Slovenski jezik / Slovene Linguistic Studies 17 (2025) | https://doi.org/10.3986/17.1.04 | COBISS: 1.01 | CC BY SA 4.0

59–91

Matic Pavlič
University of Ljubljana, Faculty of Education, Slovenia
matic.pavlic@pef.uni-lj.si | https://orcid.org/0000-0001-8248-8860

Andrej Perdih
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Slovenia
andrej.perdih@zrc-sazu.si | https://orcid.org/0000-0002-2248-9666

Lexical processing of morphologically complex Slovene words in a lexical decision task: the role of pseudowords

This study investigates how the structure of pseudowords influences the lexical decision accuracy and response time in morphologically complex Slovene words. We tested three methods of constructing pseudowords: manual modification with preserved suffixes, manual modification with altered suffixes, and algorithmic generation with preserved suffixes, using the Wuggy application (Keuleers and Brysbaert 2010) adapted to Slovene. The pseudoword types did not differ significantly by accuracy, only by response time. However, these differences did not affect the accuracy rate or response time of existing Slovene words, suggesting that morphological complexity and manual/algorithmic construction of pseudowords are not relevant factors in lexical decision.

KEYWORDS: pseudowords, lexical decision task, lexicography, Slovene, psycholinguistics

Ta študija raziskuje, kako struktura psevdobesed vpliva na pravilnost leksikalne odločitve in reakcijski čas pri slovenskih izpeljankah. Preizkusili smo tri načine tvorjenja: ročno spreminjanje z ohranjenimi priponami, ročno spreminjanje s spremenjenimi priponami in algoritemsko tvorjenje z ohranjenimi priponami; za slednje smo prilagodili in uporabili aplikacijo Wuggy (Keuleers and Brysbaert 2010). Posamezne vrste psevdobesed so se razlikovale v odzivnih časih, ne pa tudi v pravilnosti leksikalne odločitve. Vendar te razlike niso vplivale na pravilnost leksikalne odločitve in odzivni čas pri obstoječih slovenskih besedah, zato lahko rečemo, da morfološka kompleksnost oziroma ročno/algoritemsko tvorjenje prevdobesed nista ključni za leksikalno odločanje.

KLJUČNE BESEDE: psevdobesede, leksikalno odločanje, leksikografija, slovenščina, psiholingvistika

1 Introduction

Linguistics focuses primarily on the study of grammatical linguistic expressions, but native speakers may also encounter and judge expressions that are not grammatical either because of their illicit form or meaning. Native speakers judge ungrammatical expressions based on their linguistic intuition by comparing the uttered expression with expressions constructed according to the rules of their internal (i.e. mental) grammar. These judgments often prove helpful in exploring mental grammar. Moreover, such expressions are necessary in many linguistic procedures or psycholinguistic experiments to balance the task. For example, depending on the estimated grammaticality of the target construction in the traditional grammaticality judgment task, researchers try to balance the experimental responses by inserting filler expressions with different grammaticality. Similarly, in many psycholinguistic procedures, experiments must include not only grammatical stimuli, but also an approximately equal number of meaningless filler stimuli or filler stimuli with degraded grammaticality. Finally, the entries in a mental lexicon are organized according to linguistic features (meaning and form) and non-linguistic features (imaginability, familiarity, age of acquisition and frequency, etc.), so that entries that are similar in one or more of these properties are likely to be activated together or processed in a similar way. Since these features are often confounding factors in psycholinguistic experiments, researchers try to avoid them by using meaningless expressions that allow better control over the morphological, semantic and syntactic properties of the stimuli. Thus, a particular advantage of using meaningless expressions in linguistic research is that they facilitate control over a variety of other potentially interfering features of linguistic expressions that are difficult to manipulate, manage and account for. Especially in lexical decision tasks and many other research methods, it is not sufficient to simply use random meaningless sequences of sounds or letters, because it is important that the human brain does not reject these stimuli out of hand for non-linguistic reasons, but processes them as if they were part of a language. In such cases, the stimuli must be constructed in such a way that the corresponding linguistic rules are observed.

This article reports on the compilation of a list of meaningless (i.e. non-existent) but phonologically and phonotactically grammatical sound or grapheme sequences (i.e. pseudowords) for Slovene. As for the structure of pseudowords in the psycholinguistic literature for Slovene, they were investigated in the three studies listed below using grammaticality judgment and lexical decision tasks. In a pilot study, Marjanovič et al. (2013) investigated how Slovene native speakers (N = 20, mean age = 27.3 years) perceive pseudowords that either conform to or violate the Slovene agentive word formation rules in agentive nouns derived from verbs. The stimulus

set comprised pseudowords with various morphological violations, well-formed pseudowords, non-words and existing Slovene words. The participants performed an offline grammaticality judgement task. The study showed that while speakers clearly distinguished between legal and illegal forms, they did not distinguish between types of violations – they rejected all malformed pseudowords equally.

Manouilidou et al. (2016) conducted an offline grammaticality judgment task and an online lexical decision task based on the above-mentioned pilot study for Slovene (Marjanovič et al. 2013). Three groups of stimuli that violate certain word formation constraints in Slovene were used: well-formed pseudowords, existing words and non-words, all formed with a masculine nominal suffix *-ec*. 21 healthy subjects (mean age = 67.8) and 23 subjects with mild cognitive impairment (mean age = 68.6) took part in the study. Compared to the former, the latter were slower only in online lexical decision, which shows that the time pressure plays an important role: offline tasks mask some effect that online tasks reveal.

Finally, Pavlič et al. (2022) examined whether knowledge of Italian as a second language influences how Slovene speakers process non-existent words (pseudowords and non-words) in their native language, Slovene. Specifically, they showed that the Italian helps Slovenians to better distinguish between Slovene and non-Slovene phonology when they hear stimuli, especially those containing phonemes that exist in Italian but not in Slovene.

While pseudowords have already been used in psycholinguistic experiments for Slovene to investigate their morphological (Marjanovič et al. 2013, Manouilidou et al. 2016) and phonological structure (Pavlič et al. 2022), our list of pseudowords was created to obtain word prevalence data for a large part of the Slovene vocabulary (80,000 words) in a megastudy. Similar megastudies based on lexical decision experiments have already been conducted for some European languages, including Dutch (Brysbaert et al. 2016), English (Brysbaert et al. 2019), Spanish (Aguasvivas et al. 2018) and Catalan (Guasch et al. 2022). In these mega-studies, the authors used computerized algorithms to create lists of pseudowords containing several thousand elements, but without observing the internal morphological structure of the words that served as models for their pseudowords. This decision might be problematic since, in English, for example, recognizing a suffix *-er* might contribute to a higher likelihood that participants would list both the word *reporter* and the pseudoword *tiporter* as English words, compared to the morphologically simple word *report* and the pseudoword *tiport*. Consequently, the retention of word formation suffixes in pseudoword generation could contribute to the list being more word-like.

The issue of retention of internal morphological structure has been raised in Slavic languages, especially in the context of megastudies: Polish researchers (Imbir et al. 2015;

Dołżycka et al. 2022) have recently compared different aspects of pseudoword generation. They were interested in the effect of manual versus computerized algorithmic generation of pseudowords and in the effect of word class on the rating of pseudowords. They prepared two lists of stimuli and had two groups of participants rate them. Note that this was not a lexical decision task, but a grammaticality judgment study in which pseudowords were to be rated (without mixing them with existing words) on a four-point Likert scale ("*Estimate the probability that X can be a Polish word*"). They also compared pseudowords where the word ending was retained with those where the word ending was not. However, they retained the last syllable, which does not match the ending or suffix in morphologically complex words. In the Slovene words *oškodovanec*-∅ 'victim' and *pravnik*-∅ 'lawyer', for example, the ending is zero, the last syllable is *nec* and *nik*, and the suffix is -*ec* and -*nik* respectively. In the Slovene words *govorica* 'rumor' and *knjigarna* 'bookstore', on the other hand, the ending is -*a*, the last syllable is *ca* and *na*, and the suffix is -*ica* and -*arna*, respectively. Their study was therefore limited by a possible bias (only pseudowords were assessed), the direct involvement of metalinguistic capacity (grammaticality judgments rather than lexical decisions), and the retention of the last syllable, which did not consistently correspond to a derivational suffix or inflectional ending.

The aim of our study is to investigate how different types of pseudowords affect participants' performance in a lexical decision task, focusing on both response time and accuracy. Manually constructed pseudowords with retained suffixes will be compared with algorithmically generated ones to test whether human-generated forms are processed differently from machine-generated forms. The study will also investigate whether retaining or altering word formation suffixes in manually constructed pseudowords affect processing and whether pseudowords with retained suffixes appear more word-like, making it more difficult to distinguish them from existing words. These comparisons will not only form the basis for deciding how to construct the 10,000 or so pseudowords needed for a new Slovene prevalence megastudy but will also shed light on the role of morphological cues in word recognition by investigating whether retaining or altering suffixes has an impact on how strongly a pseudoword activates lexical representations. If certain suffixes make pseudowords appear more word-like, this suggests that morphological information plays a crucial role in shaping lexical access and controls decision-making processes in distinguishing existing words from non-existing words. In this way, the study not only tests the relative effects of different methods of constructing pseudowords but also contributes to a broader understanding of how morphology interacts with lexical processing mechanisms.

The following section 2 first presents the methods commonly used to construct and evaluate pseudowords. Section 3 describes the methodology used in the study

and continues with the analysis and results of Experiments 1 and 2 in Section 4. The last section 5 discusses the results.

2 Constructing pseudowords

A lexical decision task is a psycholinguistic procedure that was first described by Meyer and Schvaneveldt (1971). A participant is presented with sequences of sounds or graphemes and asked to judge whether they represent a word in the target language. In a digitally designed experiment, the participant responds by pressing a key or clicking or tapping a button. If the participants find the sequence in their mental dictionary, they select "yes", otherwise they select "no". Search for the sequence is influenced by several factors (Field 2004), including phonological form (Levelt, Roelofs and Meyer 1999; Marslen-Wilson 1987), syntactic category (Jackendoff 2002; Pulvermüller 1999), semantic features (Collins and Quillian 1969; McRae et al. 2005), frequency of use (Oldfield and Wingfield 1965; Jescheniak and Levelt 1994), lexical neighborhood density (Luce and Pisoni 1998), age of acquisition (Morrison and Ellis 1995) and, finally, morphological structure: words are linked by common morphemes (e.g., »teach«, »teacher«, »teaching«) and morphologically complex words are often decomposed during lexical access (Taft and Forster 1975; Marslen-Wilson et al. 1994). When words are presented in a sequence, the linguistic context also plays a role. The context for a particular sequence in a lexical decision task is provided by all the stimuli in the experiment. To avoid response bias, the test must also contain stimuli where the expected answer is "no". Consequently, in addition to the meaningful sequences that represent existing words, it is extremely important how non-existing sequences are structured (Longtin and Meunier 2005).

There are two kinds of non-existent sequences. If they are constructed in such a way that they violate the phonology or phonotactics of the target language, the participants in the experiment do not have to search their mental lexicon but can easily answer based on the violated rule. For example, there are no words in Slovene with the onset *#ng* (1a) and practically none with the cluster *th* (1b). When participants come across the onset *#ng* or the cluster *th*, they know immediately (i.e. without accessing their mental dictionary) that this is not a Slovene word. These examples of non-existent sequences are called non-words and are distinguished from pseudowords, i.e. non-existent sequences that are formed according to the rules of the target language, such as Slovene in (2a) or (2b).

(1a) *ngapa* (1b) *patha*
(2a) *gapa* (2b) *pata*

Pseudowords should be processed as if they were existing words. Therefore, their construction must be based on the phonological rules of a language, in particular its phonemic inventory, phonotactics and syllable structure. In their overview, König et al. (2019) list three basic methods for constructing pseudowords:

Method	Example Input	Example Output	Key Operation
Manipulation of Existing Words	(3a) table	(3b) tabla	**Substitution of one letter (e→a)**
	(3c) fear	(3d) faar	
Concatenation of Grapheme Units	(4a) str, amp, ing	(4b) stramping	**Combining high-freq trigrams**
		(4c) ingstramp	
Sub-Syllabic Manipulation	(5a) fear	(5b) fer (5c) feaer	**Changing the nucleus vowel**

TABLE 1: Basic methods for constructing pseudowords by König et al. (2019)

However, each of these methods has its own limitations, especially when they are algorithmically generated: Manipulation of word stimuli requires an understanding of permissible changes from the source word (3a/c) to maintain phonological and morphological plausibility (3b), otherwise phototactically illicit combinations may result (3d). High-frequency grapheme sequences (4a) must follow phonotactic constraints (4b), otherwise phototactically illicit combinations may result (4c). And subsyllabic modifications require knowledge of the syllable structure and the transitions between syllables (5b), as otherwise phototactically illicit combinations may result (5c).

Similarity to existing words is the most important assessment point and an important aspect of all methods. Pseudowords that are more like existing words lead to shorter response times (Dorffner and Harris 1997). If a pseudoword is too like an existing word, participants may even associate the two words with each other, leading to a priming bias (New et al. 2023), whereas a pseudoword that is too dissimilar may be processed as a non-word. Research by Barca and Pezullo (2012) has shown that existing words are unambiguously recognized, while pseudowords are ambiguous but eventually classified as non-lexical stimuli. Similarity can be measured using Levenshtein distance, i.e. by counting the minimum number of individual steps required to turn one word into another (insertions, deletions or substitutions). For example, the distance between the English words *cat* and *bat* is one because only one substitution is required, and the distance between *cat* and *cart* is also one because only one addition is required. The Levenshtein distance is now often extended to the orthographic Levenshtein distance 20 (OLD20), which determines the average Levenshtein distance to the twenty most similar words in a reference list (Yarkoni, Balota and Yap 2008). To calculate the OLD20 for a pseudoword, the algorithm first identifies the twenty

most similar words from a reference list based on their Levenshtein distance. The final OLD20 score is then determined by averaging these distances. A lower OLD20 score indicates greater similarity to existing words. A higher OLD20 score, on the other hand, indicates that the pseudoword is more different from existing words, so that it appears less word-like. The OLD20 score thus ensures that researchers select pseudowords with a comparable degree of similarity to existing words, making them useful for experimental comparisons.

The methods for generating pseudowords have become increasingly sophisticated over time. An early method of generating pseudowords was letter substitution, in which letters in existing source words were replaced to form pseudowords (Brown et al. 1987). This approach has been widely used in linguistic and psychological studies (Imbir et al. 2015 and in language-specific databases, including English (Balota et al. 2007), French (Ferrand et al. 2010) and Dutch (Keuleers and Brysbaert 2010).[1] Slowly, this method evolved into more advanced computerized algorithmic methods based on language-specific lexicons that can be used to control various properties of pseudowords, e.g. MCWord (Medler and Binder 2005), WordGen (Duyck et al. 2004), WordCreator (Trost 2002) and Wuggy (Keuleers and Brysbaert 2010). MCWord supports English only, WordGen supports English, French and Dutch and Wuggy supports Basque, Bulgarian, Dutch, English, French, German, Polish, Spanish, Turkish and now also Slovene. These tools combine sub-word units, usually syllables, to generate pseudowords that reflect the frequency distribution of letter sequences of different lengths (n-grams) in natural language (Suen 1979; Solso et al. 1979).

The Wuggy algorithm, for example, breaks down an existing word from the input source into its sub-syllabic components (onset, nucleus and coda) and systematically recombines these elements to generate new but linguistically plausible pseudowords. By preserving syllable structure and controlling segment length and transitions between letters, Wuggy generates pseudowords that are very similar to existing words in both orthographic and phonological form. In addition, Wuggy offers some customization options that allow researchers to adapt the results to specific linguistic or experimental requirements. For this reason, we decided to adapt Wuggy to Slovene and use it in our study.

[1] Judging by the limited description on the website https://aljaxus.gitpage.si/generator-nebesed/#/, this method was also used by M. Ozbič and A. Starc to create a pseudoword generator for Slovene.

3 Methodology

This section presents the materials, procedures, and participants of our study, which consisted of two experiments based on a lexical decision task. Keep in mind that the experiments were planned as a preparatory study for a mega-study in which prevalence data for a large part of the Slovene vocabulary (80,000 words) were to be collected. Due to the large number of participants needed to collect responses to so many stimuli, mega-studies such as Brysbaert et al. (2016) and Guasch et al. (2022) are conducted exclusively online and without the presence of the experimenter, which inevitably leads to a loss of experimental control: Researchers cannot monitor the participants' hardware, software or environment, which is especially problematic for time-critical tasks to record response times. The online lexical decision task is therefore also susceptible to latency and fluctuations between experimental setups. This is due to participant-side issues (distractions, multitasking and different motivations) that can affect data quality, as well as technical issues (e.g. slow browsers, intermittent connections) that can cause noise.

To test the extent to which these pitfalls can affect data quality, Ratcliff and Hendrickson (2021) repeated the lexical decision experiment of White et al. (2010) with subjects recruited from Amazon Mechanical Turk to directly compare the procedures. Overall, the results of these two experiments and four tasks show that the accuracy and response times from Amazon Mechanical Turk subjects replicate the results of experiments that provided carefully controlled in-person data collection. However, the results also revealed serious problems with the data from subjects where there were large differences in response time distributions between experimental runs. In many cases, these could be attributed to rapid guessing. With an aim, similar to Ratcliff and Hendrickson (2021), Angele et al. (2023) tested whether masked priming effects can be captured both qualitatively and quantitatively using either lab- or browser-based experimental software. The results of their online-based experiments replicated results previously established in the laboratory-based studies, suggesting that masked priming can reliably capture timed behavior across a variety of devices.

Note that online designs allow for faster and more extensive data collection and broader representation of participants, resulting in larger and more diverse samples that improve external validity (Rodd 2024). Even in traditional cognitive research, online recruitment enables the rapid collection of large data sets (Peer et al. 2017). This increased efficiency is a strong argument for moving experiments online, as it supports the much-needed scaling of laboratory-based paradigms to larger sample sizes (Hartshorne et al. 2019). Encouragingly, the typical sample size has improved somewhat over the last decade, likely due in part to increased online recruitment of

participants (Fraley et al. 2022; Sassenberg and Ditrich 2019). In reviewing these advantages, Hartshorne et al. (2019) conclude that online volunteers follow instructions and respond truthfully to a degree that matches or exceeds that of laboratory subjects. With appropriate technology standards, participant screening, and task monitoring in place, researchers can ensure reliable, valid, and scalable data collection outside the lab, even for time-sensitive tasks.

3.1 Research questions and hypotheses

In the first experiment, a within-subjects design was used to investigate whether different types of pseudowords are processed differently (in terms of their generation) by observing the accuracy and timing of participants' responses:

- **H1a:** Manually constructed pseudowords with a preserved suffix differ from algorithmically generated pseudowords with a preserved suffix (see 7a, 8a, 9a and 10a below) in terms of response time and accuracy.
- **H1b:** Manually constructed pseudowords with a retained word-formation suffix (see 7b, 8b, 9b and 10b below) differ from manually constructed pseudowords with an altered word-formation suffix (see 7c, 8c, 9c and 10c below) in terms of response time and accuracy.

The second experiment used a between-subjects design to investigate how participants that had not participated in the first experiment processed existing words by again measuring the accuracy rate and response time. To this end, we hypothesized that pseudowords with retained word-formation suffixes would appear more word-like, making it more difficult to decide on the existing words. This in turn would be reflected in longer response times and a lower accuracy compared to the version of our experiment with pseudowords with altered word-formation suffixes (Hypothesis H2).

- **H2:** Manually constructed pseudowords with a retained suffix are more word-like, making it more difficult to distinguish them from existing words in a lexical decision task; this is reflected in longer response times and a lower accuracy rate for existing words.

This hypothesis is based on models of lexical access that assume early morphological decomposition during word recognition: The presence of a valid derivational suffix causes the parser to treat the pseudoword as a potentially legitimate lexical item (Taft and Forster 1975; Rastle, Davis and New 2004). If this is true for Slovene, it has methodological consequences that suggest that the construction of pseudowords is not a neutral design decision. Instead, the degree of morphological well-formedness of the pseudowords directly affects the difficulty of the task and influences both accuracy and response latency (Keuleers and Brysbaert 2010).

3.2 Procedure

The experiment was conducted online using the web-based software environment Ibex Farm (Drummond 2007), which was extended with the PennController module (Zehr and Schwarz 2018). Prior to participating, participants gave their informed consent and completed a demographic questionnaire. This was followed by two practice trials (one pseudoword + one word), after which the words and pseudowords appeared in a random order on the screen. The participant's task was to judge for each item individually whether it was a Slovene word or not by pressing, clicking, or tapping the *C* or *M* key for *NE* 'no' or *JA* 'yes' displayed at the bottom left and right of the screen, respectively. The average duration was 3.3 minutes (SD = 0.6). Participants conducted the experiment using their own devices (i.e., computer: 40.5%, smartphone: 56.4%, and tablet: 3.1%; see Table 4) at a location of their choice and were asked to do so quickly and undisturbed. There was no time limit set for responding.

There are several theoretical and practical considerations for setting (or not setting) a time limit on a lexical decision task. Typically, the lexical decision is limited to 3–5 seconds, as the goal is to measure lexical access rather than deliberate reasoning. It has been shown that time pressure encourages automatic processing at the lexical level: Without a time limit, participants might resort to post-lexical strategies, such as consciously analyzing word structure, which can bias the results. A time limit favors the automatic activation of word representations, so that response time is a purer measure of lexical retrieval (Balota and Chumbley 1984). It also reduces variability in strategy use between participants or between trials, which improves the consistency of the data. In addition, given unlimited time, participants may bias their responses by, for example, waiting longer for difficult items or guessing pseudowords. Lexical effects, including word frequency, neighborhood density, and concreteness, are often more detectable under time-limited conditions. For example, word frequency effects may diminish or disappear when participants are allowed to think (Seidenberg et al. 1984). However, setting a time limit can also have disadvantages. It can increase the error rate, cause frustration, make the task seem unnatural or mask actual effects. Slower participants, such as younger children or older adults, slower devices, such as smartphones compared to computers, or longer stimuli could be unfairly disadvantaged. Because our experiment included various participant age groups, stimuli of different lengths, and was conducted on various devices, we decided not to set a time limit during the task and instead removed outlier responses afterward to maintain data quality.

3.3 Materials

We formed the pseudowords from existing Slovene complex words with the suffixes *-ec*, *-nik*, *-ica*, and *-arna*. The suffixes were selected such that they differed in length (two to four phonemes) and meaning (agent, experiencer, tool, theme, or location), and that there were two for the masculine and two for the feminine gender. The source words were balanced in terms of their corpus frequency according to the deduplicated version of the Gigafida 2.0 corpus (Krek et al. 2019), as shown in Table 2. Twenty words were selected for each suffix, out of these five for each word length (seven, eight, nine, ten, and eleven letters)[2] and four for each predefined frequency interval (10–99, 100–999, 1,000–9,999, and 10,000–99,000).[3]

(6)		Lexeme 'gloss'	Suffix	Ending	Phonemes	Gender
	a.	*oškodovan-ec-*∅ 'victim'	*-ec*	∅	2	m.
	b.	*prav-nik-*∅ 'lawyer'	*-nik*	∅	3	m.
	c.	*govor-ic-a* 'rumor'	*-ica*	a	3	f.
	d.	*knjig-arn-a* 'bookshop'	*-arna*	a	4	f.

TABLE 2: Source words were balanced with respect to gender and suffix length

After the list of original morphologically complex words was compiled, we began to create pseudowords (examples are presented in Table 3).

- For pseudoword set P1, we retained the onset, length, syllable structure, and word-formation suffix of the source word, and we replaced two sounds of the stem with a related sound (e.g., a voiceless stop with a voiced stop).
- For pseudoword set P2, we applied the same procedure as for set P1 but also altered the suffixes. Because we wanted to maintain the structure of the complex word in order to compare P2 with P1, we did not change the suffixes arbitrarily phoneme by phoneme, but removed the existing suffixes, created four pseudo-suffixes (*-ec* → *-es, -nik* → *nok, -ica* → *-epa* and *-arna* → *-arja*) and added them to the stems that were previously modified as described for P1. In Slovene, there are many existing word-formation suffixes which extremely limited our choice for pseudo-suffixes if we wanted to maintain the syllabic structure of the originals and adhere to the rules governing the internal structure of Slovene words.

[2] Complex words with fewer than seven and more than eleven letters are rare, and so it is impossible to create a balanced set.

[3] Intervals are loosely based on Zipf's law, according to which the value of the nth entry is often approximately inversely proportional to n. Therefore, in a frequency table of words in a text or corpus of natural language, word frequency is inversely proportional to the word rank.

- For pseudoword set P3, we used the Wuggy software (Keuleers and Brysbaert 2010), which was originally developed for English and then adapted for other languages. We adapted it for Slovene in four stages. First, a list of hyphenated Slovene words was created using the headword lists from three Slovene explanatory dictionaries: the second edition of the *Dictionary of the Slovenian Standard Language*, *eSSKJ*, and the *Growing Dictionary of the Slovene Language*, as described in Perdih et al. (2025). The word selection was limited by certain criteria, including word length, frequency in the corpus, and exclusion of proper names. The final list comprised 79,413 words. Second, we hyphenated these words with Pyphen (https://pyphen.org/), a Python module for hyphenating words using a Slovene dictionary of hyphenation patterns included in LibreOffice (these were based on Slovene TeX hyphenation patterns by Matjaž Vrečko (GPL/LGPL license; https://github.com/hyphenation/tex-hyphen/blob/master/hyph-utf8/tex/generic/hyph-utf8/patterns/txt/hyph-sl.pat.txt) and were later corrected by Mojca Miklavec, as described by Martin Srebotnjak (https://wiki.openoffice.org/wiki/Documentation/SL/Using_TeX_hyphenation_patterns_in_OpenOffice.org). After algorithmic hyphenation, we counted the occurrence of different syllables and we manually checked words that contained rare syllables (frequency < 10), thus correcting some repeating incorrect patterns (especially the hyphenation of words with an onset starting with a vowel; e.g., *abe_ce_da* → *a_be_ce_da* 'alphabet')[4] and filtering out words with non-repeating patterns ($n = 377$). Third, the words were supplemented with the corpus frequency from the deduplicated version of the Gigafida 2.0 corpus (Krek et al. 2019). Fourth, we imported the list into Wuggy and applied its algorithm to create the P3 list of pseudowords. By restricting the output in the Wuggy interface, we preserved the number of letters, sub-syllabic length, letter transition frequencies, and sub-syllabic segments (see Figure 1). In addition, the word onset and the word-formation suffix were preserved by providing a regular expression for each word (e.g., *^[p].+arna$* for *pekarna* 'bakery').

In total, 120 stimuli were created: forty existing words used as fillers (B0), forty existing morphologically complex words (B1), and three types of forty pseudowords based on these existing words (i.e., types P1, P2, and P3). In Experiment 1, we used all the different stimuli types, namely B0 + B1 + P1 + P2 + P3. In Experiment 2, we used either B0 + B1 + P1, B0 + B1 + P2, or B0 + B1 + P3.

[4] Marks syllable boundaries.

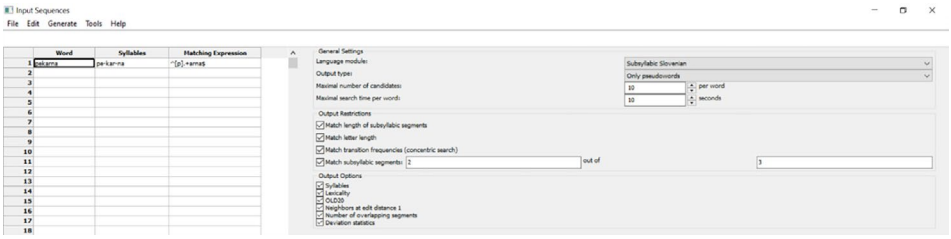

FIGURE 1: Screenshot of the Wuggy interface with all the settings

Pseudoword			Suffix	Type	Source word	Length	Onset	Frequency
(7)	*a.*	*pluvnik*	*nik*	P1	*prav-nik*	7	*p*	10,000–99,000
	b.	*pluvnok*	*nok*	P2				
	c.	*prejnik*	*nik*	P3				
(8)	*a.*	*gevolica*	*ica*	P1	*govor-ica*	8	*g*	10,000–99,000
	b.	*gevolepa*	*epa*	P2				
	c.	*gonorica*	*ica*	P3				
(9)	*a.*	*knjotarna*	*arna*	P1	*knjig-arna*	9	*k*	10,000–99,000
	b.	*knjotarja*	*arja*	P2				
	c.	*knjičarna*	*arna*	P3				
(10)	*a.*	*ohkadovanec*	*ec*	P1	*oškodovan-ec*	11	*o*	10,000–99,000
	b.	*ohkadovanes*	*es*	P2				
	c.	*odnodovanec*	*ec*	P3				

TABLE 3: Examples of pseudowords by the three generation methods (P1, P2, and P3) for all four suffixes (*-ec*, *-nik*, *-ica*, and *-arna*)

3.4 Participants

In total, we recruited 168 unique participants through personal contacts and social media for the two experiments, five of whom were excluded due to their early bilingualism, because we wanted to avoid pseudowords representing existing words in their other languages. We analyzed 163 adult Slovene native speakers (114 women, 48 men and 1 non-binary), with an average age of 33.5 years (SD = 13.3) and varying level of education. All informants participated in the survey voluntarily and anonymously, for which they were neither financially nor materially compensated.

Variable	*n*	%
Education		
Primary	2	1
Secondary vocational	0	0
Secondary technical	5	3
High school	47	29
Vocational college	4	2
Applied bachelor's	17	10
Bachelor's	51	31
Master's	11	7
Doctorate	26	16
Test application		
Computer	66	40
Smartphone	92	56
Tablet	5	3
Gender		
Male	48	29
Female	114	70
Other	1	1

TABLE 4: Participants in both experiments: education, test application, and gender

We divided the participants into four groups, and all of them received the same existing words (both filler and control words) and different sets of pseudowords (but the same number of stimuli). In within-subjects experiment 1, group G0 ($n = 61$) received all three types of pseudowords (P1 + P2 + P3). In between-subjects experiment 2, group G1 ($n = 38$) received manually prepared pseudowords with preserved suffixes (P1), group G2 ($n = 34$) received manually prepared pseudowords with non-preserved suffixes (P2), and group G3 ($n = 30$) received algorithmically created pseudowords with preserved suffixes (P3). The demographic details by group are shown in Table 5.

Experiment	Group, gender	*n*	Age	SD (age)
1	G0	61	34.9	12.8
	Other	1	43.0	NA
	Male	18	35.9	15.5
	Female	42	34.2	11.7
2a	G1	38	35.4	11.5
	Male	15	35.0	11.7
	Female	23	35.7	11.6
2b	G2	34	32.9	16.8
	Male	6	39.5	23.2
	Female	28	31.4	15.2
2c	G3	30	29.0	11.5
	Male	9	32.4	13.3
	Female	21	27.6	10.6
Total		**163**	**33.5**	**13.3**

TABLE 5: Experiment participants by age and gender

4 Analysis and results

The independent variable in the two experiments using the lexical decision task method was the generation of pseudowords. The dependent variables were response accuracy and response time for both words and pseudowords. An answer was scored as accurate if the participant identified an existing Slovene word as a word or if the pseudoword was not identified as a word. An answer was scored as inaccurate if the participant identified a non-existent Slovene word as a word or if the pseudo-word was identified as a word. Of the expected 19,560 responses, 19,554 were recorded. 288 (1.5%) were removed before analysis because their response time was two standard deviations above the average (> 4,420 ms), which was likely due to environmental interference (unlike most online lexical decision tasks, no time limit was set for the response). The total number of data points per type was comparable (mean = 1223.7; SD = 68.8): in Experiment 1 1192 for P1, 1210 for P2 and 1193 for P3, in Experiment 2 1369 for P1, 1226 for P2 and 1152 for P3. The accuracy rate of all participants' responses was above 95.7% (filler words B0: 99.5%, control words B1: 87.8%, pseudowords P1–3: 97.6%), indicating that participants were generally focused and attentive.

Modeling was performed in the open-source statistical environment R (version 4.2.0, R 2022) using the packages lme4 and lmerTest, and graphs were created using the packages ggplot2 and ggpubr. We report main effects and interactions as a series of chi-squared statistics. When results were significant, they were further modeled with mixed effects (Baayen 2008), which describe the relationship between dependent and independent variables through a linear/logistic combination of the latter. In these models, coefficients can vary with respect to one or more grouping variables and maximum random effects (i.e., participants and items: (1|ID)+(1|Item)) as long as they are justified by the design (Matuschek 2017). Ninety-five percent confidence intervals (CI) and p-values for the estimates were calculated using Laplace approximation.

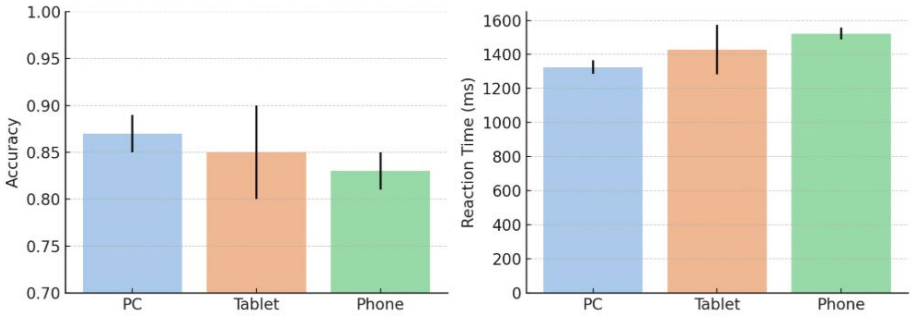

FIGURE 2: Accuracy and reaction times by device

Pairwise post-hoc comparisons were estimated using the Emmeans function of R, with p-values adjusted for multiple comparisons using the Bonferroni correction. Models were compared to their respective null models by subtracting the fixed factor and using the maximum likelihood method via R's Anova function.

Predictors such as word frequency and word length were scaled before integration into the model to improve both the numerical stability and interpretability of the model. When predictors have large ranges or different units, the optimizer can have difficulty converging, especially for models with random slopes. By z-scaling the predictor, the intercept becomes meaningful as it represents the expected result at an average measure, and the slope reflects the expected change when the measure changes by one standard deviation. Scaling also reduces the correlations between the predictors and the interaction terms, which minimizes multicollinearity and makes the effect sizes between the variables more comparable.

Before the actual analysis, we checked the effect of a device that the participants used to conduct the experiment. Using a one-way ANOVA, we tested whether device type affected response accuracy (generalized linear mixed model adjusted by maximum likelihood with the formula accuracy ~ device * type + (1|ID)+(1|item)). The effect of device type was not significant, $F(2, 162) = 1.84$, $p = 0.162$, partial $\eta^2 = 0.01$, suggesting that accuracy does not reliably differ between PCs, tablets, and phones. On the other hand, the one-way ANOVA (Linear mixed model fit by REML with the formula RT ~ device * type + (1|ID)+(1|Item)) revealed a significant effect of device type on response times, $F(2, 162) = 7.22$, $p = 0.001$, partial $\eta^2 = 0.08$. The estimated marginal means showed that responses were fastest on PCs (M = 1325 ms, SE = 40),

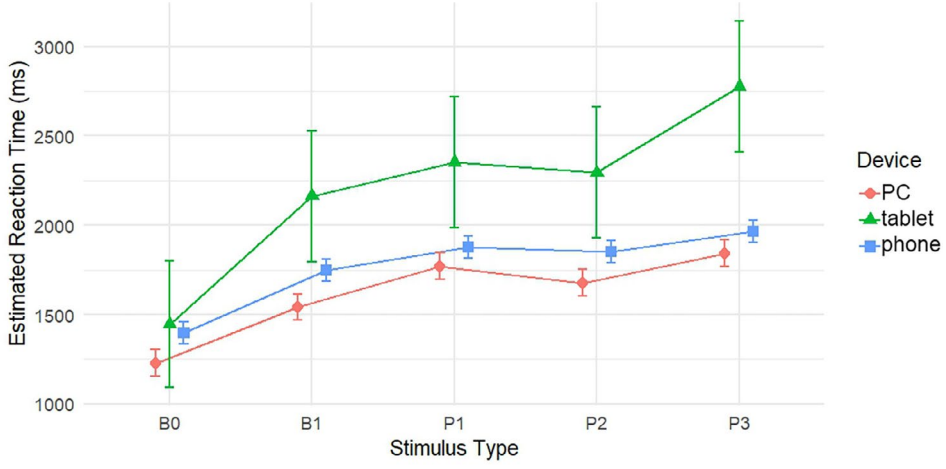

FIGURE 3: Estimated reaction times per stimulus type

followed by phones (M = 1428 ms, SE = 145) and slowest on tablets (M = 1523 ms, SE = 33). Post-hoc Tukey tests revealed that participants on phones responded significantly slower than participants on PCs ($p = 0.001$), while response times on tablets were not significantly different from those on PCs or phones (both $ps > 0.77$). Next, we wanted to see whether, despite the differences in absolute response times between devices, the pattern of relative response time differences between stimulus types was consistent. The interaction between device and stimulus type was significant, $F(8, 7163.7) = 3.35$, $p < 0.001$, but subsequent contrasts revealed that responses on all devices were fastest for B0, slower for B1, and slowest for pseudowords (P1–P3), thus maintaining the general rank order of the conditions. Only the magnitude of these differences varied between devices: while the contrasts between B1 and pseudowords and between pseudoword types were robust on PCs and phones, they were attenuated or non-significant on tablets. This indicates that the qualitative pattern of results was consistent across devices, but the strength of the pseudoword effects differed somewhat from device to device. We can now move on to our research questions. In the following section, we report on the analysis of experiments 1 and 2.

4.1 Experiment 1

In the first experiment, we were interested in how participants in group G0 (n = 61) responded when presented with the two conditions for words (B0 and B1) and the three conditions for pseudowords (P1, P2 and P3). B0 words that served as fillers, yielded the highest accuracy rate (99.0%) and the shortest response times (1,313 ms). B1, morphologically complex words with balanced frequency from low to high,

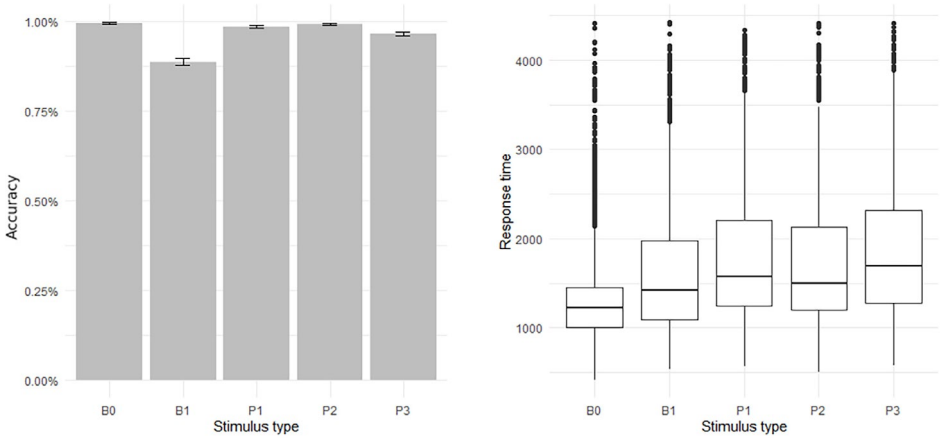

FIGURE 4: a) Accuracy rate (left), and b) response times (right) by stimulus type with SE error bars

which served as source words for the creation of pseudowords and as controls in the experiment, yielded the lowest accuracy rate (88.5%), but the second shortest response time (1,664 ms). The pseudowords were between B0 and B1 in terms of accuracy, and their response times were the longest. Pearson's chi-square tests yielded highly significant results for both accuracy rate ($\chi^2 = 354.53$, df = 4, p-value < 0.000) and response times ($\chi^2 = 10753$, df = 9684, p-value < 0.000). The results for the accuracy and response times are shown in Figure 4 and Table 6.

	Accuracy		**RT**	
Type	mean	SD	mean	SD
B0	0.99	0.08	1404	1176
B1	0.87	0.33	2123	2695
P1	0.98	0.14	2150	1814
P2	0.99	0.10	2096	2125
P3	0.96	0.19	2250	2205
SUM	**0.97**	**0.18**	**1905**	**1977**

TABLE 6: Accuracy rate and response times by stimulus type

Mean response times for words (1400 ms (B0) and 2100 ms (B1)) and pseudowords (2150–2250 ms) were higher than expected based on other studies (see Table 7) reporting mean values and using lexical decision paradigms with pseudowords and without priming in healthy adults. The mean response times in the presented studies reporting the relevant values are between 550 ms and 850 ms in younger adults; only in older people do they regularly exceed 1000 ms – with the notable exception of Roxbury et al. (2016). Also note that words are generally processed faster than pseudowords while in our study filler B0 were faster while control B1 were slower.

Researcher	**Year**	**Age group**	**RT (words)**	**RT (pseudowords)**
Tainturier	1987	17 younger adults 54 older adults	551 681	NA NA
Gold et al.	2010	17 younger adults	574	644
Lynchard and Radvansky	2012	61 younger adults 54 older adults	879 1244	NA NA
Katz et al.	2012	99 younger adults	641	814
Roxbury et al.	2016	17 younger adults 17 older adults	1187 1288	1434 1738
Manouilidou	2016	21 older adults	960	1057

TABLE 7: Response times on words and pseudowords in recent studies

We attribute the longer response times (compared to previous studies) to a combination of two effects. The first effect was due to the experimental procedure: the lack of time pressure may have led participants to take more time to respond overall. This explanation is supported by the fact that response times were prolonged across all stimulus types, not just for words or pseudowords. However, the two groups of words, namely the filler group B0 and the control group B1, unexpectedly differed in response time: for B0 it was shorter (as expected), while for B1 it was longer (unexpected) than the response time for pseudowords. Therefore, the second effect may be linked to the difference between the two groups of stimuli. To understand the difference in accuracy and response time between word types B0 and B1, we analyzed their frequency and length in the corpus, since both low frequency and greater length can prolong lexical decisions. For example, in Gold et al. (2010), who reported a mean word length of 4.7, the response time was 574 ms for words and 644 ms for pseudowords. In our experiment, the mean word length was 6.58 for type B0 and 9.0 for type B1. The two groups did not differ much in mean frequency (7.23 in B0 versus 7.16 in B1). Using Pearson's chi-square tests, we found that the B0 and B1 types did not differ significantly in frequency ($\chi^2 = 80$, df = 74, $p = 0.296$), but differed significantly in length ($\chi^2 = 45.281$, df = 6, $p < 0.000$).

	Frequency (corpus)		**Length (phonemes)**	
Word type	mean	SD	mean	SD
B0	7.23	9.01	6.58	0.98
B1	7.16	15.23	9.00	1.43
SUM	**7.20**	**12.43**	**7.79**	**1.73**

TABLE 8: Corpus frequency and length for types B0 and B1

We included both corpus frequency and word length (in phonemes) in our model, using generalized linear mixed effects fitted with maximum likelihood for accuracy (accuracy ~ type * frequency * length + (1 | ID) + (1 | item)) and a linear mixed model fitted by REML for response times (RT ~ type * frequency * length + (1 | ID) + (1 | item)).

Frequency had a positive but nonsignificant effect on accuracy and did not differ between B0 and B1 words. Word length had no significant effect on accuracy, nor did it differ between B0 and B1 word types. The interaction between word length and word frequency was also not significant.

We also modeled the effects of corpus frequency and word length (in phonemes) on response times. Word frequency had a small, nonsignificant effect on response times ($p = 0.491$) and did not differ between B0 and B1 words. Word length also had

a small, nonsignificant effect on response time (p = 0.430), and again, the effect did not differ between B0 and B1 word types. The interaction between word length and word frequency was also not significant. Notably, for the B0 word type, each additional letter was associated with an estimated increase in response time of about 52 ms, while the increase was larger for the B1 word type, at about 81 ms per additional letter. However, as neither the main effect nor the interaction reached significance, these values should be interpreted as descriptive tendencies rather than reliable effects.

4.1.1 Accuracy

Here we present the more complex generalized linear mixed model (GLMM) without frequency or length as factors to account for random effects and provide a more refined analysis. The fixed effect results reveal that B0 has the highest log-odds of accuracy (5.860), and B1 has a substantial negative impact (−3.217, $p < 0.000$). P1 (−0.9995, $p = 0.019$) and P3 (−1.848, $p < 0.000$) also show significant effects, whereas P2 (−0.195, $p = 0.684$) does not significantly differ from B0. The estimates were transformed into probabilities using the odds ratio formula (see Tables 9a–c).

Variable	Estimate	*SE*	*z*	*p*	Odds ratio	Probability
(Intercept)	5.8603	0.369	15.896	**0.000**	350.45	~1.00
B1	−3.2173	0.372	−8.648	**0.000**	0.04	0.04
P1	−0.9995	0.424	−2.356	**0.019**	0.37	0.27
P2	−0.1954	0.481	−0.407	0.684	0.82	0.45
P3	−1.8477	0.391	−4.723	**0.000**	0.16	0.14

TABLE 9a: Summary of the GLMM accuracy used in Experiment 1

Model	Value
AIC	1621.9
BIC	1670.2
logLik	−804.0
Deviance	1607.9
Residual *df*	7306.0

TABLE 9b: GLMM performance in Experiment 1

Random effects	Variance	*SD*
Item ID	1236 0.55	11116 0.74

TABLE 9c: Random effects in Experiment 1

Pairwise comparisons using the Bonferroni correction show significant differences between B0 and all other types, as well as between B1 and P1, P2, and P3. However, there are no significant differences between types of pseudowords (Table 10).

Contrast	Estimate	*SE*	*z*-ratio	*p*-value
B0–B1	−3.6220	0.329	−1.0998	**< 0.0001**
B0–P1	−2.0423	0.156	−1.3078	**< 0.0001**
B0–P2	−2.1408	0.163	−1.3162	**< 0.0001**
B0–P3	−1.6836	0.139	−12.091	**< 0.0001**
B1–P1	1.5797	0.358	4.417	**0.0001**
B1–P2	1.4812	0.360	4.113	0.0004
B1–P3	1.9384	0.350	5.534	< 0.0001
P1–P2	−0.0986	0.215	−0.459	1.0000
P1–P3	0.3587	0.198	1.809	0.7039
P2–P3	0.4572	0.203	2.250	0.2445
B1–P2	1.4812	0.360	4.113	0.0004

TABLE 10: Pairwise comparisons of accuracy rates in Experiment 1 show significant differences between words and pseudowords but not among pseudowords themselves

A comparison with the null model (which excludes type as a predictor) confirms that including type significantly improves the model's performance ($\chi^2 = 235.09$, $p < 0.000$). We further used the Akaike information criteria (AIC) to compare models with respect to both their fit and complexity. The lower AIC (1670.2 vs. 1869.7) and deviance (−921.50 vs. −803.96) in the full model indicate a better fit.

4.1.2 Response time

We tested the significance of differences in response times using a more complex GLMM fitted with restricted maximum likelihood. The variances indicate considerable variability in response times between participants and less between items, and large residuals indicate considerable unexplained variability. However, a median close to zero indicates a well-centered model with reasonable spread and few potential outliers (Tables 11a and 11b). The intercept (1330 ms) represents the baseline response time for the reference type (i.e., B0). B1 has the smallest increase, followed by P2, P1, and P3, which have the largest increase. For linear mixed models fitted with restricted maximum likelihood, the degrees of freedom are often difficult to estimate accurately, making traditional *p*-values unreliable. Instead, *t*-values are used as a measure of significance. They indicate by how many standard errors the estimated coefficient deviates from zero. The higher the absolute *t*-value, the stronger the evidence that the predictor has an influence on the dependent variable. *T*-values around 2 usually indicate statistical significance at the level of $p < 0.05$. Because all *t*-values in our model were greater than 10, this indicates that all predictors (types) had highly significant

effects on response time. When we applied the Kenward–Roger corrections to derive *p*-values from *t*-values, we confirmed the significance.

Variable	Estimate	*SE*	*t*	*p*
(Intercept)	1,330.93	49.64	26.81	**0.000**
B1	339.15	31.96	10.61	**0.000**
P1	512.57	31.99	16.02	**0.000**
P2	456.33	31.91	14.30	**0.000**
P3	597.45	31.98	18.68	**0.000**

TABLE 11a: Summary of the GLMM response time used in Experiment 1

Random effects	Variance	*SD*
Item	15733	125.4
ID	119748	346.0

TABLE 11b: Random effects in Experiment 1

Pairwise comparisons provide information on how response times differ between the types. B0 has the shortest response time (1331 ms), whereas B1 takes significantly longer to process (i.e., 339 ms longer). Similarly, P1 and P2 show even longer response times, and P3 shows the longest (1928 ms). The standard errors (*SE*) were consistently between 49.2 and 49.6, and the confidence intervals confirmed that all the differences observed were significant. Among the pseudoword types, the difference between P1 and P2 was not statistically significant ($p = 0.201$), suggesting similar processing times. However, P3 was significantly slower than both P1 ($p = 0.005$) and P2 ($p < 0.0001$).

Contrast	Estimate	*SE*	*df*	*z*-ratio	*p*-value
B0–B1	−339.2	32.0	Inf	−10.612	< **0.0001**
B0–P1	−512.6	32.0	Inf	−16.022	< **0.0001**
B0–P2	−456.3	31.9	Inf	−14.298	< **0.0001**
B0–P3	−597.5	32.0	Inf	−18.681	< **0.0001**
B1–P1	−173.4	24.3	Inf	−7.151	< **0.0001**
B1–P2	−117.2	24.2	Inf	−4.843	< **0.0001**
B1–P3	−258.3	24.2	Inf	−10.656	< **0.0001**
P1–P2	56.2	24.2	Inf	2.325	0.2007
P1–P3	−84.9	24.3	Inf	−3.492	**0.0048**
P2–P3	−141.1	24.2	Inf	−5.836	< **0.0001**

TABLE 12: Pairwise comparisons of response times in Experiment 1 show significant differences between words and pseudowords but also among P1 and P3 and P2 and P3

A comparison with the null model (which excludes type as a predictor) confirms that the inclusion of type significantly improves the model's performance ($\chi^2 = 263.6$, $p < 0.000$). The lower AIC (114477 vs. 114732) and deviance (114724 vs. 11446) in the full model indicate a better fit.

4.1.3 Intermediate discussion

In Experiment 1, we found that pseudowords were processed longer than existing words, consistent with previous literature on this topic (Barca and Pezullo 2012). Because pseudowords are not listed in participants' mental dictionaries, they must search the entire dictionary before responding. The longer-than-expected response times overall were probably due to no time limit set for answering in the experimental protocol. We also expected participants to respond more accurately to words than to pseudowords, which was true only for filler B0 words. We tentatively explained that the differences between B0 and B1 might be due to their difference in mean length. Another possible factor could be that participants were less familiar with some of the existing words in B1, leading to a drop in the accuracy rate.

Turning to the research question, we found no differences in accuracy between the different types of pseudowords. However, there were slight but significant differences in response times between the algorithmically generated pseudowords (P3) and the two types of manually constructed pseudowords (P1 and P2). From this, we conclude that manual versus algorithmic construction of pseudowords plays a role in processing pseudowords in a Slovene lexical decision task, whereas the retention of the word-formation suffix does not. Because we found differences in processing pseudowords, we conducted Experiment 2 to test whether these differences influenced the processing of existing words.

4.2 Experiment 2

In Experiment 2, we explored how the structure of pseudowords influenced the processing of existing words. In this, we could not mix the different types of pseudowords because we could not disentangle their effects. We therefore opted for a between-groups

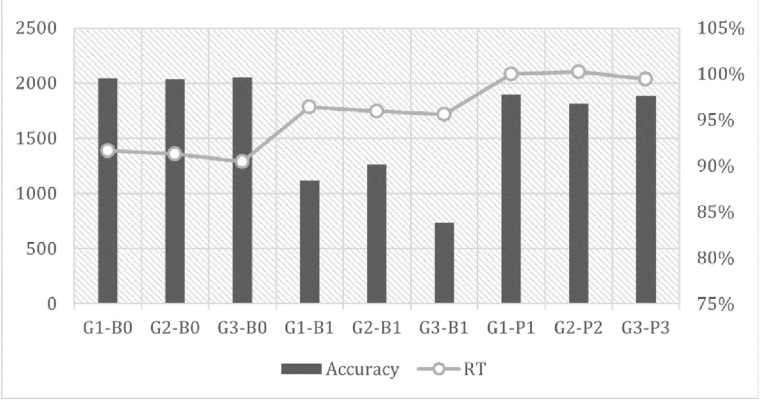

FIGURE 5: Accuracy (columns) and response times (line) by stimulus type

design in which different participants received different conditions so that each participant was only exposed to a subset of the total stimuli. Specifically, we tested three new groups of participants with the two types of words (B0 and B1) and only one type of pseudowords each (group G1 received pseudowords P1, G2 received P2, and G3 received P3). We plotted the results by both stimulus type and group (Figure 5).

4.2.1 Pseudowords

According to Pearson's chi-squared test, there was no statistically significant difference between pseudoword types (P1, P2, and P3) in terms of accuracy ($\chi^2 = 2.124$, $df = 2$, p-value = 0.346) or response times ($\chi^2 = 4213$, $df = 4162$, p-value = 0.284), as is evident from Figure 6.

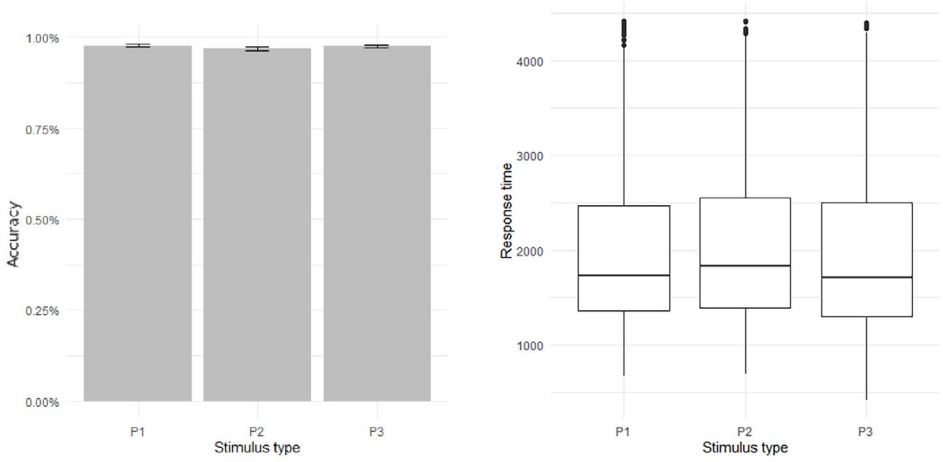

FIGURE 6: Accuracy (left) and response times (right) by pseudoword type with SE error bars

4.2.2 Words

On the other hand, according to Pearson's chi-squared test, there was a statistically significant main effect of the group in processing target words, both in terms of accuracy ($\chi^2 = 17.742$, $df = 2$, p-value = 0.0001) and response times ($\chi^2 = 5141$, $df = 4870$, p-value = 0.003), as is evident from Figure 7. Because all the participants received the same set of B0 and B1 stimuli, we attribute the effect to the different types of pseudowords they received. However, the values predicted by the model are characterized by relatively large standard errors that might signal a considerable effect of random variables (i.e., item and participant). To determine whether the main effect is due to differences between the groups, we modeled results with linear mixed effects. Again, we used Laplace's approximation to examine the accuracy and restricted maximum likelihood to examine the probability of response times.

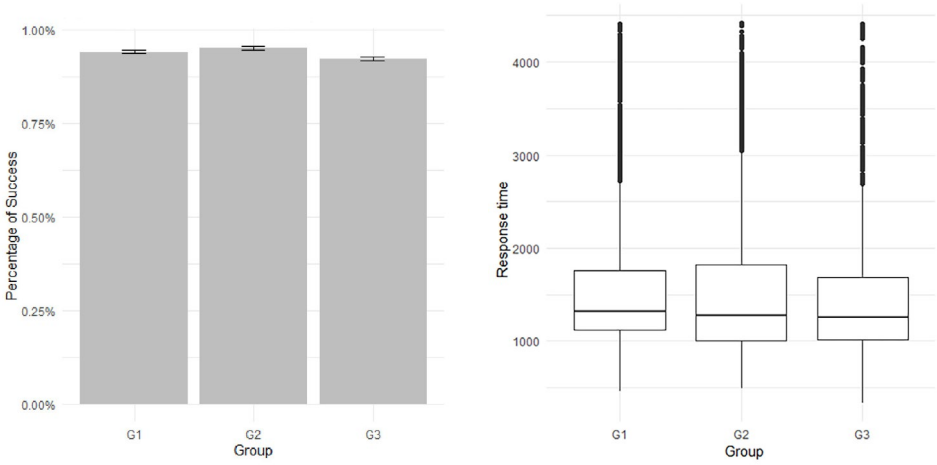

FIGURE 7: Accuracy (left) and response times (right) for existing words (B1) by group receiving different pseudoword types with SE error bars

Accuracy rate. The base estimate represents the logarithmic probability of accuracy for Group 1. The estimated effects for Group 2 (p-value = 0.88) and Group 3 (p-value = 0.08) indicate a slight decrease in probability compared to Group 1, but none of the effects reach significance. The random effects show considerable variability, especially for the items. Post-hoc pairwise comparisons using the Bonferroni correction confirm that none of the group differences are significant. Finally, when a comparison was conducted between the full model (with the group) and a null model (without the group), the AIC values were almost identical (2102.1 vs. 2101.7), suggesting that inclusion of the group does not improve the fit of the model. The likelihood-ratio test ($\chi^2(2) = 3.67$, $p = 0.160$) also confirms that the addition of the group does not significantly increase the predictive power of the model.

Response times. Comparisons between the groups show that neither Group 2 nor Group 3 differ significantly from Group 1, which is confirmed by p-values, high standard errors, and low t-values. The random effects show considerable variance at both the participant and item level, and the residuals indicate considerable unexplained variability. A comparison between the full model and a null model shows that the inclusion of the group does not improve the fit of the model ($\chi^2(2) = 0.88$, $p = 0.643$).

4.2.3 Intermediate discussion

In Between-Group Experiment 2, the results showed that the group did not significantly predict the accuracy and response times. Participants belonging to different groups (G1, G2, or G3) and receiving different pseudowords (P1, P2, or P3) processed

existing Slovene words in a similar way, in terms of both the accuracy and response times. From this we conclude that the construction of pseudowords does not play a role in processing existing words in a Slovene lexical decision task.

5 Conclusion

The motivation for our study was to explore whether pseudowords can be constructed in a systematic, computer-assisted way. We hypothesized that pseudowords generated by hand and those generated by computer might differ in their similarity to real words, which could in turn influence lexical processing. We used a lexical decision task to test how the structure of pseudowords affects the response accuracy and response time for morphologically complex Slovene words. We compiled a list of Slovene source words and balanced them in terms of their corpus frequency, length, and word-formation suffixes. We established various procedures for creating pseudowords, including the application of the Wuggy software (Keuleers and Brysbaert 2010) based on Slovene bigram chains and the manual substitution of similar phonemes with or without influence on the word-formation suffix. We included all three sets together with the Slovene source words in the first experiment. Although the three sets of pseudowords differed significantly from the words in terms of accuracy (higher) and response times (longer), they differed only partially from one another—that is, in terms of the response time only: there were no differences in the manually prepared pseudowords, regardless of whether their suffix was retained or not, but the two manually prepared sets were processed faster than the set that was created algorithmically. In the second step, we opted for a between-group design of the experiment so that each of the newly recruited participants received only one set of pseudowords. This time we were able to compare the processing of the existing words as a function of the type of pseudowords in the experiment. We found no statistically significant differences. Thus, we conclude that the construction of pseudowords with respect to their internal morphological structure and the protocol of their generation (computerized or by hand) has no effect on processing words in a lexical decision experiment. Consequently, we cannot propose concrete improvements for existing Slovene studies that rely on pseudowords.

An important question is to what extent these conclusions can be generalized to other languages: We would expect certain principles to apply more generally, since the cognitive processes underlying lexical decision – such as orthographic familiarity, phonotactic well-formedness and neighborhood density – are not language-specific (Balota and Chumbley 1984; Coltheart et al. 1977; Keuleers and Brysbaert 2010),

and the distinction between word and non-word is reliably dependent on frequency and length in many language systems (Brysbaert, Mandera and Keuleers 2018; Diependaele, Lemhöfer and Brysbaert 2013). These consistent results suggest that while the specific implementation of pseudoword generation must be tailored to the morphological properties of a particular language, the underlying mechanisms involved in the task are largely the same.

For these reasons, it is quite justified to extend our conclusions beyond Slovene. We hypothesize that while the relative usefulness of different pseudoword types may vary depending on the morphological profile of a language, the general finding that multiple construction methods can yield reliable and interpretable results should be generalized for typologically different languages.

Finally, mean response times for words (1400–2100 ms) and pseudowords (2150–2250 ms) in our study were substantially higher than values typically reported in lexical decision studies, where they generally range from 550 ms to 850 ms in younger adults and exceed 1000 ms only in older participants. Thus, the response times in our study were unusually long. We suggested two alternative and possibly cumulative explanations: (1) the absence of time pressure, which likely encouraged slower responses compared to studies with a limited response window, and (2) the relatively long stimuli. In our study, mean word length was 6.6 for type B0 and 9.0 for type B1 ($p < .001$), while in one study (Gold et al. 2010) that reported the mean length of existing words, it was 4.7. When length was included in our response time model, descriptively, each additional letter was associated with a 52 ms increase in response times for B0 words and an 81 ms increase for B1 words. However, this effect did not reach significance and should be interpreted as a tendency rather than a robust effect. Future work is needed to determine the effect of word length on response times in the lexical decision task.

Acknowledgements

This article was written as part of the research project Slovenian Word-Prevalence: An Online Mega-Study of Word Knowledge (code J6-50199) and the program Slovene Language in Synchronous and Diachronic Development (code P6-0038), funded by the Slovenian Research and Innovation Agency (ARIS).

Članek temelji na raziskovalnih podatkih, ki se hranijo na Inštitutu za slovenski jezik Frana Ramovša in so javno dostopni na povezavi <https://osf.io/ejqza/files/956eg>.

Bibliography

Aguasvivas, Jose Armando, Carreiras, Manuel, Brysbaert, Marc, Mandera, Paweł, Keuleers, Emmanuel, Duñabeitia, Jon Andoni. 2018. Spalex: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology* 9. <https://doi.org/10.3389/fpsyg.2018.02156>.

Angele, Bernhard, Baciero, Ana, Gómez, Pablo, Perea, Manuel. 2023. Does online masked priming pass the test? The effects of prime exposure duration on masked identity priming. *Behavior Research Methods*, 55/1: 151–167. <https://doi.org/10.3758/s13428-021-01742-y>.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Balota, David A., Yap, Melvin J., Hutchison, Keith A., Cortese, Michael J., Kessler, Brett, Loftis, Bjorn, Neely, James H., Nelson, Douglas L., Simpson, Greg B., Treiman, Rebecca. 2007. The English Lexicon Project. *Behavior Research Methods* 39/3: 445–459. <https://doi.org/10.3758/BF03193014>.

Balota, David A., Chumbley, James I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, 10/3: 340–357.

Barca, Laura, Pezzulo, Giovanni. 2012. Unfolding visual lexical decision in time. *PLoS One*, 7/4: e3593. <https://doi.org/10.1371/journal.pone.0035932>.

Brown, Roger et al. 1987. Letter Substitution in Pseudoword Generation. *Linguistic Research Review* 5/2: 45–67.

Brysbaert, Marc, Stevens, Michael, Mandera, Paweł, Keuleers, Emmanuel. 2016. The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance* 42. 441–458. <https://doi.org/10.1037/xhp0000159>.

Brysbaert, Marc, Mandera, Paweł, Keuleers, Emmanuel. 2018. The word frequency effect in word processing: An updated review. *Current directions in psychological science*, 27/1: 45–50.<https://doi.org/10.1177/0963721417727521>.

Brysbaert, Marc, Mandera, Paweł, McCormick, Samantha F., Keuleers, Emmanuel. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods* 51. 467–479. <https://doi.org/10.3758/s13428-018-1077-9>.

Collins, Allan M., Quillian, M. Ross. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8/2. 240–247.

Coltheart, Max, Davelaar, Eileen, Jonasson, Jon Torfi, Besner, Derek. [1]1977, 2022. Access to the internal lexicon. In: S. Dornič (ed.). *Attention and performance VI*. London: Routledge. 535–555. <https://doi.org/10.4324/9781003309734>.

Diependaele, Kevin, Lemhöfer, Kristin, Brysbaert, Marc. 2013. The word frequency effect in first-and second-language word recognition: A lexical entrenchment account. *Quarterly journal of experimental psychology*, 66/5. 843–863.<https://doi.org/10.1080/17470218.2012.720994>.

Dołżycka, Joanna Daria, Nikadon, Jan, Formanowicz, Magdalena. 2022. Constructing Pseudowords with Constraints on Morphological Features - Application for Polish Pseudonouns and Pseudoverbs. Journal of Psycholinguistic Research 51/6, 1247–1265. <https://doi.org/10.1007/s10936-022-09884-6>.

Dorffner, Georg, Harris, Catherine L. 1997. When pseudowords become words: Effects of learning on orthographic similarity priming. In: M. G. Shafto, P. Langley (eds.): *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Mahwah, New Jersey: Lawrence Erlbaum Associates. 185–190.

Drummond, Alex. 2007. Ibex Farm: A web-based experiment platform. <https://spellout.net/ibexfarm>.

Duyck, Wouter, Desmet, Timothy, Verbeke, Lieven P. C., Brysbaert, Marc. 2004. WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, 36/3. 488–499. <https://doi.org/10.3758/BF03195595>.

Ferrand, Ludovic, New, Boris, Brysbaert, Marc, Keuleers, Emmanuel, Bonin, Patrick, Méot, Alain, Augustinova, Maria, Pallier, Christophe. 2010. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42/2. 488–496. <https://doi.org/10.3758/BRM.42.2.488>.

Field, John. 2004. *Psycholinguistics: the key concepts*. London, New York: Routledge.

Fraley, R. Chris, Chong, Jia Y., Baacke, Kyle A., Greco, Anthony J., Guan, Hanxiong, Vazire, Simine. 2022. Journal N-pact factors from 2011 to 2019: evaluating the quality of social/personality journals with respect to sample size and statistical power. *Advances in Methods and Practices in Psychological Science*, 5/4. 1–17. <https://doi.org/10.1177/25152459221120217>.

Gold, Brian T., David K., Powell, Xuan, Liang, Jiang, Yang, Hardy, Peter A. 2007. Speed of lexical decision correlates with diffusion anisotropy in left parietal and frontal white matter: evidence from diffusion tensor imaging. *Neuropsychologia*, 45/11. 2439–2446. <https://doi.org/10.1016/j.neuropsychologia.2007.04.011>.

Guasch, Marc, Boada, Roger, Duñabeitia, Jon Andoni, Ferré, Pilar. 2022. Prevalence norms for 40,777 Catalan words: An online megastudy of vocabulary size. *Behavior Research Methods* 55. 3198–3217. <https://doi.org/10.3758/s13428-022-01959-5>.

Hartshorne, Joshua K. et al. 2019. The meta-science of adult statistical word segmentation: Part 1. *Collabra: Psychology*, 5/1. <https://doi.org/10.1525/collabra.181>.

Imbir, Kamil K., Spustek, Tomasz, Żygierewicz, Jarosław. 2015. Polish pseudo-words list: dataset of 3023 stimuli with competent judges' ratings. Frontiers in Psychology 6. <https://doi.org/10.3389/fpsyg.2015.01395>.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Jescheniak, Jörg D., Levelt, Willem J. M. 1994. Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of experimental psychology: learning, Memory, and cognition*, 20/4. 824–843.

Keuleers, Emmanuel, Brysbaert, Marc. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42/3. 627–633. <https://doi.org/10.3758/BRM.42.3.627>.

König, Jemma, Calude, Andreea S., Coxhead, Averil. 2020. Using Character-Grams to Automatically Generate Pseudowords and How to Evaluate Them. Applied Linguistics 41/6: 878–900. <https://doi.org/10.1093/applin/amz045>.

Krek, Simon, et al. 2020. Gigafida 2.0: the reference corpus of written standard Slovene. V: N. Calzolari (ur.): *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Marseille, France*. Paris: ELRA - European Language Resources Association. 3340-3345. <http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf>.

Levelt, Willem. J. M., Roelofs, Ardi., Meyer, Antje S. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22/1. 1–75. 10.1017/s0140525x99001776>.

Longtin, Catherine-Marie, Meunier, Fanny. 2005. Morphological decomposition in early visual word processing. *Journal of Memory and Language*, 53/1: 26–41. <https://doi.org/10.1016/j.jml.2005.02.008>.

Luce, Paul A., Pisoni, David B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19/1. 1–36.

Manouilidou, Christina, Dolenc, Barbara, Marvin, Tatjana, Pirtošek, Zvezdan. 2016. Processing complex pseudo-words in mild cognitive impairment: The interaction of preserved morphological rule knowledge with compromised cognitive ability. *Clinical Linguistics & Phonetics*, 30/1: 49–67. <https://doi.org/10.3109/02699206.2015.1102970>.

Marjanovič, Katarina, Manouilidou, Christina, Marvin, Tatjana. 2013. Word-formation rules in Slovenian agentive deverbal nominalization: A psycholinguistic study based on pseudo-words. *Slovenski jezik / Slovene Linguistic Studies* 9. 93–109. <https://hdl.handle.net/1808/11432>.

Marslen-Wilson, William D. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25/1–2. 71–102. <https://doi.org/10.1016/0010-0277(87)90005-9>.

Marslen-Wilson, William, Tyler, Lorraine K., Waksler, Rachelle, Older, Lianne. 1994. Morphology and meaning in the English mental lexicon. *Psychological review*, 101/1. 3–33. <https://doi.org/10.1037/0033-295X.101.1.3>.

Matuschek, Hannes, Kliegl, Reinhold, Vasishth, Shravan, Baayen, Harald, Bates, Douglas. 2017. Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315. <https://doi.org/10.1037/0033-295X.101.1.3>.

McRae, Ken, Ferretti, Todd R., Amyote, Liane. [1]1997, 2010. Thematic roles as verb-specific concepts. Language and Cognitive Processes, 12/2–3. 137–176. <https://doi.org/10.1080/016909697386835>.

Medler, David A., Binder, Jeffrey R. 2005. *Mcword. An on-line orthographic database of the English language*. <http://www.neuro.mcw.edu/mcword/>.

Meyer, David E., Schvaneveldt, Roger W. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90/2. 227–234. <https://doi.org/10.1037/h0031564>.

Morrison, Catriona M., Ellis, Andrew W. 1995. Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of experimental psychology: learning, memory, and cognition*, 21/1. 116–133. <https://doi.org/10.1037/0278-7393.21.1.116>.

Oldfield, Richard C., Wingfield, Arthur. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17/4: 273–281. <https://doi.org/10.1080/17470216508416445>.

Pavlič, Matic, Andreetta, Sara, Stateva, Penka, Stepanov, Arthur. 2022. Vpliv koaktivacije italijanščine kot drugega jezika na fonološko presojanje besedišča v slovenščini kot prvem jeziku. N. Pirih Svetina, I. Ferbežar (eds.): *Na stičišču svetov: slovenščina kot drugi in tuji jezik. Obdobja 41*. Ljubljana: Založba Univerze v Ljubljani. 261–270.

Peer, Eyal, Brandimarte, Laura, Samat, Sonam, Acquisti, Alessandro. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology* 70: 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>.

Perdih, Andrej, Gabrovšek, Dejan, Pavlič, Matic. 2025. Izdelava seznama besed za množično raziskavo razširjenosti slovenskih besed. *Slavistična revija*, 73/1. 121–138. <https://doi.org/10.57589/srl.v73i1.4231>.

Pulvermüller, Friedemann. 1999. Words in the brain's language. *Behavioral and Brain Sciences*, 22/2. 253–336.

Rastle, Kathleen, Davis, Matthew H., New, Boris. 2004. The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic bulletin & review* 11. 1090–1098. <https://doi.org/10.3758/BF03196742>.

Ratcliff, Roger, Hendrickson, Andrew T. 2021. Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behavior Research Methods*, 53/6. 2302–2325. <https://doi.org/10.3758/s13428-021-01573-x>.

Rodd, Jennifer M. 2024. Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language* 134. <https://doi.org/10.1016/j.jml.2023.104472>.

Roxbury, Tracy, McMahon, Katie, Coulthard, Alan, Copland, David A. 2016. An fMRI study of concreteness effects during spoken word recognition in aging. Preservation or attenuation?. *Frontiers in Aging Neuroscience* 7. <https://doi.org/10.3389/fnagi.2015.00240>.

Sassenberg, Kai, Ditrich, Lara. 2019. Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, 2/2. 107–114. <https://doi.org/10.1177/2515245919838781>.

Seidenberg, Mark S., Waters, Gloria S., Barnes, Marcia A., Tanenhaus, Michael K. 1984. When does irregular spelling or pronunciation influence word recognition? *Journal of verbal learning and verbal behavior*, 23/3. 383–404. <https://doi.org/10.1016/S0022-5371(84)90270-6>.

Solso, Robert L., Barbuto, Paul F., Juel, Connie L. 1979. Bigram and trigram frequencies and versatilities in the English language. *Behavior Research Methods & Instrumentation*, 11/5. 475–484. <https://doi.org/10.3758/BF03201360>.

Suen, Ching Y. 1979. N-gram statistics for natural language understanding and text processing. *IEEE transactions on pattern analysis and machine intelligence* 2. 164–172. <https://doi.org/10.1109/TPAMI.1979.4766902>.

Trost, Stefan. 2002. *WordCreator*. <https://www.sttmedia.com/wordcreator>.

White, Corey N., Ratcliff, Roger, Vasey, Michael W., McKoon, Gail. 2010. Using diffusion models to understand clinical disorders. *Journal of mathematical psychology*, 54/1: 39–52. <https://doi.org/10.1016/j.jmp.2010.01.004>.

Yarkoni, Tal, Balota, David, Yap, Melvin. 2008. Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15/5. 971–979. <https://doi.org/10.3758/PBR.15.5.971>.

Zehr, James, Schwarz, Florian. 2018. PennController for Internet Based Experiments (IBEX). <https://doi.org/10.17605/OSF.IO/MD832>.

Summary

In this study, we used a lexical decision task to examine how the structure of pseudowords affects accuracy and response time in processing morphologically complex Slovene words, specifically derivations. We compiled a list of Slovene source words and balanced them for corpus frequency, length,

and derivational suffixes. We then developed several procedures for generating pseudowords: using the Wuggy application (Keuleers and Brysbaert 2010) based on Slovene bigrams, as well as manual replacement of similar phonemes with or without altering the derivational suffix. All three pseudoword lists, together with the Slovene source words, were included in the first experiment. Although all three pseudoword lists differed significantly from source words in accuracy (lower) and response time (longer), they only partially differed from one another in response time: for the manually created pseudowords, there were no differences regardless of whether the suffix was preserved, while both manually created lists were processed more quickly than the algorithmically generated list. In the second phase, we repeated the experiment, but each group of participants – none of whom had taken part in the first experiment – received only one pseudoword list. This allowed us to compare the processing of source words as a function of pseudoword type. No statistically significant differences were observed. We therefore conclude that the described differences in pseudoword construction do not affect word processing in a lexical decision experiment in a highly inflectional language with rich morphology.

Leksikalno procesiranje morfološko zapletenih slovenskih besed pri testu presojanja besedišča: vloga psevdobesed

V tej študiji smo z nalogo leksikalnega presojanja preverili, kako struktura psevdobesed vpliva na uspešnost reševanja in reakcijski čas pri procesiranju morfološko zapletenih slovenskih besed, in sicer izpeljank. Sestavili smo seznam slovenskih izhodiščnih besed in jih uravnotežili glede na korpusno pogostnost, dolžino in besedotvorne pripone. Vzpostavili smo različne postopke za tvorbo psevdobesed, in sicer uporabo programa Wuggy (Keuleers & Brysbaert 2010) na podlagi slovenskih dvočrkovnih verig in ročno zamenjavo podobnih fonemov z ali brez vpliva na besedotvorno pripono. Vse tri sezname smo skupaj s slovenskimi izhodiščnimi besedami vključili v prvi eksperiment. Čeprav so se vsi trije seznami po uspešnosti (višja) in odzivnem času (daljši) bistveno razlikovali od besed, so se po odzivnem času med seboj razlikovali le deloma: pri ročno pripravljenih psevdobesedah ni bilo razlik ne glede na to, ali je bila njihova pripona ohranjena ali ne, medtem ko sta bila dva ročno pripravljena seznama obdelana hitreje kot seznam, ki je bil pripravljen strojno. V drugem

koraku smo isti eksperiment ponovili tako, da je vsaka od skupin udeležencev, ki niso bili vključeni v prvi eksperiment, prejela le en nabor psevdobesed. Zato smo lahko primerjali obdelavo obstoječih besed glede na vrsto psevdobesed. Statistično pomembnih razlik nismo ugotovili. Tako sklepamo, da opisane razlike pri pripravi psevdobesed ne vplivajo na procesiranje besed v eksperimentu leksikalnega odločanja v pregibnem jeziku z bogato morfologijo.