English Studies at NBU, 2024
Vol. 11, Issue 1, pp. 141-159
https://doi.org/10.33919/esnbu.25.1.8

pISSN 2367-5705
eISSN 2367-8704
www.esnbu.org

EXTRACTION OF ANGLICISMS FROM A CORPUS OF MACEDONIAN MAGAZINE TEXTS

Lina Miloshevska

University of Information Science and Technology
"St. Paul The Apostle", Ohrid, North Macedonia

Abstract

The present article is a description of the stages involved in compiling a specialized corpus of Macedonian magazine texts and the software tools employed to extract anglicisms from the corpus. The texts were collected from the magazine Kapital and cover two distinct periods: the years 2000 and 2020. The size of the corpus is about 2 million tokens and 141,852 types. The software employed produced word lists that later in combination with other statistical techniques produced a refined Anglicism headword list from which new anglicisms were extracted. In addition to the software tools, careful manual inspection was necessary in both the extraction and analysis stages. As a result of the research, a total of 220 completely new anglicisms have been identified. Most of these new anglicisms are not yet included in existing Macedonian dictionaries.

Keywords: Anglicisms, anglicisms extraction, corpus linguistics, corpus analysis tools

Article history:
Received: *03 December 2024*
Reviewed: *15 February 2025*
Accepted: *08 March 2025*
Published: *30 June 2025*

Citation: Miloshevska, L. (2025). Extraction of anglicisms from a corpus of Macedonian magazine texts. *English Studies at NBU, 11*(1), 141-159. https://doi.org/10.33919/esnbu.25.1.8

Lina Miloshevska is a senior lecturer at the University of Information Science and Technology "St. Paul The Apostle", North Macedonia. Her research interests are in the field of corpus linguistics, discourse analysis, language contact and language change, CALL, and ESP.

E-mail: lina.miloshevska@uist.edu.mk https://orcid.org/0000-0003-3856-5448

The present article describes a research project aimed at building a specialized corpus of Macedonian magazine texts to be used for anglicisms identification and analysis. Previous research on anglicisms has provided a methodological background for the identification and extraction of these English loan words.

To identify anglicisms, the language of the media, whether print or digital is regarded as a very convenient source since "it is representative of a wide range of registers and is highly receptive and open towards neologisms, loanwords and linguistic creativity in general" (Furiassi & Hofland, 2007, p. 347). Foreign words, anglicisms, and false anglicisms are often used for their positive connotation and their strategically communicative features, especially in eye-catching headlines (p. 347).

The media plays a significant role as a primary source for introducing anglicisms in the Macedonian language. Written texts are particularly important in the study of new loans as they provide these items more visibility in the influx of newly introduced borrowings and coinages. The fact that some new anglicisms are accompanied by their original in parenthesis, quotation marks or explained is more likely to fix them in both the passive and the active lexical repertoire of Macedonian readers. Additionally, the presence or absence of typographical resources (such as inverted commas or italics) can be interpreted as marks of novelty or foreign character. In fact, they reveal to what extent the writer considers the word should be highlighted as foreign or not. The examples from the corpus in **Appendix 1** illustrate this.

Applying corpus analysis tools to study Anglicisms in Macedonian media texts is a robust approach to understanding how English loanwords influence Macedonian language use, particularly in journalism and media discourse. Attempts at automatic and semi-automatic retrieval of anglicisms with varying degrees of success are discussed by Andersen (2005, 2011, 2012), Furiassi & Hofland (2007), Furiassi (2008), and Losnegaard & Lyse (2012).

Identification of Anglicisms

The starting point for any study of anglicisms is based on the definition of these loan words, i.e. what counts as an Anglicism, which is essential to determine in order to calculate the number and the impact (frequency) of English vocabulary on a language. Definitions vary significantly in the literature and are usually adapted to the researcher's interest. Definitions can be quite restrictive, focusing only on the most recent anglicisms (cf. Görlach 2001) or more accommodating (cf. Gottlieb 2004), including both new and older anglicisms that have long been accepted into the recipient language.

For the purpose of this paper, the definition of what constitutes an Anglicism focuses mainly on lexical items without constraints on their degree of acceptance. Moreover, no limitations are placed on whether a word is an Anglicism, Americanism, or Briticism. Consequently, the term Anglicism used in this paper covers any variant of English origin adapted or adopted (unadapted) and serves as a portmanteau term. In other words, anglicisms in this study are adapted lexical items and unadapted lexical items that clearly have an English origin (are attested in the source language) and bear English traits in their phonology, morphology, orthography, and semantics. Adapted anglicisms are words or compounds whose orthography and morphology are adapted to the recipient language system. Such items often become a productive source for new terms in the recipient language system; for example, *финишира, стратува, инвестира*, are clear loanwords adapted to the Macedonian phonological, morphological, and grammatical system. On the other hand, adopted/unadapted loans are words or compounds borrowed from English "wholesale" without much structural integration so that the expression remains recognizably English, such as *скрининг, кобрандинг, бот, бизнис*. The only intervention of the recipient language is in the phonology of the term, given the difference between English and the recipient language phonological systems. For practical reasons, the anglicisms discussed in this paper are one-word lexical units or single-unit compounds unhyphenated. Thus defined a list of anglicisms for further analysis was extracted from the corpus as explained in section 4

The KAPITAL corpus

To extract and study anglicisms in Macedonian, a corpus of magazine articles was compiled and analyzed. The corpus was created specifically for the purpose of this study and was compiled from scratch. The corpus size is 2,288,999 tokens. The corpus was extracted from two distinct time periods: the years 2000 and 2020. This proved to be crucial in identifying trends and changes in the usage and frequency of anglicisms as well as detecting new anglicisms. The corpus was extracted from a total number of 1511 articles that represent the yearly publication of Kapital in the year 2000 and 396 articles published in the year 2020. A total of 1907 articles were examined. Given the two examined time periods, i.e. the years 2000 and 2020, the corpus is divided into two sub-corpora: MK2000 (hereinafter MK2000) has 1562927 tokens, and 97255 types. Corpus MK2020 (hereinafter MK2020) has 726072 tokens and 44597 types. **Figure 1** shows the file number, token count, type count, as well as other data for MK2000. **Figure 2** shows the file number, token count, type count, and other data for MK2020.

Figure 1

File count, token count, and type count for MK2000

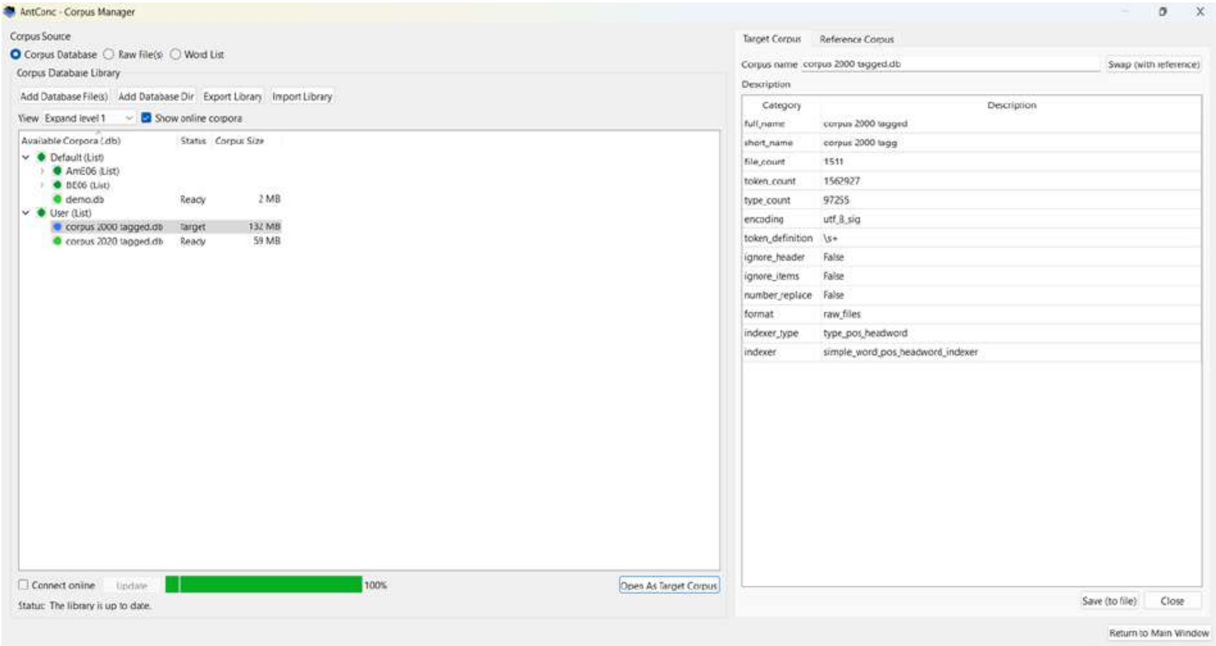
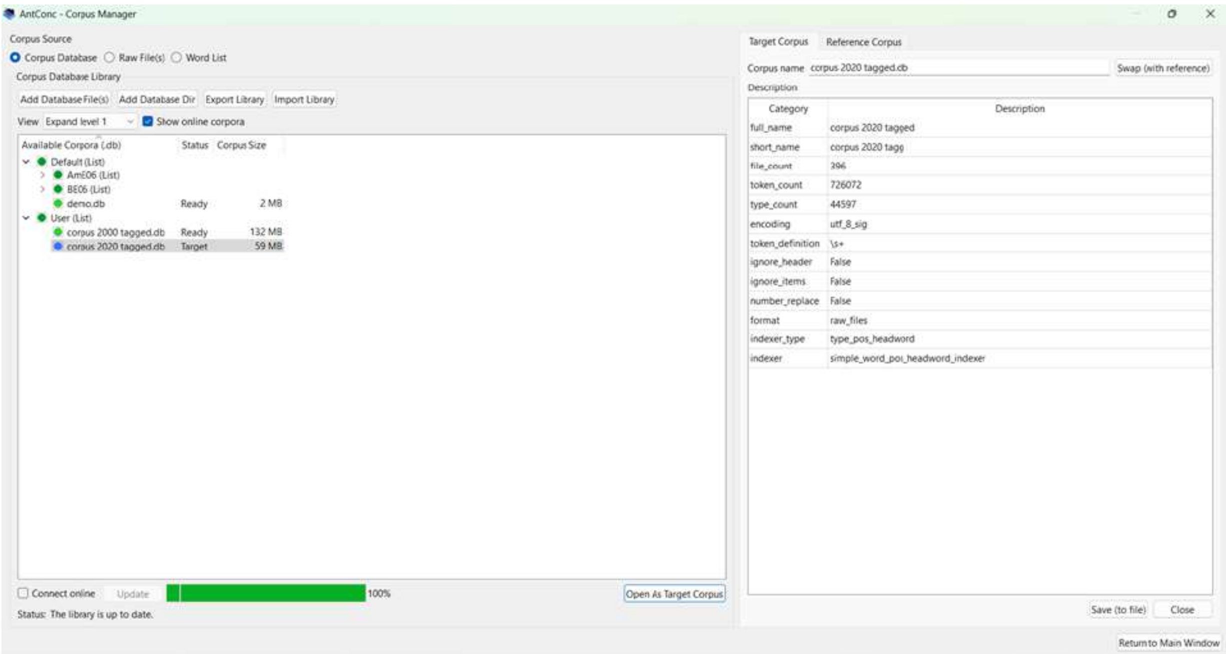


Figure 2

File count, token count, and type count for corpus MK2020



The KAPITAL magazine was selected for analysis because of the following reasons: Kapital covers topics in economics, business, politics, technology, and innovations. Many scholars (cf. Nogueroles 2017, Winter-Froemel & Onysko 2012, Khoutyz 2010) have identified such topics as major fields where anglicisms' presence in different languages is attested, due to borrowing from English and the role English plays in spreading new concepts and innovations.

The KAPITAL magazine is published in print and digital format. The texts are balanced in their methodological approach, relatively accurately proofread and edited, unlike most online publications with questionable linguistic correctness. All texts are archived in electronic format suitable for computer processing for the designated periods analyzed in the study.

Journalistic texts usually have a semi-formal style that should produce a corpus containing lexicalised anglicisms as well as more recent borrowings. Journalistic texts aim to appeal to audiences and attract more readership, so the topics chosen for publication will likely reflect current trends in the country at the time. Since English represents prestige and modernity in Macedonian, journalistic texts seem like a genre that would be receptive to anglicisms.

To ensure the representativeness of the corpus, all published articles in the years 2000 and 2020 were taken into account and included in the corpus. The texts that make up the corpus required several pre-processing steps to be made analyzable. This was a necessary procedure to create a computer-readable corpus. Texts from the year 2000 were obtained as archived (ZIP) files of entire folder structures covering publications from numbers 39-84 (46 issues) of KAPITAL. That is a total of 1511 files in separate Word files (.doc). Because these articles were written in the year 2000, the texts were in Macedonian Tms font, which is a non-UTF-8 format (non-regular font for text formation). This makes the processing of the data problematic because of global character incompatibility. Therefore, for further analysis, the raw files had to be pre-processed in two stages:

1. Conversion of texts to UTF-8 format: using a character type converter, the texts were converted to UTF-8 encoded format.
2. The UTF-8 converted files were saved in txt file format to be compatible with the software-supported text formats that will be used in the process of corpus analysis

Through this procedure, the MK2000 sub-corpus was obtained in a .txt text file format. However, further manual inspection of the content of the files revealed duplicate files, badly converted files due to file corruption, and files containing text in languages other than Macedonian (mainly English web links). These issues were dealt with, and a clean corpus of texts was obtained.

Similarly, the texts from the year 2020 were obtained in a similar format and structure from folders covering the issues from 1048 to 1088 of Kapital in 2020. That is a total of 41 issues that make a total of 396 articles. The texts were in Word (.doc) format and UTF-8 standard. The pre-processing of the texts to obtain the MK2020 sub-corpus was conducted by conversion of the text format from .doc to .txt files. In this manner, both the MK2000 sub-corpus and MK2020 sub-corpus had identical file formats ready to be analyzed using corpus software tools.

Methodology for extraction of anglicisms

As stated in the introduction, attempts for automatic and semi-automatic extraction of English loans from corpora have been undertaken by scholars examining these linguistic items while trying to take advantage of software analysis tools. The main goal is to develop data processing tools that would automatise the process of English loans identification and extraction. Attempts at the automatic retrieval of lexical false anglicisms in Italian are discussed in (Furiassi & Hofland 2007, Furiassi 2008) and (Andersen 2005, 2011, 2012) for anglicisms in Norwegian. The proposed data processing tools for the automatic extraction of English loans rely on the differences in orthography between the languages in contact and use grapheme typicality algorithms, as well as dictionary-based methods and word-formation regularity (Wohlfeld & Witalisz, 2019).

Despite the efficacy of the methods employed in reducing "noise" and minimizing human labour, they still do not seem reliable in either the identification or analysis phases. Andersen maintains that the proposed tool "does not offer the full answer as to which forms to include and which forms to leave out, but it promises a systematic and empirically based proposal of where to start looking. This will hopefully lead to a significant reduction of manual work and a radical simplification of the task of looking for the needle in the linguistic hay-stack". (Andersen 2011). Given the speed in technological advancements, such tools might be available in the future; however, at the moment, human expertise must be combined with computational procedures for the findings to be accurate.

The identification of loans which is the starting point in any analysis of Anglicisms in a recipient language "still remains the researcher's responsibility and depends on their knowledge-based human skills that have not, as yet, been successfully copied and replaced with artificial intelligence" (Wohlfeld & Witalisz 2019: 172). Therefore, for the extraction of anglicisms for analysis in this paper, a combination of corpus analysis tools and careful manual inspection is applied at each stage of the process

Because the corpus was composed for the purpose of this study and not tagged in any way, the only available option for extracting a list of anglicisms was to go through the texts and identify the loans manually. Given the size of the corpus, this method was too time-consuming and laborious. However, it seemed necessary as computerized automatic identification of loans in texts is unreliable. Another possible method was to go through a representative sample of the corpus and extract a list of Anglicisms and to use the resulting list to analyze the frequency and use of these terms in the rest of the corpus. However, this method will not capture all anglicisms in the corpus, especially those anglicisms that have very few occurrences, but still their usage might provide valuable insight on the choices of the particular Anglicism in the particular context.

In order to ensure comprehensiveness and inclusiveness, i.e. to capture all anglicisms in the corpus, the following steps were taken to extract a list of anglicisms from the corpus:

1. Using the software AntConc, a word list was made for all the words occurring in the corpus. A word list is made for MK2000 and MK2020 separately. A word list counts how many times each word occurs in the corpus and lists all the words in the corpus.
2. The 'Word list' function in AntConc is accessed by using the tab word list. Using this option, AntConc provides the following information:

- The total number of words in the corpus (word tokens)
- The total number of unique words in the corpus, which is the vocabulary size of the corpus (Word Types).
- A ranking of every unique word type by its frequency in the corpus.

The obtained two lists were merged, filtered for duplicates, cleaned, and refined to obtain unique words only. From this list, a list of anglicisms was extracted manually by looking at each word individually and including all lexical anglicisms and excluding all other words of Macedonian origin and different foreign origin. Also, proper nouns, abbreviations and acronyms, pseudo / false anglicisms (e.g. *маунтбајкинг*), direct translation from English (e.g. *бежичен* “wireless”, *безхартиена* “non-paper”, *еднорози* “unicorn”, *споделен* “share”, *широкопојасен* “broadband”, *сапуница* “soap opera”) were also excluded from the final list as these categories are not the focus of the current study. This procedure originated a refined Anglicisms word list, which is used for further analysis.

The final Anglicisms list contains 4436 anglicisms. This list will later serve to count frequencies and to extract new anglicisms. This list includes just lexical items (one-word items) and compounds (compound words unhyphenated, written as one word). This choice is made for practical reasons as the used software AntConc turned to provide inaccurate results when the word list contained hyphenated compounds, or two-word compounds separated by a space when counting frequencies. From the part of speech point of view, the anglicisms list includes nouns, verbs, adjectives, verbal nouns, and adverbs in their base form uninflected. In other words, all definite markers, feminine, masculine, neuter, and plural markers, and markers of tense are erased.

Once a list of anglicisms had been obtained as described above, AntConc software was used to determine the exact number of times each term occurred in each sub-corpus. However, before counting frequencies[1] could be performed, another procedure had to be undertaken to ensure the results' maximum accuracy.

As mentioned earlier, the raw files that constitute both corpora MK2000 and MK2020 were obtained in their original format and not tagged in any way. Tagging the corpus was necessary to ensure that different tokens of a certain type of anglicism were counted. For example, the anglicism *блог* (blog), can appear in its base form but can also appear with the definite marker *блогот*, with the plural marker *блогови,* and with the

[1] Frequencies of anglicisms are not going to be discussed in detail in this paper as the main focus here is on the procedural stages involved in the identification and extraction process of anglicisms. Analysis of frequencies is discussed in detail in a separate paper. Frequency is mentioned here briefly when deemed necessary.

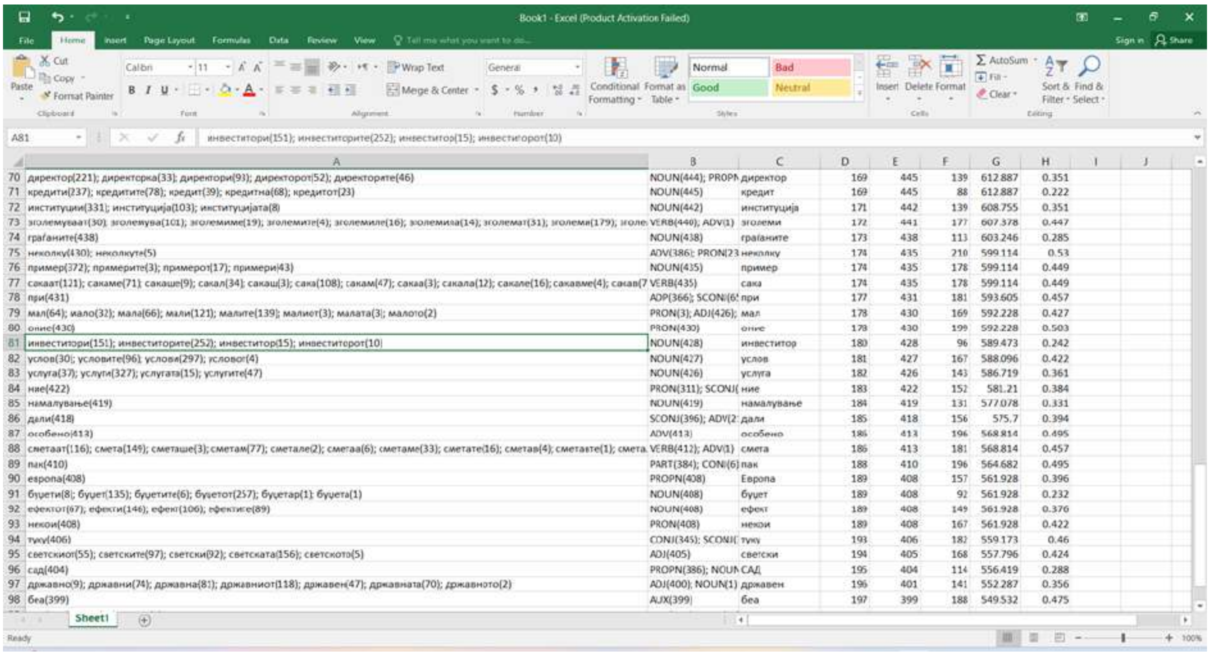
plural and definite marker *блоговите*. Without tagging the corpus, the software will not recognize that *блог, блогот, блогови, блоговите* are tokens of the same type.

To tag the corpus, the files were fed into the multi-language segmenter and Part-Of-Speech (POS) tagger software TagAnt (Anthony, 2024b), where a trained pipeline model for the Macedonian language was used to tag the words in the texts. The tagging later helped in the process of grouping the anglicisms according to their part of speech category during the phase of statistical analysis. The model "mk_core_news_lg" was used since it is the largest and most comprehensive model available. This model is part of the spaCy open-source library for natural language processing (Honnibal & Montani, 2025). It is important to note that due to the imperfection of the trained language model, some of the words appeared as duplicates and were tagged as different types of words (incorrectly tagged). The erroneous tagging was manually corrected. Moreover, due to the imperfection of the language model, the lemmatization of the words in many cases was incorrect. The errors were manually corrected by correcting the word form and the corresponding frequency counts.

The cleaned tagged files are later analysed by the software AntConc (Anthony, 2024a). The rationale behind choosing AntConc for corpus analysis in this study is based on 1. AntConc is a corpus analysis software available for free on the internet, unlike other software that require a subscription fee 2. Hence, no funds were acquired for this study; choosing a freeware was the only available choice 3. AntConc is regularly updated, and new releases are made freely available on the internet, which improves the accuracy and the performance of the software. The process of corpus analysis started with importing the tagged files into the AntConc software. AntConc provides various features to analyse a digital corpus. First, the word counter tab was used to conduct a frequency analysis of the corpus. The output of this step was a word list with all the words occurring in the corpus with their respective frequencies but now tagged and grouped according to their part of speech category. Figure 3 is a screenshot of the results obtained by AntConc correctly counting headwords. For example, the anglicisms инвеститор (investor) with all its occurrences инвеститори(151); инвеститорите(252); инвеститор(15); инвеститорот(10) is correctly counted and placed under the POS Noun with 428 occurrences in the MK2020 corpus.

Figure 3

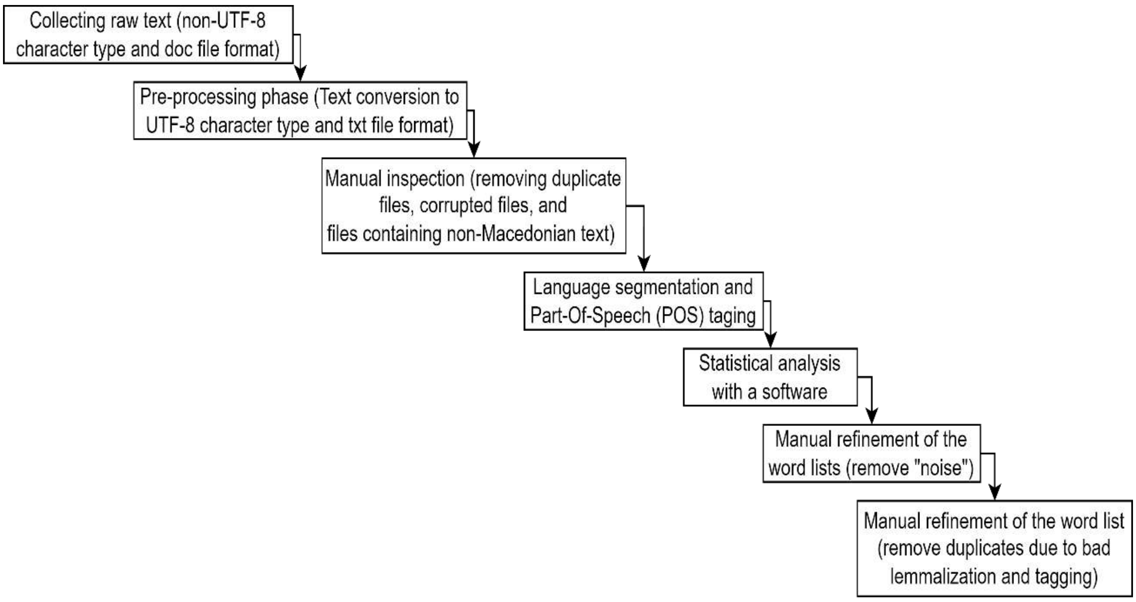
Headword list and frequency by AntConc



The next step in the analysis is performed using the selected list of anglicisms and the list of headwords obtained as displayed in Figure 3. MATLAB software was used to extract the frequencies of anglicisms grouped according to the respective headwords. In this manner, all tokens of a certain type are correctly counted. Figure 4 displays the procedural stages performed to obtain an anglicism list and, subsequently anglicisms frequencies.

Figure 4

Procedural stages applied to the corpus data from pre-processing to obtaining an accurate headword list



New anglicisms attested in MK2020

To extract new anglicisms that appeared only in the MK2020 corpus and were not attested in the MK2000 corpus we crossed and examined the frequency lists of both corpora, removed duplicates, and extracted those items that appeared only in the MK2020 corpus. The items were also checked in the corpus to verify context of occurrence. The results show that 582 anglicism lemmas are only attested in the MK2020 corpus. However, we cannot jump to the conclusion that all these 582 items are completely new words that entered the language after the year 2000. To verify which anglicisms can be labelled as new we manually inspected the data and found that many items are word forms not used in the MK2000 corpus, and not new words. For example, the anglicisms n. *девелопер* is attested in the MK2000 corpus but not the adj. *девелоперски*. Unless many written resources are inspected, there is no way to tell for sure which word-form was first borrowed into Macedonian. However, many scholars agree that usually nouns are borrowed first into the recipient language then other word forms develop using the language's word-formation mechanisms. Based on the corpus data, we can list a few examples of word forms not attested in the MK2000 corpus (**Appendix 2**). The emergence of these new word-forms in the MK2020 corpus shows the level of lexicalisation of these anglicisms. Over the 20 years' span (the time distance between the MK2000 and MK2020 corpora) these anglicisms got adapted to the Macedonian language system and through different word-formation mechanisms produced new word forms.

Out of the 582 anglicism lemmas attested in MK2020, just 220 are completely new anglicisms not attested in the MK2000. This is 0.03% of the total tokens of the MK2020. However, this percentage of new anglicisms, although true for the MK2020 cannot be generalized to the whole language. Although the number of new anglicisms seems small yet the total frequency of anglicisms in the MK2020 is higher compared to the MK2000 as corpus data show.

We assume that the truly new anglicisms in the MK2020 corpus will be related mainly to two domains 1) COVID-related terminology and 2) digital/ technological terminology. The year 2020 witnessed the global outbreak of the Covid-19 pandemic. The words used to talk about the COVID-19 pandemic are captured in the MK2020 corpus as a timestamp of the events of the time. The COVID-19 pandemic gave rise to new terminology on the one hand but also contributed to the high frequency of older terms that were infrequent in the year 2000 on the other. Examples of new COVID-related anglicisms attested only in the MK2020 corpus are given in **Appendix 3**. These Covid-related terms are specific in that they reflect the Covid crisis, and their usage has dropped significantly after the end of the pandemic.

The other domains where new anglicisms in the MK2020 corpus are attested are related to new technology, new concepts, and less frequently, political/ historical events. Unlike the COVID-19-related terms, these items will probably linger longer in the language and produce new derivates as long as the technology and the concepts they refer to are useful and needed by the speakers. New technology related anglicisms are plenty (**Appendix 4**). None of these words are attested in the MK2000 corpus so we might conclude that these are new anglicisms that recently entered the language.

New anglicisms are also attested in the MK2020 corpus that are not necessarily technology-related but refer to concepts, objects, and ideas that are recent in a wide range of domains, such as the economy, lifestyle, business transactions, sports, recreational activities, and the like (**Appendix 5**).

It is worth noting that some new anglicisms show variation in their spelling. For example, the anglicisms *блокчеин* (blockchain) appears one time as such in MK2020, and as *блокчејн* appears 4 times. The variation in transcription of blockchain is probably due to the newness of this word. The available dictionaries do not provide much help on the matter as the word is not yet indexed in Macedonian dictionaries. However, a quick Google search renders 116.000 hits with the variant *блокчејн* whereas the variant *блокчеин* renders 61.800 hits. It remains to be seen which of these two variants will gain dominance in the language. However, the transcription *блокчејн* imitates more closely the English pronunciation. If language users decide to imitate the English pronunciation more closely, we assume that this variant will be more prevalent.

The Orthography of the Macedonian Language, page 183 prescribes the "correct" transcription for English phonemes and diphthongs. According to this guidebook, the grapheme ai corresponding to the sound ei in English should be transcribed as ej or as e in Macedonian as in the examples Daisy (English) > *Дејз* (Macedonian), Twain (English) > *Твен* (Macedonian). The Orthography of the Macedonian Language states that when transferring English sounds to Macedonian, the starting point should be their pronunciation, not their graphic presentation. In other words, the transcription should be recorded in a manner corresponding to how Macedonian speakers perceive the specific English sound.

Other cases of spelling variation in new anglicisms are noticed in *бреинсторминг* n. (brainstorming) while the verb version is rendered as *брејнстормира* (to brainstorm) with a change in the vowel in the verb. Both noun and verb are not indexed in Macedonian dictionaries. This is a similar case to *блокчејн/блокчеин*, where a hesitation in the vowel choice in spelling is apparent when it comes to new anglicisms. Similarly, the case of the anglicism *меинстрим/мејнстрим* (mainstream).

Inspecting the corpus, it appears that not only new Anglicism show variation in their transcription but certain "old" anglicisms also do. In fact, some borrowings never settle on a definite visual representation and competing variants persist in the recipient language. The cases of *email* and *online* are interesting because these items have been in usage for a long time but still show spelling variations. The Anglicism *email* appears as имејл, емаил, мејл, and и-мејл. The transcription *емаил* is attested only in the MK2000 corpus (3 occurrences) while the transcription *имејл* is attested only in the MK2020 corpus (1 occurrence). *И-мејл* (2 occurrences MK2000, no occurrence MK2020), *мејл* (1 occurrence MK2000, 5 occurrences MK2020). This leads to the conclusion that email over the years has settled on two variants: either the transcription *имејл* which imitates more closely the English pronunciation or the shortened form *мејл* which for Macedonian users refers exclusively to electronic mail. When this word *имејл* is looked up in the Macedonian Digital Dictionary, the dictionary returns yet another variant *е-пошта* which is more formal. As for *online* in the MK2000 corpus , it appears as *онлајн* just once while it appears as *онлине* 7 times. Compared to the MK2020 corpus *онлине* has no occurrences at all while *онлајн* occurred 296 times.

In terms of frequency, the top ten most frequent new anglicisms among the 220 attested in MK2020 are presented in Appendix 6. These items demonstrate a very high frequency at the time of the global pandemic in both the language of the media and everyday language.

Conclusion

The extraction of new anglicisms from a corpus of Macedonian magazine articles cannot entirely rely upon automatic processing. Time-consuming and careful manual scanning must be combined with computational procedures at each stage of the analysis As to the Kapital corpus, the approach enabled the extraction of examples of anglicisms not attested in the year 2000. The software was successful in identifying and counting a refined headword list of one-word items or compounds written as single words. Although some automatic filters were added in order to eliminate the undesired 'noise' in the final word list, only further time-consuming manual scanning of such a list led to the tracing of new anglicisms. A total of 220 completely new anglicisms were identified. Among these, the most frequent are Covid-related items, while other terms have low frequency. Most of these new anglicisms are not yet included in existing Macedonian dictionaries, and although some of them are highly frequent in everyday usage, their frequency in the corpus is low. This is due to the topics tackled by the corpus and do not reflect their overall frequency in the language. To have more accurate measurements of the overall frequency of a term in the language, huge resources need to be compiled and analyzed.

Undoubtedly, computational tools were extremely useful in saving time in compiling and digitalizing the corpus, in retrieving specific items, checking the context of usage, and in collecting a preliminary list of anglicisms in Macedonian. Despite the advantages, the software analysis tools used still lack accuracy and cannot replace the insight of a linguist when handling a complex and multifaceted phenomenon such as anglicisms.

References

Anthony, L. (2024a). AntConc (Version 4.3.1) [Computer Software]. Waseda University https://www.laurenceanthony.net/software/AntConcAnthony

Anthony, L. (2024b). TagAnt (Version 2.1.1) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/TagAnt

Andersen, G. (2005). Assessing algorithms for automatic extraction of anglicisms in Norwegian texts. *Corpus Linguistics 2005*.

Andersen, G. (2011). Corpora as lexicographical basis: the case of anglicisms in Norwegian. *Methodological and Historical Dimensions of Corpus Linguistics* (Studies in Variation, Contacts and Change in English 6), ed. by P. Rayson, S.

Hoffmann & G. Leech. VARIENG. https://varieng.helsinki.fi/series/volumes/06/andersen

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. *The anglicization of European lexis*, *10*, 111-130. https://doi.org/10.1075/z.174.09and

Andersen, G. (2021). On a daily basis… a comparative study of phraseological borrowing. In R. Marti Solano & P. Ruano San Segundo (Eds.), *Anglicisms and Corpus Linguistics: Corpus-Aided Research into the Influence of English on European Languages* (pp. 13-30). Peter Lang. https://www.peterlang.com/document/1184575

Furiassi, C. G. (2008). Non-adapted Anglicisms in Italian: Attitudes, frequency counts, and lexicographic implications. In R. Fischer, & H. Pulaczewska (Eds.), *Anglicisms in Europe. Linguistic Diversity in a Global Context* (pp. 313-327). Cambridge Scholars Publishing. https://hdl.handle.net/2318/100769

Furiassi, C., & Hofland, K. (2007). The retrieval of false anglicisms in newspaper texts. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 347-363). Brill. https://doi.org/10.1163/9789401204347_020

Görlach, M. (Ed.). (2001). *A dictionary of European anglicisms: A usage dictionary of anglicisms in sixteen European languages*. Oxford University Press.

Gottlieb, H. (2004). Danish echoes of English. *Nordic Journal of English Studies, 3*(2), 39-65. https://doi.org/10.35360/njes.161

Khoutyz, I. (2010). The pragmatics of anglicisms in modern Russian discourse. In R. Facchinetti, D. Crystal, & Barbara Seidlhofer (Eds.), *From international to local English – and back again* (pp. 197-208). Peter Lang.

Losnegaard, G. S., & Lyse, G. I. (2012). A data-driven approach to anglicism identification in Norwegian. In G. Andersen (ed.), *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, Studies in *Corpus Linguistics vol. 49* (pp. 131-154). John Benjamins. https://doi.org/10.1075/scl.49.07los

Noguerolez, E. E. N. (2017). The Use of Anglicisms in Various Thematic Fields: An Analysis Based on the Corpus de Referencia del Español Actual. *ANGLICA-An International Journal of English Studies*, *26*(2), 123-149. https://doi.org/10.7311/0860-5734.26.2.08

Honnibal, M., & Montani, I. (2025). spaCy. https://spacy.io/models/mk

Winter-Froemel, E., & Onysko, A. (2012). Proposing a pragmatic distinction for lexical Anglicisms. In C. Furiassi, V. Pulcini, & F. R. González (Eds.), *The anglicization of European lexis* (pp. 43-64.). John Benjamins. https://doi.org/10.1075/z.174.06win

Mańczak-Wohlfeld, E., & Witalisz, A. (2019). Anglicisms in the National Corpus of Polish: Assets and limitations of corpus tools. *Studies in Polish Linguistics*, *14*(4), 171-190. https://doi.org/10.4467/23005920SPL.19.019.11337

Reviewers:
1. Anonymous
2. Anonymous

Handling Editor:
Stan Bogdanov, PhD
New Bulgarian University

APPENDICES

Appendix 1

- **Болдеринг** (анг. bouldering), што претставува искачување по помали природни или вештачки карпи, само со помош на прстите на качувачот и без употреба на јажиња или дополнителна опрема – уште еден голем потенцијал за развој на адреналински туризам.
- Започнуваме со утрински напиток кој секогаш има доза на **бреинсторминг** или бура од идеи што се би можеле да направиме во текот на денот.
- Протестантизам – генерално, потекнува од Макс Вебер, според кого: а) протестантската етика е значајна за економскиот успех, и б) овоземниот успех може да се интерпретира како знак на „ спасение “ и затоа толку силно се поставува како цел. Мажите се секогаш облечени во црно како знак на скромност. Зборот **воркохолик** (workaholic) потекнува од оваа култура.
- Аналитичарите на Голдман Сакс очекуваат **бенчмаркот** (цена на одредено подрачје) на Лондонската берза за следната година просечно да изнесува 22 долари за барел.

Appendix 2

Examples of word forms not attested in the MK2000 corpus
Агрегатен adj. (aggregate), адаптабилност n. (adaptability), активизам n. (activism), актуелност n. (modern/ current), дигитализација n. (digitalization), дигитализиран adj. (digital), дигитализирање n. (digitalization), интерактивност n. (interactivity), принта v. (print), принтање n. (printing), ресетирање n. (reset), роботизација n. (robotization), роботика n. (robotics), консумира v. (consume).

Appendix 3

Examples of new COVID-related anglicisms attested only in the MK2020 corpus
Асимптоматски (asymptomatic), *вирулентност* (virulence), *карантин* (quarantine), *кормορбидитет* (comorbidity), *корона* (corona), *коронавирус* (coronavirus), *коронакриза* (corona-crisis), *пандемиски* (pandemic adj.), *пандемичен* (pandemic adv.), *пик* (peak), *локдаун* (lockdown), *аеросол* (aerosol), *дезинфицира* (disinfect), *инкубаторски* (incubation-wise), *инкубација* (incubation), *инкубира* (incubate), *интубација* (intubation), *инфицира* (infect), *инфицирање* (infecting), *кластер* (cluster), *ковид* (covid), *новодијагностициран* (newly diagnosed), *посткризен* (post-crisis), *постпандемичен* (post-pandemic), *постпандемиски* (post-pandemic), *ревакцинација* (revaccination), *респиратор* (respirator), *самоизолација* (self-isolation), *самоизолира* (self-isolate), *серопеваленца* (seroprevalence), *социјализација* (socialization), *хоспитализација* (hospitalization), and the phrase *социјално дистанцирање* (social distancing).

Appendix 4

New technology related anglicisms not attested in the MK2000 corpus
Акаунт (account), алгоритам (algorithm), атачмент (attachment), баркод (barcode), бекап (backup), блог (blog), блогер (blogger), блокчејн (blockchain), бот (bot), вебинар (webinar), видеоконферндциски (video conference), видеотелеконференцциски (video teleconference), визибилност (visibility), вирален (viral), гугла (to google something/find on Google), десктоп (desktop computer), дрон (drone), емотикон (emoticon), инфлуенсер (influencer), инфографик (infographic), клауд (the cloud), кликбајт (clickbait), клип (clip), копирајтер (copywriter), логира (log in), мултитаскер (multitasker), мултитаскинг (multitasking), нанодимензија (nano-dimension), наноматеријал (nano-material), нотификација (notification), персонализација (personalization), персонализиран (personalized), плејстејшн (PlayStation), постира (post), постиран (posted), постирајќи (posting), принтскрин (print screen), реконфигурација (reconfiguration), реконфигурирање (reconfiguration), репрограмирање (reprograming), сајбербезбедност (cybersecurity), симулатор (simulator), смарт (smart), смартфон (smartphone), стартап (startup), стартапџија (someone who works/owns a startup company), стрим (stream), стриминг (streaming), таблет (tablet), твит (tweet), твита (to tweet), телеконференциски (teleconference), токен (token), торент (torrent), финтек (fintech), фотоволтаик (photovoltaic), фотоволтаичен (photovoltaic), хакер (hacker), хаштаг (hashtag), хуманоид (humanoid).

Appendix 5

New anglicisms are also attested in the MK2020 corpus
Аутгоинг (outgoing), аутсорс (outsource), аутсорсиг (outsourcing), аутсорсира (to outsource), апликабилен (applicable), болдер (boulder), болдеринг (bouldering), бајк (bike), бајкинг (biking), банџи (bungee), бекпротектор (back protector), билдинг as in тим билдинг (team-building), биопринтање (bioprinting), ботокс (botox), бреинсторминг (brainstorming), бејнстормира (to brainstorm), ветинг (vetting), вординг (wording), воркохолик (workaholic), декарбонизација (decarbonization), декарбонизиран (decarbonized), драфт (draft), енотуризам (enotourism), импакт (impact), импичмент (impeachment), индивидуализација (individualization), кетеринг (catering), исхендла (get something done/ deal with something), консумеризам (consumerism), контејнеризација (containerization), контролинг (controlling), корпоратизација (corporatization), кул (cool), лајфстајл (lifestyle), лаунџ (lounge), листинг (listing), мајндсет (mindset), манга (manga) , маунтбајкинг (mountain biking), маунтин (mountain), мејнстрим (mainstream), миленијалци (millennials), мол (mall), невромаркетинг (neuromarketing), офлејбл (off-label), рафтинг (rafting), ребрендирање (rebranding), револвинг (revolving) , рентакар (rent-a-car), рецептор (receptor), ресорт (resort), саундтрак (soundtrack), секвенционирање (sequencing), скрининг (screening), спин (spin),

спинување (spinning) , стејкхолдер (stakeholder), туроператор (tour operator), туторијал (tutorial), факторинг (factoring), фактчекинг (fact-checking), фит (fit), фриленсер (freelancer), фрирајд (freeride), фотофрејм (photoframe), хаб (hub), хедхантинг (headhunting), хеџ (hedge), џампинг (jumping), шинтоизам (Shintoism). Anglicisms in the domain of politics and/or historical events that are attested in the MK2020 corpus only are бенчмаркира (benchmark), бенчмаркирање (benchmarking), брегзит (Brexit), брегзитер (Brexiteer), моментум (momentum), фрегзит (Frexit). Many of the stated anglicisms above have not been included in Macedonian dictionaries yet. Examples of such anglicisms are бекграунд, адвертајзинг, алоцира, атачмент, бартер, бекап, билда, тим билдинг, блокчејн, болдер, болдеринг, бос, бот, бреинсторминг, ветинг, вебинар, визитинг, вирален, волатилност, вординг, воркохолик, гејзер, гугла, датабаза, девелопер, декодер, депрецијација, драјвер, интермедијарен, интерсекција, интубација, инфлуенсер, контрибуира, контролинг, копирајтер, корморбидитет, корпоратизација, лаунџ, листинг, локализација, локдаун, мајндсет, мејнстрим, микроменаџира, миленијалец, моментум, монетизира, монетизација, мултитаскер, нанодимензија, наноматеријал, невромаркетинг, нотификација, нутритивен, офлејбл, панелист, параглајдинг, парамедицински, перзистентен, принта, прудентен, рафтинг, реверзибилитет, револвинг, резистентност, репетитивен, рестарт, сајбербезбедност, саундтрак, серопреваленца, смарт, скрининг, сноубординг, спин, спинувања, стартап, стејкхолдер, стендбај, стрим, стриминг, таблет, твит, тејлоризам, телемедицина, токен, торент, тренди, туторијал, факторинг, фејк, финтек, фискализација, фриленсер, хаб, хаштаг, хедхантинг, хеџ, перспирација, фронтмен.

Appendix 6

Ten most frequent new anglicisms attested in MK2020
Корона (corona) – raw frequency of 274, кластер (cluster) – 94, стартап (start-up) – 44, ковид (COVID) – 42, коронавирус (Corona virus) – 41, локдаун (lockdown) – 34, стриминг (streaming) – 13, импакт (impact) – 9, скрининг (screening) – 8, аутсорсинг (outsourcing) – 6.