UDK 811.163.42'373.613
81'373.613:811.111
Izvorni znanstveni članak
Prihvaćeno za tisak: 2. lipnja 2025.
https://doi.org/10.22210/suvlin.2025.099.02

Mirjana Borucinsky, Irena Bogunović, Jasmina Jelčić Čolakovac
University of Rijeka, Faculty of Maritime Studies
mirjana.borucinsky@pfri.uniri.hr, irena.bogunovic@pfri.uniri.hr, jasmina.jelcic@pfri.uniri.hr

Applying Zipf's Law to English words in Croatian: A comparative study of frequency and word length

In corpus linguistics, the negative correlation between word frequency and word length is a well–documented phenomenon referred to as *Zipf's law*. This linguistic universal observed by Zipf, which posits that the length of a word is in an inverse relation to its frequency (but not necessarily proportional to), has also been confirmed by numerous studies, and its implications can be observed in different fields such as language teaching and cognitive language processing. However, there is a gap in research data when it comes to studying this phenomenon from the perspective of loanwords. Even though it has been observed that translation equivalents in Croatian generally exist for loanwords or English words that appear 5000 times or more in the Croatian corpus (*ENGRI* corpus), the question still remains why speakers of Croatian resort to using English words in such cases where first language (L1) equivalents exist. This paper examines the systematicity of the language universal that shorter words are more frequent when it comes to foreign (primarily English) words in the (Croatian) language, i.e. whether the most frequent English words are shorter than their Croatian equivalents. For the purpose of this research, the *Database of English words and their equivalents in Croatian* was examined. Results indicate that some degree of systematicity between word length and frequency can be observed, but they also highlight the need for incorporating a semantic component into the analysis. The results contribute to the theoretical discussion on language universals, and explain why Croatian users prefer English words and whether language economy is one of the reasons for the use of English words in Croatian.

1. Introduction

English words in Croatian have been studied from various perspectives. For example, in a psycholinguistic and sociolinguistic study, Ćoso and Bogunović (2017) analysed how English is perceived by speakers of Croatian, showing that there is a correlation between the frequent use of English words and higher ratings of social attractiveness. On the other hand, Kučić (2021) looked at English words in Croatian from the perspective of natural language processing and how foreign inclusions can best be extracted and identified in Croatian texts. Similarly, Bogunović (2023a), and Borucinsky and Bogunović (2022) used corpus linguistics methods to explain how to extract foreign inclusions from a corpus of Croatian, describing challenges that arise from this method. The majority of papers on English words in Croatian are from contact linguistics point of view (e.g. Muhvić–Dimanovski and Skelin Horvat 2006, 2008), contrastive linguistics, translation studies (e.g. Pritchard 1997; Ivir 1998; Pavlinušić Vilus, Bogunović and Ćoso 2022), and lexicology. For example, Jelčić Čolakovac and Borucinsky (2023) studied interlingual lexical types such as false cognates, antonomasia, and polysemy of single– and multi–word expressions in English and Croatian. However, in this paper we adopt a novel perspective on English words in Croatian that aims to establish if there is a link between foreign (English) words and language universals. For that purpose, we start with a fundamental concept that Zipf (1936, 1949) observed more than eighty years ago, and that is known as *Zipf's law of abbreviation*. This phenomenon states that more frequent words tend to be shorter, a principle that has been attested empirically across various languages in both written and spoken corpora (e.g. Teahan et al. 2000; Sigurd, Eeg-Olofsson and Van Weijer 2004; Strauss, Grzybek and Altmann 2005; Piantadosi, Tily and Gibson, 2011; Ferrer–i–Cancho and Hernández–Fernández 2013; Kanwal et al. 2017). Hence, Zipf's law is considered to be a universal property of human languages (Bentz and Ferre–i–Cancho 2016).

2. Theoretical considerations

2.1. Zipf's law of abbreviation

Even though human language is highly complex and seemingly random, there have been suggestions that there is an inherent, underlying systematicity in language. One way of testing this hypothesis is through *Zipf's law of abbreviation*, also termed the *Brevity law* (1936, 1949), which states that "the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences" (Zipf 1936: 23). In other words, what this law proposes is that higher–frequency words also tend to be shorter in length. "This law suggests that there is a correlation between the frequency of a word and its length, where more common words are typically shorter than less common ones" (Bentz and Ferre–i–Cancho 2016). Zipf's law is also attested in a number of physical and biological systems (e.g.

Newman 2005; Farmer and Geanakoplos 2006; Frank 2009). Tuđman, Mikelić and Boras (2003) studied the applicability of universal linguistic laws, such as Zipf's law to various languages, including Croatian. Although Zipf's law of abbreviation has been confirmed for various languages, their research on the distribution of types and tokens in the Croatian language shows that lexical density can vary between languages, concluding that Zipf's law is not applicable to Croatian. Thus, there is need for further research to better understand the Law's universality. It would seem that communication efficiency and economy play an important role in language production, as "information can be conveyed as concisely as possible by giving the most frequently used meanings the shortest word forms" (Piantadosi, Tily and Gibson 2011: 3526). This strategy is also known as *Zipf's principle of least effort*.

Word frequencies are one of the most basic properties of human language. A corpus inquiry into most languages will show that there are few very high frequency words that account for most of the tokens in the text. Indeed, a quick inquiry into the *English Web Corpus* (*enTenTen 21*) and the *Croatian Web Corpus* (*MaCoCu Croatian Web v. 2* 2020–2021)[1] shows that the following words are the most frequent ones:

(1) a. enTenTen 21: *the*, *and*, *of*, *two*, *a*, *in*, *is*, etc.
b. MaCoCu Cro 20–21: *i*, *u*, *je*, *na*, *se*, *da*, *za*, *su*, *od*, *a*, etc.

As can be inferred from example (1a), the most frequent words in the English corpus consist of strings that are one to three character long, a pattern also followed by the most frequent words in Croatian (1b). It should, however, be borne in mind that the words shown in (1a, b) all present grammatical or functional words, rather than lexical words. A look into lexical words and their frequencies in the two corpora, particularly nouns, reveals the following:

(2) a. enTenTen 21: *time*, *year*, *people*, *day*, *way*, *world, thing, part*, etc.
b. MaCoCu Cro 20–21: *godina, dan, vrijeme,* čovjek*, rad,* etc.

Example (2a, b) presents the most frequent nouns in the two corpora. Although the strings in (2a, b) are somewhat longer than those in (1a, b), we can still infer that the most frequent nouns in English and Croatian such as *time* (ReF = 1643.5 freq/mill[2]), *year* (ReF = 1493.2 freq/mill), *day* (ReF = 890.7 freq/mill), and *godina* (ReF = 3091.9 freq/mill), *dan* (ReF = 1639.5 freq/mill), and *rad* (ReF = 1147.3 freq/mill) are shorter than the less or least frequent words found in the corpus such as *precarious* (ReF = 2.33 freq/mill) and *commitment* (ReF = 0.32 freq/mill) in *en-*

1 *EnTenTen 21* is an all–purpose English corpus covering the largest possible variety of genres, topics, text types and web sources, counting 61.5 billion tokens. *MaCoCu Cro 20–21* was built by crawling the internet top–level domains in 2021 and 2022, and has 2.9 billion tokens. It presents the biggest and latest corpus of Croatian. The frequency lists were obtained using *Sketch Engine* (Kilgarriff et. al. 2004).

2 Relative frequency (ReF) expressed in number of occurrences per million tokens (freq/mill).

TenTen21, and *žiteljstvo* (ReF = 0.14 freq/mill), *redarstvenih* (ReF = 0.33 freq/mill), *samohodni* (ReF = 0.11 freq/mill), and *izvjestiteljica* (ReF = 1.01 freq/mill) in *MaCoCu Cro 20–21*.

Determining the absolute least frequent words can be challenging due to the noise in the corpora and the dynamic nature of language. Some of the least frequent words are: highly specialised technical terms such as *anoplocephala* and *benzyloxycarbonyl*, compounds such as *authoritarian–learning*, neologisms such as *dry–texting*, all with a relative frequency of less than 0.01 per million.

Furthermore, the frequency of words and categories tends to drop sharply. Zipf's law (1936, 1949) dictates that the amount of evidence we can obtain from the corpus as a representative language sample has a tendency to drop. The second most frequent word has about twice the frequency of the first, the third word has about a third of the frequency of the first, etc. In other words, the most common word or category will occur approximately twice as often as the next common one, etc. In some cases, different sets of empirical data (such as corpora) may deviate somewhat from this law. In this case, such a distribution is termed quasi–Zipfian (in this paper, we use the term Zipfian or Zipf's law to account for such variations). Thus, we can say that the Brevity law is a statistical tendency rather than a strict rule, as there seem to be exceptions to the rule.

Researchers have tried to answer the question as to why words "follow such precise mathematical rule" (Piantadosi 2014), especially given the fact that language production is highly complex and follows various rules of morphology, syntax, semantics, and pragmatics. One of the explanations is the principle of least effort, which can be proven by the fact that, if a word is more frequently used, it is more likely to be abbreviated (e.g. *celeb* for *celebrity*).

Other authors have sought to explain Zipf's law from the semantic point of view by looking into semantic hierarchy (e.g. Fellbaum 1998), whereas others have considered the communicative optimization principle (Mandelbrot 1962, 1966; Ferrer–i–Cancho and Solé 2003; Manin 2009). Piantadosi (2014) provided evidence that near–Zipfian word frequency distribution occurs for novel words in a language production task and made a claim about semantics strongly influencing word frequency. This is also supported by cross–linguistic studies; for example, Calude and Pagel (2011) examined 17 languages from six language families, i.e. the translations of simple, frequent words (the so–called *Swadesh list*), and they found out that word frequencies are "surprisingly robust across languages and predictable from their meaning" (cited in: Piantadosi 2014: 1116). Calude and Pagel (2011) conclude that the meaning of a word is a significant factor in determining a word's frequency. Altmann, Pierrehumbert and Motter (2011) showed that characteristic features of words and the context in which they are used influence the change in word frequency. Furthermore, Meylan and Griffiths (2021: 5) found that "word lengths are more strongly correlated with average information content than with frequency".

The frequency distribution of words is an ongoing debate in statistical linguistics and it is related to one seemingly universal property of human languages. It has implications for language teaching, language acquisition, development of reading skills and cognitive language processing (Ellis 2002).

2.2. English words in Croatian

The main reason for borrowing words from other languages is filling lexical gaps. However, other factors also play an important role in lexical borrowing. For instance, borrowed words may have narrower meaning than native words, or they may convey an additional meaning (Filipović 1986). Also, loanwords might be perceived as more neutral compared to native words, which may reflect traditional, cultural or emotional connotations (e.g. Drljača Margić 2011). Another factor that has been recognised as important is prestige (e.g. Field and Comrie 2002). Research has shown that Croatian students generally have positive attitudes towards English loanwords (Drljača Margić 2012). Another Croatian study found that the frequency of use of such words positively correlates with social desirability (Ćoso and Bogunović 2017).

Today, Croatian is most receptive to borrowing from English (e.g. Drljača Margić 2011). In part, this can be ascribed to the global status of the English language. As a result, exposure to English has been constantly increasing. For many Croatian speakers English has become their second language (L2) (e.g. Bogunović 2023b). Weinreich (1953) noted that bilinguals undergo the process of *interlingual identification*. In other words, they identify units and patterns of one language and map them onto units and patterns of the other language (Hakimov and Backus 2021). Moreover, it seems that bilingual speakers tend to use L2 words and expressions in their first language, especially if their L2 is considered prestigious (e.g. Drljača Margić 2011).

The media is an important factor in introducing new words (e.g. Muhvić–Dimanovski and Skelin Horvat 2008). English loanwords have become common in the language of Croatian media (e.g. Brdar 2010), and they vary from completely adapted (e.g. *klub*) to completely or partially unadapted (e.g. *snowboard*). Still, the use of native words is generally recommended (e.g. Hudeček and Mihaljević 2005; Halonja and Hudeček 2014). This sometimes includes using descriptions and multi–word expressions, adding a new meaning to already existing words, and calques. Research has shown that some of the proposed Croatian solutions for English loanwords are not accepted well by Croatian speakers (e.g. Patekar 2019). Multi–word expressions, in particular, have proven to be complex to use (e.g. Drljača 2006; Bogunović 2023b), which can be exemplified by Croatian translations of *software* as 'programska podrška', and *developer* as 'razvojni inženjer' (Institute of Croatian language and linguistics 2015). These examples illustrate that borrowed words are sometimes more economical compared to native equivalents, which has also been recognised as an important reason for borrowing words (e.g. Drljača Margić 2011).

This study focuses on English loanwords that retain their original, English–specific graphemes (e.g. *e–mail*, *freelancer*, *celebrity*), and they will be referred to as 'English words' (e.g. Ćoso and Bogunović 2017; Borucinsky and Bogunović 2022). Recent computational linguistic resources developed for Croatian offer insight into the use of English words in the Croatian media. One such example is the *Database of English words and their equivalents* (henceforth *EWCE*) (Bogunović, Jelčić Čolakovac and Borucinsky 2022), which is based on the *Database of English words in Croatian* (Bogunović and Kučić 2022).

The *EWCE* database presents results from combining algorithm extraction, corpus linguistics methods and manual evaluation, and it was updated using the corpus processing tool *Sketch Engine* (Kilgarriff et al. 2004) and by consulting the *hrWaC* (Ljubešić and Erjavec 2011; Ljubešić and Klubička 2014) and *ENGRI* (Bogunović et al. 2021) corpora. Corpus results were obtained via corpus query language (CQL) and manual filtering to remove noise resulting from corpus processing, duplicates, proper nouns, false cognates, false pairs, etc. The *EWCE* database contains 2982 unadapted English words and their Croatian equivalents, with absolute and relative frequencies from the *ENGRI* and *hrWaC* corpora.[3] It contains both single–word (e.g. *top*, *fan*, *online*, *e–mail*, *brand*, etc.), and multi–word expressions (e.g. *custom made*, *cost benefit*, *early access*, *duty free*, etc.) with their absolute and relative frequencies from *hrWaC* and *ENGRI* as well as assigned semantic categories.

3. The present study

As English words in Croatian have not yet been explored from this angle, in this paper we examine the link between English words and the language universal that more frequent words are shorter.

The research questions that we aim to answer are:

1. Is there a link between frequency and word length of English words in Croatian corpora?
2. Are English words used more frequently than their Croatian equivalents because they are shorter?

Bogunović (2023a) states that translation equivalents in Croatian generally exist for loanwords or English words that appear 5000 times or more in the *ENGRI* corpus (Bogunović et al. 2021), however, the question remains as to why speakers of Croatian use English words instead of Croatian equivalents in such cases. This paper examines whether the language universal that shorter words are more frequent is systematic when it comes to foreign (primarily English) words in the Croatian language, i.e., whether the most frequent English words are shorter than their Croatian equivalents.

3 The *EWCE* database is freely available on the *FigShare* platform: https://doi.org/10.6084/m9.figshare.20014712.v2.

4. Methodology

For the purpose of this research the *EWCE* database (Bogunović, Jelčić Čolakovac and Borucinsky 2022) was examined. In this paper, our focus is on single, lexical words only. This, however, presents a limitation since functional words that act as 'glue' in a text are generally more frequent than lexical words. Other limitations include homographs (e.g. the English word *net*) and the semantic categories to which the words have been assigned. The semantic categorization is based on the Croatian context in which the English words appear and does not reflect the semantic context in which these words are regularly used in English (Jelčić Čolakovac and Bogunović 2024) or contexts in which they usually appear in English corpora.

To answer the research questions, we looked at frequencies of English words and their Croatian equivalents in the *EWCE* database. Then we identified the number of characters (i.e. string length) of each English word and their Croatian equivalent. For that purpose, we used *Google Colab*[4], as shown in (3).

(3) df['characters_in_word'] = df['Entry'].str.len()
df['characters_in_equivalent'] = df['Croatian equivalent'].str.len()

To understand the relationship between English words, their frequency and length in comparison to the frequency and length of their Croatian equivalents, we divided the English words into groups (e.g. three–string words, four–string words, five–string words, etc.) and looked at frequencies and word length of the equivalents (if they existed) in the respective group. Since we are trying to find out if shorter English words are more frequent than their longer Croatian equivalents, we looked only at equivalents within one group that are longer than the English word (e.g. English word – 3–character string – Croatian equivalent 4+ character string). Then we calculated the average frequency for each group.

5. Results and discussion

This study aimed to investigate the relationship between frequency and length of English words within Croatian corpora. Specifically, we sought to determine whether English words are used more frequently than their Croatian equivalents, which could potentially be attributed to their shorter length.

5.1. Relationship between frequency and word length of English words in Croatian

Table 1 provides the first 20 examples from the dataset, showing English words and their Croatian equivalents, combined relative frequencies from *ENGRI* and

4 The code can be accessed via the following link: https://colab.research.google.com/drive/1i83bMhzoXyfm4OgoMBhA0254tUJliCo9.

hrWaC, number of characters, and semantic categories.[5] The disambiguation of various contexts and assigning of categories is based on the relative frequencies (ReF) of English words in the Croatian corpus (cf. Jelčić Čolakovac and Borucinsky 2023). The semantic categories are described in detail in Jelčić Čolakovac and Bogunović (2024).

English word	ReF (*ENGRI + hrWaC 2.2*)	String length of English word	Croatian equivalent	ReF (*ENGRI + hrWaC 2.2*)	String length of Croatian equivalent	Semantic category
real	132.82	4	*stvaran*	131.24	7	OTHER
web	115.29	3	*mreža*	397.64	5	ICT
show	106.02	4	*emisija*	242.28	7	TV
blog	95.00	4	*mrežni dnevnik*	0.01	14	ICT
post	82.49	4	*objava*	146.87	6	ICT
fan	74.50	3	*obožavatelj*	65.07	11	SPORT
link	64.02	4	*poveznica*	22.52	9	ICT
online	59.96	6	*mrežni*	25.52	6	ICT
e–mail	57.79	6	*e–pošta*	35.14	7	ICT
mail	48.43	4	*pošta*	59.39	5	ICT
net	44.42	3	*mreža*	397.64	5	ICT
jazz	42.66	4	*džez*	0.80	4	MUSIC
summit	41.58	6	*sastanak na vrhu*	4.01	16	BUSINESS
top	29.12	3	*vrhunski*	162.51	8	OTHER
brand	28.60	5	*brend*	63.32	5	COMMERCE
party	25.90	5	*zabava*	122.87	6	PEOPLE
break	25.14	5	*pauza*	49.59	5	ART
topic	24.40	5	*tema*	592.36	4	OTHER
gay	24.08	3	*gej*	8.13	3	PEOPLE
laptop	22.54	6	*prijenosno računalo*	7.85	19	TECH

Table 1. Dataset of English words and Croatian equivalents with ReF, number of characters, and semantic category

5 The complete dataset can be downloaded from: https://osf.io/4kbmc/?view_only=4c0f281ddb1946c68f99986e5576b630.

The most common lengths of English words are 4–, 5–, 6– and 7–character words, which collectively account for 65% of all words in the dataset (see Figure 1). Words with 2– and 3–characters constitute only 4% of all the words in the dataset. Interestingly, 8–character words are found in 11%, while 9–character words make up a notable 9%. Moreover, 11% of words consist of strings longer than 10 characters. The longest English words in the dataset are *state–of–the–art* and *entrepreneurship*, each consisting of 16 characters. The former falls into the category of multi–word expressions, while the latter belongs to a highly specialised register. Other longer words include *crowdsourcing* and *benchmarking*, both of which are compounds. As for the Croatian words, Table 1 reveals that translation equivalents for English words sometimes consist of two or even three words (e.g. *mrežni dnevnik* for 'blog', *sastanak na vrhu* for 'summit', *prijenosno računalo* for 'laptop'), resulting from morpho–syntactic differences between the two languages.

Analysis of string length data for Croatian words in Figure 1 suggests that 2– and 3–character Croatian equivalents are rather uncommon (only 3%), with 4–character words comprising 8% of the dataset and the majority of other words belonging to the category of 5– to 8–character words (53% in total). It appears that 9– and 10–character words are common as well, with 9% and 8% respectively. In addition, 11– to 15–character words constitute 14% of the dataset, followed by 16– to 20–character words (4%) and 21– to 37–character words (1%).

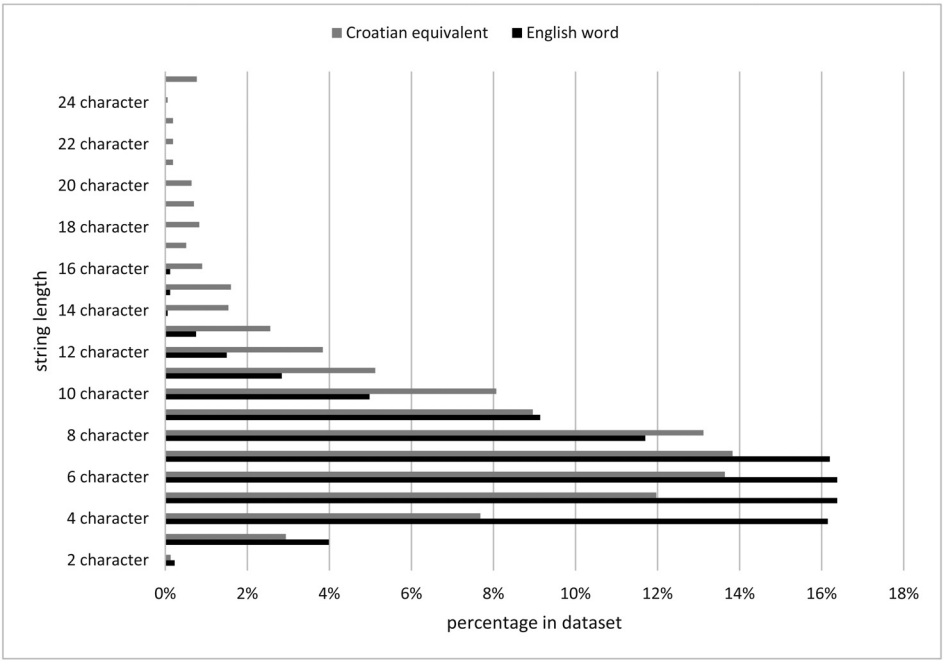

Figure 1. String length of English words and Croatian equivalents and their percentage in the dataset

The analysis confirms that higher–frequency English words in Croatian corpora are typically shorter in length, thereby supporting Zipf's law, which predicts an inverse correlation between word length and frequency. However, as previously noted, this law presents a general tendency rather than an absolute rule, with exceptions such as *state–of–the–art* and *triple(–)double*. Figure 2 shows the relationship between string length and average frequency in English and Croatian.

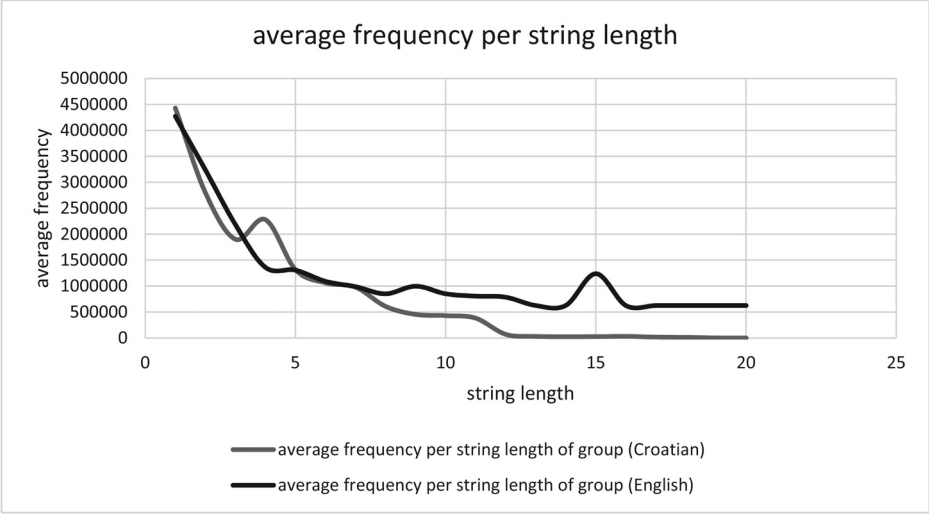

Figure 2. Average frequency per string length for English words and Croatian equivalents

For word lengths of 1 to 5 characters, both languages show a sharp decline in average frequency as string length increases, implying that shorter lexical words are more prevalent in natural language use. For longer strings, the rate of decline becomes less pronounced, indicating that longer words appear less frequently, but still remain consistent. Particularly interesting is the frequency distribution of 6–character words, which becomes more unpredictable. This category shows greater frequency dispersion than other string lengths (e.g. the word *online* has a ReF of 599.59, while the next word in the group, *golden*, has a ReF of 137.77 – a reduction of nearly two thirds). Furthermore, within the same category of 6–character words, frequencies range from below 1 (e.g. *rubber*, *circle*) to around 20 per million tokens (e.g. *cinema*).

Similarly, Croatian equivalents adhere to Zipf's law, with shorter words appearing more frequently than longer words. However, Croatian has more variation in frequency, which can be attributed to its specific morpho–syntax. The type–token ratio is different between English and Croatian, with Croatian having a higher number of types than English (cf. Tuđman, Mikelić and Boras 2003).

5.2. Comparison of English words and their Croatian equivalents

Table 2 provides a comparative analysis of word length and frequency between English words and their Croatian equivalents. It shows the average frequency of English words consisting of 3 characters, 4 characters, 5 characters, etc. and the average frequency of the equivalents in each group consisting of N+ characters in the target language (Croatian). Only equivalents that are longer than their English counterparts have been taken into consideration.

English word string length	Average frequency of English word	Average frequency of equivalent
3	9.34	3617.64
4	5.86	1364.15
5	2.32	1133.03
6	2.46	858.35
7	1.36	552.45
8	1.12	529.97
9	0.82	124.08
10	1.52	352.09
11	0.64	28.82
12	0.77	149.90
13	0.18	2.07

Table 2. Average frequencies of English words and their equivalents per word characters

As can be inferred from Table 2, average frequencies of Croatian equivalents in each group are much higher than the average frequencies of English words. For instance, 3–character English words have an average frequency of 9.34, whereas their Croatian equivalents (which are longer than 3 characters) have an average frequency of 3617.74. Even for longer strings (e.g. 9–character), Croatian equivalents tend to exhibit higher frequencies than the English words. These findings suggest that, while English words may be shorter, their Croatian equivalents are more frequent in the overall corpus. Consequently, the hypothesis that English words are used more frequently than their Croatian equivalents due to the fact that they are shorter, is not directly supported. Hence, to prove the hypothesis, other criteria are needed, such as semantic category, context and informativity. Especially problematic are words with the same equivalent in Croatian (like *net* and *web* 'mreža'), which accounts for the equivalent's high frequency.

English word	ReF (*ENGRI + hrWaC 2.2*)	String length of English word	Croatian equivalent	ReF (*ENGRI + hrWaC* 2.2)	String length of Croatian equivalent
abs	0.4559	3	*trbušnjaci*	3.8733	10
fax	11.1066	3	*faks*	32.9229	4
hot	4.9338	3	*vruć*	62.2608	4
oil	2.4489	3	*ulje*	203.207	4
old	12.6553	3	*star*	907.08	4
sex	14.741	3	*seks*	122.425	4
sis	2.8473	3	*seka*	8.9785	4
aid	1.3059	3	*pomoć*	665.732	5
big	19.8455	3	*velik*	4133.35	5
fly	1.2499	3	*letan*	0.8526	5
job	3.1326	3	*posao*	1306.56	5
log	1.4303	3	*zapis*	45.1543	5
net	**44.4231**	**3**	***mreža***	**397.642**	**5**
one	5.4871	3	*jedan*	4937.04	5
raw	1.3441	3	*sirov*	30.0573	5
sun	4.9259	3	*sunce*	163.154	5
tax	0.475	3	*porez*	194.543	5
try	0.1785	3	*proba*	29.92	5
web	**115.297**	**3**	***mreža***	**397.642**	**5**
add	0.5008	3	*dodati*	696.995	6
bet	1.3498	3	*oklada*	5.2527	6
bid	0.3008	3	*ponuda*	357.459	6
top	29.1238	3	*vrhunski*	162.509	8

Table 3. Representation of 3–character English words and their Croatian equivalents with their respective ReF

5.3. The importance of semantic categories and context

Semantic categorization is an important factor for frequency distribution. As previously mentioned, English words such as *net* and *web* share the same Croatian equivalent, which inflates the frequency of the Croatian equivalent and introduces bias into the results. In addition, certain semantic categories (e.g. ICT, Business) contain a higher number of borrowed English words, which, in turn, affects their frequency patterns in Croatian corpora. Furthermore, Croatian equivalents are of-

ten multi–word expressions, which influences string length, but does not necessarily impact usage frequency in the same way as single word equivalents.

6. Conclusion

The phenomenon of English words across world languages has been extensively studied from various research perspectives. In this paper we examined the phenomenon from the perspective of a specific language universal based on Zipf's law (1936, 1949) which states that shorter words have a tendency to appear more frequently in use. For this purpose, we used an existing database of English words in Croatian (*Database of English words and their equivalents in Croatian*, Bogunović, Jelčić Čolakovac and Borucinsky 2022) to ascertain whether the universal, which posits the length of a word is in an inverse relation to its frequency, can also be applied to English words and their Croatian equivalents. We divided the single English words according to the number of characters in each word (3–, 4–, 5–, etc.) and looked only at equivalents that are longer than the English word. In line with Zipf's law, our results revealed more frequent English words also tend to be shorter or have a shorter string length, with exceptions such as hyphenated English compounds (e.g. *state–of–the–art*). The frequency distribution of Croatian equivalents also lent support to Zipf's law; no shorter equivalents were registered for 2–character words and the average frequency for a particular group of English words was much lower than the average frequency of the equivalent. The data collected in the analysis show the need for further investigation into the phenomenon of English words in Croatian, but with the inclusion of the semantic component as a relevant factor (as was evident with the words *web* and *net* from our analysis). The data can, nevertheless, contribute to further discussions on language universals in the context of language borrowing, with future studies focusing on potential interactions between factors such as language economy, language prestige and word frequency.

Two major limitations were encountered in the implementation of this study. First, it was challenging to apply universal linguistic laws to Croatian, which is one of the languages with its own specificities. Secondly, the choice of word length (in characters) as the primary variable might have oversimplified the complexity of borrowing motivations.

Future work will include an analysis of homographs and multi–word expressions, as well as grouping of words into semantic categories to explore whether frequency–length patterns differ across domains. Additionally, as borrowing patterns differ across registers (e.g. formal vs. informal), future work should also investigate how academic texts differ from social media content in terms of word length and frequency.

Funding

The study outlined in this paper has been supported in part by the Croatian Science Foundation (HRZZ) under project number UIP–2019–04–1576.

References

Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter (2011). Niche as a determinant of word fate in online groups. *PLoS ONE* 6(5): e19009, http://doi.org/10.1371/journal.pone.0019009

Bentz, Christian, and Ramon Ferrer–i–Cancho (2016). 'Zipf's law of abbreviation as a language universal'. Bentz, Christian, Gerhard Jäger, and Igor Yanovich, eds. *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tubingen: University of Tübingen, 1–4, https://doi.org/10.15496/publikation–10057

Bogunović, Irena (2023a). A corpus–based approach to English loanwords: Introducing the Database of English loanwords in Croatian. *Fluminensia:* časopis *za filološka istraživanja* 35(2): 437–460, http://doi.org/10.31820/f.35.2.1

Bogunović, Irena (2023b). Engleske riječi u hrvatskome: Jezično posuđivanje i dvojezična leksička obrada. *Suvremena lingvistika* 49(96): 251–280, http://doi.org/10.22210/suvlin.2023.096.04

Bogunović, Irena, and Mario Kučić (2022). The database of English words in Croatian.xlsx. Figshare, https://doi.org/10.6084/m9.figshare.20014364.v1

Bogunović, Irena, Jasmina Jelčić Čolakovac, and Mirjana Borucinsky (2022). The database of English words and their Croatian equivalents. Figshare. Dataset, https://doi.org/10.6084/m9.figshare.20014712.v2, accessed on June 10, 2024

Bogunović, Irena, Mario Kučić, Nikola Ljubešić, and Tomaž Erjavec (2021). Corpus of Croatian news portals *ENGRI* (2014–2018), Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1416

Borucinsky, Mirjana, and Irena Bogunović (2022). Crpljenje engleskih riječi iz korpusa hrvatskoga jezika. *Fluminensia:* časopis *za filološka istraživanja* 34(2): 435–461, https://doi.org/10.31820/f.34.2.13

Brdar, Irena (2010). Engleske riječi u jeziku hrvatskih medija. *LAHOR:* časopis *za hrvatski kao materinski, drugi i strani jezik* 2(10): 217–232

Calude, Andreea S., and Mark Pagel (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366(1567): 1101–1107, https://doi.org/10.1098/rstb.2010.0315

Ćoso, Bojana, and Irena Bogunović (2017). Person perception and language: A case of English words in Croatian. *Language and Communication* 53: 25– 34, https://doi.org/10.1016/j.langcom.2016.11.001

Drljača Margić, Branka (2012). Croatian university students' perception of stylistic and domain–based differences between Anglicisms and their native equivalents. Dąbrowska, Marta, Justyna Leśniewska, and Beata Piątek, eds. *Languages, Literatures and Cultures*

in Contact: English and American Studies in the Age of Global Communication, Volume 2: Language and Culture. Krakow: Tertium, 109–126

Drljača Margić, Branka (2011). Leksički paralelizam: je li opravdano govoriti o nepotrebnim posuđenicama (engleskoga podrijetla)?. *Fluminensia:* časopis *za filološka istraživanja :* časopis *za filološka istraživanja* 23(1): 53–66

Drljača, Branka (2006). Anglizmi u ekonomskome nazivlju hrvatskoga jezika i standardnojezična norma. *Fluminensia:* časopis *za filološka istraživanja* 18: 65–85, https://hrcak.srce.hr/6224

Ellis, Nick C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition* 24(2): 143–188

Farmer, J. Doyner, and John Geanakoplos (2006). Power laws in economics and elsewhere. *Santa Fe Institute Tech Report*

Fellbaum, Christiane (1998). A semantic network of English: The mother of all WordNets. *Computers and the Humanities* 32: 209–220

Ferrer–i–Cancho, Ramon, and Antoni Hernández–Fernández (2013). The failure of the Law of Brevity in two new world primates. Statistical Caveats. *Glottotheory* 4(1): 45–55

Ferrer–i–Cancho, Ramon, and Ricard V. Solé (2003). Optimization in complex networks. Pastor–Satorras, Romualdo, Miguel Rubi, and Albert Diaz–Guilera, eds. *Statistical Mechanics of Complex Networks. Lecture Notes in Physics*, vol 625. Berlin, Heidelberg: Springer, https://doi.org/10.1007/978–3–540–44943–0_7

Field, Frederic, and Bernard Comrie, (2002). *Linguistic borrowing in bilingual contexts*. Amsterdam: John Benjamins Publishing Company, https://doi.org/10.1075/slcs.62

Filipović, Rudolf (1986). *The beginnings of lexicography in Croatia.* Hartmann, Reinhard Rudolf Karl, eds. The History of Lexicography. Amsterdam: John Benjamins Publishing Company, 65–73, https://doi.org/10.1075/sihols.40.08fil

Frank, Steven A. (2009). The common patterns of nature. *Journal of Evolutionary Biology* 22(8): 1563–1585

Hakimov, Nikolay, and Ad Backus (2021). Usage–based contact linguistics: Effects of frequency and similarity in language contact. *Journal of Language Contact* 13(3): 459–481

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

Hudeček, Lana, and Milica Mihaljević (2005). Nacrt za višerazinsku kontrastivnu englesko–hrvatsku analizu. *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 31: 107–151, https://hrcak.srce.hr/9381

Institute of Croatian language and linguistics (2015). Bolje je hrvatski! http://bolje.hr /, accessed on June 30, 2024

Ivir, Vladimir (1998). Linguistic and communicative constraints on borrowing and literal translation. Beylard–Ozeroff, Ann, Jana Králová, Barbara Moser–Mercer, eds. *Translators' Strategies and Creativity*. Amsterdam – Philadelphia: John Benjamins, 137–144

Jelčić Čolakovac, Jasmina, and Irena Bogunović (2024). Putting languages into perspective: A comprehensive database of English words and their Croatian equivalents. *Crossroads. A Journal of English Studies* 45: 62–81, https://doi.org/10.15290/CR.2024.45.2.04

Jelčić Čolakovac, Jasmina, and Mirjana Borucinsky (2023). In the melting pot of web–crawled texts: The challenges of extracting English words and phrases from Croatian corpora. *International Journal of Applied Linguistics* 34(1): 166–182, https://doi.org/10.1111/ijal.12485

Kanwal, Jasmeen, Jenny Smith, Jennifer Culbertson, and Simon Kirby (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition* 165: 45–52

Kilgarriff, Adam, Pavel Rychlý, Pavel, Smrž, and David Tugwell (2004). Itri–04–08 The Sketch Engine. Williams, Geoffrey, and Sandra Vessier, eds. *EURALEX Proceedings*. Université de Bretagne–Sud, Faculté des lettres et des sciences humaines: Euralex – European Association for Lexicography, 105–115

Kučić, Mario (2021). Creating a web corpus using GO. *Proceedings of the 44th International Convention MIPRO 2021*. Croatian Society for Information, Communication and Electronic Technology, 1931-1933, https://doi.org/10.23919/MIPRO52101.2021.9597093

Ljubešić, Nikola, and Tomaž Erjavec (2011). HrWaC and slWaC: Compiling web corpora for Croatian and Slovene. Habernal, Ivan, ed. *Text, speech and dialogue, lecture notes in computer science*. Berlin, Heidelberg: Springer, 395–402

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

Mandelbrot, Benoit (1962). On the theory of word frequencies and on related Markovian models of discourse. Jakobson, Roman, ed. *Structure of Language and its Mathematical Aspects*. Rhode Island: American Mathematical Society, 190–219

Mandelbrot, Benoit (1966). Information theory and psycholinguistics: A theory of word frequencies. Lazarsfield, Paul Felix, and Neil W. Henry, eds. *Readings in Mathematical Social Sciences*. Cambridge, MA: M.I.T Press, 151 –168

Manin, D. Yu (2009). Mandelbrot's model for Zipf's law: Can Mandelbrot's model explain Zipf's law for language?. *Journal of Quantitative Linguistics* 16(3): 274–285

Meylan, Stephan C., and Thomas L. Griffiths (2021). The challenges of large-scale, web–based language datasets: Word length and predictability revisited. *Cognitive Science* 45(6): e12983, https://doi.org/10.1111/cogs.12983

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

Muhvić–Dimanovski, Vesna, and Anita Skelin Horvat (2006). O riječima stranoga podrijetla i njihovu nazivlju. *Filologija* 44–47: 203–215, https://hrcak.srce.hr/22242

Newman, Mark E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics* 46(5): 323–351

Patekar, Jakob (2019). Prihvatljivost prevedenica kao zamjena za anglizme. *Fluminensia:* časopis *za filološka istraživanja* 31(2): 143–179, https://doi.org/10.31820/f.31.2.17

Pavlinušić Vilus, Eva, Irena Bogunović, and Bojana Ćoso (2022). Students' strategies for translating most frequent English loanwords in Croatian. *Rasprave:* Časopis *Instituta za hrvatski jezik i jezikoslovlje* 48 (2): 547–570, https://doi.org/10.31724/rihjj.48.2.7

Piantadosi, Steven T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin and review* 21: 1112–1130

Piantadosi, Steven T., Harry Tily, Harry, and Edward Gibson (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9): 3526–3529, http://dx.doi.org/10.1073/pnas.1012551108

Pritchard, Boris (1997). On Anglicisms in Maritime Croatian. *Studia Romanica et Anglica Zagrabiensia : Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb* 42: 321–338

Sigurd, Bengt, Mats Eeg-Olofsson, and Joost Van Weijer (2004). Word length, sentence length and frequency–Zipf revisited. *Studia linguistica* 58(1): 37–52

Strauss, Udo, Peter Grzybek, and Gabriel Altmann (2005). Word length and word frequency. Grzybek, Peter, ed. *Contributions to the Science of Text and Language*. Dordrecht: Springer, 277–294

Teahan, William John, Yingying Wen, Rodger McNab, and Ian H. Witten (2000). A compression–based algorithm for Chinese word segmentation. *Computational Linguistics* 26(3): 375–393

Tuđman, Miroslav, Mikelić, Nives, and Damir Boras (2003). Vocabulary size prediction of Croatian texts. *Proceedings of the 25th International Conference Information Technology Interfaces ITI*, 223–228, https:doi.org/10.1109/ITI.2003.1225349

Weinreich, Uriel (1953). *Languages in contact*. New York: Linguistic Circle of New York

Zipf, Georg (1936). *The Psychobiology of Language*. London: Routledge

Zipf, Georg (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison–Wesley

Primjena Zipfova zakona na engleske riječi u hrvatskom jeziku: usporedba čestote i dužine riječi

Engleske riječi u hrvatskom jeziku proučavalo se iz različitih perspektiva, ponajviše sa stajališta kontaktne lingvistike, kontrastivne lingvistike, translatologije i leksikologije. No, u ovom radu engleske riječi promatramo iz nove perspektive, točnije pokušava se istražiti postoji li veza između čestote i dužine engleskih riječi u hrvatskom jeziku te koriste li se engleske riječi češće od svojih prijevodnih istovrijednica u hrvatskome upravo zato jer su kraće. U tu svrhu testirana je hipoteza da su češće riječi i kraće, a na kojoj se temelji Zipfov zakon (1936., 1949.). U nastojanju da se utvrdi imaju li riječi s većom čestotom tendenciju da budu kraće, provedeno je korpusno ispitivanje mrežnog korpusa engleskog jezika (*enTenTen 21*) i mrežnog korpusa hrvatskog jezika (*MaCoCu Croatian Web* v. 2 (2020.–2021.). Analiza je pokazala da se najčešće riječi u engleskom korpusu sastoje od jednog do tri znaka, a da najučestalije riječi u hrvatskom također slijede taj obrazac. Kako bismo vidjeli vrijedi li isto i za engleske riječi u hrvatskom, analizirali smo čestote iz *Baze engleskih riječi i njihovih ekvivalenata u hrvatskom* (Bogunović, Jelčić Čolakovac and Borucinsky, 2022). U sljedećem koraku ustanovili smo broj znakova (tj. duljinu niza) svake engleske riječi i svake hrvatske istovrijednice, podijelili engleske riječi u skupine (npr. riječ s tri niza, riječ s četiri niza, itd.) te zatim analizirali duljinu prijevodnih istovrijednica (ukoliko iste postoje u *Bazi*) za svaku pojedinu skupinu engleskih riječi. Rezultati su pokazali da je najčešća duljina niza engleskih riječi 4, 5, 6 i 7 znakova, što čini 65% svih riječi u skupu podataka. Što se tiče hrvatskih riječi, prijevodne istovrijednice se ponekad sastoje od dvije ili čak tri riječi (npr. 'mrežni dnevnik' za *blog*), što je rezultat morfosintaktičkih razlika između dvaju

jezika. Podaci o duljini niza za hrvatske riječi sugeriraju da su hrvatske istovrijednice od 2 i 3 znaka prilično rijetke (samo 3%), dok riječi s 4 znaka čine 8% podataka, a većina ostalih riječi pripada kategoriji od 5 do 8 znakova (ukupno 53%). Nadalje, analiza prosječnih čestota istovrijednica u svakoj skupini ukazala je da su one znatno veće od prosječnih čestota engleskih riječi. S obzirom da je to očekivano u korpusu hrvatskih tekstova, smatramo da su za dokaz hipoteze potrebni i drugi kriteriji, poput semantičke kategorije i informativnosti. Zaključujemo da podaci prikupljeni analizom ukazuju na potrebu daljnjeg istraživanja fenomena engleskih riječi u hrvatskom jeziku, ali uz uključivanje semantičke komponente kao utjecajnog čimbenika.

Ključne riječi: Zipf's law, frequency, language universals, loanwords, English words, Croatian
Key words: Zipfov zakon, čestota, jezične univerzalije, posuđenice, engleske riječi, hrvatski jezik