the highest consolidated level. The revised Pillar 3 disclosures are underpinned by five guiding principles that draw on lessons learned from the 2007-2009 financial crisis: comprehensibility, comprehensiveness, meaningfulness/usefulness, consistency over time, and comparability. The frequency of data disclosure varies between quarterly, semi-annually, and annually, depending on the nature of the request. For example, disclosures regarding the linkages between financial statements and regulatory exposures are generally annual; disclosures regarding capital composition are semi-annual; and disclosures of risk-weighted assets are quarterly. BCBS sets globally applicable standard, which is then implemented by the countries via applicable legislative procedure. In case of the European Union (EU), the Basel Framework is implemented via Capital Requirements Regulation and Capital Requirements Directive. The Capital Requirements Regulation is directly applicable to all banks across the EU, while the Capital Requirements Directive is implemented via national transposition into legal framework of EU Member States. The disclosure requirements are an integral part of the Capital Requirements Regulation, based on which technical standards including specific disclosure templates are prepared by the European Banking Authority.

Commercial banks in CEE (Central and Eastern Europe) countries have many specifics. Among them, the predominant ownership by large multinational groups, their legal form as entities whose shares are not traded on capital markets (and hence different information requirements from stakeholders), and business models based on depository clients. The depositors of these banks represent a very important group of stakeholders, where alongside depositors having covered deposits up to EUR 100.000 (mostly physical persons), there is a significant group of not uninsured depositors (legal persons such as companies and other organizations). However, there is a lack of empirical studies on their behavior and interest in using Pillar 3 information. Information on usage of disclosed information by the targeted audiences is important as feedback to the regulators on usefulness of the set standards and if these help to contribute to market discipline, as happened during the last financial crisis. However, to achieve the regulatory objectives, the market discipline mechanism must be effective and used in accordance with the expectations of the regulators. In CEE countries, there is a lack of studies evaluating Pillar 3 disclosures based on the relevance of the content to key stakeholders from commercial banks. Therefore, several studies focus on the analysis of interest in the disclosure of information aimed at a specific type of interested parties (uninsured depositors) in a foreign bank whose shares are not traded on the capital markets. The importance of this group is also supported by the fact that, for example in Slovakia, almost half of deposits in bank accounts are uninsured deposits, and a similar situation can be expected in other CEE countries.

The disclosures of Pillar 3 represent the implementation of effective market discipline from the regulators' point of view. Various authors point out the benefits of the implementation of Pillar 3 disclosures as an effective supervisory market discipline tool. To support the transparency and efficiency of markets, regulatory authorities try to ensure that the financial documents disclosed to stakeholders are readable and understandable. This research aims to create indicators of user preferences on commercial banks' web portals and to investigate whether the complexity and readability of mandatory published information impact user preferences. Since this information is among highly specialized economic texts, it will be interesting to identify which metrics of complexity and readability can be used in the context of user preferences. The motivation for this experiment was to build on previous research (Munk et al., 2021c; Pilková, Munk, Benko, et al., 2021) and to examine the content of disclosure information in the context of the complexity of the text in more detail.

The structure of this paper is as follows. The related work section focuses on the importance of information disclosure and how other authors address the readability of finance texts. The materials & methods section describes the used

dataset and the research methodology applied in detail, providing a comprehensive description of each step. The subsequent section focuses on the research results and analyzes user preferences in combination with text complexity and readability. The Discussion and Conclusion sections summarize the findings.

## 2. Related Work

From the regulators' perspective, disclosures in Pillar 3 imply the application of effective market discipline. Several authors highlight the benefits of using Pillar 3 disclosure as an effective tool for market surveillance discipline. Pillar 3 reduces information asymmetry (Niessen-Ruenzi et al., 2015; Parwada et al., 2013), increases the security of the banking system (Vauhkonen, 2012), and quarterly disclosure is proper for investors (Parwada et al., 2013). A study (de Araujo & Leyshon, 2016) suggests differences in the relevance of disclosed information, as depositors and creditors are most responsive to information such as bank assets, off-balance sheet items, and ratings related to other banking activities. Stakeholders prefer quantitative rather than qualitative information, which can positively influence bank risk-taking (Fernandes et al., 2021). Website postings can serve as a timely information medium and form of communication that is accessible to a wide range of stakeholder groups (Akbar & Deegan, 2021). Therefore, the sensitive reaction of stakeholders to negative disclosures can also trigger withdrawals in an inefficient bank (Faria-e-Castro et al., 2017) and can reduce the distance to failure (Del Gaudio et al., 2020). Furthermore, the authors (Oliveira, Lima Rodrigues, & Craig, 2011) found that the reputational contribution of bank management can influence stakeholders' perceptions of risk disclosure manipulation. Although investors and customers welcome increased information disclosure, care must be taken to ensure that the information disclosed brings appropriate value.

Previous research focused mainly on investigating the behavior of stakeholders in the banking institution's portals. The data source was the log files of banking institutions (Munk et al., 2021b). One of the results was the creation of a methodology describing the estimation of the probability of stakeholder access to web categories related to Pillar 3 (Munk et al., 2021a). The main objective of the article (Munk et al., 2021c) was to evaluate the behavior and interests of stakeholders on the Web portals of a commercial bank who were interested in disclosing information during and after turbulent times in a country belonging to the CEE region. The focus of the research was on data analysis during and after the crisis to identify key types of information of interest to stakeholders, and based on the results, it was possible to optimize the policy of publishing the given information. A detailed analysis of the probability of stakeholder access during the weeks to the Web portal with information disclosure of the commercial bank showed that the results correspond to the results of the previous quarterly analysis (Munk et al., 2017). The greatest interest of stakeholders in information related to Pillar 3 was during the first quarter of the year, where the period around the 10th week was mainly identified, which was the week with the greatest interest in the given web categories. Based on this, it can be concluded that the frequency of mandatory quarterly information disclosure is not necessary for market discipline. It would be sufficient to publish this information annually, ideally in the year's first weeks. The research (Pilková, Munk, Blažeková, et al., 2021) aimed to evaluate the interest of stakeholders in two groups of information disclosure: Pillar 3 disclosure requirements and Pillar 3 related from 2009-2012. The second objective was to evaluate robustness by verifying the results using two approaches based on different time variables: week and quarter from 2009 to 2012. The first quarters of 2009-2010, during the global financial crisis, had a significant impact on the quantity of extracted frequent itemsets of web categories. This is also confirmed by the weekly analysis, where the first quarters achieved the highest interest in the investigated web categories. On the basis of the results, it can be argued that quarterly information disclosure is not necessary for market discipline. Annual disclosure information would be sufficient, ideally in the year's first weeks. The research in the article (Pilková, Munk, Benko, et al., 2021) is a follow-up to previous research with a focus on investigating the behavior of stakeholders on the web portals of a banking institution in connection with Pillar 3 information disclosure. The article examines how the behavior of stakeholders has changed on the web portals of another banking institution in Slovakia during the years 2016-2018. As a result of the research, the downward trend of interest in information related to Pillar 3 continued. The main contribution of the mentioned research was the identification of interesting areas that can increase stakeholders' awareness of Pillar 3 information. And it was summarised in the following recommendations (Pilková, Munk, Blažeková, et al., 2021): improve standardization, i.e. harmonization of information disclosure by national authorities of requirements and requirements for information disclosure at the EU level (Pillar 3 and national requirements); increase comparability of published information by creating one standard template; reduce the frequency of Pillar 3 disclosures given the low interest of stakeholders in quarterly disclosures; the obligation to use English as a single language for the publication of information.

The results of previous research (Munk et al., 2021c; Pilková, Munk, Benko, et al., 2021) were the motivation for the integration of readability and complexity of texts in further research. The complexity and readability of professional

texts are evaluated using automatic metrics by several authors. Gunning (2003) presents more than a hundred metrics of text complexity, but only a few of them are used. At the same time, several are used to determine the basic characteristics of the text, such as the length of the sentence, the frequency of part-of-speech, etc. (Awan et al., 2021; Sadeek Quaderi & Varathan, 2024). Most of the research is mainly directed at the field of learning English as a second language, which, however, tends to examine less specialized texts (Maqsood et al., 2022). This section focuses primarily on those metrics that have been used in the context of professional texts, ideally related to banking or finance. Ehara (2022) focused on investigating the readability of introductory computer science texts. In his research, he compared BERT classification with conventional readability metrics such as Flesch-Kincaid Grade Level (Kincaid et al., 1975), ARI (Senter & Smith, 1967), Coleman-Liau Index (Coleman & Liau, 1975), Flesch Reading Ease (Flesch, 2016), Gunning Fog Index (Gunning, 2003), LIX (Björnsson, 1968), SMOG Index (McLaughlin, 1969), RIX (Anderson, 1983) a Dale-Chall Index (Chall & Dale, 1995). Ehara (2022) analyzed texts extracted from GitHub (software manuals), ACL Anthology, and PubMed (scientific articles), selecting only abstracts. Ehara (2022) proposed a dictionary-based readability metric and compared it with conventional metrics. The research results show a higher correlation than conventional metrics. While scientific texts are unreadable for intermediate students, and on the other hand, software manuals are mostly readable for them. Ehara (2021) conducted similar research on economic news texts. However, it only focused on comparing the BERT-based approach and the dictionary. The results showed that most of the texts were readable for intermediate students, but 2.4% were not understandable for them.

Das (2014) focused on investigating text analysis in finance, examining the text of his own scientific article. Das (2014) also deals with the influence of text sentiment on stock market development over time, with results showing a high correlation between sentiment and stock market price development. When examining financial texts, Das (2014) used the metrics Gunning Fog Index, Flesch-Kincaid Grade Level, Flesch Reading Ease, Coleman-Liau Index. He compared his results with Loughran & McDonald (2014), who argued that conventional readability metrics do not work in the case of financial texts because, e.g. the Gunning Fog Index is based on sentence length and word complexity. However, the length of sentences is complicated when calculating financial texts. Das (2014) showed that the results of all the investigated metrics are roughly similar to the Gunning Fog Index for the examined publication. Buzarna-Tihenea (2020) investigated several characteristics of texts from the field of economic sciences, focusing on lexical density, keywords, most frequent words, and readability. The examined text came from an EU regulation aimed at raising awareness of social responsibility in businesses. The Buzarna -Tihenea, (2020) showed that the lexical density corresponds to the technical text, which ranks it among the more complex texts. The results of the readability metrics confirmed the previous statement, as several metrics achieved scores representative of the level of undergraduate or graduate students. Lesmy et al. (2023) examine the development of the readability of approximately 200000 Item 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations) sections from 10-K reports of public companies in the USA during the period 1996-2022. For investors, these reports serve as management's view of the company's financial conditions and other factors affecting the company's performance. The results of the Gunning Fog Index score showed that in 1996, the readability was 17.6, and in 2022, the texts scored 20.19. The research concludes that during these years the readability of the examined documents gradually deteriorated.

Hayo et al. (2019) studied European Central Bank (ECB) press conferences, specifically the impact of the linguistic and content complexity of the opening statement on trading behavior in financial markets. To date, central banks of all major economies hold regular press conferences following meetings of their monetary policy committees. Each press conference begins with a prepared opening statement and ends with a discussion with journalists. Hayo et al. (2019) investigated whether the higher complexity of central bank communication causes financial markets to delay trading and whether a generally less complex closing discussion can moderate this effect. The authors used transcripts of press conferences and the Flesch-Kincaid grade level metric to analyze the complexity of the transcribed text. The results showed that the average score for all the introductory statements is 15.4, which represents a rather difficult complexity of the texts. On the other hand, the average closing discussion score reached a level of less than 11, which represents 4 years less study. For comparison, the authors also implemented other readability metrics that confirmed the results: Flesch Reading Ease, Gunning Fog Index, SMOG, Coleman-Liau Index, ARI.

More and more authors are dealing with the readability or complexity of mandatory published information. Linsley & Lawrence (2007) demonstrated that the level of readability of risk disclosures was difficult, even very difficult, for UK companies. Jia & Li (2022) suggested a positive association between the presence of risk committees and the readability of risk management reports for Australian companies. Guay et al. (2015) focused on investigating the complexity of financial statements, using ReadIndex in the case of readability, which consisted of the scores of the metrics: Flesch-Kincaid Grade Level, LIX, RIX, Gunning Fog Index, ARI and SMOG. The results of the analysis showed that there is a high correlation between all readability metrics and with ReadIndex. The benefit of the research

is the finding that even though complex financial statements negatively affect the information environment, some companies try to mitigate these effects by voluntarily disclosing additional information. Moreno & Casasola (2016) focused on analyzing the readability of annual reports in Spanish, using the modified Flesch Reading Ease metric, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, LIX, and RIX. The authors provided an overview of the most widely used readability metrics for accounting texts in English-speaking countries (mainly Flesch Reading Ease and Gunning Fog Index). The research results confirmed the results of other authors focused on annual reports in English: Annual reports in Spanish show signs of being more difficult to read (Moreno & Casasola, 2016). Aymen et al. (2018) investigate the influence of the readability of financial information on the behavior of financial analysts. They focused on 88 French companies with the expectation that readable annual reports would provide homogeneous, simple, clear, and legible information that is understandable to all investors. The prerequisite for excessively complex annual reports is the need for financial analysts to process and interpret these reports. The Gunning Fog Index and Flesch Reading Ease were used as readability metrics. The results (Aymen et al., 2018) showed that the Gunning Fog Index marked the examined texts as difficult to read, but on the contrary Flesch Reading Ease considered them easier. The main result of the article is that the number of financial analysts influences the readability of annual reports. Colliard & Georg (2023) proposed a framework for distinguishing different dimensions of regulatory complexity, including six measures that can be applied to texts. This is an initial proposal that will be subject to further experiments. Smailović et al. (2018) conducted a case study on linguistic characteristics in combination with sentiment for annual reports (10-K) of public companies. The article's main contribution is that, despite differences in the characteristics of the reports, the sentiment estimate remains essentially the same. This indicates that the authors of the annual reports try to keep the information content neutral. On the other hand, some linguistic characteristics have a strong link to the company's financial status. It turned out that companies in financial distress use wording that expresses more doubt. Smailović et al. (2018) justify this by saying that despite the fact that the creators of annual reports manage to keep the sentiment neutral, they may have less conscious control over the modal words that capture these linguistic characteristics. Toerien & du Toit (2024) focused on risk reporting in South Africa. The authors focused on annual reports collected for the years 2005-2021, when several standards regarding information disclosures based on the practice in the EU were introduced in the country. They examined the texts using the Flesch-Kincaid and Gunning Fog Index readability metrics and various metrics that focus on word count, sentence length, and the like. The results indicate (Toerien & du Toit, 2024) that published reports have low readability. The introduction of standards that are supposed to increase the readability and reduce the complexity of published messages has proven ineffective, and it is necessary to reconsider the form of information that could reach a larger number of users. They consider problems with determining complex words, sentence structure and length, as well as the lack of context, to be a limitation of the research. Toerien & du Toit (2024) recommend the inclusion of additional metrics to examine the complexity of disclosed information.

Based on the reviewed literature, this article aimed to analyse the complexity and readability of text related to Pillar 3 information published on the websites of commercial banks and investigate its impact on the behavior of stakeholders.

## 3. Materials & Methods

The research is based on two data sources. One source is a log file obtained from the web portal of a banking institution. The second source is documents extracted from the web portal based on the documents visited from the log file.

### 3.1 Methods

In this part, the methodology will be described only in the form of steps; a more detailed description will follow in the next part of the article. The methodology (Figure 1) was inspired by several researches (Munk et al., 2021c, 2021a; Munkova et al., 2021; Pilková, Munk, Benko, et al., 2021; Yao et al., 2017).

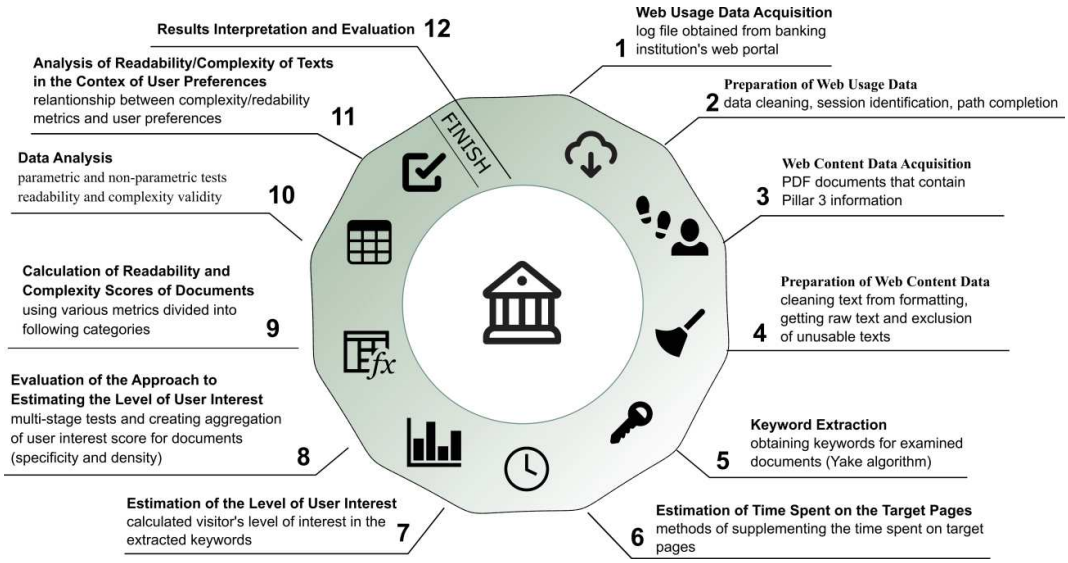



**Fig. 1 - Methodology workflow diagram**

### 3.2 Acquisition and Preparation of Web Usage Data

The research is based on two data sources. The first source is a log file obtained from the banking institution's web portal. The second source is documents extracted from the web portal based on the log file. The log file contained accesses to the web portal throughout 2018 and underwent a data preparation phase consisting of data cleaning, user/session identification, and path completion (Munk et al., 2021a). During the pre-processing phase, other necessary variables such as *Category* and *Subcategory* were created, which were used to connect the web parts of the portal accessed by visitors. The taxonomy of the web portal, based on which the web parts were divided, can be found in Table 1. Priority was given to information related to Pillar 3 information; therefore, the log file was reduced to sessions that contained at least one access to one of the examined categories. The log file modified in this way consisted of 265 216 records.

**Table 1 - Taxonomy of the web portal**

| Category | Subcategory |
|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ |
| /Pilar3 disclosure requirements/ | /information_about_bank/ |
| /Pillar3 related/ | /annual_reports/ |
| /Pillar3 related/ | /financial_reports/ |
| /Pillar3 related/ | /covered_bonds/ |
| /Pillar3 related/ | /information_for_investors_except_shareholders/ |
| /Pillar3 related/ | /information_for_shareholders/ |

### 3.3 Acquisition and Preparation of Data about Web Content

Investigation of the impact of the complexity and readability of the texts on the accesses of stakeholders to the web portal required the extraction of the texts accessed by stakeholders. The examined web portal of the banking institution discloses information related to Pillar 3 in the format of PDF documents. Direct links to the given documents were extracted from the log file, and most of the documents were retrieved using a web crawler. Since it was a log file from 2018, not all documents were accessible, but more than 90% of the visited documents were extracted. More than half of the documents were in English, with the fact that there was also an official translation into Slovak. Documents containing only images, tables or graphs and duplicate language versions were excluded from the selection of

documents (a total of 178 documents were removed). Some Slovak documents did not have an accessible English version, so their English translation was created using the Google Translate machine translation system (97 documents were translated). 226 documents divided into individual categories according to taxonomy were used for text analysis (Table 1). Plain text was extracted from these documents using a PDF OCR tool, which was then used to analyze the complexity of the text and search for user interest.

### 3.4 Keyword Extraction

The next stage involved the extraction of keywords necessary for analyzing the behavior of web visitors (Yao et al., 2017). Keywords for individual documents were extracted using the Yake algorithm (Campos et al., 2020). The given algorithm achieves better results than standard keyword identification techniques such as TF-IDF, TextRank, KP-Miner, or Rake (Campos et al., 2020). TF-IDF is a frequently used statistical measure that determines the significance of a keyword relative to its significance in a single document and across all documents in the entire corpus. However, experiments have shown that in the case of professional texts, tools such as Yake, KEA, or KP-Miner achieve better results than TF-IDF (Sarwar et al., 2021; Sarwar & Noor, 2021). Since this experiment focuses on documents within the banking field, the Yake tool was chosen to extract keywords. For each document examined, 100 keywords were extracted, and based on the achieved score, those keywords meeting the selected threshold for each document subcategory were chosen. The threshold was determined based on the average score extracted for documents within the given subcategory. Keywords scoring higher than the average were excluded from further analysis (lower score indicates greater importance of the keyword). Table 2 presents the frequency of keywords extracted for individual categories, along with the number of documents examined. This table includes the total of all keywords for the documents, thus also counting duplicate keywords shared among documents within the same subcategory.

**Table 2 - Frequency of documents extracted and keywords**

| Category | Subcategory | No. documents | No. keywords |
|---|---|---|---|
| /Pilar3 disclosure requirements/ | /financial_statement/ | 54 | 1367 |
| /Pilar3 disclosure requirements/ | /information_about_bank/ | 45 | 1554 |
| /Pillar3 related/ | /annual_reports/ | 17 | 1034 |
| /Pillar3 related/ | /financial_reports/ | 18 | 1134 |
| /Pillar3 related/ | /covered_bonds/ | 85 | 5939 |
| /Pillar3 related/ | /information_for_investors_except_shareholders/ | 2 | 120 |
| /Pillar3 related/ | /information_for_shareholders/ | 5 | 312 |

### 3.5 Estimation of Time Spent on the Target Pages

An important parameter to consider when examining the readability of a text is the time estimate of the so-called target (content) page. In the session identification phase using the Reference Length method (Kapusta et al., 2012; Munk et al., 2015), the web portal pages in the session are divided into navigational and content pages. The content pages are the documents under examination but determining the exact time that visitors spend on them is no possible. Therefore, various methos of estimating this time have been proposed. As the time window was set at 3600 seconds (60 minutes) in the data preparation phase, one method was based on this given time window. Another approach involved using the text readability metric, specifically the reading time (*reading_time*) of individual documents, either by supplementing the average reading time with a document from a given category or by supplementing the specific reading time for each document. Four methods of supplementing the time spent on target (content) pages were proposed and new variables were created in the dataset:

- original approach without adding times (*length*),
- addition of 3601 seconds for all target pages in sessions (*length3601*),
- addition of the average reading time for documents from the examined categories (*lengthRT_cat*),
- addition of a specific reading time for individual documents (*lengthRT_doc*).

Since the length of time spent on the page contributes to the overall procedure of estimating the level of user interest, the entire procedure was performed for all four content page time supplementation approaches.

### 3.6 Estimation of the Level of User Interest

The visitor's level of interest in the extracted keywords was calculated from the time spent on the page (Yao et al., 2017):

$$time_u(c_j, k_i) \begin{cases} \frac{length_u(c_j)}{m}, & \text{if } k_i \text{ in } c_j \\ 0, & \text{if } k_i \text{ not in } c_j \end{cases},$$

where $length_u(c_j)$ denotes the length of time visitor $u$ spends on the web part of category $c_j$ that contains the extracted keywords $\{K_1, K_2, \dots, K_i\}$, for the total number of keywords $m$ in all categories.

Total time $sum_u$, that the website visitor spends on a certain keyword $K_i$ in $KT_u$, is calculated as follows:

$$sum_u(c_j, k_i) = \begin{cases} \sum_{i=j}^{f} time_u(c_j, k_i), & \text{if } c_j \text{ in } KT_u \\ 0, & \text{if } c_j \text{ not in } KT_u \end{cases}.$$

The aim of the given procedure is to predict access to a certain keyword which is based on information about visitors' navigation. If the user has a word that is interesting or important to him/her, then in that case the user repeatedly visits some pages containing the word on which he/she is likely to spend more time than on other pages. A model called UISM (User Interest Structure Model) was used (Yao et al., 2017), it is a model combined with web content data, web structure data and web usage data and includes all web domains. The UISM model is defined as:

- set of states: $Q = \{q_1, q_2, \dots, q_n\}$, initial state starts at $q_1$, each $q_i$ represents a web category,
- set of keywords: $K = \{k_1, k_2, \dots, k_n\}$, $K$ contains all keywords from all web categories $Q$,
- the state transition probability $P_1(q \rightarrow \acute{q})$ between two categories is defined as follows:

$$P_1(q \rightarrow q') = \frac{count\ (q \rightarrow q')}{count\ (q)},$$

where $(q \rightarrow q')$ denotes that the user first visited category $q$ and then category $q'$. Yao et al. (2017) presented two different approaches of UISM to the calculation of $P_1$. According to the structure of the Web portal, they distinguished the so-called vertical and horizontal structure. In the case of a vertical structure, $count\ (q \rightarrow q')$ represents the number of sessions in which the user visited category $q'$ immediately after category $q$. So, both categories must not only be in the same session but must be visited directly after each other. On the other hand, in the case of the horizontal structure approach, $count\ (q \rightarrow q')$ represents the number of sessions in which both categories $q$, $q'$, are found, but they are not limited by having to directly follow each other.

In all states $q$, there is a distribution probability $P_2\ (k_i|q)$ for each keyword $k_i$ of $K$:

$$P_2\ (k_i|q) = \frac{\sum_{u=1}^{N} sum_u(q, k_i)}{\sum_{u=1}^{N}(\sum_{m=1}^{M} sum_u(q, k_i))},$$

which is referred to as the hidden Markov model observation symbol probability and is calculated from the total time spent by the user on a given keyword of a given web category. It can also be defined as the probability of interest in a keyword.

For a session $S^l$ ($l$ represents the length of the session) and the interest of the user, it is possible to express the level of interest of the user in the given keyword $R\ (k|S_l)$, the calculation of which is as follows:

$$R\ (k|S^l) = P_1(q_{start} \rightarrow q_1) \times P_2(k|q_1) \times P_1(q_1 \rightarrow q_2) \times P_2(k|q_2) \times \dots \times P_1(q_{l-1} \rightarrow q_l) \times P_2(k|q_l)\,.$$

If $R(k|S^l)$ is greater than or equals to $C$ – the confidence threshold value, then $R(k|S^l)$ is an interesting session because users with the same interest can access the session categories. The confidence threshold is set according to Yao et al. (2017) in the interval $10^{-3}$ to $10^{-7}$. In the case of the hidden Markov model, the threshold value ranges from $10^{-5}$ to $10^{-10}$.

### 3.7 Evaluation of the Approach to Estimating the Level of User Interest

Based on the above-mentioned procedure, a data matrix was created containing the horizontal/vertical levels of user interest (UIH and UIV) for individual extracted keywords from the examined documents. The aim of the research is to link the complexity of the texts with the interest in the visited web categories. For this reason, it was necessary to

identify a suitable threshold value of reliability, an approach to the level of user interest (horizontal or vertical) and an approach to estimating of the time spent on the target page.

Descriptive statistics results (Table Appendix A) showed that supplementing the estimate of time spent on target pages has no effect on user interest. Therefore, an approach with time supplementation based on reading time was chosen. This approach suggests that if the user searched for the given document, they probably spent some time on it, and this time can be precisely represented by the time spent reading the document. Based on multi-stage tests, differences in the threshold value of reliability in determining the level of interest were identified. In the case of a horizontal approach to the estimation of user interest, based on the results of the Cochrane Q tests, the global null hypothesis (stating that the confidence limit has no effect on the number of generated interesting keywords) is rejected at a significance level of 0.001 (Cochran Q Test: $N = 13054$, $Q = 737.7514$, $df = 4$, $p < 0.001$). Similar results were also obtained for the vertical approach (Cochran Q Test: $N = 13054$, $Q = 114.4615$, $df = 4$, $p < 0.001$).

The results (Table 3) show that the confidence threshold is interesting from the point of view of multi-stage tests only in the case of the horizontal approach, where four homogeneous groups were identified, while statistically significant differences were identified between levels $10^{-3}$, $10^{-4}$ and others. In the case of the vertical approach, two homogeneous groups with statistically significant differences between grades of $10^{-4}$, $10^{-5}$, $10^{-6}$, $10^{-7}$ and $10^{-3}$ were identified.

**Table 3 - Multi-stage tests to identify differences in confidence threshold for horizontal (a) and vertical (b) approaches to estimating user interest**

| (a) | 1's (%) | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| UIH_10-3_rtdoc | 1.33 | **** | | | |
| UIH_10-4_rtdoc | 1.49 | **** | | | |
| UIH_10-5_rtdoc | 2.38 | | **** | | |
| UIH_10-6_rtdoc | 2.99 | | | **** | |
| UIH_10-7_rtdoc | 3.29 | | | | **** |

| (b) | 1's (%) | 1 | 2 |
|---|---|---|---|
| UIV_10-3_rtdoc | 0.08 | | **** |
| UIV_10-4_rtdoc | 0.31 | **** | |
| UIV_10-5_rtdoc | 0.31 | **** | |
| UIV_10-6_rtdoc | 0.31 | **** | |
| UIV_10-7_rtdoc | 0.38 | **** | |

Note: **** $p > 0.05$

The results showed that the vertical approach to user interest estimation yields the least useful information, since the examined documents are rarely found in consecutive sequences in sessions. For this reason, only the horizontal approach was selected for further analysis, where documents are located in sessions regardless of the order of visitation. Since the authors' proposed user interest estimation procedure (Yao et al., 2017) was based on keywords and their interest based on a confidence threshold value, it was not possible to evaluate user interest in specific documents. Therefore, aggregations were designed to represent users' interest in individual documents, regardless of the reliability threshold. Based on the results, the horizontal approach was chosen, which had better discriminant power than the vertical approach. An analogous procedure can also be applied to web subcategories, categories, or other investigated web parts, not only documents. Each keyword characterizes the given document differently, and therefore the aggregations attempt to consider the weight for each document. Aggregations were created based on the following characteristics, and the following weights were created:

- User interest based on specificity of keywords in a document: $UIH_{s_i} = \sum_{k=1}^{K_i} specificity_k * UIH_k$, where *UIH* is estimation of user interest for the given keyword *k* and $specificity_k = \ln \frac{n_i}{N}, i = 1, \ldots, N$, where *N* is the total number of documents examined, *n* is the number of documents that contain a given keyword,
- User interest based on density of keywords in a document: $UIH_{d_i} = \sum_{k=1}^{K_i} density_k * UIH_k$, where *UIH* is estimation of user interest for the given keyword *k* and $density_k = \frac{n_i}{N}, i = 1, \ldots, N$, where *N* is the total number of examined documents, *n* is the number of documents that contain a given keyword.

As an example, how to picture the aggregations can be mentioned by the keyword "client". The given word is found in 83 documents out of 226 examined documents. The natural logarithm of the share of these numbers gives a weight of -1.00169. The estimated UIH user interest value for this keyword is 0.002064. Multiplying these two values yields a specificity of -0.0020675 for the examined keyword. Similarly, in the case of density, the result in that case is 0.000758. The value of the weight is higher if the keyword is more general that means it is contained in more documents. Based on the absolute values of specificity, the results are similar, but the scores are two opposite extremes. In this way, different weights were obtained for individual keywords in the documents, which were then aggregated for each document.

Support, entropy (Shannon, 1948) and the number of sessions in which a document is the target page (variable *target*) were calculated as other indicators of user preferences. It was based on identified sessions using the Reference Length method (Kapusta et al., 2012; Munk et al., 2015). In the case of the support variable, the support of the given

documents was calculated in the identified sessions. Support expresses interest in the documents visited and serves as a reference to the experiment, where the attempt is to come up with additional indicators of user preferences that should correlate with a valid criterion (support). One such indicator is the entropy of sessions, where emphasis was placed on the composition of individual sessions. The disorder was represented by the fact that the visitor to the web portal visited different pages from different subcategories during his session, while the entropy of the session was calculated based on the entropy defined by (Shannon, 1948): $entropy_s = -\sum_{x \in X} p(x) \log_n p(x)$, where $n$ is the number of pages in session $s$ and $p(x)$ is the probability of page $x$ occurring in the session. Therefore, if the entropy was equal to 1, it means that the sessions contained websites from different categories. If the entropy is equal to 0, then all pages in the session came from one category, and thus the user was looking for targeted information from the given category. Entropy was calculated for each session for each document, and the result was the average entropy for each document.

The second indicator was the *target* variable, which represented the number of sessions in which the document was the target page. A target page is a page whose time the user spends on the page is greater than the threshold time (Kapusta et al., 2012), that is, its content is interesting for the visitor and it could be said that it is the goal of his search. In the case of the presented experiment, the visitors' target was the content of the examined documents.

### 3.8 Calculation of Readability and Complexity Scores of Documents

The complexity or readability of the text plays a significant role in text analysis. Various metrics of text complexity and readability help determine whether texts are suitable for the target group or not. In addition to the style of the text, its language also plays a crucial role. Quantitative measures of text complexity focus mainly on the characteristics of words themselves and their occurrence in sentences and paragraphs. Word-level analysis belongs to the first level of analysis, where the word length itself (number of characters) can indicate the extent to which the reader has to decode the word. Monosyllabic words are easier to decode than polysyllabic ones. However, word frequency alone is not a complete metric, as the context in which words are found can increase complexity. Over the years, the word list has been expanded to include more than three thousand words (Chall & Dale, 1995), and authors have used this list to determine readers' effort based on how many words are not on the list. Therefore, a text containing a large number of words not on the list indicates a more complex text. The second level of quantitative measures of readability is sentence length (number of words) and related characteristics (Kintsch, 1974). In the presented experiment, a larger number of metrics were used with the motivation to identify those that will best capture the complexity of the documents related to Pillar 3 disclosures. The metrics were divided into several categories for better clarity. In the case of some categories there is an overlap (this is also caused by the overlap of categories); therefore some metrics were used as if duplicated, but in that case they were calculated using a different tool. All metrics were implemented using Python or external tools were used (Cvrček et al., 2020; Lu, 2010, 2011, 2012; Lu & Ai, 2015).

List of the metrics examined and their division into groups:

- Text characteristics [char]: It is a group of metrics that is focused on the basic characteristics of the text, such as frequencies, mean, and median of tokens (Gray & Leary, 1935). The most used are the number of tokens, sentences, characters, and then the number of unique tokens.
- Readability [read]: Conventional readability metrics were created mainly to replace outdated metrics. Most of them are based on the students' grade level. The Flesch-Kincaid grade level metric is based on grade levels in the United States (Kincaid et al., 1975). Sometimes it is also understood as the number of years of education required to understand the given text: $Flesch - Kincaid = 0{,}39 * \left(\frac{total\ words}{total\ sentences}\right) + 11{,}8 * \left(\frac{total\ syllables}{total\ words}\right) - 15{,}59$. The metric is focused on the length of the sentence, rather than the length of the word. The score of the Gunning Fog Index for metric represents the year of study in the American educational system and is among the most used metrics in today's linguistics (Spiers et al., 2017). The minimum requirement to calculate the metric is the selection of a part of the text that contains at least one hundred words, then calculate the score as follows: $index = 0{,}4 * \left[\left(\frac{words}{sentences}\right) + 100 * \left(\frac{complex\ words}{words}\right)\right]$, where *complex words* are such words that consist of three or more syllables. Texts aimed at a wider audience should have an index of less than 12. Other readability metrics include the Coleman-Liau index (Coleman & Liau, 1975), Automated Reliability Index (Senter & Smith, 1967), SMOG (McLaughlin, 1969), and Flesch reading ease (Flesch, 2016).

- Lexical variation [lex_var]: It refers to the range of a reader's vocabulary and how it is reflected in their use of language (Malvern et al., 2004). One of the basic metrics of lexical variation is the number of different words (NDW), which is used to measure a child's language development (Klee, 1992; Miller, 1991). However, the disadvantage of this metric is its dependence on the length of the language sample, as it cannot compare samples of different lengths. One possibility is to shorten the sample to a uniform length based on the shortest sample (Thordardottir & Weismer, 2001). Malvern et al. (2004) opposed sample truncation as a waste of useful data and, therefore, proposed two standardization methods. In both cases, a set of sub-samples of equal size is randomly selected from the study sample and the NDW for these sub-samples is averaged to approximate the expected NDW value. In one method, each subsample consists of a standard number of words randomly selected from the study sample. In the second method, each subsample contains a standard number of consecutive words from the study sample with a random starting point. McClure (1991) focused on different ratios of selected word types (number of verb types, nouns, adjectives, adverbs, and modifier types, which is a combination of adjectives and adverbs), with the same denominator (number of lexical words).
- Lexical richness [lex_rich]: Refers to the range and variety of vocabulary in the text under study (McCarthy & Jarvis, 2007). Lexical richness is used in combination with lexical variation, density, and diversity. Lexical richness tells us about the number of different expressions in the text and the variety of the vocabulary. Among the most used metrics is the Type-token ratio (TTR): $TTR = \frac{T}{N}$, where *T* is the number of words and *N* is the total number of words in the examined text (Templin, 1957). The disadvantage of this metric is the reduction of the ratio depending on the increase of the examined patterns (Arnaud, 1992). Some authors state that lexical variation and diversity are similar properties. For this reason, some metrics in both categories were presented in the experiment, while they were calculated using different tools. Other similar metrics are mainly a modification of the original TTR, such as Root TTR (RTTR) (Guiraud, 1960), Corrected TTR (CTTR) (Carroll, 1964), Bilogarithmic TTR (LogTTR) (Herdan, 1964), Uber Index (Dugast, 1979) and normalized TTR (zTTR) (Cvrček & Chlumská, 2015).
- Lexical diversity [lex_div]: It is basically the range and variety of vocabulary used by an author in a text, taking into account writing quality, vocabulary knowledge, general characteristics, and socioeconomic status (McCarthy & Jarvis, 2007). Several authors take lexical diversity as analogous to lexical variation or richness. In this case, the attention was on metrics that explicitly focus on lexical diversity, that is, diversity, despite the fact that most are inspired by the TTR metric. The Measure of Textual Linguistic Diversity (MTLD) is a metric where the text is divided into segments, and a TTR score is calculated for each, and the length of the text is a variable that depends on the TTR value as the segments expand. Each segment ends when the TTR value reaches 0.72 (McCarthy, 2005). Other metrics used were Hypergeometric distribution diversity (HD-D) (McCarthy & Jarvis, 2007), Herdans' lexical diversity measure (Herdan, 1964), Dugasts' lexical diversity measure (Dugast, 1979), and Maass' lexical diversity measure (McCarthy & Jarvis, 2007).
- Lexical sophistication [lex_sop]: It is also called lexical rareness and measures the proportion of relatively rare or advanced words in texts. Linnarud (1986) and Hyltenstam (1988) both calculated the lexical sophistication as $LS1 = \frac{N_{slex}}{N_{lex}}$, where $N_{slex}$ is the number of sophisticated lexical words and $N_{lex}$ is the total number of lexical words in the text. Both authors focused the metric on English second language learners, while Linnarud (1986) defined sophisticated lexical words as English words learned by students from grade 9 and above in the Swedish school system. Laufer (1994) created the Lexical Frequency Profile model, which focuses on the proportion of word types in a text in combination with a list of the first 1000 most frequent words, the second 1000 most frequent words, and a list of university words (Xue & Nation, 1984). The created model also offers a metric of lexical sophistication: $LS2 = \frac{T_s}{T}$, where $T_s$ s the number of sophisticated parts of speech and *T* is the total number of parts of speech in the text (Wolfe-Quintero et al., 1998). Another approach to lexical sophistication was metrics: verb sophistication metric, corrected verb sophistication (Wolfe-Quintero et al., 1998).
- Expert metrics [expert]: LIX is included among the metrics of readability (Björnsson, 1968), but it is one of the less widely used. It was created primarily for Swedish texts, but was successfully applied regardless of the language being studied, because standard text characteristics are used for its calculation: $LIX = \frac{words}{periods} + \frac{long\ words * 100}{words}$, where *words* is the number of words, *periods* represents the number of punctuation marks defined as dots, colon or uppercase first letter, and *long words* is the number of long words (more than 6

characters). According to the results of (Anderson, 1983) the resulting score represents the level of education, while a value greater than 55 indicates highly specialized texts suitable for university students and graduates. Anderson (1983) came up with an optimization of the LIX metric and called it RIX (Rate Index). He defined it as follows: $RIX = \frac{long\ words}{sentences}$, while a score higher than 7.2 represents the high difficulty of the text as with LIX. Anderson (1983) demonstrated that LIX and RIX correlate almost perfectly with each other (r = 0.99). O'Hayre (1966) proposed the LINSEAR Write metric, which is based on syllable counting. The metric score calculation is as follows: $LW = \frac{\frac{easy\ words}{sentences} + 2 * hard\ words}{sentences}$, where *hard words* are the number of words that contain more than two syllables and, on the contrary, *easy words* are words that consist of two or less syllable. The resulting score represents the school level for which the text is intended, a value of 13-16 represents a university student, and 17+ a university graduate. All these metrics are also used in the examination of economic texts and, therefore, were included in a group together. The Gunning Fog Index, which is also often used as a relevant metric for economic texts, was included in the group of readability metrics to verify whether there are differences between these metrics.

- Syllable [syl]: These are similar metrics as in the case of the group of text characteristics, but in this case the metrics only capture features related to the number of syllables. A problem with these metrics may be that they may not work in some languages.
- Part of speech ratio [pos_ratio]: Tagging or morphological annotation is the assignment of a lemma and a tag (morphological mark) to each token in the text. Each tag consists of letters of the Latin alphabet, numbers, and symbols, while a set of individual characters forms one tag for one token. When examining the readability of the text, it is sufficient to identify the type of word in the token. Since this is a time-consuming operation, the automatic tagging tool Stanza was used in this case (Qi et al., 2020). The tool was selected based on the results achieved in the experiment in the case of English texts, where it reached more than 99% success in determining the part of speech word type (Qi et al., 2018). The ratios were extracted for the following word types: numerals, spaces, nouns, adpositions, determiners, proper nouns, adjectives, verbs, coordinating conjunctions, punctuations, adverbs, auxiliaries, particles, pronouns, subordinating conjunctions, interjections, symbols, and other tags.
- Other characteristics: There are many readability and complexity metrics, some of them could not be included in the above groups due to the different range of scores achieved, and therefore the last group containing various metrics was created. Here, for example, Size of file in kB, which in economic texts tends to be an indicator of the complexity of the text, were included, as well as Reading time (Demberg & Keller, 2008), which is an essential metric in the experiment representing the length of reading the text in seconds. Cvrček et al. (2020) developed the QuitaUp tool, which is intended for quantitative stylometric text analysis. They implemented several metrics already mentioned into the tool, but also various others that were included in this category: h-point, frequency of hapaxes, entropy, verb distance, activity, descriptiveness, average length of tokens, thematic concentration, secondary thematic concentration (Cvrček et al., 2020). Among the other metrics, two self-designed metrics EAWL and EAWL_unique were also included.

A metric of the complexity of the texts was proposed, the motivation of which was the specialization in economic texts. The emergence of the metric was inspired by the Economic Academic Word List (EAWL), which consists of 887 words that are most often found in economic texts (O'Flynn, 2019). The EAWL is similar to the Academic Word List (AWL), but is more suitable for economic texts, as it is newer and was created as an extension of the New General Service List (NGSL), which was created in 2013. Based on a validation study using economics articles in the British Academic Written English corpus, EAWL gives 5.5% word coverage in texts and about 90% coverage for a combination of NGSL and EAWL. A final difference is that EAWL contains fewer word forms than AWL. EAWL contains only inflected forms or variant spellings of words, rather than entire groups of words, which means that, although it has more words than AWL (887 compared to 570), it has fewer word forms overall (1763 compared to 3112). The score of the proposed EAWL metric is calculated as follows: $EAWL = \frac{eawl\ words}{words}$, where $eawl\ words$ is the number of all words of the text found in the EAWL dictionary and $words$ is the total number of words in the text. As an alternative, a metric optimization was proposed where only unique words were examined: $EAWL\_unique = \frac{eawl\ words_{unique}}{words_{unique}}$, where $eawl\ words_{unique}$ is the number of unique words of the text found in the EAWL dictionary and $words_{unique}$ is the total number of unique words in the text.

After completing all the steps, the result was a dataset containing the examined documents, an estimate of the level of interest in the documents, time characteristics, the number of sessions in which the document is the target page, and the entropy of the sessions, along with characteristics based on the complexity and readability of the text.

## 4. Results

The analysis of the dependency between the level of user interest and the readability/complexity of the text consisted of several steps. In the first step, it was necessary to evaluate which of the aggregation levels of user interest best describes the examined documents in combination with time, other user-oriented characteristics (target and entropy) and support (valid criterion) (Table 4). The results of the aggregations compared to the support variable, which expresses interest in the visited documents and serves as a valid criterion, pointed out that the number of sessions in which the document is the target page, the entropy of the sessions, and the level of user interest based on density of keywords in a document are relevant (Table 4).

**Table 4 - Correlations between the support variable and the UIH and other variables**

| support & | N | r | r2 | t | p |
|---|---|---|---|---|---|
| target | 226 | 0.9835*** | 0.9673*** | 81.422 | 0.0000 |
| entropy_mean | 226 | -0.8887*** | 0.7899*** | -29.017 | 0.0000 |
| Sum(DensityKW*UIH) | 226 | 0.1997** | 0.0399** | 3.050 | 0.0026 |
| Sum(SpecificityKW*UIH) | 226 | -0.0527 | 0.0028 | -0.789 | 0.4309 |

Note: **** $p > 0.05$

The examined categories are not evenly represented, so they are represented by different numbers of documents, which naturally causes large differences in interval estimates. Due to the aggregation of data into documents and subsequent summarization into broader content categories (subcategories), differences in variability were naturally identified in some cases, for this reason was required an agreement of parametric (analysis of variance) and non-parametric tests (Kruskal-Wallis test) when testing of global null statistical hypotheses.

It is assumed that the indicators of user preferences (the number of sessions in which the document is the target page, the entropy of the sessions and the level of user interest based on specificity and density of the keywords in the document) will be relevant, in terms of the differentiation and measure of explanation of web portal accesses.

The global null hypothesis is rejected at the significance level of 0.001 (target: $F(6,219) = 8.206$, $p < 0.001$; $H(6, N = 226) = 100.751$, $p < 0.001$), which claims that there is no statistically significant difference in user preferences expressed by the number of sessions in which a document is the target page subcategories (Table 4). Similarly, the global null hypothesis is rejected at the significance level of 0.001 (entropy_mean: $F(6,219) = 17.000$, $p < 0.001$; $H(6, N = 226) = 111.818$, $p < 0.001$) also in the case of user preferences expressed by entropy of sessions (Table 1).

In the case of user preferences expressed by the level of interest, the global null hypothesis is rejected at the significance level of 0.001 (Sum(SpecificityKW*UIH): $F(6,219) = 11.925$, $p < 0.001$, $H(6, N = 226) = 30.596$, $p < 0.001$; Sum(DensityKW*UIH): $F(6,219) = 102.661$, $p < 0.001$; $H(6, N = 226) = 25.797$, $p < 0.001$), which claim that there is no statistically significant difference in level of user interest based on specificity as well as density of keywords in a document among the examined content subcategories.

The results of parametric and non-parametric procedures are consistent and can be considered robust.

Through multi-stage test (Table 5), statistically significant differences ($p < 0.05$) and homogenous groups ($p > 0.05$) were identified.

The content subcategories *covered-bonds*, *information-for-shareholders*, *financial-statement*, *financial-reports,* and *information-about-bank* (Table 5a) represent one homogeneous group in terms of user preferences expressed by the number of sessions in which the document is the target page ($p > 0.05$). In the case of the number of sessions in which the document is the target page, the statistically significant highest user preferences (Table 5a) were shown by the subcategory *information-for-investors-except-shareholders* ($p < 0.05$) followed by the subcategory *annual-reports* ($p < 0.05$). Similar results were obtained for user preferences expressed by the entropy of sessions (Table 5b), where the subcategory *information-for-investors-except-shareholders* ($p < 0.05$) showed the statistically significantly lowest entropy of sessions, followed by the subcategory *annual-reports* ($p < 0.05$). The remaining subcategories (*information-about-bank, financial-statement, financial-reports, information-for-shareholders, covered-bonds*) together form one homogeneous group in terms of entropy of sessions ($p > 0.05$).

In terms of user preferences expressed by the level of interest (Table 5c, 5d), the results are similar when considering user interest based on specificity and density of keywords in a document.

In the case of specificity (Table 5c), two homogeneous groups were identified. Based on the absolute values (Table 5c), the categories *information-for-investors-except-shareholders* and *information-for-shareholders*, which together form a homogeneous group, showed the highest level of user interest based on specificity of keywords in a document ($p > 0.05$). In contrast, the lowest level of interest was identified when considering the level of user interest based on specificity of keywords in a document for the subcategories *annual-reports, covered-bonds, financial-reports, financial-statement,* and *information-about-bank*, while together they form one homogeneous group in terms of level of interest ($p > 0.05$).

In the case of density, the category *information-for-investors-except-shareholders* ($p < 0.05$) showed the statistically significantly highest level of interest (Table 5d), followed by the category *information-for-shareholders* ($p < 0.05$). The remaining categories (Table 5d) together form a homogeneous group in terms of level of interest ($p > 0.05$).

Within the identified homogeneous groups, parametric and non-parametric estimates (ordinal statistics) resemble, and the results achieved can be considered robust. An interesting result was obtained by non-parametric estimates in the case of the number of sessions in which the document is the target page (Table 5a) and entropy (Table 5b), where the *annual-reports* category achieved the highest mean rank value in the case of target and lowest mean rank value in the case of entropy (based on the absolute values of specificity, it can be seen that the results are similar, but the scores are opposite), while in the case of the other investigated indicators, it did not reach such high values.

**Table 5 - Multi-stage tests for a) target, b) entropy_mean, c) Sum(SpecificityKW*UIH), d) Sum(DensityKW*UIH)**

| a) subcategory \ target | Mean | Mean Rank | 1 | 2 | 3 |
|---|---|---|---|---|---|
| covered-bonds | 4.259 | 63.994 | **** | | |
| information-for-shareholders | 5.400 | 64.400 | **** | | |
| financial-statement | 18.759 | 135.093 | **** | | |
| financial-reports | 19.722 | 130.278 | **** | | |
| information-about-bank | 20.600 | 144.967 | **** | | |
| annual-reports | 89.235 | 200.147 | | **** | |
| information-for-investors-except-shareholders | 161.500 | 161.750 | | | **** |

| b) subcategory \ entropy_mean | Mean | Mean Rank | 1 | 2 | 3 |
|---|---|---|---|---|---|
| information-for-investors-except-shareholders | 0.996 | 58.500 | | **** | |
| annual-reports | 0.998 | 27.353 | | | **** |
| information-about-bank | 0.999 | 73.867 | **** | | |
| financial-statement | 1.000 | 90.519 | **** | | |
| financial-reports | 1.000 | 110.139 | **** | | |
| information-for-shareholders | 1.000 | 158.600 | **** | | |
| covered-bonds | 1.000 | 165.665 | **** | | |

| c) subcategory \ Sum(SpecificityKW*UIH) | Mean | Mean Rank | 1 | 2 |
|---|---|---|---|---|
| information-for-investors-except-shareholders | -0.090 | 10.500 | | **** |
| information-for-shareholders | -0.087 | 40.500 | | **** |
| annual-reports | -0.017 | 76.647 | **** | |
| covered-bonds | -0.016 | 108.841 | **** | |
| financial-reports | -0.010 | 96.361 | **** | |
| financial-statement | -0.007 | 131.241 | **** | |
| information-about-bank | -0.006 | 134.478 | **** | |

| d) subcategory \ Sum(DensityKW*UIH) | Mean | Mean Rank | 1 | 2 | 3 |
|---|---|---|---|---|---|
| covered-bonds | 0.000 | 109.488 | **** | | |
| annual-reports | 0.001 | 150.235 | **** | | |
| financial-reports | 0.001 | 127.722 | **** | | |
| financial-statement | 0.001 | 104.574 | **** | | |
| information-about-bank | 0.001 | 99.122 | **** | | |
| information-for-shareholders | 0.004 | 186.600 | | **** | |
| information-for-investors-except-shareholders | 0.030 | 225.500 | | | **** |

Note: **** $p > 0.05$

The new proposed indicators representing user preference (number of sessions in which the document is the target page and session entropy) achieved the same discriminative power, with a very large statistically significant correlation being achieved in both cases (Table 4, Table 5a, 5b). Both indicators also turned out to be interesting from the point of view of the level of explanation of the accesses to the Web portal. The number of sessions in which the document is the target page explains 97% of the support variability, and the session entropy explains 79% of the support variability (Table 4).

It is assumed that the highest performance indicator (in terms of differentiation and measure of explanation of web portal accesses) indicator of user preferences in terms of time spent on the page will be the time spent on the page taking into account the reading time of the document.

The global null hypothesis is rejected at the significance level of 0.001 (lengthRT_doc_mean: $F(6, 219) = 3.538$, $p < 0.001$; $H(6, N = 226) = 122.387$, $p < 0.001$), which claims that there is no statistically significant difference in user preferences expressed by the time spent on the page taking into account the reading time of the document among the investigated content subcategories (Table 6). In the case of additional time variables is the global null hypothesis not rejected ($p > 0.05$). This means that there have been identified differences in time values between the examined content subcategories.

**Table 6 - Correlations between the support variable and the length variables**

| support & | N | r | r2 | t | p |
|---|---|---|---|---|---|
| lengthRT_doc_mean | 226 | 0.6719*** | 0.4514*** | 13.577 | 0.0000 |
| lengthRT_cat_mean | 226 | 0.3826*** | 0.1464*** | 6.197 | 0.0000 |
| length_mean | 226 | 0.2375*** | 0.0564*** | 3.660 | 0.0003 |
| length3601_mean | 226 | -0.0320 | 0.0010 | -0.479 | 0.6327 |

Note: **** $p > 0.05$

Through multi-stage tests (Table 7), statistically significant differences ($p < 0.05$) and homogeneous groups ($p > 0.05$) were identified.

The content subcategories *covered-bonds, information-for-shareholders, financial-reports, financial-statement, information-about-bank,* and *information-for-investors-except-shareholders* (Table 7) represent one homogeneous group in terms of user preferences expressed by time spent on the page taking into account the reading time of the document ($p > 0.05$). In this case, the subcategory *annual-reports* (Table 7) showed the highest user preference ($p < 0.05$). Statistically significant differences were identified (Table 7) between *annual-reports* and *information-for-shareholders, covered-bonds, financial-statement, financial-reports*.

**Table 7 - Multi-stage tests for lengthRT_doc_mean**

| subcategory \ lengthRT_doc_mean | Mean | Mean Rank | 1 | 2 |
|---|---|---|---|---|
| information-for-shareholders | 1.279 | 102.400 | **** | |
| covered-bonds | 2.413 | 56.718 | **** | |
| financial-statement | 3.372 | 134.944 | **** | |
| financial-reports | 3.700 | 126.944 | **** | |
| information-about-bank | 7.079 | 155.356 | **** | **** |
| information-for-investors-except-shareholders | 9.051 | 141.500 | **** | **** |
| annual-reports | 23.109 | 204.235 | | **** |

Note: **** $p > 0.05$

### 4.1 Readability and Complexity Validity

Subsequently, it was necessary to focus on the validity of the readability and complexity metrics, which were divided into ten groups. In several studies focused on economic texts, several readability metrics are used, with some authors preferring the Gunning Fog Index and, on the other hand, some preferring LIX or RIX (Ebaid, 2023; Guay et al., 2015; Moreno & Casasola, 2016). Therefore, the LIX, RIX, and LinsearWrite group of expert metrics was created. The Gunning Fog Index metric has been included among the readability metrics that are often used together. The correlation was calculated between the vectors of the metric group and the expert metric group, whose variable vector [expert] represents a valid criterion. A statistically significant degree of dependence was demonstrated between the investigated vectors of the metrics groups against the reference vector of the expert metrics group (Table 8). The results (Table 8) show that all categories are statistically significant, but the highest degree of dependence with the expert metrics group is achieved by the readability metrics group (including the Gunning Fog Index) and the group of

text characteristics metrics. This suggests that readability metrics and expert metrics can jointly point to similar features of complexity and readability of financial texts. Therefore, it does not matter that the Gunning Fog Index metric was not included among the expert metrics group, in principle readability metrics could also be chosen as a valid criterion.

**Table 8 - Canonical Analysis of vectors: [expert] & [variables]**

| [expert] & [variables] | Canonical R | df | Chi-square | p |
|---|---|---|---|---|
| [read] | 0.9996*** | 21 | 2124.700 | 0.0000 |
| [char] | 0.9995*** | 48 | 1987.000 | 0.0000 |
| [pos_ratio] | 0.9784*** | 54 | 1056.300 | 0.0000 |
| [synt] | 0.9393*** | 12 | 521.790 | 0.0000 |
| [lex_rich] | 0.8486*** | 57 | 616.180 | 0.0000 |
| [lex_sop] | 0.7977*** | 18 | 384.400 | 0.0000 |
| [lex_div] | 0.7861*** | 18 | 296.930 | 0.0000 |
| [lex_var] | 0.7455*** | 42 | 389.000 | 0.0000 |
| [syl] | 0.6810*** | 15 | 198.660 | 0.0000 |

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The last group of metrics [other] is diverse in terms of measuring text complexity, so it cannot be approached as a vector of consistent metrics. Therefore, the comparison was performed using multiple analysis in combinations of variable vs. vector of expert metrics group. All correlations were statistically significant in the case of the metrics of the group [other] (Table 9). In the case of metrics *other_DCR, other_qui_q, other_qui_d, other_eawl_unique, other_qui_vd, other_qui_ent*, a high degree of dependence was achieved ($R > 0.5$), in the case of metrics *other_RT, other_eawl, other_qui_hpoint, other_qui_tc, other_qui_stc*, a medium degree of dependence was achieved ($R > 0.3$) and in the case of *other_Size_in_kB* a small degree of dependence was achieved ($R > 0.1$). However, in all cases, the multiple statistical coefficient between the individual variables and the vector of the expert metrics group was statistically significant at the significance level of 0.001, except for the last, which was statistically significant at the significance level of 0.01.

**Table 9 - Correlation Analysis of variables: variable & [expert]**

| variable & [expert] | Multiple R | df1 | df2 | F | p |
|---|---|---|---|---|---|
| other_DCR | 0.6991*** | 3 | 222 | 70.748 | 0.0000 |
| other_qui_q | 0.6256*** | 3 | 222 | 47.585 | 0.0000 |
| other_qui_d | 0.6256*** | 3 | 222 | 47.585 | 0.0000 |
| other_eawl_unique | 0.5966*** | 3 | 222 | 40.903 | 0.0000 |
| other_qui_vd | 0.5856*** | 3 | 222 | 38.629 | 0.0000 |
| other_qui_ent | 0.5519*** | 3 | 222 | 32.408 | 0.0000 |
| other_RT | 0.4918*** | 3 | 222 | 23.604 | 0.0000 |
| other_eawl | 0.4421*** | 3 | 222 | 17.977 | 0.0000 |
| other_qui_hpoint | 0.4291*** | 3 | 222 | 16.696 | 0.0000 |
| other_qui_tc | 0.3662*** | 3 | 222 | 11.462 | 0.0000 |
| other_qui_stc | 0.3007*** | 3 | 222 | 7.355 | 0.0001 |
| other_Size_in_kB | 0.2607** | 3 | 222 | 5.396 | 0.0013 |

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

As a result of the validity of the readability and complexity metrics, all groups of metrics are usable. The newly proposed *other_eawl* metric, based on a dictionary of economic words, achieved a statistically significant moderate correlation. The *other_eawl_unique* metric derived from it, which considered only unique words, even reached a statistically significant large correlation with the vector of expert metrics group. These metrics have been shown to have the potential to describe the readability of economic texts.

### 4.2 Analysis of Readability/Complexity of Texts in the Context of User Preferences

The main goal of the research was to determine the relationship between complexity and readability metrics and user preferences (number of sessions in which the document is the target page, entropy of sessions, time spent on the page taking into account the reading time of the document, and the level of user interest based on density of keywords in the document). A dependency analysis was performed for each group of metrics in combination with user preference indicators. A multiple analysis was performed for a group of metrics, and then a one-dimensional analysis was calculated for individual metrics from the given group in combination with user preferences indicators. Due to the

presence of redundancy in the vectors [read], [pos_ratio], [lex_rich], and [syl], stepwise analysis was used to estimate multiple correlation coefficients, while the parameters of stepwise forward analysis were set so that as many variables of the vector as possible were included in the estimate (*F to enter* = 0.0001, *F to remove* = 0).

In the case of the created group of lexical variation metrics [lex_var] (Table 10), the highest statistically significant relationship was identified with the entropy of sessions and the smallest with the level of user interest based on density of keywords in the document, while it is statistically insignificant. The vector of lexical variation metrics is significantly correlated with the number of sessions in which a document is the target page, the entropy of sessions, and the time spent on the page, taking into account the reading time of the document. Multiple coefficients are statistically significant at the significance level of 0.001. Within the [lex_var] group, 14 metrics were used (Table 10) and in the case of 7 metrics a statistically significant dependency was identified with entropy of sessions and for 1 metric with the number of sessions in which a document is the target page at the significance level of 0.001, for 5 metrics a statistically significant dependency was identified with entropy of sessions and for 7 metrics with the number of sessions in which a document is the target page and 3 metrics with the time spent on the page taking into account the reading time of the document at the significance level of 0.01, and for 3 metrics a statistically significant dependence was identified with the number of sessions in which a document is the target page and for 3 metrics with the time spent on the page taking into account the reading time of the document at the significance level of 0.05.

In the case of the metrics of the [lex_var] group, a statistically insignificant relationship was identified with the level of user interest based on density of keywords in a document (Table 10). In the case of the NDW metric, the highest statistically significant relationship was identified with the entropy of sessions ($r$ = -0,4), the number of sessions in which a document is the target page ($r$ = 0,4) and the time spent on the page taking into account the reading time of the document ($r$ = 0,2). This is the basic metric of this category and talks about the level of language development of the child and refers to the extent of the reader's vocabulary (Malvern et al., 2004). The results indicate that the higher the value of NDW, the lower the entropy of the sessions (and vice versa), which means that the stakeholders who search for the examined documents in the sessions should have a larger vocabulary, and therefore the sessions are more organized, that is, they know where they want to find the information. These results are also confirmed by the relationship with the remaining two indicators of user preferences. The higher the value of NDW, the more time the stakeholder spends on the Web portal, and the greater the number of sessions in which the document is the target page. It indicates that expert documents are more likely to be visited by more expert stakeholders because they understand them.

**Table 10 - Dependency analysis of variables: selected user-oriented variables & [lex_var]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [lex_var]** | 0.231 | 0.467*** | 0.474*** | 0.486*** |
| **Multiple R2 [lex_var]** | 0.053 | 0.219*** | 0.225*** | 0.237*** |
| **Adjusted R2 [lex_var]** | 0.000 | 0.167*** | 0.173*** | 0.186*** |
| **r lex_var_lca_ndw** | -0.043 | 0.228** | 0.345*** | -0.397*** |
| **r lex_var_lca_ndwz** | 0.028 | -0.210** | -0.182** | 0.123 |
| **r lex_var_lca_ndwerz** | -0.053 | 0.087 | 0.141* | -0.208** |
| **r lex_var_lca_ndwesz** | 0.000 | -0.073 | -0.067 | 0.000 |
| **r lex_var_lca_uber** | -0.014 | 0.069 | 0.140* | -0.208** |
| **r lex_var_lca_lv** | 0.054 | -0.098 | -0.181** | 0.242*** |
| **r lex_var_lca_vv1** | 0.047 | -0.111 | -0.206** | 0.267*** |
| **r lex_var_lca_svv1** | -0.037 | -0.112 | 0.080 | -0.210** |
| **r lex_var_lca_cvv1** | -0.047 | -0.111 | 0.078 | -0.198** |
| **r lex_var_lca_vv2** | 0.080 | -0.145* | -0.192** | 0.240*** |
| **r lex_var_lca_nv** | 0.057 | -0.086 | -0.173** | 0.233*** |
| **r lex_var_lca_adjv** | 0.037 | -0.157* | -0.217** | 0.279*** |
| **r lex_var_lca_advv** | 0.001 | -0.181** | -0.154* | 0.209** |
| **r lex_var_lca_modv** | 0.069 | -0.154* | -0.189** | 0.240*** |

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The results of the analysis for the other eight groups of metrics can be found in Appendix B. Based on the results achieved, groups of metrics that achieved the highest relationship with the examined indicators of user preferences were identified.

**Readability/complexity of texts in comparison to the level of user interest**

The metrics of the [pos_ratio] group achieved the highest degree of dependence with the level of interest of the user based on density of keywords in the document (*Multiple R* = 0.848; *Multiple R2* = 0.720; *Adjusted R2* = 0.697), with the multiple coefficients being statistically significant at the significance level of 0.001 (Appendix B, Table 4). Within the [pos_ratio] group, 18 metrics were used and for 3 metrics (*pos_ratio_SCONJ, pos_ratio_INTJ, pos_ratio_PROPN*) a statistically significant dependence was identified at the significance level of 0.001, for 1 metric (*pos_ratio_VERB)* a statistically significant dependence was identified at the significance level of 0.01 and for 3 metrics (*pos_ratio_NUM, pos_ratio_CCONJ, pos_ratio_X*) a statistically significant dependence was identified at the significance level of 0.05. Interesting in this group were the results for the metrics that describe the ratio of proper nouns and verbs. From the group of other metrics [other], the multiple coefficients were shown to be statistically significant at the significance level of 0.01 for the *eawl_unique* metric with the level of user interest based on density of keywords in a document (Appendix B, Table 11). The results showed that the level of user interest based on density of keywords in a document, is mainly related to the ratio of numerals, the ratio of proper nouns, the ratio of verbs, and the ratio of unique economic words (eawl_unique). It can be assumed that a higher number of proper nouns ($r = 0.4$) and verbs ($r = 0.2$) can indicate a higher level of user interest based on density of keywords in a document. In contrast, a lower frequency of numbers ($r = -0.2$) in the text shows a higher level of user interest based on density of keywords in a document (Appendix B, Table 11). Likewise, the proposed metric of the ratio of unique economic words can capture the level of user interest based on density of keywords in a document. This may be because the generated keywords are mainly proper names, which are also found in the list of economic words. From the perspective of interpretating complexity metrics, it is appropriate to utilize the *eawl_unique* metric, even for laymen in the field of readability, for several reasons:

- It does not necessitate the use of a third-party tool, which would require time-consuming morphological annotation to identify the frequency of parts of speech;
- The list of economic words can be further expanded, thereby constantly improving the accuracy of the metric;
- Economic words are comprehensible to financial experts; however, they may be less comprehensible to ordinary stakeholders. It has been observed that a higher proportion of these words indicates a higher interest of users (r = 0.2), suggesting that financial experts are primarily interested in the given documents.

For example, in the case of the *information-for-shareholders-not-investors* subcategory, the document "slovakiavub_presentation_for_investors_062018" had an above-average interest level of 0.03029 and an eawl_unique value of 0.08, which means 8% of the unique economic words in the document. On the other hand, the *annual-reports* subcategory document named "vubannualreport14" had a below-average interest level of 0.00001 and an eawl_unique value of 0.02, so a lower percentage of economic words may reduce interest in that document among the finance experts.

**Time Spent on Page vs. Readability/Complexity**

The groups of metrics [char] and [lex_rich] achieved the highest degree of dependence with the time spent on the page taking into account the reading time of the document ([char]: *Multiple R* = 0.566; *Multiple R2* = 0.321; *Adjusted R2* = 0.269; [lex_rich]: *Multiple R* = 0.558; *Multiple R2* = 0.311; *Adjusted R2* = 0.258), while the multiple coefficients are statistically significant at the significance level of 0.001 (Appendix B, Table 3, 6). Within the [char] group, 16 complexity metrics were used, and for 5 metrics (*char_n_tokens, char_n_unique_tokens, char_n_characters, char_n_sentences, char_qui_tok*) a statistically significant dependence was identified at the significance level of 0.001, for 1 metric (char_qui_type) a statistically significant dependence was identified at the significance level of 0.01 and for 2 metrics (*char_token_length_mean, char_token_length_median*) a statistically significant dependence was identified at the significance level of 0.05. In the case of the [lex_rich] group, 19 complexity metrics were used, and for 2 metrics (*lex_rich_qui_mamr100, lex_rich_qui_mamr500*) a statistically significant dependence was identified at the significance level of 0.01 and for 1 metric (*lex_rich_qui_hapax*) a statistically significant dependence was identified at the significance level of 0.05. The results (Appendix B, Table 3) showed that the time spent on the page taking into account the reading time of the document is significantly related to the number of word units of the text, such as the number of characters ($r = 0.3$), tokens ($r = 0.2$) and unique tokens ($r = 0.2$). This was also confirmed by the results of the lexical richness group (Appendix B, Table 6), where both metrics depict the number of different word types in the text that turned out to be interesting ($r = -0.2$). From the group of metrics [other], it turned out that the multiple coefficients are statistically significant at the significance level of 0.001 for the *other_Size_in_kB* and *other_RT* metrics with the time spent on the page taking into account the reading time of the document (Appendix B, Table 11). Correlation coefficients are statistically significant at the level of significance of 0.01 for *other_DCR,*

*other_qui_hpoint, other_qui_tc* and *other_qui_stc* metrics with time spent on page taking into account document reading time. In the case of the metrics of the [other] group, it is positive that the reading time metric ($r$ = 0.3) is related to the time spent on the page taking into account the reading time of the document. The results (Appendix B, Table 3, 6, 11) showed that the longer the time spent on the page taking into account the reading time of the document is related to the longer length of the texts, which is also confirmed by the metrics on the reading time and the size of the given document ($r$ = 0.3). An interesting finding is that a lower rate of different word types ($r$ = -0.2) in texts increases the time spent on the pages.

For example, in the case of the *annual-reports* subcategory, the document "ar_2017_en_final_web" had an above-average level of average time spent on pages of 188 seconds, and the metric values for the document achieved an above-average score: number of tokens = 106317, number of sentences = 3378, reading time = 7484 seconds.

**Number of Sessions where the Document is the Target Page vs. Readability/Complexity**

The metrics of the [char] group achieved the highest degree of dependence with the number of sessions in which a document is the target page (*Multiple R* = 0.628; *Multiple R2* = 0.394; *Adjusted R2* = 0.348), with the multiple coefficients being statistically significant at the significance level of 0.001 (Appendix B, Table 3). Within the [char] group, 16 complexity metrics were used, and for 6 metrics (*char_n_tokens, char_n_unique_tokens, char_n_characters, char_n_sentences, char_qui_tok, char_qui_type*) a statistically significant dependence was identified at the significance level of 0.001 and for 2 metrics (*char_token_length_mean, char_token_length_median*) a statistically significant dependence was identified at the 0.05 significance level. The results (Appendix B, Table 3) showed that the number of sessions in which the document is the target page, like the time spent on the page, is mainly related to the frequency of word units of the text, such as the number of characters ($r$ = 0.4), tokens ($r$ = 0.4) and unique tokens ($r$ = 0.4). From the group of metrics [other], multiple coefficients were shown to be statistically significant at the significance level of 0.001 for the *other_Size_in_kB, other_RT*, and *other_qui_hpoint* metrics with the number of sessions in which a document is the target page (Appendix B, Table 11). Correlation coefficients are statistically significant at the significance level of 0.01 for the *other_qui_ent* metric, and a statistically significant dependence was identified at the significance level of 0.05 for the *other_DCR* metric (Appendix B, Table 11). A longer text length represented by sentence length ($r$ = 0.3) and a greater number of characters, tokens, and even unique tokens in a document have an impact on a greater number of sessions in which the document is the target page. These results (Appendix B, Table 11) are also confirmed by metrics from the [other] group, either file size ($r$ = 0.4) or reading time ($r$ = 0.4). Similarly, the h-point metric scores higher for larger texts ($r$ = 0.3). The text entropy metric ($r$ = 0.2) also confirms the results achieved, the greater the diversity of the vocabulary in the document is identified, the greater the number of sessions in which the document is the target page is found in the sessions.

For example, in the case of the *information-about-bank* subcategory, the document "pillar-iii_15_12_en" had an above-average level of sessions in which the document is a landing page of 35, and the metric values for the document scored above-average: number of tokens = 34367, number of sentences = 3952, reading time = 2526 seconds, h-point = 54 and text entropy = 9.59.

**Session Entropy vs. Readability/Complexity**

The metrics of the groups [char] and [pos_ratio] achieved the highest degree of dependence with the entropy of sessions ([char]: *Multiple R* = 0.646; *Multiple R2* = 0.418; *Adjusted R2* = 0.373; [pos_ratio]: *Multiple R* = 0.630; *Multiple R2* = 0.397; *Adjusted R2* = 0.348), while multiple coefficients are statistically significant at the significance level of 0.001 (Appendix B, Table 3, 4, 11). Within the [char] group, 16 complexity metrics were used, and for 8 metrics (*char_token_length_mean, char_token_length_median, char_n_tokens, char_n_unique_tokens, char_n_characters, char_n_sentences, char_qui_tok, char_qui_type*) a statistically significant dependence was identified at the significance level of 0.001, for 4 metrics (*char_sentence_length_mean, char_sentence_length_median, char_syllables_per_token_mean, char_syllables_per_token_std*) a statistically significant dependence was identified at the significance level of 0.01 and for 1 metric (*char_sentence_length_std*) a statistically significant dependence was identified at the significance level of 0.05. In the case of the [pos_ratio] group, 18 complexity metrics were used, and for 5 metrics (*pos_ratio_PROPN, pos_ratio_NUM, pos_ratio_PUNCT, pos_ratio_AUX, pos_ratio_INTJ*) a statistically significant dependence was identified at the significance level of 0.001 and for 1 metric (*pos_ratio_ADJ*) a statistically significant dependence was identified at the significance level of 0.05. The results (Appendix B, Table 3) showed that the session entropy is also related to the abundance of characters ($r$ = -0.5), tokens ($r$ = -0.4) and unique tokens ($r$ = -0.4). This is demonstrated more closely by the metrics from the [pos_ratio] group, where a higher share of proper names ($r$ = -0.4) in the document reduces the entropy of the sessions, that is, the stakeholder searches for linked information in the sessions (Appendix B, Table 4). From the group of

metrics [other], multiple coefficients were shown to be statistically significant at the significance level of 0.001 for the *other_Size_in_kB, other_RT, other_qui_hpoint*, and *other_qui_ent* metrics with entropy of sessions. Correlation coefficients are statistically significant at the significance level of 0.01 for the *other_DCR* metric with entropy of sessions (Appendix B, Table 11). The results (Appendix B, Table 11) showed that a lower entropy of sessions relates to a longer length of texts, confirmed by the metrics reading time ($r$ = -0.5) and Size_in_kB ($r$ = -0.2), and a larger document entropy ($r$ = -0.3), which represents the measure of different word types in the document.

For example, in the case of the *financial-reports* subcategory, the document "semi-annual-financial-report-for-the-year-2012" had an above-average session entropy level = 0.99990, and the metric values for the document scored below average: number of tokens = 30354, number of sentences = 396, reading time = 1188 seconds, h-point = 44 and text entropy = 8.64.

## 5. Discussion

The aim of this research was to create indicators of user preferences on the Web portals of commercial banks and to investigate whether the complexity and readability of disclosure information have an impact on the user preferences. During the experiment, two text complexity metrics, eawl and eawl_unique, were created, which focus on economic texts. The main contribution of the experiment can be considered indicators of user preferences at three levels in terms of web portal accesses: entropy of sessions, number of sessions in which the document is the target page, and levels of user interest based density of keywords in the document, and in terms of time: time spent on the page taking into account reading time document. Access-oriented indicators were compared with support, which was used as a reference in this case. Each of the user preference indicators has been shown to also capture the complexity of documents using different complexity or readability metrics for each indicator. The group of metrics of text characteristics, such as different frequencies of tokens, sentences, characters, achieved the highest degree of dependence with user preference based on the entropy of sessions, the number of sessions in which the document is the target page, and the time spent on the page taking into account the reading time of the document. The length of the text plays a significant role in its complexity and readability. Longer documents contain more information and therefore, according to the results achieved, are more interesting to stakeholders, and they prefer them over short documents. From the point of view of the indicator of the level of user interest based on density of keywords in the document, the proposed metric eawl_unique, which was created with the intention of describing the complexity of economic texts, proved to be interesting.

An interesting result was provided by non-parametric estimates in the case of the number of sessions in which the document is the target page, where the *annual-reports* category obtained the highest value, while in the case of the other indicators investigated (entropy of sessions and level of user interest based on specificity and density of keywords in a document), it did not reach such high values. Similarly, it also achieved the highest value in terms of the indicator of the time spent on the page, taking into account the reading time of the document. Annual report of 2017 was the highest visited document (1208 accesses), but also other *annual-reports* documents had above average access rate in comparison to documents from other categories. Annual reports provide a comprehensive information on the financial performance of banks and is prepared as a combination of explanatory text and quantitative information. This format might provide at least partial explanation why the use of this document is higher compared to others where more technical knowledge is necessary to understand the presented data.

The results of the complexity and readability of the documents for the categories (Pillar3 related and Pillar3 disclosure requirements), have shown that more complex texts are in the *Pillar3 disclosure requirements* category (Table 11). This was analyzed using the complexity/readability metrics that have achieved an association with the user preference indicators. Ten complexity metrics were used, and for two metrics (*pos_ratio_PROPN, other_eawl_unique*) a statistically significant dependence was identified at the significance level of 0.001 and for two metrics (*expert_LW, expert_lix*) a statistically significant dependence was identified at the 0.05 significance level with *Pillar3 disclosure requirements* category. The standard readability metrics like LW, LIX, and Gunning Fog Index achieved mean scores that represent texts intended for university graduate level readers. A lower score (representing 7$^{th}$ or 8$^{th}$ grade) was achieved in the case of Coleman Liau Index but this could be caused by that the metric does not consider complex sentences where a text with simple sentences, but complex concepts may yield a misleading score. The remaining text complexity metrics focus on characteristics as ratio of proper nouns, number of sentences, variety of unique words (including economic terminology), and document reading time. These characteristics indicate more text complexity and worse readability in documents within the *Pillar3 disclosure requirements* category.

**Table 11 - Comparing two independent samples (category) for selected complexity/readability metrics**

| Category | Pillar3 related | Pillar3 disclosure requirements |
|---|---|---|
| | Mean | Mean |
| expert_LW | 26.467 | **27.542*** |
| expert_lix | 65.647 | **69.953*** |
| read_coleman_liau_index | 7.442 | **8.399** |
| read_gunning_fog | 21.958 | **23.615** |
| pos_ratio_PROPN | 0.061 | **0.078***** |
| char_qui_type | 3315.184 | **3418.159** |
| char_n_sentences | 796.301 | **879.508** |
| other_RT | 1612.151 | **1800.210** |
| other_qui_ent | 9.145 | **9.236** |
| other_eawl_unique | 0.033 | **0.046***** |

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, marked represent higher scores achieved

A more detailed look at the subcategories offers more insight into the readability of examined documents. Similar text readability/complexity metrics as before were used (Table 12a,b,d). The readability metrics from the expert (LW, LIX) and readability group (Gunning Fog Index) showed that despite all of the subcategory documents are very difficult texts, documents from *financial-reports* are the most difficult. On the other hand, based on the subcategory *information-for-shareholders* and *annual-reports* contain fewer complex texts. From the point of view of user preferences, in the case of Gunning Fog Index and LIX metric the user interest based on the entropy of sessions is higher the more complex are the documents. This means that the visitor looking for more complex documents had issues to find the required information. The documents with higher LW metric score show a higher time spent on the documents by the stakeholders.

A very low score was achieved in the case of Coleman Liau Index for the *financial-reports* subcategory, that is out of ordinary (Table 12c). This could be caused by that the metric does not consider complex sentences where a text with simple sentences, but complex concepts may yield a misleading score. The results of this metric will be omitted in this case.

The subcategory *information-for-investors-except-shareholders* should be omitted in the case of text characteristic metrics (Table 12e-j) as it contained only 2 documents, but the scores are kept for clarity. The results of the ratio of proper nouns show that based on readability (Table 12a,b,d) more complex texts contain fewer proper nouns (Table 12e). The analysis of user preferences shows that user interest of documents is very high for documents with higher ratio of proper nouns, and this means subcategories *information-about-bank* and *annual-reports*. The subcategory *annual-reports* achieved the highest mean scores for almost all of the rest of the readability/complexity metrics (ratio of proper nouns, number of sentences, variety of unique words, and document reading time).

The last metric (Table 12j) measuring the ratio of unique economic words in documents showed that the subcategory *annual-reports* contained fewer unique economic words despite that the texts were longer. This shows that documents from *annual-reports* are longer but have better readability than other examined documents. Similar results were obtained also for the subcategory *information-about-bank*. On the other hand, documents of the subcategory *financial-reports* are considered as more complex and harder to read, contain fewer proper nouns, lower variety of unique words and more unique economic words.

**Table 12 - Summary results for examined subcategories and for selected complexity/readability metrics**

| subcategory | expert_LW |
|---|---|
| **(a)** | **Mean** |
| information-for-shareholders | 19.774 |
| annual-reports | 25.291 |
| covered-bonds | 26.240 |
| information-about-bank | 26.586 |
| financial-statement | 27.751 |
| information-for-investors-except-shareholders | 28.175 |
| financial-reports | 29.933 |

| subcategory | expert_lix |
|---|---|
| **(b)** | **Mean** |
| information-for-shareholders | 56.039 |
| financial-statement | 57.700 |
| annual-reports | 57.705 |
| information-for-investors-except-shareholders | 60.024 |
| information-about-bank | 61.137 |
| covered-bonds | 72.982 |
| financial-reports | 91.991 |

| subcategory | read_coleman_liau_index |
|---|---|
| (c) | Mean |
| financial-reports | 2.546 |
| covered-bonds | 4.592 |
| financial-statement | 10.458 |
| annual-reports | 10.518 |
| information-about-bank | 10.740 |
| information-for-shareholders | 11.247 |
| information-for-investors-except-shareholders | 11.482 |

| subcategory | read_gunning_fog |
|---|---|
| (d) | Mean |
| information-for-shareholders | 16.920 |
| annual-reports | 17.387 |
| financial-statement | 17.886 |
| information-for-investors-except-shareholders | 19.000 |
| information-about-bank | 19.483 |
| covered-bonds | 25.824 |
| financial-reports | 33.946 |

| subcategory | pos_prop_PROPN |
|---|---|
| (e) | Mean |
| financial-reports | 0.030 |
| financial-statement | 0.053 |
| covered-bonds | 0.056 |
| information-for-shareholders | 0.071 |
| annual-reports | 0.086 |
| information-about-bank | 0.098 |
| information-for-investors-except-shareholders | 0.205 |

| subcategory | char_qui_type |
|---|---|
| (f) | Mean |
| information-for-shareholders | 816.600 |
| information-for-investors-except-shareholders | 1236.000 |
| covered-bonds | 1885.224 |
| financial-reports | 1918.056 |
| information-about-bank | 4018.200 |
| financial-statement | 4551.074 |
| annual-reports | 7518.706 |

| subcategory | char_n_sentences |
|---|---|
| (g) | Mean |
| information-for-investors-except-shareholders | 142.500 |
| information-for-shareholders | 207.600 |
| financial-reports | 310.333 |
| covered-bonds | 353.929 |
| information-about-bank | 1107.178 |
| financial-statement | 1127.426 |
| annual-reports | 2206.412 |

| subcategory | other_RT |
|---|---|
| (h) | Mean |
| information-for-investors-except-shareholders | 351.650 |
| information-for-shareholders | 399.586 |
| covered-bonds | 814.979 |
| financial-reports | 915.834 |
| financial-statement | 2130.450 |
| information-about-bank | 2153.961 |
| annual-reports | 4456.584 |

| subcategory | other_qui_ent |
|---|---|
| (i) | Mean |
| information-for-shareholders | 7.374 |
| information-for-investors-except-shareholders | 8.602 |
| financial-reports | 8.633 |
| covered-bonds | 8.673 |
| information-about-bank | 9.478 |
| financial-statement | 9.757 |
| annual-reports | 10.149 |

| subcategory | other_eawl_unique |
|---|---|
| (j) | Mean |
| financial-statement | 0.018 |
| annual-reports | 0.026 |
| information-about-bank | 0.037 |
| covered-bonds | 0.041 |
| financial-reports | 0.068 |
| information-for-investors-except-shareholders | 0.068 |
| information-for-shareholders | 0.069 |

From the point of view of readability and complexity of the text, it turned out that all categories of readability/complexity metrics are statistically significant, while the highest degree of dependence with professional metrics is achieved by readability metrics (among others, the Gunning Fog Index, which serves for several authors as an indicator of the readability of economic texts) and metrics of text characteristics. This suggests that readability metrics and expert metrics can jointly point to similar features of complexity and readability of financial texts. More than half of the documents examined (151 out of 226) achieved a Gunning Fog Index score representing college graduate level readability (*index* > 17), indicating highly specialized documents. Similar results were also achieved for the LIX metric (*index* > 56), which identified more than half of the documents (157 out of 226) as professional texts. However, in the combination of readability/complexity metrics with user preference indicators, in the case of the Gunning Fog Index metric, multiple coefficients are statistically significant at the significance level of 0.01 with only entropy of sessions. For the other indicators, the multiple coefficients are statistically insignificant. In the experiment, 110 complexity/readability metrics were implemented divided into 10 groups. The results of the multiple

analysis identified only one metric (*pos_ratio_PROPN*) for which the multiple coefficients were statistically significant for all user preference indicators. This means that from the point of view of user preferences, the share of proper names in documents increases the interest in these documents from stakeholders.

The results obtained confirm the results of previous studies (Munk et al., 2021c; Pilková, Munk, Benko, et al., 2021). This research followed up on previous findings of very low interest in Pillar 3 information from commercial bank stakeholders operating in Central and Eastern Europe. From the results of the surveys so far, it can be deduced that the group of clients interested in this information in this type of commercial banks are those who have uninsured deposits in the bank (legal entities, individuals with deposits above EUR 100 thousand). However, as the results of our research show, these clients are more interested in less demanding and more readable texts such as annual reports, and as Pillar 3 and other Pillar documents are more complex and harder to read, they are less interested in them. This leads to an important conclusion for regulators: increasing the interest in this information in this type of banks requires finding ways of presenting it that make it less complicated and more readable.

## 6. Conclusions

Regulatory bodies are trying to promote the transparency and efficiency of financial markets; therefore, they are trying to ensure that the financial documents published to stakeholders are readable and understandable. The results of this study indicated that, despite the fact that published financial documents show a high level of complexity, some groups of stakeholders show interest in them. In combination with readability and complexity metrics, user preferences indicators were designed, which can point to the accesses of web portals together with the already proven support method. From the point of view of interpretability, the indicator of the number of sessions in which the document is the target page acts as the best choice. If a document is found in more sessions as the target page, then it is clear that there is considerable interest in this document and, therefore, in the information in it. In the examined period for 2018, from this point of view, the most interesting document was from the subcategory *annual-reports* (Pillar3 related) turned out to be the document of the annual report for 2017. From the point of view of the subcategory *information-about-bank* (Pillar3 disclosure information), was among the most interesting documents pillar-iii_30_09_sj (Disclosure of information about the bank as of September 30, 2013). The main contribution of the study is the detailed methodology proposed to obtain user preference indicators and their combination with readability/complexity metrics. The limitations of the research are in the number of documents extracted, which was 226, of which only a few documents could be extracted in some examined subcategories. Working with older log files brings limitations in that the web portal has changed since then, and some documents are no longer accessible. One possibility is to use archival records from websites that store a trace from a given period and a given web portal. However, even this does not always guarantee that it will be possible to extract the given documents. Future research will focus on comparing indicators of user preferences and document readability over a period of several years, taking into account turbulent periods and the revision of disclosure information.

**Acknowledgements**

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and Slovak Academy of Sciences (SAS) under the contract No. VEGA-1/0734/24.

**References**

Akbar, S., & Deegan, C. (2021). Analysis of corporate social disclosures of the apparel industry following crisis: an institutional approach. *Accounting and Finance*, *61*, 3565–3600. https://doi.org/10.1111/acfi.12712

Anderson, J. (1983). Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, *26*(6), 490–496.

Arnaud, P. J. L. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics*. Palgrave Macmillan.

Awan, M. D. A., Kajla, N. I., Firdous, A., Husnain, M., & Missen, M. M. S. (2021). Event classification from the Urdu language text on social media. *PeerJ Computer Science*, *7*, e775. https://doi.org/10.7717/peerj-cs.775

Aymen, A., Sourour, B. S., & Badreddine, M. (2018). The effect of annual report readability on financial analysts behaviour. *Journal of Economics, Finance and Accounting*, *5*(1), 26–37. https://doi.org/10.17261/Pressacademia.2018.782

BIS Connect. (2019). *Pillar 3 framework - Executive Summary*. https://www.bis.org/fsi/fsisummaries/pillar3_framework.pdf

Björnsson, C. H. (1968). *Lasbarhet*. Bokforlaget Liber.

Buzarna-Tihenea, A. (2020). An Analysis of Written Texts in the Economic Field. Case Study. *"Ovidius" University Annals, Economic Sciences Series*, *XX*(2), 259–265.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.

Chall, J. S., & Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*, 283–284.

Colliard, J.-E., & Georg, C.-P. (2023). Measuring Regulatory Complexity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3523824

Cvrček, V., Čech, R., & Kubát, M. (2020). *QuitaUp – nástroj pro kvantitativní stylometrickou analýzu*. Czech National Corpus and University of Ostrava. https://korpus.cz/quitaup/

Cvrček, V., & Chlumská, L. (2015). Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguistics*, *39*(3), 309–325. https://doi.org/10.1007/s11185-015-9151-8

Das, S. R. (2014). Text and Context: Language Analytics in Finance. *Foundations and Trends® in Finance*, *8*(3), 145–261. https://doi.org/10.1561/0500000045

de Araujo, P., & Leyshon, K. I. (2016). The impact of international information disclosure requirements on market discipline. *Applied Economics*, *49*(10), 954–971. https://doi.org/10.1080/00036846.2016.1208361

Del Gaudio, B. L., Megaravalli, A. V., Sampagnaro, G., & Verdoliva, V. (2020). Mandatory disclosure tone and bank risk-taking: Evidence from Europe. *Economics Letters*, *186*, 108531. https://doi.org/10.1016/j.econlet.2019.108531

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Dugast, D. (1979). *Vocabulaire et stylistique: Théâtre et dialogue*. Slatkine-Champion.

Ebaid, I. E.-S. (2023). IFRS adoption and the readability of corporate annual reports: evidence from an emerging market. *Future Business Journal*, *9*(1), 80. https://doi.org/10.1186/s43093-023-00244-x

Ehara, Y. (2021). To What Extent Can English-as-a-Second Language Learners Read Economic News Texts? *Proceedings of the Third Workshop on Economics and Natural Language Processing*, 62–68. https://doi.org/10.18653/v1/2021.econlp-1.9

Ehara, Y. (2022). Neural Language Model-based Readability Assessment of Computer Science Introductory Texts for English-as-a-Second Language Learners. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 1698–1704.

Faria-e-Castro, M., Martinez, J., & Philippon, T. (2017). *Runs versus Lemons: Information Disclosure and Fiscal Capacity*. https://doi.org/10.3386/w21201

Fernandes, C., Farinha, J., Vitorino Martins, F., & Mateus, C. (2021). The impact of board characteristics and CEO power on banks' risk-taking: stable versus crisis periods. *Journal of Banking Regulation*. https://doi.org/10.1057/s41261-021-00146-4

Flesch, R. (2016). *How to Write Plain English*. University of Cantenbury. https://web.archive.org/web/20160712094308/http:/www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

Gray, W. S., & Leary, B. E. (1935). *What Makes a Book Readable: With Special Reference to Adults of Limited Reading Ability*. University of Chicago Press.

Guay, W. R., Samuels, D., & Taylor, D. J. (2015). Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2564350

Guiraud, P. (1960). *Probl`emes et m´ethodes de la statistique linguistique*. Springer.

Gunning, T. G. (2003). The Role of Readability in Today's Classrooms. *Topics in Language Disorders*, *23*(3), 175–189.

Hayo, B., Henseler, K., & Rapp, M. S. (2019). Complexity of ECB Communication and Financial Market Trading. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3434451

Herdan, G. (1964). *Quantitative linguistics*. Butterworths.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development*, *9*(1–2), 67–84. https://doi.org/10.1080/01434632.1988.9994320

Jia, J., & Li, Z. (2022). Risk management committees and readability of risk management disclosure. *Journal of Contemporary Accounting & Economics*, *18*(3), 100336. https://doi.org/10.1016/j.jcae.2022.100336

Kapusta, J., Munk, M., & Drlík, M. (2012). Cut-off time calculation for user session identification by reference length. *2012 6th International Conference on Application of Information and Communication Technologies, AICT 2012 - Proceedings*. https://doi.org/10.1109/ICAICT.2012.6398500

Kincaid, P. J., Fishburne Jr., R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel*.

Kintsch, W. (1974). *The Representation of Meaning in Memory* . Lawrence Erlbaum. https://doi.org/10.4324/9781315794563

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*, *12*, 28–41.

Laufer, B. (1994). The Lexical Profile of Second Language Writing: Does It Change Over Time? *RELC Journal*, *25*(2), 21–33. https://doi.org/10.1177/003368829402500202

Lesmy, D., Muchnik, L., & Mugerman, Y. (2023). Lost in the FOG: Growing Complexity in Financial Reporting – A Comparative Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4542676

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. CWK Gleerup.

Linsley, P. M., & Lawrence, M. J. (2007). Risk reporting by the largest UK companies: readability and lack of obfuscation. *Accounting, Auditing & Accountability Journal*, *20*(4), 620–627. https://doi.org/10.1108/09513570710762601

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, *69*(4), 1643–1671. https://doi.org/10.1111/jofi.12162

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. *TESOL Quaterly*, *45*(1), 36–62.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230511804

Maqsood, S., Shahid, A., Tanvir Afzal, M., Roman, M., Khan, Z., Nawaz, Z., & Aziz, M. H. (2022). Assessing English language sentences readability using machine learning models. *PeerJ Computer Science*, *7*, e818. https://doi.org/10.7717/peerj-cs.818

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [PhD. Thesis]. The University of Memphis.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*, *2*, 141–154.

McLaughlin, H. G. (1969). SMOG Grading - a New Readability Formula. *Journal of Reading*, *12*(8), 639–646.

Miller, J. F. (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research in child language disorders: A decade of progress* (pp. 211–220). Pro-Ed.

Moreno, A., & Casasola, A. (2016). A Readability Evolution of Narratives in Annual Reports. *Journal of Business and Technical Communication*, *30*(2), 202–235. https://doi.org/10.1177/1050651915620233

Munk, M., Benko, Ľ., Gangur, M., & Turčáni, M. (2015). Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. *E+M Ekonomie a Management*, *18*(3), 144–159. https://doi.org/dx.doi.org/10.15240/tul/001/2015-3-013

Munk, M., Pilkova, A., Benko, L., & Blažeková, P. (2017). Pillar 3: market discipline of the key stakeholders in CEE commercial bank and turbulent times. *Journal of Business Economics and Management*, *18*(5), 954–973. https://doi.org/10.3846/16111699.2017.1360388

Munk, M., Pilkova, A., Benko, L., Blazekova, P., & Svec, P. (2021a). Methodology of stakeholders' behaviour modelling based on time. *MethodsX*, *8*, 101570. https://doi.org/10.1016/j.mex.2021.101570

Munk, M., Pilkova, A., Benko, Ľ., Blazekova, P., & Svec, P. (2021b). Pillar 3–Pre-processed web server log file dataset of the banking institution. *Data in Brief*, *39*, 107672. https://doi.org/10.1016/j.dib.2021.107672

Munk, M., Pilkova, A., Benko, L., Blazekova, P., & Svec, P. (2021c). Web usage analysis of Pillar 3 disclosed information by deposit customers in turbulent times. *Expert Systems with Applications*, *185*, 115503. https://doi.org/10.1016/j.eswa.2021.115503

Munkova, D., Munk, M., Benko, Ľ., & Hajek, P. (2021). The role of automated evaluation techniques in online professional translator training. *PeerJ Computer Science*, *7*, e706. https://doi.org/10.7717/peerj-cs.706

Niessen-Ruenzi, A., Parwada, J. T., & Ruenzi, S. (2015). Information Effects of the Basel Bank Capital and Risk Pillar 3 Disclosures on Equity Analyst Research An Exploratory Examination. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2670418

O'Flynn, J. A. (2019). An Economics Academic Word List (EAWL): Using online resources to develop a subject-specific word list and associated teaching-learning materials. *Journal of Academic Language and Learning*, *13*(1).

O'Hayre, J. (1966). *Gobbledygook has gotta go*. U.S. Government Printing Office.

Oliveira, J., Lima Rodrigues, L., & Craig, R. (2011). Voluntary risk reporting to enhance institutional and organizational legitimacy. *Journal of Financial Regulation and Compliance*, *19*(3). https://doi.org/10.1108/13581981111147892

Parwada, J. T., Ruenzi, S., & Sahgal, S. (2013). Market Discipline and Basel Pillar 3 Reporting. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2443189

Pilková, A., Munk, M., Benko, Ľ., Blažeková, P., & Kapusta, J. (2021). Pillar 3: Does banking regulation support stakeholders' interest in banks financial and risk profile? *PLOS ONE*, *16*(10), e0258449. https://doi.org/10.1371/journal.pone.0258449

Pilková, A., Munk, M., Blažeková, P., & Benko, Ľ. (2021). Web usage analysis: Pillar 3 information assessment in turbulent times. In M. Z. Abedin, K. Hassan, P. Hajek, & M. M. Uddin (Eds.), *The Essentials of Machine Learning in Finance and Accounting* (p. 24). Routledge. https://doi.org/10.4324/9781003037903

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170. https://doi.org/10.18653/v1/K18-2016

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Sadeek Quaderi, S. J., & Varathan, K. D. (2024). Identification of significant features and machine learning technique in predicting helpful reviews. *PeerJ Computer Science*, *10*, e1745. https://doi.org/10.7717/peerj-cs.1745

Sarwar, T. Bin, & Noor, N. M. (2021). An Experimental Comparison of Unsupervised Keyphrase Extraction Techniques for Extracting Significant Information from Scientific Research Articles. *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational*

*Science and Information Management (ICSECS-ICOCSIM)*, 130–135. https://doi.org/10.1109/ICSECS52883.2021.00031

Sarwar, T. Bin, Noor, N. M., Miah, M. S. U., Rashid, M., Farid, F. Al, & Husen, M. N. (2021). Recommending Research Articles: A Multi-Level Chronological Learning-Based Approach Using Unsupervised Keyphrase Extraction and Lexical Similarity Calculation. *IEEE Access*, *9*, 160797–160811. https://doi.org/10.1109/ACCESS.2021.3131470

Senter, R., & Smith, E. (1967). *Automated Readability Index*.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Smailović, J., Žnidaršič, M., Valentinčič, A., Lončarski, I., Pahor, M., Martins, P. T., & Pollak, S. (2018). Automatic Analysis of Annual Financial Reports: A Case Study. *Computación y Sistemas*, *21*(4). https://doi.org/10.13053/cys-21-4-2863

Spiers, H., Amin, N., Lakhani, R., Martin, A. J., & Patel, P. M. (2017). Assessing Readability and Reliability of Online Patient Information Regarding Vestibular Schwannoma. *Otology & Neurotology*, *38*(10), e470–e475. https://doi.org/10.1097/MAO.0000000000001565

Templin, M. (1957). *Certain language skills in children: Their development and interrelationships*. The University of Minnesota Press.

Thordardottir, E. T., & Weismer, S. E. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders*, *36*(2), 221–244. https://doi.org/10.1080/13682820118239

Toerien, F. E., & du Toit, E. (2024). Fighting through the Flesch and Fog: the readability of risk disclosures. *Accounting Research Journal*, *37*(1), 39–56. https://doi.org/10.1108/ARJ-03-2023-0094

Vauhkonen, J. (2012). The Impact of Pillar 3 Disclosure Requirements on Bank Safety. *Journal of Financial Services Research*, *41*(1–2), 37–49. https://doi.org/10.1007/s10693-011-0107-x

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication*, *3*, 215–229.

Yao, Z., Wang, X., & Luan, J. (2017). Using Hidden Markov Model to Predict the Web Users' Linkage. *Journal of Residuals Science & Technology*, *14*(3), 554–565. https://doi.org/10.14355/jrst.2017.1403.053

## APENDIX A

**Table 1 - Descriptive statistics of variables determining user interest estimation**

| Variable | Valid N | Mean | Trimmed mean 20.00% | Winsorized mean 20.00% | Grubbs Test Statistic | p-value | Median | Mode |
|---|---|---|---|---|---|---|---|---|
| **UserInt_Hor_orig** | 8621 | 0.000129 | 0.000000 | 0.000000 | 17.217750 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Ver_orig** | 8621 | 0.000004 | 0.000000 | 0.000000 | 29.904120 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Hor_3601** | 8621 | 0.000129 | 0.000000 | 0.000000 | 17.217750 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Ver_3061** | 8621 | 0.000004 | 0.000000 | 0.000000 | 29.904120 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Hor_rt** | 8621 | 0.000129 | 0.000000 | 0.000000 | 17.217750 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Ver_rt** | 8621 | 0.000004 | 0.000000 | 0.000000 | 29.904120 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Hor_rtdoc** | 8621 | 0.000129 | 0.000000 | 0.000000 | 17.217750 | 0.000000 | 0.000000 | 0.000000 |
| **UserInt_Ver_rtdoc** | 8621 | 0.000004 | 0.000000 | 0.000000 | 29.904120 | 0.000000 | 0.000000 | 0.000000 |

| Variable | Frequency of Mode | Sum | Min | Max | Lower Quartile | Upper Quartile | Std.Dev. | Coef.Var. |
|---|---|---|---|---|---|---|---|---|
| **UserInt_Hor_orig** | 918.000000 | 1.110830 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.001033 | 801.803000 |
| **UserInt_Ver_orig** | 8544.000000 | 0.030516 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000059 | 1680.645000 |
| **UserInt_Hor_3601** | 918.000000 | 1.110830 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.001033 | 801.803000 |
| **UserInt_Ver_3061** | 8544.000000 | 0.030516 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000059 | 1680.645000 |
| **UserInt_Hor_rt** | 918.000000 | 1.110830 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.001033 | 801.803000 |
| **UserInt_Ver_rt** | 8544.000000 | 0.030516 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000059 | 1680.645000 |
| **UserInt_Hor_rtdoc** | 918.000000 | 1.110830 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.001033 | 801.803000 |
| **UserInt_Ver_rtdoc** | 8544.000000 | 0.030516 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000059 | 1680.645000 |

## APENDIX B

**Table 1 - Dependency analysis of variables: selected user-oriented variables & [expert]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [expert]** | 0.141 | 0.253** | 0.267*** | 0.352*** |
| **Multiple R2 [expert]** | 0.020 | 0.064** | 0.071*** | 0.124*** |
| **Adjusted R2 [expert]** | 0.007 | 0.052** | 0.059*** | 0.112*** |
| **r expert_LW** | 0.008 | 0.163* | 0.091 | -0.069 |
| **r expert_lix** | -0.049 | -0.074 | -0.074 | 0.199** |
| **r expert_rix** | -0.004 | 0.005 | 0.019 | 0.087 |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 2 - Dependency analysis of variables: selected user-oriented variables & [read]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [read]** | 0.243* | 0.235 | 0.296** | 0.377*** |
| **Multiple R2 [read]** | 0.059* | 0.055 | 0.087** | 0.142*** |
| **Adjusted R2 [read]** | 0.038* | 0.029 | 0.062** | 0.118*** |
| **r read_flesch_reading_ease** | 0.040 | 0.069 | 0.076 | -0.195** |
| **r read_MAR** | -0.050 | 0.028 | 0.051 | -0.056 |
| **r read_flesch_kincaid_grade** | -0.047 | -0.095 | -0.094 | 0.220** |
| **r read_smog** | -0.005 | -0.042 | -0.031 | 0.132* |
| **r read_gunning_fog** | -0.050 | -0.098 | -0.096 | 0.222** |
| **r read_automated_readability_index** | -0.032 | -0.075 | -0.068 | 0.192** |
| **r read_coleman_liau_index** | 0.075 | 0.141* | 0.142* | -0.263*** |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 3 - Dependency analysis of variables: selected user-oriented variables & [char]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [char]** | 0.374** | 0.566*** | 0.628*** | 0.646*** |
| **Multiple R2 [char]** | 0.140** | 0.321*** | 0.394*** | 0.418*** |
| **Adjusted R2 [char]** | 0.074** | 0.269*** | 0.348*** | 0.373*** |
| **r char_token_length_mean** | 0.072 | 0.139* | 0.139* | -0.260*** |
| **r char_token_length_median** | 0.074 | 0.143* | 0.151* | -0.279*** |
| **r char_token_length_std** | 0.150* | 0.048 | -0.012 | -0.014 |
| **r char_sentence_length_mean** | -0.048 | -0.103 | -0.098 | 0.224** |
| **r char_sentence_length_median** | -0.064 | -0.129 | -0.105 | 0.227** |
| **r char_sentence_length_std** | -0.054 | 0.004 | -0.057 | 0.164* |
| **r char_syllables_per_token_mean** | 0.050 | 0.124 | 0.105 | -0.217** |
| **r char_syllables_per_token_std** | 0.086 | 0.118 | 0.105 | -0.217** |
| **r char_n_tokens** | -0.038 | 0.246*** | 0.364*** | -0.362*** |
| **r char_n_unique_tokens** | -0.039 | 0.241*** | 0.371*** | -0.427*** |
| **r char_proportion_unique_tokens** | 0.133* | 0.011 | -0.076 | 0.058 |
| **r char_n_characters** | -0.048 | 0.297*** | 0.422*** | -0.453*** |
| **r char_n_sentences** | -0.058 | 0.258*** | 0.377*** | -0.439*** |
| **r char_qui_tok** | -0.050 | 0.296*** | 0.420*** | -0.450*** |
| **r char_qui_type** | -0.045 | 0.223** | 0.349*** | -0.404*** |
| **r char_qui_atl** | -0.060 | 0.113 | 0.061 | -0.128 |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 4 - Dependency analysis of variables: selected user-oriented variables & [pos_ratio]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [pos_ratio]** | 0.848*** | 0.459*** | 0.473*** | 0.630*** |
| **Multiple R2 [pos_ratio]** | 0.720*** | 0.210*** | 0.223*** | 0.397*** |
| **Adjusted R2 [pos_ratio]** | 0.697*** | 0.146*** | 0.160*** | 0.348*** |
| **r pos_ratio_SPACE** | 0.020 | -0.011 | 0.044 | -0.024 |
| **r pos_ratio_NOUN** | 0.071 | 0.054 | 0.083 | -0.074 |
| **r pos_ratio_ADP** | -0.018 | 0.011 | 0.037 | -0.014 |
| **r pos_ratio_DET** | -0.012 | 0.029 | 0.018 | -0.120 |
| **r pos_ratio_PROPN** | 0.360*** | 0.155* | 0.211** | -0.352*** |
| **r pos_ratio_NUM** | -0.152* | -0.095 | -0.165* | 0.247*** |
| **r pos_ratio_VERB** | 0.178** | 0.073 | 0.033 | -0.110 |
| **r pos_ratio_CCONJ** | 0.161* | -0.043 | -0.065 | 0.008 |
| **r pos_ratio_ADJ** | -0.017 | -0.022 | -0.055 | 0.142* |
| **r pos_ratio_PUNCT** | -0.100 | 0.112 | 0.186** | -0.263*** |
| **r pos_ratio_ADV** | -0.027 | 0.006 | 0.012 | -0.085 |
| **r pos_ratio_AUX** | -0.072 | 0.143* | 0.173** | -0.265*** |
| **r pos_ratio_PART** | 0.078 | -0.084 | -0.064 | 0.123 |
| **r pos_ratio_PRON** | 0.060 | 0.060 | 0.089 | -0.121 |
| **r pos_ratio_SCONJ** | 0.510*** | -0.071 | 0.027 | -0.085 |
| **r pos_ratio_X** | 0.159* | -0.071 | -0.077 | 0.062 |
| **r pos_ratio_SYM** | 0.116 | 0.079 | 0.131* | -0.098 |
| **r pos_ratio_INTJ** | 0.453*** | 0.026 | 0.125 | -0.253*** |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 5 - Dependency analysis of variables: selected user-oriented variables & [synt]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [synt]** | 0.282** | 0.246** | 0.184 | 0.242** |
| **Multiple R2 [synt]** | 0.079** | 0.060** | 0.034 | 0.058** |
| **Adjusted R2 [synt]** | 0.063** | 0.043** | 0.016 | 0.041** |
| **r synt_dependency_distance_mean** | -0.037 | -0.030 | -0.022 | 0.122 |
| **r synt_dependency_distance_std** | -0.035 | -0.002 | -0.016 | 0.059 |
| **r synt_prop_adjacent_dependency_ relation_mean** | 0.152* | 0.057 | 0.064 | 0.012 |
| **r synt_prop_adjacent_dependency_ relation_std** | -0.064 | -0.071 | -0.062 | -0.026 |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 6 - Dependency analysis of variables: selected user-oriented variables & [lex_rich]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [lex_rich]** | 0.221 | 0.558*** | 0.520*** | 0.573*** |
| **Multiple R2 [lex_rich]** | 0.049 | 0.311*** | 0.270*** | 0.329*** |
| **Adjusted R2 [lex_rich]** | 0.009 | 0.258*** | 0.221*** | 0.274*** |
| **r lex_rich_TTR** | 0.068 | -0.073 | -0.169* | 0.220** |
| **r lex_rich_RTTR** | -0.031 | 0.064 | 0.223** | -0.342*** |
| **r lex_rich_CTTR** | -0.031 | 0.064 | 0.223** | -0.342*** |
| **r lex_rich_MSTTR** | 0.007 | 0.030 | 0.064 | -0.089 |
| **r lex_rich_MATTR** | 0.027 | 0.034 | 0.066 | -0.089 |
| **r lex_rich_rhapax** | 0.077 | -0.120 | -0.186** | 0.231*** |
| **r lex_rich_lca_ttr** | 0.064 | -0.077 | -0.179** | 0.232*** |
| **r lex_rich_lca_msttr** | 0.074 | 0.038 | 0.033 | -0.094 |
| **r lex_rich_lca_cttr** | -0.035 | 0.120 | 0.222** | -0.297*** |
| **r lex_rich_lca_rttr** | -0.034 | 0.120 | 0.222** | -0.297*** |
| **r lex_rich_lca_logttr** | 0.040 | -0.065 | -0.135* | 0.167* |
| **r lex_rich_qui_ttr** | 0.071 | -0.084 | -0.186** | 0.233*** |
| **r lex_rich_qui_hapax** | -0.035 | 0.168* | 0.294*** | -0.361*** |
| **r lex_rich_qui_rhap** | 0.078 | -0.101 | -0.179** | 0.222** |
| **r lex_rich_qui_mattr100** | 0.039 | 0.023 | 0.041 | -0.108 |
| **r lex_rich_qui_mattr500** | 0.025 | 0.015 | 0.040 | -0.088 |
| **r lex_rich_qui_zttr** | 0.012 | 0.004 | 0.063 | -0.139* |
| **r lex_rich_qui_mamr100** | 0.005 | -0.183** | -0.059 | 0.078 |
| **r lex_rich_qui_mamr500** | -0.065 | -0.188** | 0.037 | -0.040 |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 7 - Dependency analysis of variables: selected user-oriented variables & [lex_sop]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [lex_sop]** | 0.094 | 0.251* | 0.242* | 0.357*** |
| **Multiple R2 [lex_sop]** | 0.009 | 0.063* | 0.059* | 0.127*** |
| **Adjusted R2 [lex_sop]** | 0.000 | 0.037* | 0.033* | 0.103*** |
| **r lex_sop_lca_ld** | 0.034 | 0.027 | 0.008 | -0.038 |
| **r lex_sop_lca_ls1** | -0.006 | 0.082 | -0.028 | 0.003 |
| **r lex_sop_lca_ls2** | -0.021 | -0.039 | -0.007 | -0.039 |
| **r lex_sop_lca_vs1** | 0.045 | -0.063 | -0.182** | 0.182** |
| **r lex_sop_lca_cvs1** | -0.042 | 0.041 | 0.176** | -0.322*** |
| **r lex_sop_lca_vs2** | -0.042 | 0.047 | 0.179** | -0.319*** |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 8 - Dependency analysis of variables: selected user-oriented variables & [lex_div]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [lex_div]** | 0.239* | 0.287** | 0.311*** | 0.452*** |
| **Multiple R2 [lex_div]** | 0.057* | 0.082** | 0.097*** | 0.205*** |
| **Adjusted R2 [lex_div]** | 0.031* | 0.057** | 0.072*** | 0.183*** |
| **r lex_div_MTLD** | 0.000 | -0.064 | 0.003 | 0.000 |
| **r lex_div_HD_D** | 0.009 | 0.188** | 0.229** | -0.339*** |
| **r lex_div_Herdan** | 0.049 | -0.091 | -0.164* | 0.191** |
| **r lex_div_Summer** | 0.032 | -0.074 | -0.107 | 0.109 |
| **r lex_div_Dugast** | 0.039 | 0.011 | 0.110 | -0.191** |
| **r lex_div_Maas** | 0.011 | -0.005 | -0.091 | 0.156* |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 9 - Dependency analysis of variables: selected user-oriented variables & [lex_var]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [lex_var]** | 0.231 | 0.467*** | 0.474*** | 0.486*** |
| **Multiple R2 [lex_var]** | 0.053 | 0.219*** | 0.225*** | 0.237*** |
| **Adjusted R2 [lex_var]** | 0.000 | 0.167*** | 0.173*** | 0.186*** |
| **r lex_var_lca_ndw** | -0.043 | 0.228** | 0.345*** | -0.397*** |
| **r lex_var_lca_ndwz** | 0.028 | -0.210** | -0.182** | 0.123 |
| **r lex_var_lca_ndwerz** | -0.053 | 0.087 | 0.141* | -0.208** |
| **r lex_var_lca_ndwesz** | 0.000 | -0.073 | -0.067 | 0.000 |
| **r lex_var_lca_uber** | -0.014 | 0.069 | 0.140* | -0.208** |
| **r lex_var_lca_lv** | 0.054 | -0.098 | -0.181** | 0.242*** |
| **r lex_var_lca_vv1** | 0.047 | -0.111 | -0.206** | 0.267*** |
| **r lex_var_lca_svv1** | -0.037 | -0.112 | 0.080 | -0.210** |
| **r lex_var_lca_cvv1** | -0.047 | -0.111 | 0.078 | -0.198** |
| **r lex_var_lca_vv2** | 0.080 | -0.145* | -0.192** | 0.240*** |
| **r lex_var_lca_nv** | 0.057 | -0.086 | -0.173** | 0.233*** |
| **r lex_var_lca_adjv** | 0.037 | -0.157* | -0.217** | 0.279*** |
| **r lex_var_lca_advv** | 0.001 | -0.181** | -0.154* | 0.209** |
| **r lex_var_lca_modv** | 0.069 | -0.154* | -0.189** | 0.240*** |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 10 - Dependency analysis of variables: selected user-oriented variables & [syl]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [syl]** | 0.098 | 0.389*** | 0.503*** | 0.524*** |
| **Multiple R2 [syl]** | 0.010 | 0.151*** | 0.253*** | 0.275*** |
| **Adjusted R2 [syl]** | 0.000 | 0.140*** | 0.243*** | 0.265*** |
| **r syl_SC** | -0.050 | 0.293*** | 0.421*** | -0.454*** |
| **r syl_LC** | -0.046 | 0.301*** | 0.428*** | -0.458*** |
| **r syl_LetC** | -0.049 | 0.297*** | 0.423*** | -0.455*** |
| **r syl_PolyC** | -0.055 | 0.263*** | 0.390*** | -0.429*** |
| **r syl_MonoC** | -0.042 | 0.307*** | 0.432*** | -0.459*** |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$

**Table 11 - Dependency analysis of variables: selected user-oriented variables & [other]**

| | Sum(DensityKW*UIH) | lengthRT_doc_mean | target | entropy_mean |
|---|---|---|---|---|
| **Multiple R [other]** | - | - | - | - |
| **Multiple R2 [other]** | - | - | - | - |
| **Adjusted R2 [other]** | - | - | - | - |
| **r other_Size_in_kB** | 0.033 | 0.259*** | 0.387*** | -0.320*** |
| **r other_RT** | -0.049 | 0.296*** | 0.422*** | -0.454*** |
| **r other_DCR** | -0.005 | 0.228** | 0.164* | -0.210** |
| **r other_qui_hpoint** | -0.041 | 0.193** | 0.322*** | -0.372*** |
| **r other_qui_ent** | -0.047 | 0.113 | 0.213** | -0.306*** |
| **r other_qui_vd** | -0.126 | 0.114 | -0.044 | 0.068 |
| **r other_qui_q** | 0.086 | -0.113 | -0.028 | 0.064 |
| **r other_qui_d** | -0.086 | 0.113 | 0.028 | -0.064 |
| **r other_qui_tc** | 0.117 | 0.221** | -0.002 | 0.000 |
| **r other_qui_stc** | 0.116 | 0.218** | -0.062 | 0.078 |
| **r other_eawl** | -0.030 | 0.033 | -0.007 | -0.070 |
| **r other_eawl_unique** | 0.197** | -0.102 | -0.053 | 0.082 |

Note: *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$