https://doi.org/10.31820/f.35.2.1

Irena Bogunović

A CORPUS-BASED APPROACH TO ENGLISH LOANWORDS: INTRODUCING THE DATABASE OF ENGLISH LOANWORDS IN CROATIAN

Irena Bogunović, University of Rijeka, Faculty of Maritime Studies
irena.bogunovic@pfri.uniri.hr orcid.org/0000-0002-2956-7014

izvorni znanstveni članak
UDC 811.163.42'373.45:811.111
rukopis primljen: 20. ožujka 2023; prihvaćen za tisak: 2. listopada 2023.

Unadapted English loanwords have become part of informal communication in many languages, including Croatian. Their use is often motivated by the lack of adequate native equivalents, exposure to English through the media, but also by the prestigious status of the English language. A vast body of research has been dedicated to lexical borrowing, especially from English. At the same time, corpus analyses have mostly been conducted on smaller, ad hoc corpora. Therefore, the goal of this paper is to present the database of English loanwords in Croatian. The database was developed by algorithmic and manual classification of words from the Corpus of Croatian news portals, ENGRI, and provides a list of 9,452 unadapted English loanwords together with the data on their absolute and relative frequencies. The analysis showed that most loanwords (75.85%) appear less than 50 times, while a total of 44.78% of words appear 10 times or less. The biggest drop in the number of loanwords is observed in the categories of occurrence above 500, while only 27 words appear 5,000 times or more. The most frequent English loanword in the corpus is 'show' with 80,805 occurrences, which is 0.0122% of all words in the corpus. The analysis of loanwords that occur more than 5,000 times showed that most of them have Croatian translation equivalents, which confirms the role of the media in the introduction of new words.

In addition to providing an insight into the occurrence of English loanwords in Croatian, this database also represents a valuable contribution to Croatian computational linguistics resources and enables future experimental research by providing the data on word frequency.

Keywords: *English loanwords; Croatian; lexical borrowing; database; corpus*

1. Introduction

Globalization and the development of new forms of communication have enabled constant information exchange. As a result, new concepts are constantly being introduced. Many languages do not have adequate words for these concepts, so speakers sometimes borrow foreign words. As the global language, English has become the dominant donor language for many languages, including Croatian (Drljača Margić 2011). The influence of English has been observed in many languages worldwide (e.g., Greenall 2015; Kay 1995; Pulcini et al. 2012) and in different functional styles (e.g., Alvarez-Mellado 2020; Čepon 2017; Mihaljević 2003) and domains (e.g., Matić 2017; Mykytka 2017), especially the media (e.g., Alvarez-Mellado 2020; Brdar 2010; Núñez Nogueroles 2016). The media has been recognized as an important factor in the shaping of a language and introducing new words (e.g., Drljača Margić 2009; Muhvić-Dimanovski and Skelin Horvat 2008). This is especially true for digital media, given that it has almost completely replaced print media (e.g., Twenge et al. 2019).

The media also plays an important role in language exposure (e.g., Drljača Margić 2011). English has become the dominant language across different aspects of our life such as business, education, and entertainment (e.g., Brannen et al. 2014; Drljača Margić 2011; Gluszek and Hansen 2013). Research has shown that Croatian speakers are daily exposed to English through various informal activities, such as surfing the internet or gaming (e.g., Bogunović and Jelčić Čolakovac 2019). Moreover, it is perceived as prestigious (e.g., Crystal 2003; Rüdiger 2018), and prestige has been recognized as one of the most important reasons for borrowing (e.g., Field 2002). Due to the prestigious status of English, knowledge of English is associated with a better social status and a better life (e.g., McKenzie 2010). Research has also shown that Croatian students generally have positive attitudes toward English loanwords (e.g., Drljača Margić 2012, 2014) and

that the use of unadapted English loanwords positively correlates with social attractiveness (Ćoso and Bogunović 2017).

Because of its prestigious status, the tendency of words borrowed from English to adapt to the recipient language is reduced (e.g., McKenzie 2010). As a result, many English loanwords are used in their unadapted form. Typically, borrowed words are categorized depending on their inclusion into the recipient language (e.g., Kay 1995; Međeral 2016) or the degree of their adaptation to the language's orthographic, phonological, and morphological rules (e.g., Görlach 2002; Entlová and Mala 2020). For example, Görlach (2002) describes three categories of borrowed words: fully adapted words, words of limited use, and new words (calques or loanwords). Međeral (2016) divides borrowed words into five groups. The first group includes foreign words, i.e., words that retain original orthographic features (e.g., 'snowboard', n., sg.) and, in the process of morphological adaptation, take native affixes (e.g., *snowboardi*, *n., pl*.). The second group includes orthographically adapted words with atypical phonological features (e.g., *čips,* 'chips'). The fourth group consists of words that have completely adapted to the recipient language (e.g., *tenis*, 'tennis'). Finally, the fifth group are adapted words that are no longer perceived as foreign (e.g., *klub,* 'club') and calques (e.g., *neboder*, 'skyscraper'). Sometimes the words borrowed from English are not English in origin (e.g., 'television'), but can still be considered anglicisms (Filipović 1990) because they were borrowed from English.

Borrowed words have long been a subject of discussion among Croatian linguists, who generally recommend the use of native words (e.g., Halonja and Hudeček 2014; Hudeček and Mihaljević 2005; Institute of Croatian language and linguistics 2015). Croatian solutions for English loanwords usually include multi-word expressions and descriptions, existing words with new meanings, new words and calques. However, it seems that some of these solutions are not well accepted among Croatian speakers (e.g., Patekar 2019). This is especially evident in certain domains, like show business and information technology (e.g., Drljača Margić 2012, 2014). Multi-word expressions and descriptions are often complex to use (e.g., Drljača 2006; Škifić and Mustapić, 2012), as exemplified by *programska podrška* 'software' and *razvojni inženjer* 'developer' (Institute of Croatian language and linguistics 2015). Using an existing word and giving it a new meaning can result in insufficient precision (Drljača 2006), as in *spravica* (eng. small device) for 'gadget' (Institute of Croatian language and linguistics 2015). Finally, the process of introducing a new word or calque is

usually too slow (e.g., Muhvić-Dimanovski and Skelin Horvat 2008). These reasons show that English loanwords are also used because, in some cases, they are more economical compared to native alternatives.

2. Tools and resources for English loanwords

English loanwords have been investigated from the perspective of contact linguistics (e.g., Filipović 1990), descriptive linguistics (e.g., Mykytka 2017), sociolinguistics (e.g., Ćoso and Bogunović 2017; Drljača Margić 2012; Greenall 2005), psycholinguistics (e.g., Bogunović 2017; Pavlinušić Vilus, Bogunović and Ćoso 2022), and computational linguistics (e.g., Alvarez-Mellado 2020; Serigos 2017). To gain an in-depth understanding of the phenomenon, all of these approaches need to be taken into consideration. Most Croatian studies rely on selectively chosen English loanwords (e.g., Drljača 2006; Ćoso and Bogunović 2017; Patekar 2019) or small-scale, domain-specific corpora (e.g., Brdar 2010; Hudeček and Mihaljević 2005). One of the reasons behind that is the fact that Croatian computational linguistic tools and resources are still relatively underdeveloped (e.g., Tadić 2003; Tadić et al. 2012). Only a few resources for English words are available in Croatian. Aside from traditional dictionaries of anglicisms (e.g., Filipović 1990; Görlach 2002), some new resources have been developed. For example, new words, including some English words, are listed in an online dictionary of neologisms (Muhvić-Dimanovski et al. 2016). The website *Bolje je hrvatski!* (Institute for Croatian language and linguistics 2015) selectively records the intake of foreign words into Croatian and proposes Croatian equivalents, while *Kontekst.io* (Kontekst.io n.d) searches the Croatian Web Corpus, *hrWaC* (Ljubešić and Erjavec 2011; Ljubešić and Klubička 2014) to provide the information about word frequency.

None of the above-mentioned sources offers a systematic and detailed insight into which English words are used in Croatian and how frequently. To answer that question, research in other languages has mainly relied on corpus-based searches. Currently, there are several corpora available for Croatian, some of which are The Croatian National Corpus (Tadić 2009), The Croatian Language Repository (Brozović Rončević and Ćavar 2008), The Croatian Web Corpus, *hrWaC* (Ljubešić and Erjavec 2011; Ljubešić and Klubička 2014) and The Corpus of Croatian News Portals, *ENGRI* (Bognović et al., 2021; Bogunović and Kučić 2021; Kučić, 2021). The Croatian Web Corpus, *hrWaC*, is considered the most comprehensive corpus of the

Croatian language. It includes texts representative of the standard language (e.g., official websites) as well as texts from blogs, advertisements, user comments, discussions, etc. Its greatest advantage is its size and the variety of texts, and the main disadvantage is that it has not been updated since 2014. The Corpus of Croatian News Portals, ENGRI (Bogunović et al., 2021; Bogunović and Kučić 2021; Kučić 2021) consists of 2,395,735 texts collected from the 12 most popular Croatian news portals (Reuters Institute for the Study of Journalism 2021) published between 2014 and 2020. Its advantage resides in the newer data, but it is smaller than hrWaC, with texts mostly from informal and publicist style.

To extract English loanwords from corpora, researchers have used different methods: some authors relied on manual search (e.g., Luján García 2017; Núñez Nogueroles 2016), while others used some of the existing computational linguistic tools or developed new ones (e.g., Alex 2005, 2008; Andersen 2012; Losnegaard and Lyse 2012). For example, assuming that the number of results obtained by a Google search can indicate language membership, an unsupervised system for recognition of English loanwords in German was developed using lexical databases and data available on the Internet (Alex 2005). One disadvantage of this approach is that it may not be suitable for languages which are, like Croatian, under-represented on the Internet. Another approach could be to use lexicon lookup in combination with character N-grams (e.g., Furiassi and Hofland 2007). However, this could be problematic for languages with underdeveloped computational linguistic resources, such as Croatian (e.g., Tadić 2003; Tadić et al. 2012).

To avoid the above-mentioned problems, supervised machine learning methods can be used in combination with N-grams (e.g., Alvarez-Mellado 2020; Andersen 2012; Castro et al. 2016; Serigos 2017). One disadvantage of this approach is ground truth data that needs to be collected for algorithm training. In Croatian, such a dataset cannot be obtained from a list of English words, because English loanwords sometimes occur with Croatian affixes (e.g., *snowboardi*), and some words occur in both languages as interlingual homographs (e.g., *more* 'sea'; *love* 'chase', v. 3rd person pl.; *car* 'emperor').

3. The present study

This study focuses on words borrowed from English and used in orthographically unadapted form. Literature search yielded several different

terms for such words: raw anglicisms (e.g., Kavgić 2013), English loanwords (e.g., Görlach 2002), foreign words (e.g., Međeral 2016). This paper does not aim at resolving terminological issues, so the term English loanwords will be used for such words.

As shown above, English loanwords have been thoroughly explored from various perspectives using different approaches, methods, and theoretical frameworks. At the same time, the data on which English loanwords occur in Croatian and how frequently has only recently become available. Such data enable a better understanding of the reasons behind the use of English loanwords (e.g., the lack or inadequacy of native equivalents). Additionally, word frequency seems to be the most potent of all factors that affect the word recognition process (Murray and Forster 2004), so the data on the frequency of English loanwords represents a valuable resource in psycholinguistic research. Such data should also be taken into consideration in language policy and planning because it offers an insight into speakers' communication needs and preferences. Thus, the main goal of this study is to present a corpus-based database of unadapted English loanwords in Croatian with their frequencies.

4. Methodology

The ENGRI corpus (Bogunović et al., 2021; Bogunović and Kučić 2021; Kučić 2021) was chosen because it contains the most recent data, and the texts are representative of the language of the media, which has been recognized as an important factor in introducing new words (e.g., Drljača Margić 2009, 2011; Muhvić-Dimanovski and Skelin Horvat 2008).

The classification algorithm was trained and tested on a manually labeled dataset, built from 60,000 randomly selected words from the ENGRI corpus. The words were then manually classified by three independent evaluators, all anglicists, as 'Croatian', 'Croatian and English', 'English' and 'non-Croatian and non-English'. A total of 55,395 words were unanimously evaluated and included in the final dataset.

The initial algorithmic classification resulted in 1,373,309 words. All words occurring less than twice were excluded due to low frequency. Furthermore, words with less than two letters (e.g., 'I', 'a') were also excluded as the focus was on content words, which are considered more 'borrowable' compared to function words (e.g., Tadmor et al. 2010). This reduced the number of words to 616,672. The word list was further cross-checked with the

Croatian morphological lexicon, CML (Institute of Croatian language and linguistics and Faculty of Humanities and Social Sciences, University of Zagreb 2005), and the manually labeled dataset to eliminate Croatian words classified as English by the algorithm. After eliminating all words containing three or more of the same letters in a row, the number of words was reduced to 326,838. Using the classifier, all words classified as non-English, unless manually labeled as English, were excluded. This resulted in a list of 47,080 words.

The following step was to obtain standard and non-standard lemmas for each word using *A CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian, and Bulgarian* (Common Language Resources and Technology Infrastructure, Slovenia, CLARIN.SI n.d.). The lemmas were again cross-checked with the CML, to exclude any remaining Croatian lemmas. The classifier was used for standard and non-standard lemmas to eliminate all words classified as non-English unless labeled as English in the manually labeled dataset. All words with standard lemmas beginning with 'al-' (e.g., *al-shamal*) or ending with '-hr/-com' (webpages) were also removed. Next, words starting with double letters, with the exception of 'aardvark', 'eerie', 'eek', 'eel', 'llama', 'ooh', 'oops', 'ooze', and 'oozing' (Gormandy White n.d.), were filtered out. As already mentioned, this study focuses on content words, so stop words (e.g., articles, pronouns, conjunctions, prepositions) were also excluded.

The remaining words were then reduced to the automated lemmas, which resulted in a total of 34,728 words. The next step included manual classification and cleansing, resulting in 15,751 words. Finally, the words were manually reduced to their lemmas and cross-checked with Google Dictionary.

5. Results

The final database of English loanwords contains 9,452 unadapted English loanwords as well as their absolute and relative frequencies obtained from the ENGRI corpus. The whole database is publicly available at https://figshare.com/articles/dataset/The_database_of_English_words_in_Croatian_xlsx/20014364. The words are listed according to their frequency as well as alphabetically.

The absolute frequency of 4,233 words is equal to or less than 10, meaning that 44,78% of all English loanwords from the database fall into this category. Most of them (1,197) occur only three times in the corpus,

while words whose absolute frequency is 10 occur 241 times. The number of words in each frequency category with 3-10 occurrences in the ENGRI corpus is shown in Table 1.

Table 1. *Number of words per frequency category (3-10)*

Frequency category	Number of words	Examples
3	1,197	dislike, outwear, leap, multiplay, bliss
4	808	jukebox, hillbilly, hairy, eyelash, eternity
5	559	jacket, inner, handsome, gunpoint, dusk
6	452	waterproof, upbeat, totally, sunlight, sharp
7	371	toxic, sweat, rename, outstanding, onset
8	325	unfair, timeline, seafood, poison, oldie
9	280	wildlife, spammer, relay, perk, incredibly
10	241	underwater, renew, lifespan, overboost

The distribution of categories of words from the frequency category 3-10 are illustrated in Figure 1.

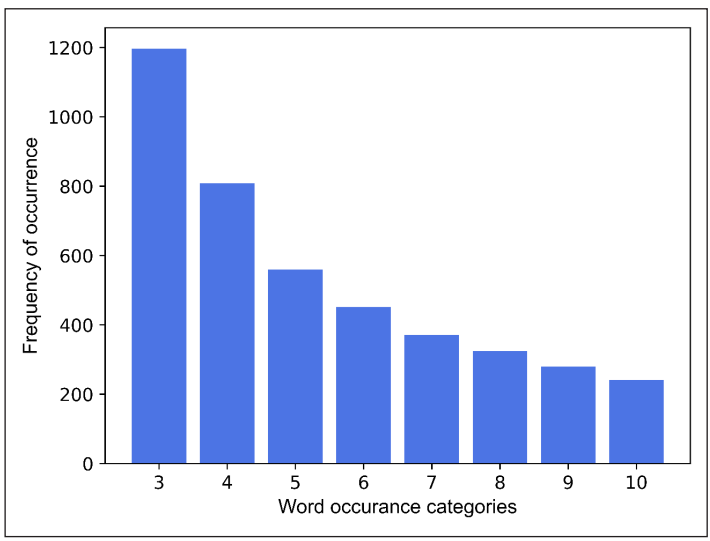

Figure 1. *Distribution of word frequency categories 3-10*

Most of the remaining English loanwords (31,07%) occur between 11 and 50 times. The words that occur between 51 and 20,000 are divided into the following frequency categories: 51-100, 101-500, 501-1,000, 1,001-5,000, 5,001-10,000, and 10,001-20,000. Only three words occur between 20,001 and 30,000 times: 'rock' (22,958), 'web' (27,045) and 'online' (28,246), while the most frequent English loanword in Croatian is 'show', with 80,805 examples found in the ENGRI corpus, which makes 0.0122% of all words from the corpus. The described word categories (11-80,805) are shown in Table 2.

Table 2. *Number of words per frequency category (11-80,805)*

Frequency category	Number of words	Examples
11-50	2,937	spotlight, duckface, gamepad, boring
51-100	807	prank, soon, jeep, hack, bypass, boyfriend
101-500	1,071	backhand, holiday, fight, swap, guy, workout
501-1,000	193	bike, outdoor, developer, friendly, jackpot
1,001-5,000	184	snowboard, grill, story, catering, beach
5,000-10,000	15	start-up, fitness, blog, selfie, event
10,000-20,000	8	e-mail, reality, shop
20,001 and 30,000	3	rock, web, online
80,805	1	show

The distribution of the above-described frequency categories is presented in Figure 2.

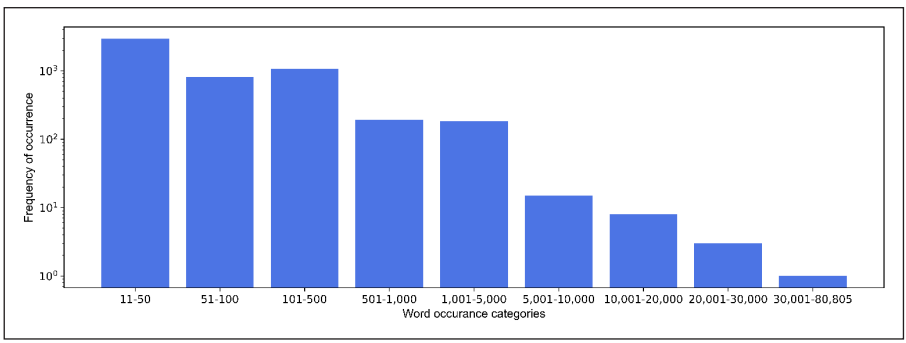

Figure 2. *Distribution of word frequency categories 11-80,805*

The list of 100 most frequent English loanwords with their absolute frequencies is given in the Appendix. Most of these words could be categorized as domain-nonspecific or general. Based on their use in Croatian, other loanwords could be categorized as follows: lifestyle, music, sports, technology, showbusiness, information technology, business, social media/gaming, media, politics, and medicine. Of course, the words can have multiple meanings, or they could be used across different domains (e.g., 'club' as a nightclub or a football club). The domain categories of 100 most frequent English loanwords from the ENGRI corpus are shown in Table 3.

Table 3. *The domains of 100 most frequent English loanwords*

Domain category	Examples
general	gay, live, medium, play, face, street, art, stand, multiple, open
lifestyle	party, style, wellness, make-up, vintage
music	rock, jazz, blues
sports	play-off, tiebreaker, fitness, football, team, triple-double
showbusiness	show, reality, celebrity, trailer, blockbuster, stage
IT	online, web, e-mail, file, link
technology	smartphone, tablet, laptop, gadget
business	start-up, outsource, offshore, lease
social media/ gaming	stream, selfie, post, blog, tweet
media	mainstream, news
politics	summit, spin
medicine	tumour

Considering a significant drop in the number of loanwords with an absolute frequency above 5,000, Croatian equivalents for these loanwords were searched in various sources (Croatian language portal 2006; Google translator; Glosbe, hr; Institute of Croatian language and linguistics 2015). The results are presented in Table 4.

Table 4. *English loanwords with a frequency higher than 5,000 with Croatian equivalents*

English word	Croatian equivalent	English word	Croatian equivalent
show	emisija, šou	**gay**	homoseksualna osoba
online	mrežni	**selfie**	sebić
web	mreža	**summit**	sastanak na vrhu
rock	-	**post**	objava
mail	pošta	**play-off**	doigravanje
break, n.	pauza	**live**	uživo
party	zabava	**blog**	mrežni dnevnik
e-mail	e-pošta	**tie-breaker**	-
reality	stvarnost	**medium**	srednji
jazz	džez	**fitness**	-
shop	trgovina	**blues**	-
press	tisak	**start-up**	-
event	događanje	**bin**	kanta, kutija
ring	ring		

6. Discussion

Some loanwords from the database (e.g., 'summit', 'vintage', 'benefit') originate from other languages. Some authors (e.g., Filipović 1990) argue that even words that are not English in origin, but were borrowed from English (proximal etymology) can be termed anglicisms. However, such issues were not in the focus of study as it presents the results obtained from algorithmic classification, and is intended for further analyses and research. Also, the database by no means represents a final product and is not a definite representation of data on English words in Croatian. Further efforts will be made to update the database and incorporate new data.

The total word count in the database is 9,452, with 44,78% of words occurring 10 or fewer times in the ENGRI corpus, while 31,07% of the words occur between 11 and 50 times. Despite their low frequency of occurrence, some of these words do not have adequate native equivalents

(e.g., 'multiplay', 'gamepad'). One possible explanation could be that they occur in very specific contexts (e.g., gaming), which are not expected to be broadly represented in the ENGRI corpus, given that it contains texts from most popular Croatian news portals. On the other hand, not all words from this frequency category are domain specific. Words like 'satisfaction' and 'boring' have well-established native equivalents, so their low frequency of occurrence is not surprising. However, the fact that they are still used could be related to prestige (e.g., Ćoso and Bogunović 2017; Field 2002).

The most significant decline in the number of loanwords can be observed after 500 occurrences, when the number of words per category drops under 200 (e.g., 'bike', 'outdoor', 'developer'), similarly as in the group of loanwords between 1,001 and 5,000 (e.g., 'snowboard', 'catering', 'beach'). Again, both frequency categories contain loanwords with (e.g., 'bike', 'beach') and without single-word native translations (e.g., 'developer', 'snowboard'). On the one hand, the more frequent use of loanwords which do not have single-word Croatian equivalents may reflect the problem of economy (e.g., Drljača 2011), indicating the need for less complex native solutions. On the other hand, English loanwords are used even when single-word Croatian words exist and are widely used (e.g., 'bike', 'beach'). While this could in part be attributed to the prestigious status of English (Crystal 2003) and/or exposure through the media (e.g., Drljača 2011), it is also possible that these loanwords are used as parts of multiword units (e.g., 'bike sharing', 'after beach party'), which cannot be directly translated into Croatian. Another option would be to use descriptions, which can be very complex to use in practice (e.g., Drljača 2006; Škifić and Mustapić 2012).

Only 27 loanwords occur more than 5,000 times. One explanation could be that these loanwords are used to fill in lexical gaps in cases when native words either do not exist or if they do, the speakers might not be fully satisfied with them (e.g., Patekar 2019). The data on native equivalents shows that five English loanwords from this group ('rock', 'tie-breaker', 'fitness', 'blues', 'start-up') do not have Croatian translations. Interestingly, the Institute of Croatian language and linguistics (2015) offers a translation for 'start-up company' (*razvojna tvrtka*), where *razvojna* is an adjective and cannot be used as a translation when 'start-up' is used as a noun. This is due to the fact that English allows nouns to modify other nouns, which is not the case in Croatian. The fact that these 27 words are most frequent English loanwords in the ENGRI corpus clearly shows there is a need for Croatian equivalents.

Multiword expressions are proposed as equivalents for three English loanwords (*homoseksualna osoba* 'gay', *sastanak na vrhu* 'summit', *mrežni dnevnik* 'blog'). The complexity of such solutions might justify the use of English words in some cases (e.g., Škifić and Mustapić 2012). The complexity of these expressions is also reflected in their (in)ability to form other word categories. For example, an adjective can be derived from 'summit' but not from its Croatian equivalent *sastanak na vrhu* (Drljača Margić 2011).

The remaining 19 words have single-word Croatian translations. Equivalents for three English loanwords are adapted forms of English words: *šou* 'show', *džez* 'jazz', and *ring* 'ring'. Hereby it should be emphasized that the word 'ring' is used in the domain of sports (e.g., 'boxing ring'). When English words enter the Croatian media, they might eventually adapt to Croatian as a result of the influence of the media on the intake of new words (e.g., Drljača Margić 2009; Muhvić-Dimanovski and Skelin Horvat 2008). Some authors advocate a more flexible approach, according to which the Croatian language should be more open to English loanwords, if they can easily adapt to its rules (e.g., Peti-Stantić 2013).

One example of creating a calque is *sebić* 'selfie'. While in many cases, the process of introducing new words is too slow can take years (e.g., Muhvić-Dimanovski and Skelin Horvat 2008), in this case *sebić* (Halonja and Hudeček 2014) was proposed only a year after 'selfie' was selected word of the year by Oxford Dictionaries (Reuters 2013) based on the spike in popularity. Given the role of the media in shaping of a language (Drljača 2009; 2011), it seems that the fact that the English loanword is still widely used in the Croatian media could have reduced the likelihood of the proposed Croatian equivalent to be accepted by Croatian speakers.

Croatian word *e-pošta* 'e-mail' illustrates the adaptation of an existing word to convey a new meaning. Although borrowed from English, the prefix 'e-' has been recognized as a very productive way of word formation in Croatian (e.g., Halonja and Mihaljević 2012). The results show that the English word 'e-mail' is still frequently used in the Croatian media. One possible explanation is that even though Croatian speakers prefer native words in formal contexts, they seem to have positive attitudes towards English loanwords in informal contexts, especially in Information technology and internet-related domains (e.g., Matić 2017; Rüdiger 2018).

Native translations of the remaining 14 English loanwords are the existing Croatian words, whose meaning has been broadened to include new

meanings (semantic borrowing). Thus, it can be observed that these English loanwords are used in a different or a narrower sense compared to their Croatian equivalents. In other words, they are examples of restriction of meaning (Filipović 1986). Aside from the adapted form *šou*, another Croatian equivalent for 'show' is *emisija*, and it is used for any type of TV or radio show, while 'show' and its adapted form *šou* are used for artistic and entertainment performances (Croatian Language Portal 2006). These definitions indicate that there was a need for a new word. Due to the fact that the word occurs frequently in the media, it was eventually adapted to Croatian. While the use of the English form could partly be attributed to prestige, it should also be noted that 'show' occurs in multiword units which describe a specific type of show (e.g., 'talk show', 'reality show'). Similarly, the native equivalent for the word 'reality' is *stvarnost,* but the English word is typically used to refer to a 'reality show'. The described example, where an element of an original English phrase is omitted, is termed ellipsis (Filipović 1986; Fabijanić 2010). As described above, such multiword units often cannot be directly translated into Croatian, which might explain why English words are used instead.

The word 'shop' is another example of an ellipsis, as it is commonly used to refer to a 'webshop'. Since 'web' and 'online' also have native translations, it can be assumed that these words are frequently used on the internet, where English is the dominant language (e.g., Gluszek and Hansen 2013). The use of the internet has been recognized as one of the main informal activities which enable spontaneous vocabulary acquisition (e.g., Godwin-Jones 2019; Peters 2018). In other words, the speakers could be more exposed to English words like 'web' and 'online' than their Croatian equivalents, which could explain the frequent use of these loanwords in the corpus. The use of English loanwords that have native equivalents could also be related to the prestigious status of English and language attitudes, as research has shown that more frequent use of English loanwords is related to higher scores on the social attractiveness dimension (Ćoso and Bogunović 2017).

Even though show business (e.g., Drljača Margić 2014) and information technology (e.g., Matić 2017; Rüdiger 2018) have often been emphasized as domains with a lot of English loanwords, most frequent loanwords found in this study can be categorized as domain-nonspecific. As the loanwords were analyzed independently and out of context, it is also possible that these words were parts of multiword units. However, the context could reveal more about how these words are used in Croatian.

Taken together the results of this study support previous findings about the influence of English on Croatian, especially on the lexical level (e.g., Brdar 2010; Matić 2017). The use of English loanwords can in part be attributed to the prestigious status of English (e.g., Ćoso and Bogunović 2017; Field 2002), and the lack of adequate native equivalents (e.g., Drljača 2006; Patekar 2019). The analysis of most frequent English loanwords with respect to the availability and form of Croatian translation equivalents highlights the role of the media in introduction of new words and shaping of a language in general (e.g., Drljača Margić 2009; Muhvić-Dimanovski and Skelin Horvat 2008).

Finally, it is necessary to emphasize that the current database represents the starting point for further analyses and development. Some of the limitations of this study include the possibility of human error in the process of manual classification and cleansing. The questions raised in this paper, such as etymological issues, also deserve future attention. More detailed elaboration regarding the availability of native equivalents would certainly give interesting results. Also, an analysis of context, including occurrence in multiword units, would help in understanding some of the findings of this study.

7. Conclusions

This study presents the corpus-based database of English loanwords in Croatian. The database contains 9,452 unadapted English loanwords as well as their absolute and relative frequencies. The majority of English loanwords, 75,85%, occur less than 50 times. This could imply that these loanwords are used in very specific contexts, or that they have well-established native equivalents.

The most significant decline in the number of loanwords is observed after 500 occurrences, when the number of loanwords per category drops under 200, while only 27 loanwords occur more than 5,000 times. The most frequent English loanword is 'show' with 80,805 occurrences, which makes 0.0122% of all words from the corpus. The analysis of loanwords that occur more than 5,000 times shows that most of them have native equivalents. Their use could be motivated by the inadequacy of Croatian translation equivalents to fulfill the speakers' communication needs, exposure to the English language, and prestige. Moreover, it is necessary to emphasize the role of the media in shaping of a language.

The database provides an insight into which unadapted English loanwords are used in Croatian, which can lead to a better understanding of the reasons behind the use of such words. Additionally, the data on the frequency of English loanwords represents a valuable resource in an interdisciplinary study of English loanwords.

Funding

This work was funded by the Croatian Science Foundation [UIP-2019-04-1576].

References

Alex, Beatrice (2005) "An unsupervised system for identifying English inclusions in German text", *43. Proceedings of the Annual Meeting of the Association for Computational Linguistics*, eds. Chris Callison-Burch and Stephen Wan, Michigan, pp. 133–138.

Alex, Beatrice (2008) "Comparing Corpus-based to Web-based Lookup Techniques for Automatic English Inclusion Detection", *Proceedings of the 6th International Conference on Language Resources and Evaluation,* eds. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias, Marrakech, pp. 2693–2697.

Alvarez-Mellado, Elena (2020) "An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines", *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, Marseille, pp. 1–8.

Andersen, Gisle (2012) "Semi-automatic approaches to Anglicism detection in Norwegian corpus data", *The Anglicization of European Lexis,* eds. Cristiano Furiassi, Virginia Pulcini and Félix Rodríguez González, Amsterdam-Philadephia, pp. 111–130.

Bogunović, Irena (2017) *Cross-language priming: Evidence from Croatian-English bilinguals with different second language proficiency levels*, doctoral thesis, University of Zagreb.

Bogunović, Irena, Kučić. Mario (2021) *Corpus of Croatian news portals ENGRI (2014-2021)*, available at https://repository.pfri.uniri.hr/islandora/object/pfri%3A2156, accessed 20 December 2021.

Bogunović Irena, Kučić Mario, Ljubešić Nikola and Erjavec Tomaž (2021) *Corpus of Croatian News Portals ENGRI (2014-2018)*. Slovenian lan-

guage resource repository CLARIN.SI, available at https://www.clarin.si/repository/xmlui/handle/11356/1416. accessed 20 December 2021.

Brannen, Mary Y., Rebecca Piekkari, Rebecca, Tietze, Susanne (2014) "The multifaceted role of language in international business: Unpacking the forms, functions, and features of a critical challenge to MNC theory and performance", *Journal of International Business Studies,* 45, pp. 495–507.

Brdar, Irena (2010) "Engleske riječi u jeziku hrvatskih medija", *Lahor,* 10, Zagreb, pp. 217–232.

Brozović Rončević, Dunja and Ćavar, Damir (2008) "Hrvatska jezična riznica kao podloga jezičnim i jezično povijesnim istraživanjima hrvatskoga jezika", *Ohrid: XIV. međunarodni slavistički kongres,* eds. Berdnardina Petrović and Marko Samardžija, Hrvatsko filološko društvo/Hrvatska sveučilišna naklada, Zagreb, pp. 173–186.

Common Language Resources and Technology Infrastructure, Slovenia (n.d.) *A CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian, and Bulgarian*, available at https://pypi.org/project/classla, accessed 11 February 2021.

Castro, Dayvid, Souza, Ellen, De Oliveira, Adriano L. I. (2016) "Discriminating between Brazilian and European Portuguese National Varieties on Twitter Texts", *5th Brazilian Conference on Intelligent Systems*, Recife, pp. 265–270.

Croatian Language Portal (2006) *Show*, available at https://hjp.znanje.hr/index.php?show=search, accessed 15 December 2021.

Crystal, David (2003) *English as a global language*. Cambridge University Press, New York.

Čepon, Slavica (2017) "Anglicizmi v poslovni nomenklaturi turistinih podjetij v Sloveniji", *Revija za ekonomske in poslovne vede,* 2, Novo Mesto, pp. 35–49.

Ćoso, Bojana, Bogunović, Irena (2017) "Person perception and language: A case of English words in Croatian", *Language and Communication,* 53, pp. 25–34.

Drljača, Branka (2006) "Anglizmi u ekonomskome nazivlju hrvatskoga jezika i standardnojezična norma", *Fluminensia,* 18, 1, Rijeka, pp. 65–85.

Drljača Margić, Branka (2009) "Latentno posuđivanje u hrvatskome i drugim jezicima – posljedice i otpori", *Rasprave Instituta za hrvatski jezik i jezikoslovlje,* 35, Zagreb, pp. 53–71.

Drljača Margić, Branka (2011) "Leksički paralelizam: Je li opravdano govoriti o nepotrebnim posuđenicama (engleskoga podrijetla)?" *Fluminensia*, 23, 1, Rijeka, pp. 53–66.

Drljača Margić, Branka (2012) "Croatian university students' perception of stylistic and domain-based differences between Anglicisms and their native equivalents", *Languages, Literatures and Cultures in Contact: English and American Studies in the Age of Global Communication, Volume 2: Language and Culture,* eds. Marta Dąbrowska, Justyna Leśniewska and Beata Piątek, Tertium, Krakow, pp. 109–126.

Drljača Margić, Branka (2014) "Contemporary English influence on Croatian: A university students' perspective", *Language Contact Around the Globe, Proceedings of the LCTG3 Conference,* eds. Amei Koll-Stobbe and Sebastian Knospe, Peter Lang, Wien, pp. 73–92.

Entlová, Gabriela, Mala, Eva (2020) "The occurrence of anglicisms in the Czech and Slovak lexicons", *Xlinguae,* 13, 2, Nitra, pp. 140–148.

Fabijanić, Ivo (2010) "Reinterpretacija elipse u formiranju anglizama", *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 36, 2, Zagreb, pp. 251–273.

Field, Fredrich W. (2002) "*Linguistic borrowing in bilingual contexts",* Benjamins Publishing Company, Amsterdam.

Filipović, Rudolf (1986) *Teorija jezika u kontaktu*, Školska knjiga, Zagreb

Filipović, Rudolf (1990) *Anglicizmi u hrvatskom ili srpskom jeziku: porijeklo – razvoj – značenje*, Jazu – Školska knjiga, Zagreb.

Furiassi, Cristiano, Hofland, Knut (2007) "The retrieval of false anglicisms in newspaper texts", *Corpus Linguistics 25 Years On,* ed. Roberta Facchinetti, Brill / Rodopi, Amsterdam-New York, pp. 347–363.

Gluszek, Agata, Hansen, Karolina (2013) "Language attitudes in the Americas", *The Social Meanings of Languages, Dialects, and Accents: An International Perspective,* eds. Marko Dragojevic, Howard Giles and Bernadette M., Peter Lang, New York, pp. 26–44.

Godwin-Jones Robert (2019) "Contributing, Creating, Curating: Digital Literacies for Language Learners", *Language Learning and Technology,* 19, 3, Honolulu, pp. 8–20.

Gormandy White, Mary (n.d.) *Words with double letters*, available at https://grammar.yourdictionary.com/word-lists/words-with-double-letters.html, accessed 11 February 2021.

Görlach, Manfred (2002) *An Annotated Bibliography of European Anglicisms*, Oxford: Oxford University Press.

Greenall, Annjo K (2005) "To translate or not to translate: Attitudes to English loanwords in Norwegian", *The Consequences of Mobility,* eds. Bent Preisler, Anne Fabricius, Hartmut Haberland, Sussane Kjærbeck and Karen Risager, Roskilde University, Roskilde, pp. 212–226.

Halonja, Antun, Hudeček, Lana (2014) "Pokloni mi svoj *selfie"*, *Hrvatski jezik,* 2, Zagreb, pp. 26–27.

Halonja, Antun, Mihaljević, Milica (2012) "Nazivi sa sastavnicom e- u hrvatskome jeziku", Rasprave *Instituta za hrvatski jezik i jezikoslovlje,* 38, 1, Zagreb, pp. 55–86.

Hudeček, Lana, Mihaljević, Milica (2005) "Nacrt za višerazinsku kontrastivnu englesko-hrvatsku analizu", *Rasprave Instituta za hrvatski jezik i jezikoslovlje,* 31, 1, Zagreb, pp. 107–151.

Institute of Croatian language and linguistics (2015) *Bolje je hrvatski!*, available at http://bolje.hr/, accessed 30 January 2022.

Institute of Croatian language and linguistics, Faculty of Humanities and Social Sciences, University of Zagreb (2005) *Croatian morphological lexicon*, available at http://hml.ffzg.hr/hml/, accessed 23 February 2021.

Kavgić, Aleksandar (2013) "Intended communicative effects of using borrowed English vocabulary from the point of view of the addressor: Corpus-based pragmatic analysis of a magazine column", Jezikoslovlje 14, 2-3, Osijek, pp. 487–499. https://hrcak.srce.hr/112204.

Kay, Gillian (1995) "English loanwords in Japanese", *World Englishes,* 14, 1, pp. 67–76.

Kontkst.io. (n.d.) available at https://kontekst.io/, accessed 11 February 2021.

Kučić, Mario (2021) "Creating a Web Corpus Using GO". *44. Proceedings of the International Convention MIPRO 2021,* ed. Karolj Skala, Institute of Electrical and Electronics Engineers, Rijeka, pp. 1931–1934.

Losnegaard, Gyri S., Lyse, Gunn Inger (2012) "A data-driven approach to anglicism identification in Norwegian", *Exploring Newspaper Lan-*

guage: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian, ed. Gisle Andersen, John Benjamins, Amsterdam, pp. 131–154.

Luján García, Carmen (2017) "Analysis of the presence of Anglicisms in a Spanish internet forum: some terms from the fields of fashion, beauty, and leisure", *Alicante Journal of English Studies,* 30*,* Alicante, pp. 281–305.

Ljubešić, Nikola, Tomaž Erjavec (2011) "hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene", *Text, Speech and Dialogue, Lecture Notes in Computer Science,* eds. Ivan Habernal and Vaclav Matousek, Springer, Berlin/ Heidelberg, pp. 395–402.

Ljubešić, Nikola, Klubička, Filip (2014) "{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian", *Proceedings of the 9th Web as Corpus Workshop,* eds. Felix Bildhauer and Roland Schäfer, Association for Computational Linguistics, Gothenburg, pp. 29–35.

Matić, Daniela (2017) "Perception of the English element in the scientific register of Croatian ICT university educational material with graduate ICT students", *Jezikoslovlje,* 18, 2, Osijek, pp. 319–345.

McKenzie, Robert M (2010) *The Social Psychology of English as a Global Language: Attitudes, Awareness, and Identity in the Japanese Context*, Springer, New York.

Međeral, Krešimir (2016) "Jezične bakterije – pomagači ili štetočine u jezičnome organizmu?", *Hrvatski jezik,* 3, Zagreb, pp. 1–10.

Mihaljević, Milica (2003) *Kako se na hrvatskome kaže WWW? Kroatistički pogled na svijet računala*. Hrvatska Sveučilišna naklada, Zagreb.

Muhvić-Dimanovski, Vesna, Skelin Horvat, Anita (2008) "Contests and nominations for new words – why are they interesting and what do they show", *Suvremena lingvistika,* 65, Zagreb, pp. 1–26. https://hrcak.srce.hr/25183.

Muhvić-Dimanovski, Vesna, Skelin Horvat, Anita, Hriberski, Diana (2016) *Rječnik neologizama u hrvatskome jeziku, available at* www.rjecnik.neologizam.ffzg.unizg.hr, accessed 12 November 2020.

Murray, Wayne S., Forster, Kenneth I. (2004) "Serial mechanisms in lexical access: The rank hypothesis", *Psychological Review,* 111, 3, pp. 721–756.

Mykyatka, Iryna (2017) "The Influence of English on the Spanish Register of Photography: An Empirical Study", *ESP Today,* 5, 1, Belgrade, pp. 68–90.

Núñez Nogueroles, Eugenia (2016) "Anglicisms in CREA: A Quantitative Analysis in Spanish Newspapers", *Language Design,* 18, Barcelona, pp. 215–242.

Patekar, Jakob (2019) "Prihvatljivost prevedenica kao zamjena za anglizme", *Fluminensia,* 31, 2, Rijeka, pp. 143–179.

Pavlinušić Vilus, Eva, Bogunović, Irena, Ćoso Bojana (2022) "Lexical access to unadapted English loanwords in Croatian: Evidence from translation priming", *ExLing 2022 Paris: Proceedings of 13th International Conference of Experimental Linguistics*, ed. Botinis A., ExLing Society, Paris, pp. 125–128.

Peters, Elke (2018) "The effect of out-of-class exposure to English language media on learners' vocabulary knowledge", *International Journal of Applied Linguistics,* 169, 1, pp. 142–168.

Peti-Stantić, Anita (2013) "Domaće je (naj)bolje", *Javni jezik kao poligon jezičnih eksperimenata,* ed. Barbara Kryžan-Stanojević, Srednja Europa, Zagreb, pp. 39–51.

Pulcini, Virginia, Furiassi, Cristiano, Gonzales, Felix R. (2012) "The lexical influence of English on European languages: From words to phraseology". *Anglicization of European Lexis*. eds. Virginia Pulcini, Cristiano Furiassi and Felix R. Rodrigues, John Benjamins, Amsterdam – Philadelphia, pp. 1–27.

Rüdiger, Sofia (2018) "Mixed Feelings: Attitudes towards English loanwords and their use in South Korea", *Open Linguistics,* 4, pp. 184–198.

Reuters (2013) "*Selfie" beats "twerk" as word of the year,* available at https://www.reuters.com/article/books-oxford-selfie-id-USL5N0J343520131119, accessed 15 March 2022.

Reuters Institute for the Study of Journalism (2021) *Reuters Institute Digital News Report*, available at https://www.digitalnewsreport.org/, accessed 28 January 2021.

Serigos, Jacqueline. R. L (2017) *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish,* doctoral thesis, University of Texas at Austin.

Škifić, Sanja, Mustapić, Emilija (2012) "Anglizmi i hrvatsko računalno nazivlje kroz prizmu jezičnog konflikta i jezične ideologije, *Jezikoslovlje,* 13, 2, Osijek, pp. 809–839.

Tadić, Marko (2003) *Jezične tehnologije i hrvatski jezik*, Exlibris, Zagreb.

Tadić, Marko (2009) "New version of the Croatian National Corpus", *After Half a Century of Slavonic Natural Language Processing,* eds. Dana Hlaváčková, Aleš Horák, Klara Osolsobě and Pavel Rychlý, Brno, pp. 199–205.

Tadić, Marko, Brozović-Rončević, Dunja, Kapetanović, Amir (2012) *Hrvatski jezik u digitalnom dobu*, Springer, Heidelberg.

Tadmor, Uri, Haspelmath, Martin, Taylor, Bradley (2010) "Borrowability and the notion of basic vocabulary", *Diachronica,* 27, 2, pp. 226–246.

Twenge, Jean M., Martin, Gabrielle N., Spitzberg, Brian H. (2019) "Trends in U.S. Adolescents' media use 1976-2016: The rise of the digital media, the decline of TV and the (near) demise of print", *Psychology of Popular Media,* 8, 4, pp. 329–345.

Appendix

Table A.1 *100 most frequent English loanwords in the ENGRI corpus.*

Lemma	Freq.	Lemma	Freq.	Lemma	Freq.
show	80805	**wellness**	4322	**grand**	2609
online	28246	**band**	4297	**make-up**	2594
web	27045	**brand**	4265	**laptop**	2570
rock	22958	**match**	4265	**news**	2523
mail	14846	**art**	4176	**blockbuster**	2403
break	14545	**beauty**	3914	**techno**	2400
party	14260	**gem**	3914	**smoothie**	2397
e-mail	12727	**hat-trick**	3875	**monitor**	2368
reality	12048	**stream**	3874	**crossover**	2347
jazz	11545	**look**	3817	**screenshot**	2285
shop	11181	**smart-phone**	3779	**open**	2279
press	10435	**football**	3726	**ecstasy**	2224
event	9113	**street**	3595	**file**	2203
ring	8193	**cool**	3506	**roll-up**	2195
gay	7951	**style**	3487	**master**	2184
selfie	7740	**house**	3370	**vintage**	2153
summit	7718	**roll**	3345	**outsource**	2140
post	6799	**tweet**	3235	**fast**	2133
play-off	6779	**dance**	3156	**offshore**	2125
live	6532	**game**	3109	**ten**	2098
blog	6503	**stand**	3103	**benefit**	2086
tiebreaker	5841	**punk**	3089	**gadget**	2085
medium	5568	**pizza**	2996	**link**	2060
fitness	5497	**spin**	2935	**lease**	2052
blues	5300	**triple-double**	2914	**stage**	2046
start-up	5231	**team**	2911	**line**	2044
bin	5081	**celebrity**	2880	**miss**	2031
tumour	4782	**trailer**	2850	**roam**	2005
outfit	4765	**tablet**	2713	**business**	1961
mainstream	4760	**craft**	2688	**fair**	1951
fan	4718	**soul**	2667	**funk**	1946
food	4677	**country**	2626	**club**	1913
play	4509	**multiple**	2614	**use**	1869
face	4478				

SAŽETAK

Irena Bogunović

KORPUSNI PRISTUP ENGLESKIM POSUĐENICAMA: BAZA ENGLESKIH RIJEČI U HRVATSKOME

Neprilagođene engleske posuđenice postale su dio neformalne komunikacije u mnogim jezicima, uključujući i hrvatski. Njihova je uporaba često motivirana nedostatkom odgovarajućih domaćih riječi, izloženošću engleskom jeziku kroz medije, ali i prestižnim statusom engleskog jezika. Jezično je posuđivanje česta tema jezikoslovnih istraživanja, posebice posuđivanje iz engleskog. Dosadašnji su rezultati uglavnom temeljeni na analizama manjih, *ad hoc* korpusa. Stoga je cilj ovoga rada predstaviti Bazu engleskih riječi u hrvatskome. Baza je nastala kao rezultat algoritamske i ručne klasifikacije posuđenica iz Korpusa novinskih portala ENGRI te donosi popis 9,452 neprilagođenih engleskih posuđenica i podatke o njihovoj pojavnosti u korpusu. Analiza dobivenih podataka pokazala je da se većina riječi (75,85%) pojavljuje manje od 50 puta, dok se ukupno 44,78% posuđenica pojavljuje 10 ili manje puta. Najveći pad u broju posuđenica primjećuje se u kategorijima pojavnosti iznad 500, dok se samo 27 posuđenica pojavljuje 5,000 puta ili više. Najčešća engleska posuđenica u navedenom korpusu je *show,* a pojavljuje se 80,805 puta, što je 0.0122% svih posuđenica u korpusu. Analiza posuđenica koje se pojavljuju više od 5,000 puta pokazala je da većina njih ima domaće prijevodne istovrijednice, što potvrđuje ulogu medija u uvođenju novih riječi. Osim što pruža uvid u pojavnost engleskih posuđenica u hrvatskome, ova baza predstavlja i doprinos hrvatskim računalno-jezikoslovnim resursima te omogućuje podatke potrebne za eksperimentalna istraživanja.

Ključne riječi: *engleske posuđenice; leksičko posuđivanje; baza; korpus*