Blaž Pažon, Vesna Bregar, and Klaudija Lukman

Faculty of Arts, University of Ljubljana, Slovenia
blaz.pazon@gmail.com, vesbregar@gmail.com, klaudijalkmn1@gmail.com

UDC: 811.163.6'242:81'27
811.163.6'371.611'367.622.22:159.947.2
DOI: 10.19090/slm.5.2

COMPETING DISTRIBUTIONS OF NOMINAL DIMINUTIVE SUFFIXES IN SLOVENIAN: DO NATIVE SPEAKERS' PREFERENCES REFLECT CORPUS DATA?

Abstract: The purpose of our experiment was to study the relationship between corpus data and native speakers' preferences regarding word form variation. Specifically, we investigated how well corpus frequency distributions of masculine nominal diminutive suffixes would be reflected in the preferences of native speakers of Slovenian when choosing suffixes to form diminutives. Based on similar research conducted by Makarova (2014), we expected native speakers' preferences to be in line with corpus distributions. Our stimuli were 33 masculine noun bases, and the diminutive suffixes observed were grouped into four categories -(*č*)*ek*, -*ič*, -*ičekk*, and -(*e*)*c*. We consulted Gigafida 2.0, the largest corpus of written Slovenian, to obtain frequencies of attestations for each of our target nouns with each of the diminutive suffixes. The most attested diminutive suffix category for each noun was dubbed its expected suffix. With an offline experiment we collected responses from 70 native speakers of Slovenian, who were tasked with transforming nouns in their base form into diminutives within sentence contexts. Our goal was to test whether native speakers would assign the expected suffix to the noun bases more often than some alternative suffix. The results indicated the number of expected responses exceeded the number of non-expected responses. The difference, however, was subtle, and closer examination of individual noun bases revealed a considerable variability in the shares of expected responses. Further analyses investigated potential predictors of suffix choices, including animacy, length, and expected suffix. Linear regression analyses showed the expected suffix category significantly predicted the proportion of expected responses. Notably, the suffixes -(*č*)*ek* and -(*e*)*c* emerged as significant predictors of the share of expected responses. The reasons for this variation, including the influence of the corpus characteristics, are discussed, and potential avenues for future research are suggested.

Key words: word form variation, nominal diminutives, diminutive suffixes, language corpora, native speaker preferences

1. Introduction

In cases of linguistic variation, where multiple forms with a similar meaning compete for usage, speakers are faced with the decision of which form to use. Their choices can be influenced by various factors, such as slight nuances in meaning, environment (phonological, morphological and syntactic) and frequency (Baayen et al., 2013). These rival forms either occur in free variation (where both variants can be used in the same context) or in complementary distribution (meaning only one of the forms is used in a given context). Researching how rival forms behave in various environments can provide us with new insights into the general relationship between form and meaning. Slavic languages are known to have especially rich morphology and inflection, which includes many cases of variation. Therefore, they can represent a source of numerous interesting research questions.

What inspired us to conduct this study on Slovenian, a language whose morphological variation has not been extensively researched using a similar method, was Makarova's work on Aktionsart in the Russian language. The term Aktionsart refers to a subtype of aspect which modifies verbal meaning by adding quantitative or qualitative characteristics (Makarova, 2016). As Makarova (2016) pointed out, in Russian different affixes can be attached to verbs to express lower intensity (attenuative Aktionsart) (see Makarova, 2014) or a single, instantaneous action (semelfactive Aktionsart) (see Makarova, 2009). However, in many cases, the attenuative, as well as the semelfactive form of a particular verb can be constructed through two

different affixes, which is supported by corpus data, e.g., the verb base *morozit'* 'to freeze' is attested with both *pri-* and *pod-* resulting in the attenuative *primorozit'*, or *podmorozit'*, meaning 'to freeze slightly' (Makarova, 2014). The presence of such variation might indicate that the two affixes or verb forms have a competing status in the native speakers' mental representations. Furthermore, it was found that some verbal stems occurred more frequently with one of the two affixes, while others were attested almost equally with both. In other words, some verbs more clearly preferred one affix, while for others a high degree of affix variation was attested.

Makarova's attenuative experiment, made accessible in the form of a dataset (2014), and which they elaborated on in a book chapter (2016), was most influential for our study, so we will describe it in more detail. After collecting corpus frequencies of the attenuative prefixes *pri-* and *pod-* for each target verb and categorizing the verbs by their frequency and the rate of prefix variation, informants were presented with a cloze test task, which required them to assign prefixes to verbs in a sentence context. The sentences were sourced from the corpus. One of the things Makarova (2016) was interested in observing was whether the prefixes used by informants would match the prefixes used with the verbs in the corpus. A statistical analysis confirmed that the majority of native speaker responses indeed employed the prefixes expected based on corpus data. For verbs almost exclusively attested in the corpus with either *pri-* or *pod-* most of the responses were target responses, i.e., *pri-* responses for *pri-* verbs and *pod-* responses for *pod-* verbs. A similar pattern was observed for verbs that exhibited greater prefix variation in the corpus. However, the frequency of opposite responses (e.g., *pri-* responses for *pod-* verbs) was much higher for these verbs compared to the exclusively *pri-* or *pod-* verbs, which is of course expected. If the mental representation or status of a verb allows greater prefix variation, one would expect native speakers to use different prefixes more, which would also be reflected in the corpus. And this is exactly what this comparison of different types of verbs indicated.

1.1. Slovenian diminutives

The morphological rivalry in Slovenian that we observed was that of nominal diminutive suffixes. We decided to focus on diminutives based on the observation that diminutive formation is very common and productive in Slovenian, while its form variation is similar to that of semelfactives and attenuatives in Russian. For instance, Vidovič Muha (1995, as cited in Vidovič Muha, 2011) analyzed five editions of SSKJ (The Dictionary of Standard Slovenian Language) with around 100,000 headwords, 1,514 of which were masculine nouns modified with affixes. She found that the most common affix modifications were diminutives. While Sicherl (2012, 2013) acknowledges that in Slovenian native speakers opt for either analytic or synthetic formations, sometimes even both, Sicherl and Žele (2011), as well as Sicherl (2013), note that diminutiveness is primarily expressed synthetically through affixation. Kavčič (2021) demonstrated that nominal diminutive formation represents a productive model. Moreover, based on some evidence from Gigafida 2.0 (Gigafida 2.0, 2023; Krek, 2020), the reference corpus of written Slovenian, her analysis indicates that synthetic formation of diminutives is more common than analytic formation. We narrowed our focus on masculine nominal diminutives, as this grammatical gender exhibits the greatest variation.

The prototypical semantic meaning of a diminutive is denotative, referring to the small size of the referent (Schneider, 2003, as cited in Sicherl, 2012), where the feature [small] is added to the base to indicate that the referent is smaller in size compared to an average representative of its category. However, diminutives often denote other features, such as the age of the referents, relative evaluation, or the speaker's emotional attitude towards the referent, with either positive or negative connotation (Sicherl, 2012). In Slovenian, diminutives are most commonly nouns, but they can also be adjectives, adverbs or even interjections and numerals (Sicherl, 2012).

Nominal diminutives can be derived with several suffixes, however their exact classification and number varies from author to author. For example, according to Toporišič (2000), there are six diminutive suffixes for the masculine gender: *-c*, *-ec* [ɔc], *-ic*, *-ič*/*-ìč*, *-ek*/*-èk* [ɔk], *-ček* [ɔk]. These can also have a derogatory or endearing connotation, and for the latter meaning four additional suffixes are listed, but these

are generally rarer and less productive. Vidovič Muha (1995, as cited in Vidovič Muha, 2011) on the other hand, lists the following six suffixes: -(*e*)*k*, -*č*(*e*)*k*, -(*e*)*c*, -*ič*, -*ic* and -*et*, where -(*e*)*k* is the most common suffix, representing over 50% of all masculine diminutive headwords in SSKJ (The Dictionary of Standard Slovenian Language).

1.2. Experimental and corpus studies on diminutives and morphological variation

Diminutiveness has already been extensively studied in the Slavic languages. Here we list only a few examples. Böhmerová (2011) analyzed the suffixal diminutive morphemes in Slovak and found that there are several different diminutive suffixes which are not equally frequent and productive, and similarly to Slovenian, some noun bases alternate in the usage of different diminutive suffixes, while others do not allow alternative options.

Bosanac et al. (2009) conducted field-work with native speakers of Croatian. The participants in the experiment were asked to write down the meanings of several diminutives in order to study the semantics of this category. The results obtained were categorized by six semantic/pragmatic features: 'small', 'affectionate', 'contextualized', 'lexicalized', 'pejorative', and, interestingly, 'large' as well. However, the feature 'large' was very rare and usually attributed by children.

Another interesting work was the comparative study of diminutives in Polish and Bulgarian, conducted by Manova and Winternitz (2011). They investigated double and triple diminutives in the two languages, and what constraints govern their formation. They also commented on the reliability of different language resources, which is also relevant for our work. They suggested that since diminutive forms are typically colloquial and highly productive, they are often not found in dictionaries or in corpora, which makes verifying their forms additionally challenging (Manova and Winternitz, 2011). We predicted that we would come across similar problems since we used a corpus of written Slovenian.

While we were unable to find research specifically comparing corpus and experimental data regarding diminutive formation, a study by Bermel et al. (2018) used a similar approach. They compared corpus frequencies with native speakers' judgements, but focused on nominal case form variation in Czech. Using a gap-filling and an acceptability rating task they obtained participants' judgments on the usage of rival forms of masculine inanimate nouns in the genitive and locative case (e.g., *čtvrtku* and *čtvrtka* are both acceptable genitive case forms of 'Thursday', while *ledě* and *ledu* are both locative forms of 'ice'). Results of both tasks correlated with corpus frequencies, i.e., the more frequent one of the forms was in the corpus compared to its rival, the likelier it was for participants to use it and the more favorably they ranked it.

2. The present study

Since different sources disagree about the exact number of male diminutive suffixes in Slovenian, we had to determine which of them we would use in our study and how we would classify them. We opted for a categorization into four separate groups of suffixes: -(*č*)*ek*, -*ič*, -*ičekand* -(*e*)*c*, which were selected based on our expectations of which suffixes would be most common and productive, as well as which of the suffixes were in competition when it came to the same noun stem.

We decided to group -*ek* and -*ček* together due to the fact that the latter is in many cases essentially the first form combined with palatalization of the last consonant in the stem (e.g., *otrok* 'child' becomes *otroček* 'small child') or is a double diminutive combining -(*e*)*c* with -*ek*, where the latter again causes palatalization of -*c*- /ts͡/ to -*č*- /tʃ͡/ (e.g., *čevelj-čeveljc-čeveljček* 'shoe' and two of its diminutives, where the intermediate –(*e*)*c* form can be seen) (Kavčič, 2021). Although they are not considered allomorphs by some (see Toporišič, 2000), we observed they often behave as such, as nouns can often take on one suffix, but not the other (e.g., *listek* 'small leaf' and *čeveljček* 'small shoe' are acceptable, but *listček* and *čevljek* are not). This is not always the case (e.g., *sin* 'son' allows both *sinek* and *sinček*), however nouns accepting both varieties are rare, and we decided not to use them in our study. We also categorized -*ec* and -*c* together

assuming that they were in a complementary distribution, as the *-e-* represents a schwa sound that is omitted in certain contexts.

Another common feature that had to be taken into consideration was double diminutiveness. In Slovenian, diminutive forms of a noun often occur with two diminutive suffixes rather than one, as seen with *-ček*. In the case of masculine diminutive forms, the suffix *-ič* is commonly combined with the suffix *-ek*, such as in *konjiček* 'little horse' or *fantiček* 'little boy'. We found that in many cases the double diminutive form is more frequently used than the form with a singular suffix, which is why we chose to treat the double diminutive *-iček* as a separate, productive suffix category.

One thing to note is that the choice of diminutive suffix can be influenced by the speaker's dialectal background, as evidenced by Koletnik's (2018) comparison of nominal diminutives in the Prlekija dialect dictionaries with SSKJ. While we acknowledge this fact, incorporating this factor into our experimental design would have introduced additional complexity. Given that the focus of our study was not to investigate this aspect, we did not intend to analyze variations in diminutive suffix choices based on participants'' dialects. Consequently, we decided to limit the demographic information obtained from our participants only to their region of origin, simply to illustrate the representativity of our sample. The regions of origin we chose were historical regions of Slovenia, as they roughly correspond to the borders of Slovenian dialects and are also intuitive enough for native speakers to know and report.

2.1. Main hypothesis

Our primary motivation was to undertake a conceptual replication of Makarova's (2014) work on attenuatives to test the applicability of a similar approach to another Slavic language and to a different morphological phenomenon. Another general aim was to compare two widely used methods in linguistics: corpus studies and experimental studies.

Considering that corpora gather varied texts in a specific language, serving as extensive repositories of language use, one might anticipate native speakers' preferences regarding word form choice in an experiment to reflect corpus frequencies. In our scenario, we anticipate that native speakers will mostly employ the diminutive suffixes for each noun that are most frequent or prevailing in the corpus. In an experimental setting, however, there are of course particularities affecting native speakers' behavior such as reduced spontaneity, learning effects and metalinguistic demands posed by the task. While we attempted to somewhat rectify these issues by randomizing the stimulus order for each participant and instructing the participants to rely on their intuition as native speakers, the impact of these factors on the participants' performance persists and must not be overlooked.

Nevertheless, we expect an overall alignment between the two approaches. In line with Makarova's (2014, 2016) findings, we hypothesize that our participants would mostly use the corpus-dominant or expected forms, as we termed them. Replicating Makarova's (2014) analysis, we compared two groups of responses: expected responses and other responses employing non-dominant suffixes, hypothesizing that the share of expected responses would exceed the share of other responses.

2.2. Variation across nouns and possible predictors

Given that the nouns we selected varied in terms of features that have been found to affect speakers' choices from a psycholinguistic perspective, such as animacy and word length, we decided to analyze these factors as predictors of the share of expected responses received.

One of the examined predictor variables was animacy – a grammatical feature that is common cross-linguistically and can affect morphology. For instance, animacy in some cases determines affix choice in languages such as Basque (Santazilia, 2013), where the addition of the morpheme *-ga*(*n*) distinguishes animate from inanimate nouns in some cases, and Persian (Sedighi, 2004), where the plural marking affix is *-an* for animate and *-ha* for inanimate nouns. In Slavic languages, including Slovenian, we find animate nouns exhibiting genitive-accusative syncretism while inanimate nouns are characterized by nominative-

accusative syncretism (Igartua & Santazilia, 2018). There is even evidence of animacy influencing the development of morphological complexity in some languages (Igartua & Santazilia, 2018).

Another variable we considered as a predictor was the noun bases' syllable number as a measure of word length. We found less cross-linguistic evidence for word length or syllable number influencing morphology. There is some allomorphy in Greek, where the deverbal nominal suffix is different when attaching to a monosyllabic or to a polysyllabic root, manifesting as *-simo* or *-ma*, respectively (Drachman et al., 1995; Drachman, 2005). We considered syllable number because word length is a variable commonly checked for its influence on linguistic processing. In our case, this meant contrasting 16 monosyllabic with 17 disyllabic noun bases in the share of expected responses received. Among all the nouns exhibiting diminutive suffix rivalry we considered, none had more than two syllables. Perhaps this alone could be indicative of a connection between syllable number and the degree of suffix variation a noun allows.

Finally, we observed to what degree the percentage of expected responses obtained could be predicted by a noun's expected suffix, i.e., the suffix most common in the corpus for that noun. An example of why this could serve as a predictor is the suffix –(*e*)*c*, which is the least common of the suffixes observed and is semantically more ambiguous, being used to refer to a great size or age as well, and to form nouns indicating an agent or an association with some characteristic (Vidovič Muha, 2011). While the suffix was corpus-dominant for some nouns, the participants may have used a more semantically transparent suffix in its stead. If that were the case, the nouns with the expected suffix –(*e*)*c* could have received less expected responses and the suffix category –(*e*)*c* could be a predictor of a smaller proportion of expected responses.

3. Method

3.1. Participants

We collected responses from 70 participants, 50 of whom were female (71.4%), 18 were male (25.7%), while 2 participants did not wish to identify with the first two options. The mean age was 24.3 years (*SD* = 4.58). While we did not obtain information on the participants' field of study or work, we assume most of them were students of social sciences and humanities based on where we shared our questionnaire and the participants' age, as 70.0% of them were in the age group 21-25. In terms of highest achieved education level, 51.43% of the participants had at least a bachelor's degree, while the rest had a high school degree – the latter were presumably all undergraduate students. Most of the participants were from the Styria (Štajerska) region and from Ljubljana, while not even one was from the Slovene Istria (Slovenska Istra) region (see Table 1).

	Region of origin								
	Styria	Ljubljana	Upper Carniola	Goriška	Inner Carniola	Lower Carniola	Prekmurje	Carinthia	Slovene Istria
N (%)	26 (37.1%)	20 (28.6%)	11 (15.7%)	5 (7.1%)	2 (2.9%)	3 (4.3%)	2 (2.9%)	1 (1.4%)	0 (0.0%)

Table 1. Number and percentage of participants by region of origin

3.2. Materials

As we intended to observe participants' suffix choice when forming nominal diminutives, our stimuli were (masculine) nouns in their base form without any diminutive morphology. We thus began by compiling a list of masculine nouns which we believed would have at least two possible diminutive forms. To compile an initial list of appropriate candidate stimuli, we consulted the dictionary and the corpus. The dictionary consulted was SSKJ – the Dictionary of the Standard Slovenian Language (*Slovar slovenskega knjižnega jezika*), available online (Fran Ramovš Institute of the Slovenian Language, n.d.), while the corpus used was

Gigafida 2.0. It is an up-to-date corpus of written Slovenian, and with over a billion words it is the largest Slovenian corpus (Gigafida 2.0, 2023).

Initially, our list consisted of 68 nouns. We then selected a subset of 46 nouns which met our selection criteria. Our selection criteria pertained to our grouping of the diminutive suffixes into four categories: *-ek*/*-ček*, *-ič*, *-ičekˌ* and *-c*/*-ec*. We considered a noun as attested with a particular suffix if it occurred with the suffix at least 5 times in the corpus – a threshold we chose arbitrarily. Each noun had to be attested in the corpus with at least two diminutive suffixes in order to be selected.

Another vital criterion was that at least two of the possible diminutive forms were used widely enough with a diminutive meaning and were not overshadowed by some secondary meaning. The semantics of the diminutives and whether they presented an issue for us was determined through our intuition and by observing the corpus data. A prime example of problematic secondary meanings is the noun *boben* 'drum' and its two possible diminutive forms – *bobenček* and *bobnič*. Both are acceptable and well attested in the corpus, however, they are both used to mean 'eardrum.' This secondary meaning is dominant compared to the meaning of a small drum. This sort of lexicalization of diminutives is quite common in Slovenian (Kavčič, 2021). Such a discrepancy between the associated meaning of the noun base and the diminutive form could influence the task of diminutive formation for our participants in an undesirable manner, so nouns of this type were excluded. Note that the mere existence of a secondary meaning did not present a problem from our perspective if the diminutive meaning was still common enough for at least two diminutive suffix categories. For instance, one diminutive of the noun *vrt* 'garden' is *vrtec*, which is acknowledged as a diminutive in SSKJ, but is nearly always used to mean 'kindergarten,' as can be observed in the corpus as well. This does not present an issue, however, as there are other diminutives in the form of *vrtek*, *vrtič*, and *vrtiček*, which are used to mean 'small garden' and have no secondary meanings.

After narrowing down our list of nouns according to these criteria, we were left with 46 nouns. We wished to narrow our list further, as we wanted the nouns exhibiting strong morphological rivalry as our stimuli. We made our selection by considering the frequency of a noun's usage in its diminutive form, as well as the ratio between the frequencies of the most frequently occurring suffix and the second most frequently occurring suffix. The former was termed 'total frequency' and it was calculated as the sum of *F1* and *F2*, where *F1* was the frequency of the most common suffix associated with a particular noun and *F2* was the frequency of the second most common suffix. The latter was termed 'competition ratio' and it was calculated as *F1* divided by *F2*. A low competition ratio (nearing 1) thus indicated that the frequencies of the two most common suffixes are fairly similar, while a high ratio indicated that one suffix was disproportionately more common than the other. Frequency was deemed important as we wanted nouns which are commonly used and heard in their diminutive form by our participants. The competition ratio was perhaps of even greater importance, as we wanted our participants to be faced with a decision regarding the suffix usage whenever they encountered a noun. If the competition ratio was high, meaning one suffix was disproportionately dominant over the others, we would expect that suffix to almost always be used by our participants, which would not be as informative for our research question.

Nouns were ranked based on their total frequency and competition ratio, and a mean of the two rankings was calculated. Thirty-three of the highest-ranking nouns based on the averaged ranking were chosen as our final target nouns. We illustrate this selection process with the noun *list* 'sheet of paper'/ 'leaf', which was our highest-ranking noun. The most common diminutive suffix attested with this noun was *-ič* with 12,465 attestations. The second most common suffix for *list* was *-ek* with 11,103 attestations. These frequencies represented *F1* and *F2* respectively. The total frequency of diminutives for the noun *list* was thus 23,568, while the competition ratio was 1.12. It ranked first in terms of its total frequency and second in terms of its competition ratio, meaning that its average rank was 1.5. This resulted in it being our highest-ranking noun, as mentioned earlier.

It is important to note that one of the main purposes of obtaining lemma frequencies of diminutive forms in the corpus was to determine which suffixes were most common in use and thus expected to be used for a particular noun. For each noun, the frequencies determined what we termed the expected suffix and the non-expected suffix(es). Returning to *list*, *-ič* would be its expected suffix as the most common suffix, while *-ek* (or any other suffix for that matter) would be categorized as a non-expected suffix. Note that we

did not use the term ‘unexpected’ suffix, as other suffixes may be just as acceptable and non-surprising, but we needed a term to distinguish them from the expected suffix category.

The noun stimuli were contextualized in sentences, which were mostly taken from the corpus, mimicking the approach of Makarova (2014). The sentences were sometimes adapted slightly, and in rare cases a sentence was made up, when the corpus did not supply us with an appropriate sentence for our task. There were no exact limitations to the length of our sentences, but all of them had at most two clauses and were syntactically not compvery complex. We provide one of our items (1) as an example:

(1) *Iz* *luknje* *je* *pokukal* *črv.*
out hole-GEN be-AUX peek-PAST worm-NOM
‘A worm peeked out of the hole.’

Each sentence contained precisely one of our target noun bases, and each target noun was featured in precisely one sentence. The target nouns were all presented in their dictionary form, without any case morphology, to exclude the effect of case on suffix choice. This meant they were all presented in the nominative case, exemplified by the noun *črv* ‘worm’ in (1), but inanimate nouns could additionally be presented in the accusative case, as in Slovenian this case form happens to match the nominative case form in the case of inanimate nouns only. Each sentence also contained one or two filler nouns, such as the noun *luknja* ‘hole’ in (1). These could be a noun of any gender, as long as it was not one of our target nouns and it was easily convertible into a diminutive. This meant that it had to be a concrete noun that did not sound unnatural in its diminutive form, e.g., using *luknja* ‘hole’ in its diminutive form *luknjica* ‘little hole’ sounds natural.

3.3. Procedure

We intended to observe whether native speakers of Slovenian would predominantly assign diminutive suffixes that are also expected suffixes for a given noun. Their task was thus to transform nouns in their base form into their diminutive form within the given sentence context. The thirty-three sentences were presented to our participants in the form of a questionnaire shared online mostly among our peers and their acquaintances through snowball sampling. Only native speakers of Slovenian above the age of 18 were allowed to participate. There was no upper age limit for participation.

All sentences were presented on the same page, and their order was randomized for each participant. They could scroll up, alter their responses, and reread the instructions at any time. The exact instructions were that they had to rewrite all the sentences, transforming all nouns into their diminutive forms. Additionally, the participants were instructed not to consult the dictionary or other people for help, but instead rely on what sounds natural to them, irrespective of what they think the diminutive would be in standard Slovene. Included in the instructions was also an example of a sentence (2a) and its transformation (2b):

(2) a. *Sova* *je* *za* *sabo* *pustila* *le* *pero.*
owl be-AUX behind itself leave-PAST only feather
‘The owl only left a feather behind.’
b. *Sovica* *je* *za* *sabo* *pustila* *le* *peresce.*
owl-DIM be-AUX behind itself leave-PAST only feather-DIM
‘The (little) owl only left a (small) feather behind.’

After obtaining responses from participants, we observed only the suffixes chosen for the target nouns, disregarding any alterations in the noun base. Similarly to Makarova’s (2014) approach, we compared the cumulative number of expected responses with the number of non-expected responses across all nouns. The expected responses were ones containing the suffix which was most common in the corpus for the given

noun, while any other suffix was counted among the non-expected responses. A 1-sample proportions test with continuity correction was performed comparing the share of both categories.

We conducted linear regression analyses to explore possible predictor variables of the percentage of expected responses. The categories of animacy (animate vs. inanimate), syllable number (monosyllabic vs. disyllabic) and the expected suffix category were considered as independent variables. As percentages are bound between 0 and 1, rendering them problematic as a dependent variable in a linear regression context, we performed a log transformation. For this purpose we needed to remove the noun *kavelj* 'hook' with 0.00% of expected responses, the log transformed value of which is undefined. When considering the expected suffix variable, we created dummy variables and then fitted a linear regression model separately for each suffix. Since multiple hypotheses were tested in this case, the *p*-values were adjusted using the Holm method.

4. Results

4.1. Main hypothesis

Out of the 2,294 valid responses we received across all nouns and participants, 1,234 contained an expected suffix based on corpus data, while the remaining 1,060 employed some other, non-expected suffix (Figure 1). A 1-sample continuity corrected proportions test proved the difference between these two shares to be statistically significant ($\chi^2 = 13.05$; $df = 1$; $p < .001$). The difference between the proportions, however, is not very pronounced, as expected responses constitute only 53.8% of all responses, indicating a considerable usage of alternative suffixes. The effect size was evaluated using Cohen's *h* and yielded a value of 0.15, indicating it is not substantial enough to even be considered a small effect. Although there is a noticeable trend suggesting that participants' responses lean towards the corpus-dominant suffix, the difference between expected and non-expected responses is nuanced, indicating the intricacies involved in suffix choice.

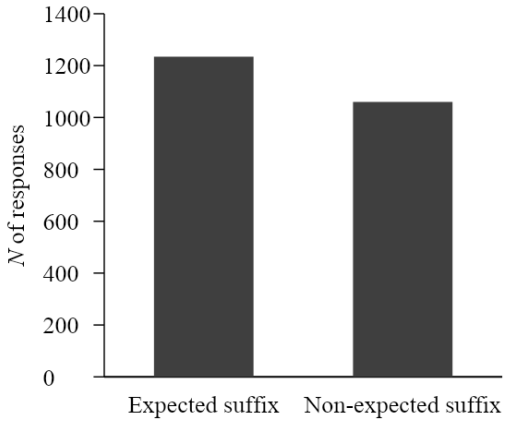

Figure 1. Number of responses containing an expected or a non-expected suffix

Among our target nouns we observed great variation in the distribution of the percentage of expected responses received, illustrated in Figure 2, which helps to explain the minuscule effect size. This variation called for further analysis.

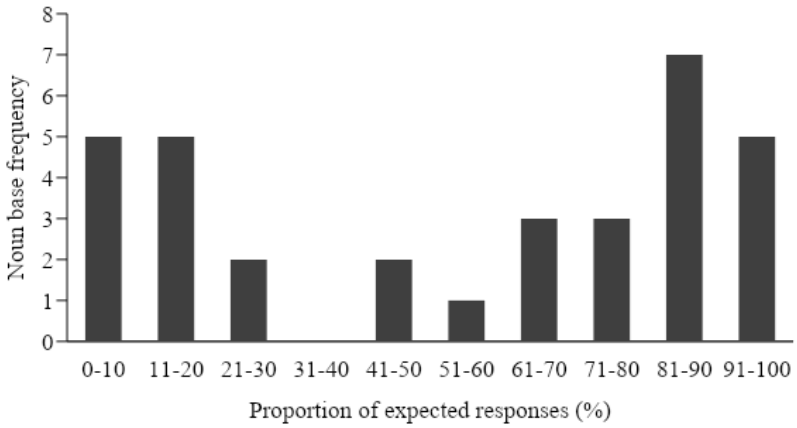

Figure 2. Frequency of noun bases across various classes of percentage of expected responses

4.2. Linear regression analysis

As our target nouns exhibited great variation in the share of expected responses received, we explored possible predictor variables of this outcome. In Table 2 we present the output of our linear regression models with log transformed percentages of expected responses as the dependent variable and animacy, syllable number and expected suffix as independent variables. In addition to the *p*-values, which were Holm-adjusted for the expected suffix models, we provide the adjusted R^2 statistic as a measure of goodness of fit.

Model	Predictor variable	*F*	*df*	*p*	$R^2_{adjusted}$
1	Animacy	0.32	1, 30	.58	-.02
2	Syllable number	0.85	1, 30	.36	-.00
	Expected suffix				
3	-(*č*)*ek*	6.89	1, 30	<.05	.16
4	-*ič*	1.19	1, 30	.28	.01
5	-*ičekk*	5.15	1, 30	.06	.12
6	–(*e*)*c*	55.31	1, 30	<.001	.64

Table 2. Outcomes of linear regression models with animacy, syllable number and expected suffix as predictors of the proportion of expected responses received

Animacy and syllable number were not predictive of the percentage of expected responses. Among the expected suffix categories, the suffixes -*ič* and -*ičekk* were not significant predictors, with -*ičekk* just falling short of statistical significance, –(*č*)*ek* marginally exceeded the threshold of significance as a predictor, while –(*e*)*c* was a significant predictor, accounting for 63.6% of the variance in the proportion of expected responses.

In Table 3 we provide a comparison of the diminutive suffixes in terms of their productivity among our target nouns, as indicated by corpus frequencies and the frequency of usage by our participants. The statistics highlight clear differences among the suffixes and could help explain their strength as predictors.

Suffix	Noun bases attested with suffix	Attestations in corpus	Frequency among participants' responses	min(p_{exp})	max(p_{exp})
-(*č*)*ek*	20	29,524	1283	79.7	100.0
-*ič*	23	25,947	196	11.8	66.2
-*ičekk*	18	41,264	725	13.0	90.0
-(*e*)*c*	14	3,700	90	0.0	11.4

Table 3. Diminutive suffixes compared by frequency of attestations, frequency among participants' responses, and the range of expected responses in percentages

Note. min(p_{exp}) refers to the lowest, and max(p_{exp}) to the highest percentage of expected responses among nouns in the category.

5. Discussion

The main aim of our study was to conceptually replicate Makarova's (2014) approach with Russian attenuatives, focusing on diminutive formation in Slovenian. The finding in her study, that native speakers' preferences in the choice of rival forms were in line with the expected form based on the corpus, was a result we anticipated as well. We hypothesized expected forms would represent a much larger share of responses than alternative forms.

Given the results, we cannot decisively reject the null hypothesis, that the overall number of expected responses and the overall number of non-expected responses are the same. While the proportions test showed a statistically significant difference between the percentages of the two groups, the effect size measure indicated that the difference is marginal. The small difference reaching significance level was likely due to the large sample size. The outcome of our experiment exhibited a pattern in line with our hypothesis, i.e., there was an overall tendency to use expected suffixes over non-expected, however, the difference between the proportions of the two categories was not as stark as initially anticipated.

Regarding the aim to compare two different approaches in linguistics, we detected a minor alignment between the two methods. This was unlike Makarova's (2014) study, where the data from the two approaches significantly coincided. We discuss possible reasons for the inconclusive alignment between corpus and experiment data in our study and the implications of these findings.

5.1. Expected suffixes as predictors

We begin by addressing the suffixes –(*č*)*ek* and –(*e*)*c*, which proved as significant predictors in our regression models and which lie at opposite ends of the spectrum of expected responses.

In addition to being very productive according to corpus data, the suffix –(*č*)*ek* is notable for being the most commonly assigned suffix in our experiment. Its high productivity was already reported by Vidovič Muha (2011), as she found –(*e*)*k* was used in over 50% of all masculine diminutive headwords in the SSKJ dictionary. Additionally, she noted the suffix is almost exclusively used to denote a small size of the referent or the speaker's affectionate attitude towards the referent, where the two meanings are usually not clearly separated (Vidovič Muha, 2011). Being transparent in meaning could contribute to it being one of the more preferred diminutive suffixes for Slovene speakers. This could explain why the share of expected responses was so high for noun bases which appeared most frequently with this suffix in the corpus.

While –(*č*)*ek* was associated with a high percentage of expected responses, the suffix category –(*e*)*c* was a predictor in the opposite direction. The six nouns with the expected suffix –(*e*)*c* were the lowest ranking nouns in terms of the proportion of expected responses. Given this outcome, the suffix ought to be examined in greater detail.

Firstly, the suffix –(*e*)*c* seems to be the least productive out of the four suffix categories we defined. As illustrated in Table 4, this suffix was the least frequently employed suffix by our participants and was the least common in terms of corpus attestations. At least for the nouns we observed, it seems to be less preferred as a diminutive suffix.

Vidovič Muha (2011) noted that in addition to being a diminutive suffix, the suffix *-ec* is also used to denote a great size or age of something, and this can also carry a negative connotation. Some examples are *dedec* 'older man' and *bogatinec* 'rich man'. But Vidovič Muha emphasized that nouns of this type are losing their pejorative/large meaning, and instead gaining a dialectal or archaic quality. Furthermore, the suffix is also used to derive nouns from a verb, e.g., *ropotec* 'wooden ratchet' from *ropotati* 'to rattle', usually to denote that someone or something is the agent, e.g., *mislec* 'thinker', or that it has a certain characteristic, e.g., *belec* 'white person', or some other association, e.g., *Kranjec* 'Carniolan (the one from Carniola)'. Because the meaning of the suffix *-ec* is clearly not very transparent, another, more semantically

transparent diminutive suffix is usually added to it when deriving diminutives, thus creating double diminutives. An example of this is the derivation of *hribček* 'small hill', by first adding *-ec* to *hrib* 'hill', and then the suffix *-ek*, which triggers the palatalization of *-c-* into *-č-* (Vidovič Muha, 2011).

We touch upon the suffix *-ček,* which we categorized together with *–ek*. In the literature, it has been analyzed as a separate diminutive suffix (Toporišič, 2000: 185; Vidovič Muha, 2011) or an allomorph of the suffix *-ek* in some cases (Vidovič Muha, 2011). Kavčič's (2021) analysis claimed it is neither of the two under any circumstance, but rather a double diminutive, which combines the suffix *–(e)c* and the suffix *-ek*, where the latter triggers the aforementioned palatalization of *-c-* /ts/ into *-č-* /t͡ʃ/. As mentioned, Vidovič Muha (2011) acknowledged this type of double diminutive derivation in cases like *hribček*, but what distinguishes the claim by Kavčič (2021) is that all *-ček* diminutives are formed this way and that the intermediate form only containing *–(e)c* is always implicitly present in the formation of the double diminutive, even if it is not overtly manifested in the language. An example from Kavčič (2021), featured also in Vidovič Muha (2011), where this intermediate form exists is the aforementioned *hrib-hribec-hribček* ('hill' and its diminutives, respectively). An example of a non-existent intermediate form is *balon-*balonc-balonček* ('balloon' and its diminutives, respectively) (Kavčič, 2021), which Vidovič Muha (2011) would have analyzed as the suffix *-ček*, an allomorph of *-ek*, attaching to noun bases ending in a sonorant.

To summarize, the suffix *-ec* exhibits a lack of semantic transparency (Vidovič Muha, 2011) and double diminutive derivation with *–(č)ek* may arise as a result of that (Kavčič, 2021; Vidovič Muha, 2011). This is exemplified by our data, as most of the non-expected responses received by *–(c)ec* nouns employed the suffix *–ček*, and this could be the leading cause of the low numbers of expected responses among *–(e)c* nouns.

5.2. Experimental approach compared to corpus data

We now address the specific limitations surrounding the two methods used, and their possible influence on our results. While corpora, as large linguistic datasets, allow for large-scale analyses of language use, their characteristics, by which we mean mostly their source material, can limit this use to a certain domain or time. In the case of Gigafida 2.0 (Krek et al., 2020), it is a written corpus, and while it is quite up to date with the most recent sources written in 2018, it is important to bear in mind it is a corpus of standard Slovene. It was specifically filtered for non-standard usage, as it was intended to be used for research on standard Slovene (Krek et al., 2020). Diminutives, however, can be highly colloquial in nature, and due to their productivity in Slavic languages the frequencies of various diminutive forms can be under- or misrepresented in written corpora (Manova & Winternitz, 2011).

Perhaps this discrepancy between standard Slovene corpus data and non-standard registers could also help explain why the suffix *–(e)c* was not assigned as often as expected. The frequency of forms employing this suffix might have been overestimated in the corpus. The double diminutive forms using *–ček* derived from *–(e)c* might already be more prevalent in everyday speech, but this fact is not (yet) reflected in the standard varieties of the language. Some written sources, mostly works of fiction, might also use a distinct language variety intentionally including more archaic words or word forms, thus contributing to the gap between information in written corpora and information obtained in experiments.

It is of course necessary to note that we did not record colloquial speech in our experiment. Our task was in written form with most stimuli gathered and adapted from the corpus. Attempting to replicate Makarova's approach (2014), our task was set up to ensure the same context as in the corpus and with that encourage the use of the expected suffix. Yet in many cases this was not the observed result. Designing a written task to motivate the usage of a similar form as in the source text did not prove to be enough in our case. In addition to the characteristics of the corpus, the characteristics of the experiment need to be considered.

Our instructions to the participants were to trust their native speaker intuition irrespective of what the standard form should be. Some participants might have taken this to heart to a greater degree and utilized forms similar to what they would use in a colloquial setting. But as all participants differ, some were likely influenced by the instructions to a lesser extent. Perhaps with some participants' less expected responses,

there was a metalinguistic component at play. For instance, they might have chosen *–ček* over *–(e)c* to solidify the diminutive meaning, knowing the task was to form diminutives and that they were being observed in an experimental setting. Instructions are instrumental in guiding participants' behavior, and as with any offline task, metalinguistic (or more generally metacognitive) factors are impossible to avoid. These are just some of the particularities that accompany experimental approaches and distinguish them from corpus studies.

In any case, the results indicate that the relationship between corpus and experimental data is not straightforward. As with any linguistic phenomenon, diminutive formation, too, is complex and influenced by various factors, which we could not fully explore or control in our study.

5.3. Limitations and future directions

Future experimental studies employing a similar design should improve on the task. While it is generally good to provide a sentence context for the stimuli, as well as the filler nouns, it was perhaps unnecessarily demanding for our participants to rewrite the entire sentence with each item. Providing the sentence context but then instructing participants only to write down the diminutives of the nouns would have made participation less taxing. It would have also minimized the number of errors, such as not recognizing a word as a noun or overlooking the noun altogether, and thus not attaching a diminutive suffix to it.

Additionally, each target noun could have been used in multiple sentence contexts to avoid any bias that a single sentence context might impose. Using data from the study on attenuatives, Makarova (2016) analyzed the relative influence of frequency, morphology, and lexical context on the use of attenuative affixes. And while her analysis indicated lexical context to be least predictive of the use of attenuatives, it is important to note that the sentence context can sometimes influence affix choice. We did not consider this, nor did we perform any similar analysis. However, future studies ought to account for this potential factor.

Another aspect we did not consider was what share of the corpus attestations of diminutive forms was actually employing the form in a diminutive use, as opposed to a non-diminutive use of the word form. We acknowledged that diminutive forms can be further lexicalized and gain secondary meanings, but we did not thoroughly analyze to what extent the diminutives of our nouns were actually used in a diminutive sense in the corpus. Additionally, some diminutives are used as proper nouns, e.g., geographical features and town names like *Gradec* 'Graz' (rarely in the sense of 'small castle'), and, albeit rare with our sample of nouns, family names like *Grozdek* (more commonly used for 'small grape'). Future research should analyze the prevalence of diminutive use for the diminutive forms when preparing stimuli, or perhaps consider it when analyzing the results.

Using a similar approach with different corpora as a reference point for the prevalence of rival forms could also provide additional insight into what text genre or language variety better aligns with native speaker preferences in an experimental setting.

Finally, as with most experiments, a greater representativeness of the sample of participants would have been desirable. Future studies could thus include a more diverse sample in terms of the dialects spoken by the participants, as well as their age. The participants' dialect and age could then also be analyzed as predictor variables.

6. Conclusion

In our study we employed both a corpus analysis and an experimental approach, two commonly used methods in linguistics, to assess their alignment regarding the preferred diminutive suffix of each noun. We attempted to replicate Makarova's (2014) study on attenuatives, anticipating participants' responses to generally reflect corpus tendencies. In contrast with her study, our results indicated only a subtle alignment between the two approaches. Further analysis of our data showed the expected suffix of a noun to be a relevant factor in the share of expected responses received. Specifically, -(*e*)*c* and to a lesser extent –(*č*)*ek* emerged as predictors. These findings were discussed in the light of the characteristics and limitations of the two approaches, stressing the importance of their consideration in a study's design. We thus suggest further research explores different corpora, and improves on the task and sample, to better assess the complex process of diminutive formation.

References

Baayen, R. H., Endresen, A., Janda, L. A., Makarova A., & Nesset, T. (2013). Making choices in Russian: pros and cons of statistical methods for rival forms. *Russian Linguistics, 37*(3), 253–291. doi: https://doi.org/10.1007/s11185-013-9118-6

Bermel, N., Knittl, L., & Russell, J. (2018). Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguistics and Linguistic Theory, 14*(2), 197–231. doi: https://doi.org/10.1515/cllt-2016-0032

Böhmerová, A. (2011). Suffixal Diminutives and Augmentatives in Slovak. A Systemic View with Some Cross-Linguistic Considerations. *Lexis, 6*. doi: https://doi.org/10.4000/lexis.429

Bosanac, S., Lukin, D., & Mikolić, P. (2009). *A Cognitive Approach to the Study of Diminutives: The Semantic Background of Croatian Diminutives.* University of Zagreb, Faculty of Humanities and Social Sciences.

Drachman, G., Kager, R., & Malikouti-Drachman, A. (1995). Greek Allomorphy: An Optimality Account. In M. Dimitrova-Vulchanova & L. Hellan (Eds.), *Papers from the First Conference on Formal Approaches to South Slavic Languages, October 1995* (pp. 345–361).

Drachman, G. (2005). A note on 'shared' allomorphs. *Journal of Greek Linguistics, 6*(1), 5–38. doi: https://doi.org/10.1075/jgl.6.04dra

Fran Ramovš Institute of the Slovenian Language. (n.d.). *SSKJ: Slovar slovenskega knjižnega jezika [Dictionary of the standard Slovenian language]*. Retrieved from https://www.fran.si/

Gigafida 2.0: Corpus of Written Standard Slovene, viri.cjvt.si/gigafida, accessed on January, 2023.

Igartua, I., & Santazilia, E. (2018). How animacy and natural gender constrain morphological complexity: Evidence from diachrony. *Open Linguistics, 4*(1), 438–452. doi: https://doi.org/10.1515/opli-2018-0022

Kavčič, K. (2021). *Typology of Slovene and English Diminutive Suffixes in the Framework of Distributed Morphology* [Master's Thesis, University of Ljubljana]. Repository of the University of Ljubljana. https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=130542

Koletnik, M. (2018). Samostalniške manjšalnice v prleških narečnih slovarjih. *Jezikoslovni Zapiski, 23*(1), 23–39. doi: https://doi.org/10.3986/JZ.23.1.6841

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3340–3345). European Language Resources Association.

Makarova, A. (2009). *Psycholinguistic evidence for allomorphy in Russian semelfactives*. Master's thesis, University of Tromsø. https://munin.uit.no/handle/10037/2377

Makarova, A. (2014). *Variation in pri- and pod- attenuatives in Russian* (Version 2.0) [Data set]. DataverseNO. doi: https://doi.org/10.18710/TP3TIY

Makarova, A. (2016). Variation in Russian verbal prefixes and psycholinguistic experiments. In T. Anstatt, A. Gattnar & C. Clasmeier (Eds.), *Slavic languages in psycholinguistics: Chances and challenges for empirical and experimental research* (pp. 113–133). Narr Francke Attempto.

Manova, S., & Winternitz, K. (2011). Suffix order in double and multiple diminutives: with data from Polish and Bulgarian. *Studies in Polish linguistics, 6*(1), 115–138.

Santazilia, E. (2013). Noun Morphology. In M. Martínez-Areta (Ed.), *Basque and Proto-Basque: Language-Internal and Typological Approaches to Linguistic Reconstruction [= Mikroglottika 5]* (pp. 223–281). Peter Lang.

Sedighi, A. (2004). Animacy: The overlooked feature in Persian. In M.-O. Junker, M. McGinnis & Y. Roberge (Eds.), *Proceedings of the 2004 annual conference of the Canadian Linguistic Association* (pp. 1–12). https://cla-acl.ca/pdfs/actes-2004/Sedighi-CLA-2004.pdf

Sicherl, E. (2012). Slovene nominal diminutives and their English equivalents: A comparison. *ELOPE: English Language Overseas Perspectives and Enquiries, 9*(2), 53–63. doi: https://doi.org/10.4312/elope.9.2.53-63

Sicherl, E. (2013). Diminutive Nouns and Verbs in Slovene Compared to Their English Equivalents. *Slovene Linguistic Studies, 9*, 145–162. doi: https://doi.org/10.17161/SLS.1808.11435

Sicherl, E., & Žele, A. (2011). Nominal diminutives in Slovene and English. *Linguistica, 51*(1), 135–142. doi: https://doi.org/10.4312/linguistica.51.1.135-142

Toporišič, J. (2000). *Slovenska slovnica* (4th ed.). Obzorja.

Vidovič Muha, A. (2011). *Slovensko skladenjsko besedotvorje.* Znanstvena založba Filozofske fakultete Univerze v Ljubljani.