Ksenija Mišić[a.b], Sara Anđelić[b], Lenka Ilić[b], Dajana Osmani[b], Milica Manojlović[b], and Dušica Filipović Đurđević[a,b]

[a]Department of Psychology, Faculty of Philosophy, University of Belgrade
[b]Laboratory for Experimental Psychology, Faculty of Philosophy, University of Belgrade
ksenija.misic@f.bg.ac.rs

UDC 811.163.41'373.4
81'37::159.956.2
612.821.8
DOI: 10.19090/slm.5.1

AN OPEN DATABASE OF SENSES FOR SERBIAN POLYSEMOUS NOUNS, VERBS, AND ADJECTIVES[1]

Abstract. The goal of this study was twofold. On the one hand, we aimed to obtain the number of senses estimations for polysemous nouns, verbs, and adjectives, as well as frequencies of occurrence for each sense based on native speakers' intuitions. From that, we derived multiple quantifications to describe each word. We determined what the senses of each word are by allowing participants to list everything they recall a word could denote. We categorised the responses and counted the senses per word. On the other hand, we aimed to investigate the reliability of native speakers' intuitions. The novelty in our approach was that multiple researchers who categorised the responses were also treated as participants in the study. In this paper, we present the database and explore the multiple-coder approach. This process generated data on various levels that could be useful for psycholinguistic research. We share both qualitative data – participants' individual responses and classification into senses by our coders, and quantitative data – number of senses counts, sense probabilities, entropy, redundancy, familiarity, and concreteness measures. All data are freely available at https://osf.io/ukcrg/.

Key words: lexical ambiguity, polysemy, open database, number of senses, entropy, redundancy, dictionary, semantic intuitions, part of speech

1. Introduction

From the very beginning, researchers of the mental lexicon posed the question of how words with multiple interpretations i.e. ambiguous words, were represented (Rubenstein et al., 1970). Particular focus was given to polysemes, i.e. words with multiple related senses, such as the word *paper* (writing paper / wrapping paper / scientific paper), and homonyms, i.e. words with multiple unrelated meanings, such as the word *bank* (financial institution / side of the river). Numerous studies addressed factors that influence the processing of polysemes and homonyms (for reviews see Eddington & Tokowicz, 2015; Rodd, 2020). Here, we focus on polysemous words. For example, paper as a material used for writing is semantically related to paper used for gift-wrapping, and the two are associatively related to the published paper. Despite significant advances in understanding the processing of polysemous words, the enumeration of the senses and the estimation of their frequencies remained a challenge for linguists and psycholinguists alike.

This chapter presents an attempt to estimate the number of senses (NoS) and to describe the distribution of sense probabilities for a group of Serbian polysemous nouns, verbs, and adjectives. We approached this problem from a psychological point of view, aiming to count senses and estimate their frequencies based on semantic intuitions of native speakers, whilst being aware that our classification of polysemous words may not perfectly align with the current linguistic classifications of ambiguous words.

[1] This research is funded by the Ministry of Education and Science of the Republic of Serbia (grant numbers: 179033 and 179006), and the Ministry of Education, Science and Technological Development of the Republic of Serbia (grant numbers: 451-03-9/2021-14/ 200163, and 451-03-68/2022-14/200163).

More novel, we also shine a spotlight on the variability among the speakers' intuitions and we do so by approaching the coders (researchers who performed the coding) as participants.

1.1. Describing senses

There are several approaches that can be used to estimate NoS and sense probabilities, such as relying on the currently available dictionaries, speakers' intuitions, or computational methods to extract data from the corpus. The empirical tests revealed that each of these methods has its advantages or disadvantages (for a detailed review see Filipović Đurđević & Kostić, 2017). In the following text, we present the main points.

The most recent approach relies on computational models based on the distributional hypothesis (Harris, 1954) and vector-based semantics to extract various descriptions of ambiguity from the corpus. The earliest distributional models, such as HAL (Lund & Burgess, 1996) and LSA (Landauer & Dumais, 1997) were not able to separate senses from each other but instead produced a blend of senses for each word, obscuring the information on NoS, sense frequencies, and sense relatedness. However, when second-order co-occurrence vectors were implemented (Schütze, 1998), sense separation could be achieved. These models led to a series of measures based on corpus data that describe the ambiguity of words (Filipović Đurđević et al., 2009; Hoffman et al., 2013; Jones et al., 2012). Some of the later developments, such as the TOPIC model (Griffiths et al., 2007; see also BEAGLE; Jones & Mewhort, 2007), could extract variation in semantics within ambiguous words, as well. However, some did not take into account the subtleties incurred by variation in meaning (e.g., GloVe; Pennington et al., 2014). A more successful model for word sense disambiguation was based on the WordNet database (Miller, 1995; for Serbian: Krstev et al., 2004). The structure of the database offers the possibility for more precise disambiguation. The technological development that followed led to models based on neural networks (predict models as opposed to count models; Baroni et al., 2014; Mandera et al. 2017) such as word2vec models CBOW and skip-gram (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) and FastText (Mikolov et al., 2019). These models are usually referred to as bag-of-words models, which were not developed in a way that could separate different senses of the words as the vectors they produce are not interpretable. Some measures could be extracted, but little specific information is available. Finally, the most recent type of models is based on transformers – also called Large Language Models (LLM). Instances of LLMs such as BERT (Devlin et al., 2019; for Serbian: Ljubešić & Lauc, 2021), ELMo (Peters et al., 2018; for Croatian: Ulčar & Robnik-Šikonja, 2019), or their more famous counterpart GPT (OpenAI, 2023), learn from the sequential context that the word appears in. Compared to previous models that bundled all contexts together, this led to a more precise insight into each sense of the ambiguous word. This contextual learning should be able to aid disambiguation and consequently obtain information on the properties of ambiguous words. However, the results are not straightforward. Whereas some research found that polysemy detection, differentiation between polysemous words regarding NoS (Yenicelik et al., 2020), and even sense disambiguation is possible within transformer models (Soler & Apidianaki, 2021), some claim that structures that are present in these pre-trained models are so seamless that information extraction is impossible (Wiedemann et al., 2019). The issue with these models is that the more advanced and better at disambiguation, the more technically complex they are. For example, BERT has between 100 and 400 million parameters, depending on the version. This leads to high complexity so the information regarding ambiguity can be difficult to obtain and can require both advanced coding knowledge and high-performing computing equipment. This can make computational techniques somewhat inaccessible for researchers working in psycholinguistics with linguistic or psychological training.

The earliest method of determining the number of senses relied on the available dictionaries (Gernsbacher, 1984; Jastrzembski, 1981; Klein & Murphy, 2001; Rice et al., 2019; Rodd et al., 2002). Within this approach, the NoS was determined by counting the descriptions listed under the given dictionary entry, whereas the relatedness among senses was determined by analysing the structure of descriptions which was imposed by the lexicographers. For example, dictionaries traditionally list all polysemous words' senses within one entry, whereas homonymous meanings are listed as separate entries. This enables simple but useful and easily accessible binary information (related/unrelated). Additionally, within a single entry, based on relatedness, the senses are enumerated in a cluster-like fashion (e.g. 1a, 1b, 2a, 2b etc). There are multiple advantages to using dictionaries. For example, they are typically available in many languages and bring information that is easily accessed. Additionally, this information is provided by highly trained experts in lexicography and is thus rooted in relevant linguistic knowledge. However, using dictionaries to estimate ambiguity comes with disadvantages, as well. For example, the available dictionaries are rarely exhaustive and sufficiently up-to-date to compete with speakers' mental representations, especially for languages with a small population of speakers. Although the languages with large enough speaker populations (such as Mandarin Chinese or English) are better represented in this respect, the dynamics of language change surpasses the speed of updating the dictionaries (Gernsbacher, 1984; Lin & Ahrens, 2005). This is further exacerbated in languages with smaller pools of speakers and consequently fewer linguistic resources, limiting the languages in which ambiguity can be researched.

To overcome the limitations of available dictionaries, the researchers introduced a method which relied directly on native speakers' intuitions. Within this approach, the information about different senses is collected from native speakers. Sometimes, the dictionary descriptions would serve as the starting point and the participants could provide familiarity ratings for each of them and introduce new senses/meanings in case the list was not comprehensive enough (Armstrong et al., 2012). In other studies, participants would sort phrases into categories that could either be predefined by the researchers or added by the participants (Lopukhina et al., 2018). The least directive approach is the one where participants are presented with a word and are instructed to list all of the senses that they can remember – the total meaning metric approach (Azuma, 1996; Filipović Đurđević & Kostić, 2017; Millis & Button, 1988). In the following step, the collected responses are categorized by the researcher. The NoS is represented by the number of categories, whereas the sense frequency is represented by the number of participants who listed the given category (and transformed to relative frequencies, i.e. probabilities). The question of dominance of one sense (the highest probability) or the balance of all probabilities (equity) proved to be of theoretical and empirical interest (Filipović Đurđević & Kostić, 2023; Foraker & Murphy, 2012; Gilhooly & Logie, 1980; Klepousniotou et al., 2008; Twilley et al., 1994).

In spite of the advantage of introducing new information rooted in contemporary language use, the total meaning metric approach is very time-consuming both for participants and for researchers (Armstrong et al., 2012: 1107). The task is demanding for participants, as they need to list many senses for many words. In this process, due to fatigue, they might not report all the senses that they know (especially the senses with lower frequencies). Despite that, a small number of participants (i.e., 20 per word) can produce a total of thousands of responses across the word sample, making the job of detecting categories for each word and classifying responses very time-consuming for researchers and prone to inconsistencies. Finally, this approach does not offer information on sense relatedness. To get more informative data, researchers must conduct a separate study, where participants judge the relatedness of categories, adding more time and effort to the already arduous task of response classification. However, one of the weakest points of this approach is the process of sense categorization which is at risk of subjective biases.

With all this in mind, we can conclude that the estimation of the number of senses and their frequencies remains an open question, especially for languages with less available resources such as Serbian.

1.2. Current study

In this study we set two broad goals. On the one hand, we aimed to create and share a freely available database of speakers' intuitions on the senses of Serbian polysemous nouns, verbs, and adjectives. In addition to collecting the speakers' intuitions, we aimed to provide the categories of similar descriptions, thus estimate the number and relative frequencies of senses. The approach of total meaning metric (i.e. collecting senses by prompting native speakers to freely list all of the familiar senses) was successfully employed previously in various studies concerning ambiguity phenomena, in languages including Serbian (Azuma & Van Orden, 1997; Filipović Đurđević, 2019; Filipović Đurđević & Kostić, 2017, 2023; Hino & Lupker, 1996; Klepousniotou & Baum, 2007; Mišić & Filipović Đurđević, 2022a; Millis & Button, 1989). The predictive power of various numerical descriptions of polysemy, such as NoS, entropy, and redundancy of the sense probability distribution has been demonstrated in previous psycholinguistic studies (Filipović Đurđević & Kostić, 2008; 2023; Filipović Đurđević, Đurđević & Kostić, 2009; Rodd et al., 2002). However, many questions in relation to the process of obtaining the values of the numerical descriptions remained open. The first question within the first goal that we will address in this paper is the choice of strategy in creating the categories of senses. We will compare the estimations obtained by relying on dictionary categories with those that are obtained by allowing the speakers' intuitions to guide the creation of novel categories. By doing so, we will contribute to the ongoing debate on the best practices (Armstrong et al., 2012; Lopukhina et al., 2018). The second question that we will address is the relation of different PoS categories with respect to the effects of strategies in ambiguity estimation. We will therefore compare nouns, verbs and adjectives. There are several theoretical reasons why the inclusion of multiple PoS is important. Most of the findings on polysemy, and ambiguity in general, were observed on a mix of noun and verb lemmata, or the meanings of the words were from two different PoS. In most of the influential papers in the field, this was rarely explicitly reported in the methods section and was mostly visible through stimuli lists (Azuma & Van Orden, 1997; Hino & Lupker, 1996; Klepousniotou, 2002; Klepousniotou et al., 2008; Millis & Bution, 1989; Rodd et al., 2002; Twilley et al., 1994). Occasionally, the issue of PoS was raised as an issue for future research in empirical work (Klepousniotou et al., 2008; Lopukhina et al., 2018) and in systematic reviews (Eddington & Tokowicz, 2015), where it is stated that this is both a methodological issue and an issue of theoretical interest. Prior research in Serbian controlled for this, exploring only polysemous and homonymous nouns (Anđelić & Filipović Đurđević, 2023; Filipović Đurđević, 2019; Filipović Đurđević & Kostić, 2017; 2023; Mišić & Filipović Đurđević, 2022a; 2022b). In the current research, we will address the generalizability of those findings to adjectives and verbs. To summarize, within the first broad goal, we will collect speaker intuitions, and categorize them in two ways – by strictly abiding by the dictionary and by allowing the addition of new categories. Based on the two categorization strategies we will derive NoS, Entropy, and Redundancy of sense probability distribution of polysemous nouns, verbs, and adjectives.

Our second goal is to investigate the reliability of the estimations in relation to the subjectivity bias in sense categorization. In other words, we want to address the variability in sense categories that is incurred by the semantic intuition of the person performing the categorization. We will therefore compare the numerical descriptions obtained from categorizations conducted by different researchers (coders). According to linguistic theories, polysemous senses would be harder to categorize due to their mutual relatedness (Gortan-Premk, 2004). Some even claim that unlike homonyms, which are represented as separate meanings in the mind, polysemes are merely contextual variations of a single mental representation (Frisson & Pickering, 1999; Frisson, 2009). However, some claim that both polysemous senses and homonymous meanings are stored as separate representations (Klein & Murphy, 2001; 2002). Therefore, the lack of consensus among the coders would also have theoretical consequences. Some previous research suggests that there is a high level of coherence among the coders (Twilley et al., 1994; Azuma, 1996). However, there are studies that suggest the opposite (Armstrong et al., 2012). Therefore, in this study, we will not only compare the categorizations based on the dictionary and those based on speakers' intuitions but also compare the categorizations based on the intuitions of different speakers.

2. Method

2.1. Initial word selection

Our goal was to select 100 polysemous words per part-of-speech category – 100 nouns, 100 verbs, and 100 adjectives. We chose 100 for two reasons. One reason was to collect a number of words that would allow sufficient variation in NoS, word, frequency, word length, etc. Another reason was to keep the further necessary work feasible for the team. The first step then was to sample a larger number of words from the existing databases and/or a dictionary, for our final samples to have satisfactory distributions of the number of senses and word frequency. For nouns, we relied on an open database of 2100 Serbian nouns (Popović

Stijačić, 2021; Popović Stijačić & Filipović Đurđević, in preparation) which contains norms on multiple lexical variables, aiming to cross information between that dataset, and the dataset we present in this paper. For verbs and adjectives, such a database was not available. Those words were selected from Matica Srpska's Dictionary of the Serbian Language (Vujanić et al., 2007), one of the most up-to-date dictionaries available for Serbian. Having in mind that all senses of polysemous words are presented within one dictionary entry (compared to homonyms where each meaning constitutes a separate entry), we set the criterion of selecting only words with a single entry and two or more listed senses within one entry. Such is the word *bronza* (eng. *bronze*) presented in Table 1. The second selection criterion was the familiarity of the word – we aimed to avoid words that were unfamiliar to the average speaker. In the first pass, the authors selected nouns, verbs, and adjectives with at least two senses, taking notes of the sense count from the dictionary (dictionary NoS). Following the criteria, we sampled 474 nouns, 860 verbs, and 420 adjectives. For each word, the data on lemma frequency was taken from the Frequency Dictionary of the Serbian language (Kostić, 1999).

Lemma		Sense description
bronze	Sense 1	An alloy of copper and tin (possibly with traces of other metals)
	Sense 2	Bronze bell, usually put on the cattle

Table 1. Example of a polysemous word with two senses adapted and translated from the Matica Srpska's *Dictionary of the Serbian Language*

In the next step, 100 words were selected from each PoS category. The limit of 100 words was decided upon to keep the subsequent steps feasible for the team, as many steps required manual coding. During the selection, we made an effort to achieve as large as possible a range of dictionary NoS and lemma frequencies within each PoS category, while keeping the ranges comparable across PoS categories. As these two measures are usually correlated (words more frequently in use usually carry multiple senses; Zipf, 1945), the final word selection had to be done in such a way as to avoid this correlation. Therefore, we created bins of NoS values (words with NoS between 10 and 20 being rarer within our selections constituted one bin) and selected a range of frequency values for each value of NoS. For example, in a selection of words that have three senses, we selected words with frequencies 1, 2, 3, 5, 6, 7, 9, 10, 11, and 12 (if we assume an imaginary frequency scale that ranges from 1 to 12). This wider range of frequencies per NoS value allowed for avoiding correlation between NoS and frequency within our three groups of words. Still, on average, there are no large numerical differences between average frequencies per NoS value.

Using this method, we selected 100 nouns and verbs, whereas for adjectives, our final selection contained 108 words (to enable counterbalancing in three lists according to grammatical gender). During the final verb selection, there was an error which led to the inclusion of a word with a frequency value 9.7 standard deviations away from the mean. This word was removed from the analyses, reducing the verb sample to 99 words. Descriptive values of all three group samples are presented in Table 2; information per word is presented in Supplementary material (available at https://osf.io/ukcrg/).

	PoS	M	SD	Min	Max	n
Dictionary NoS	adjectives	6.56	3.81	2	20	108

	PoS	M	SD	Min	Max	n
	nouns	5.78	3.20	2	17	100
	verbs	5.71	2.70	2	12	99
Word frequency	adjectives	231.11	270.85	13	1238	108
	nouns	364.12	494.05	11	2387	100
	verbs	206.84	263.80	2	1526	99

Table 2. Frequency and Dictionary NoS values for the three final samples of words

2.2. Sense production task

As we aimed to collect a list of senses known by the population of Serbian speakers, our final selection of 100 nouns, 100 verbs, and 108 adjectives was presented to groups of participants. Each PoS word selection was split in half (50 verbs/nouns or 54 adjectives per list) to avoid the participants' task being too long. The sense production task was conducted following Azuma's (1996) and Filipović Đurđević and Kostić's (2017) procedure of total meaning metric. We instructed participants to read the word carefully and then list all meanings of the word they could recollect. To illustrate the task, the instructions included an example, but we avoided giving detailed instructions, as we wished to allow participants to freely write whatever came to their minds. In addition to this, participants provided word familiarity and concreteness ratings, both on a 7-point Likert scale prior to listing senses. Ratings were given prior to the sense production task to avoid considering various word senses to affect both ratings. Note that the ratings were given for the lemma and not for each particular sense (see Filipović Đurđević & Kostić, 2017; Anđelić & Filipović Đurđević, 2023). Descriptive values of average familiarity and average concreteness ratings and the number of participants per word are presented in Table 3. A full list of listed senses and instructions (in Serbian) is available in the Supplementary material (https://osf.io/ukcrg/).

	PoS	M	SD	Min	Max	n
Word concreteness	adjectives	5.06	0.94	2.53	6.80	108
	nouns	5.11	1.39	1.59	6.94	100
	verbs	5.12	0.77	3.00	6.53	99
Word familiarity	adjectives	6.58	0.29	5.75	7.00	108
	nouns	5.62	0.75	3.71	6.89	100
	verbs	6.57	0.30	5.42	7.00	99

Table 3. Descriptive values of familiarity and concreteness ratings, the number of listed senses per word and the number of words rated

2.3. Categorising senses

The following step in this study was the estimation of the number of senses as well as the frequency of each of the senses, based on the participants' responses. In order to complete this step, we first needed to categorize the responses into coherent groups (senses). The participants provided a total of 17,596 responses in the sense production task across all three PoS. In order to make the classification feasible in a reasonable amount of time, multiple coders were involved in the process of categorization. The six authors were divided into two groups: main coders and control coders. The main coders team consisted of four members. Each of the four received one-quarter of each PoS word list. In order to avoid a single coder only coding words with few senses and another coding only words with many, we made four lists with numerically roughly equal ranges of dictionary NoS. Each member of the main coders' team classified only one of the lists. However, as the lists did not overlap, there was a risk of inconsistencies among coders. To control for this, two control coders were distributed across the aforementioned four lists. Each of the control coders coded the same number of words per PoS as each main coder (25 for nouns and verbs, and 27 for adjectives). Each of the control coders could either be paired with just one main coder, or with one main and the other control coder. This was decided to test the correlation between two control coders. An illustrative coding scheme is presented in Table 4. Due to some data loss for verbs, some words needed to be coded again, swapping some data from Main coders with data from Control coder 2.

List 1		List 2	
Word	Coders	Word	Coders
Word 1	MC 1, CC 1, CC 2	Word 11	MC 3, CC 1, CC 2
Word 2	MC 1	Word 12	MC 3
Word 3	MC 1, CC 1	Word 13	MC 3, CC 1
Word 4	MC 1	Word 14	MC 3
Word 5	MC 1, CC 2	Word 15	MC 3, CC 2
List 3		List 4	
Word	Coders	Word	Coders
Word 6	MC 2, CC 1, CC 2	Word 16	MC 4, CC 1, CC 2
Word 7	MC 2	Word 17	MC 4
Word 8	MC 2, CC 1	Word 18	MC 4, CC 1
Word 9	MC 2	Word 19	MC 4
Word 10	MC 2, CC 2	Word 20	MC 4, CC 2

Table 4. An example coding scheme. MC 1, 2, 3, and 4 are main coders, and CC 1 and 2 are control coders

The classification was performed in two steps. For the first pass-through, coders placed each response into a category defined in the dictionary. This was done for all words in this set. Let us take the example from Table 1, the word *bronze*, with two senses. All responses that referred to the sense “An alloy of copper and tin (possibly with traces of other metals)” were placed in that category. However, no participants listed anything that could be placed in the “Bronze bell, usually put on cattle” category. Therefore, the remaining responses were not placed in any category. Henceforth, we refer to this as *dictionary categories* or DC. In the second step, for all *remaining* uncategorized responses coders could suggest additional categories that appear in the responses but are not listed in the dictionary. Also, in this step, a novel sense category could be created based on a group of responses that could fit a dictionary category but are estimated by the subjective criterion of our coders to clearly constitute a separate, well-defined group in the semantic intuitions of our participants. In the example from Table 1, for the word *bronze*, responses provided by the participants that were not previously categorized clearly contained more senses known by the participants. Additional categories that were “discovered” in the responses were: the colour bronze, darker skin tone (occasionally referred to in Serbian as bronze tan), and a bronze medal in sports competitions (the word in the exact same form as the metal is used to refer to a bronze medal). This was the second set of categories, which is henceforth called *dictionary + added categories* or DAC. Both steps of classification were conducted using Categor 1.2 software (Filipović Đurđević & Đurđević, 2021). The two sets of resulting categories of senses (DC and DAC), together with the frequency of each sense (number of responses per category) for both pass-throughs are available (in Serbian) in the Supplementary material (https://osf.io/ukcrg/).

2.4. Estimating the number of senses and sense frequencies

Following Azuma (1996) and Filipović Đurđević and Kostić (2017) we determined the number and frequency of senses. The number of senses was estimated by counting the number of previously created sense categories with a non-zero number of responses. This was done both for DC and DAC sets. Dictionary categories that did not appear in the responses were discarded as senses unfamiliar to the speakers. Additionally, in order to control for idiosyncrasies, we applied another strategy of counting only the senses that were listed by at least 10% of the participants. Henceforth, we refer to this group of measures as *corrected measures*, as opposed to *uncorrected* measures, without the 10% correction being applied. Therefore, we created five estimations of the number of senses: raw dictionary-based count, dictionary count based on the descriptions elicited from the speakers, i.e., based on speakers' semantic intuitions (*DC uncorrected* and *DC corrected*), dictionary count enriched by additional categories based on the coders' semantic intuitions (*DAC uncorrected* and *DAC corrected*).

Correspondingly, frequencies of the senses were calculated based on how many participants' responses were present in a category. For example, if five participants listed something as a sense of a word, that sense would have a frequency of five. From these frequencies we calculated probabilities. These were calculated by dividing sense frequency by the sum of all sense frequencies for that word.

Finally, following Filipović Đurđević and Kostić's (2017) measures of uncertainty, describing sense probability distributions were calculated from previously calculated probabilities. First, we calculated Shannon's entropy ($H = -\sum_{i=0}^{n} p_i \times \log p_i$; Shannon, 1948). Entropy encompasses both the NoS and the description of the sense probability distribution. Redundancy ($T = 1 - \frac{H}{\log NoS}$) was calculated to exclude information about NoS, and just leave the information on how (un)equal the sense probabilities are.

3. Results

3.1. Descriptives of all measures

All measures described here are calculated for both DC and DAC, corrected and uncorrected for main and control coders. Descriptive values of all presented measures are presented in Appendix A due to the size of the table. Values for particular words can be found in the Supplementary material (https://osf.io/ukcrg/).

3.2. Correlations among coder-based polysemy measures

We performed reliability analyses by calculating correlation coefficients between the estimations for each group of coders for every measure. Since our aim was to create a set of measures describing words with multiple senses, correlations will provide an index of the reliability of the measures coming from the Main coders group.

The distribution of correlations between three coder groups (Main coders, Control coder 1, and Control coder 2) is presented in Figure 1. The figure reveals that the values of correlation coefficients tend to group towards higher positive values for all measures. This suggests a general consistency between groups of coders. However, there are some correlations that are zero or numerically close to zero. We assume that these point to inconsistencies between the speaker's intuitions regarding sense classification. Small differences between all correlations and just the significant correlations further reveal that even most of the low correlations are a significant further point to the information value of those low correlations.

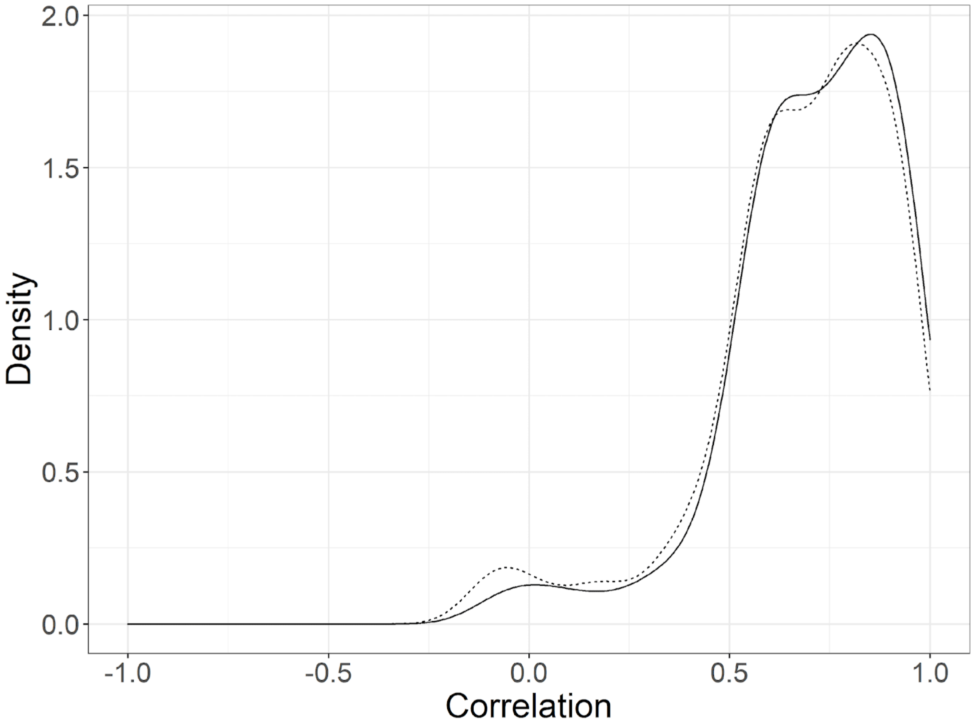

Figure 1. Density plot of correlation values between groups of coders. The dotted line represents all correlations; the full line represents only significant correlations

To understand where smaller correlations occur and to get a more detailed picture of the differences among measures, we present individual correlations in Figure 2. Overall, we observed that correlations are generally higher for coding situations with higher restrictions in coding. For example, DC categorization can be considered a higher restriction, since coders only classify in predetermined categories. DAC categorization allowed coders more degrees of freedom to create new categories leading to variation between individual coders.

First, Control coders 1 and 2 overlapped on just a few words, making those correlations less informative and most likely to be non-significant. We present them for complete results, but those correlations are based on 12 and 15 data points, whereas all other correlations are based on at least 25 data points (see Table 4). In the case of verbs, it is important to note that due to some data loss, Control Coder 2 was moved to Main coder data, in order to complete the set. This led to verbs having the most non-significant correlations.

Comparing PoS we see that correlations are mostly similar across all three groups and that most differences are due to other variations in the data, such as DC/DAC distinction, or whether counts have been corrected or not. We observed a small drop in correlations when a 10% correction was introduced when calculating measures. A more significant drop is observable when comparing DC and DAC measures. This may be due to each coder having more freedom in adding categories based on listed senses that could not be classified into any of the categories from the dictionary.

Finally, differences in correlation intensity between measures reveal that NoS was likely estimated better than sense probabilities, especially in cases of classifying into DAC. Most non-significant or very low correlations occurred for redundancy. This measure describes only the shape of the sense probability distribution, excluding the number of events (i.e., NoS) from the measure. With NoS correlations being higher we may conclude that high entropy correlations are driven by the NoS component, whereas probabilities drive the correlations down. This leads us to conclude that there is consensus between coders in the number of senses of a word, but estimating sense probabilities is more difficult and likely less stable over time and among participants.

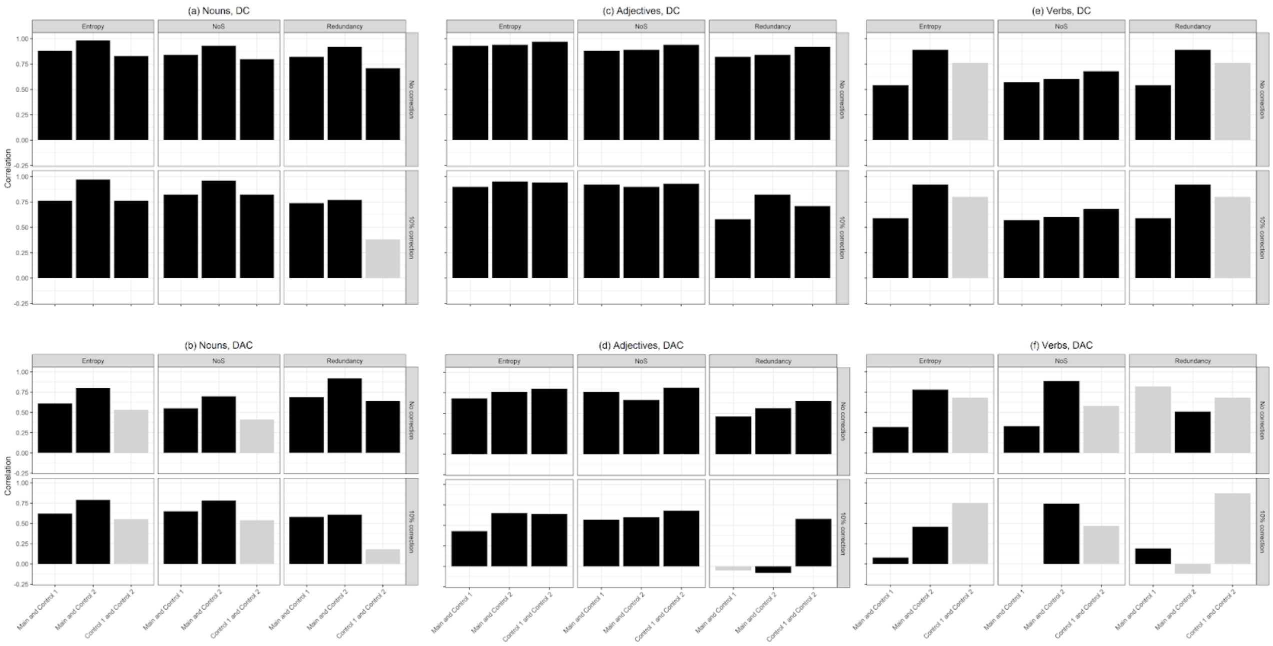

Figure 2. This figure presents individual correlations between each group of coders: Main Coders and Control Coder 1, Main Coders and Control Coder 2, and Control Coder 1 and Control Coder 2. Black bars represent significant correlations and light grey bars non-significant correlations. Each panel is denoted by the letters A-F and consists of three columns for each measure: Entropy, NoS, and Redundancy, respectively, along with two rows. The top row represents uncorrected measures, and the bottom row represents measures in which senses that were mentioned by fewer than 10% of the participants were excluded from the counts and calculations

3.3. Correlations between main coders’ NoS counts and dictionary NoS counts

Next, we compared the native speaker-based measures to the dictionary ones. We selected only the NoS measures from the main coders’ team as NoS is the only measure directly comparable to the counts obtained from the dictionary. Each of the four estimations that were based on participant responses and coders categorizations were compared to dictionary NoS – DC uncorrected, DC corrected, DAC uncorrected, and DAC corrected. We present the five NoS measure distributions in Figure 3.

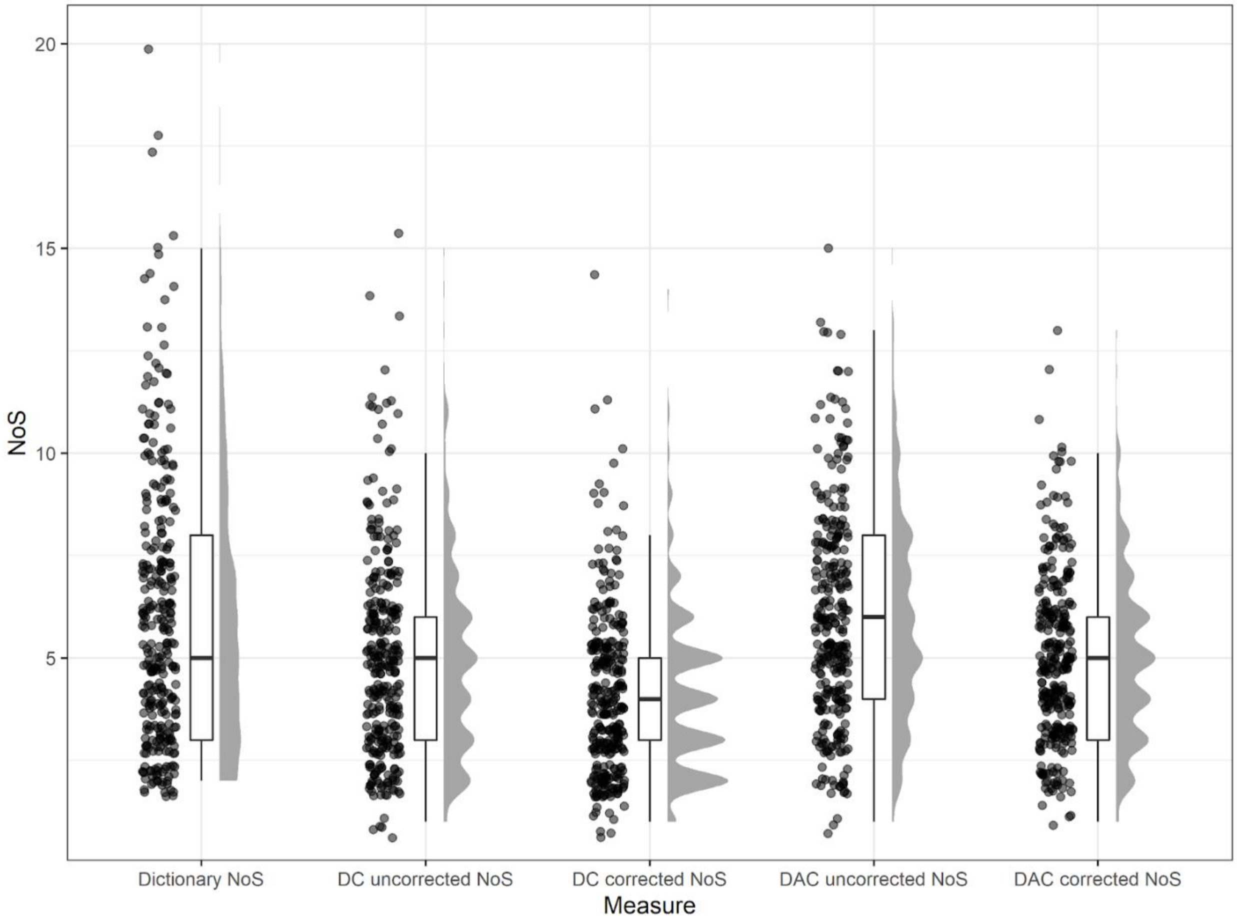

Figure 3. Distributions of dictionary NoS, DC uncorrected NoS, DC corrected NoS, DAC uncorrected NoS, and DAC corrected NoS. Boxplots represent the interquartile range between the 25th (lower) percentile and the 75th (upper) percentile. The line represents the median. On the left of each boxplot, scatterplots are presented, and on the right, the density plots of the same data

The figure reveals some narrowing of the interquartile range. In other words, NoS demonstrates the tendency to group toward the lower NoS values. We inspected correlations of four participant-based measures and dictionary NoS (Table 5). Correlations between dictionary NoS and both uncorrected measures were moderate. When the 10% correction was applied, the correlations slightly dropped, pointing to larger deviations from dictionary NoS.

We further explored this using a general linear model (GLM) in which we compared all four participant-based NoS measures to dictionary NoS. The model was significant overall (F(4, 1483) = 30.54, *p* < .001). In Table 5, we can see the deviations of the four measures from the dictionary NoS. All measures but one were significantly different from the dictionary NoS, with averages becoming smaller. This further supports what is observed in Figure 4, and in the correlations in Table 5.

NoS	Correlations with Dictionary NoS	General linear model			
		Estimate	S.E.	t-value	*p*-value
Dictionary	-	5.65	.13	42.48	<.001
DC uncorrected	.59***	-1.64	.19	-8.82	<.001
DC corrected	.49***	-0.69	.19	-3.73	<.001
DAC uncorrected	.54***	-0.77	.19	-4.12	<.001
DAC corrected	.45***	.18	.19	.98	.32

Table 5. Correlations between dictionary NoS, and DC uncorrected NoS, DC corrected NoS, DAC uncorrected NoS, and DAC corrected NoS with GLM-based comparisons
Note. *** – p < .001

4. Discussion

Starting from 18,000 responses rooted in native speakers' semantic intuitions on 307 Serbian polysemous words belonging to three part-of-speech categories – nouns, verbs, and adjectives mean, we estimated the number of senses and the sense probability distributions. We categorized the collected descriptions following the dictionary and the semantic intuitions of the researchers who performed the coding (categorization) procedure. Based on this categorization, we calculated the number of senses and sense frequencies, and based on sense frequencies we derived entropy and redundancy of the sense probability distribution. Finally, we compared categories produced by different coders thus evaluating the reliability of the obtained categories, but also introducing some novel insights on the nature of polysemy.

We believe that our work has three major contributions. On the one hand, we provide quantitative descriptions of lexical ambiguity for a group of Serbian nouns, verbs, and adjectives. On the other hand, we provide the community with materials for future insights into semantic intuitions on the meaning of polysemous words. Finally, we provide insights into the dual role of the researchers performing the categorization of semantic intuitions of native speakers.

4.1. Quantitative description of lexical ambiguity

We estimated the number of senses, entropy, and redundancy of the sense probability distribution. Each of the measures was calculated for categories listed in the dictionary and for categories modified by the coders (authors). Further, each measure was additionally presented after being corrected for idiosyncratic responses (responses provided by less than 10% of participants). Correlations revealed moderate similarity to the dictionary NoS. However, similarities became more prominent after the 10% correction. This further confirms that, although there is some overlap between dictionaries and participant-based measures, the latter may present another source of information. This is further supported by the GLM results, which point to the fact that NoS obtained from participants is lower on average than dictionary NoS.

First and foremost, these estimations will serve researchers interested in the processing of polysemy, as they will enable them to study the effects of polysemy on various processing measures. However, they will also be useful to researchers generally using these words as stimuli and trying to match them for various lexical features. The facilitatory effect of polysemy on word recognition has been well-documented (Filipović Đurđević & Kostić, 2008; 2023; Filipović Đurđević et al., 2009; Klepousniotou et al., 2008; Mišić & Filipović Đurđević, 2022a; Rodd et al., 2002), making this variable an important lexical feature that should be controlled in psycholinguistic studies.

4.2. Semantic intuitions on word sense descriptions

We also provide raw responses as produced by the participants. The descriptions of polysemous words that we collected are elicited from naive native speakers and are thus based on their untrained semantic intuitions. As such, they represent an important tool for the evaluation of the dictionary collections of descriptions. Our study showed that some of the senses known by participants are also listed in the dictionary. However, there are senses that are known to participants but are not listed in the dictionary, as is apparent in Figure 5, and from the overall drop in the number of senses when comparing dictionary-based and participant-based measures. This is the case with some novel usages of words (e.g., *bronze* as the medal – *They won the bronze!*). At the same time, there are senses that are listed in the dictionary but are not recalled by any of the participants. This is the case with some archaic usages of the words (e.g., *bronze* as the bell on a cow). Our finding fits well with the findings of other similar studies conducted in other languages (Lin & Ahrens, 2005) and represents an important source of information on the nature of language change. Additionally, it can serve as the starting point for fine-tuning Large Language Models and machine learning in general. This will allow keeping the assumed semantic representations based on human knowledge but allow exploring this on larger language samples.

4.3. Semantic Intuitions on Word Sense Categories

Finally, we introduced the novelty of considering the researchers (authors of this study) as performers of two roles. On the one hand, they were the researchers who coded/categorized the native speakers' intuitions on interpretations of polysemous words. However, at the same time, they themselves were participants who relied on their own intuitions on how the elicited descriptions of senses should be grouped.

Our analyses of the correlations among the measures derived from per-coder categorizations of the same word senses provide interesting insights into the nature of polysemy. Our results reveal that the estimations of the number of senses were very reliable, whereas the estimations of entropy and redundancy oscillated across coders. This indicates that the coders tended to produce a similar number of categories, but the precise distribution of sense descriptions across categories was not the same. For example, one coder would place many senses in category A, and few senses in category B, whereas another coder would distribute the senses equally between the two categories. This would result in both coders ending with two categories, but the first coder would produce a distribution of high redundancy, whereas the second coder's distribution would be of low redundancy. Consequently, the correlation coefficient for the number of senses would be high for the two coders, whereas the correlation coefficient for the redundancy measures estimated by the two coders would be low. This is exactly what we observed. Our finding of stability of the number of senses estimation is in accordance with Twilley et al. (1994) and Armstrong et al. (2012). Unlike our study, Armstrong et al. also observed the stability of sense frequency estimation. It should be noted that they focused only on the dominant, i.e., the most frequent meaning. We believe that a shift of focus to the remaining meanings would reduce the level of reliability. In fact, the conclusion laid out by Twilley et al. supports our claim, as they found a modest overlap of approximately 60-70% among the coders.

This finding is informative of the reliability of different estimations of polysemy. We conclude that the number of senses is more reliable than the more fine-grained measures that are based on sense probability distributions. However, this finding is also informative of the nature of polysemy, as it shows how variable the precise grouping of the categories of polysemous senses is. This finding is in accordance with some recent views of polysemy as a dynamic phenomenon, as opposed to word senses seen as fixed categories. According to such approaches, words are tools that are adaptable to the given situation (e.g., construction integration model; Kintsch, 1988; 1998; Kintsch & van Dijk, 1978; Schmalhofer et al., 2002). Also, it is in accordance with some recent empirical work that showed immense sensitivity of the human semantic system to recent exposures (Curtis et al., 2022; Hulme et al., 2019; Mak et al., 2023; Parker et al., 2023).

5. Conclusion

To conclude, this study mainly set out to quantify lexical ambiguity for a sample of Serbian nouns, verbs, and adjectives. We offer the full dataset of measures, as well as some control measures from additional coders and raw data. While quantitative measures can be used to predict different processing measures such as response times, eye movements, or neural measures, additional datasets were created along the way that can be used for e.g. qualitative analyses. For example, in the sense production task participants gave approximately 18,000 responses that are available for qualitative analyses, particularly applying other classification methods. Alternatively, these can be used as a potential starting point in using large language models to estimate sense probability distributions. Also, within the confines of the currently presented classification, additional analyses based on available data in our dataset can be done for a more in-depth understanding of variation between polysemous words' representations, both for our participants as well as for the coders.

References

Anđelić, S. & Filipović Đurđević, D. (2023). *Similarity of Sensorimotor Experience as a Measure of the Relatedness Among the Meanings of Ambiguous Words*. ESCOP 2023 - 23rd Conference of the European Society for Cognitive Psychology. https://escop2023.org/images/schedule/escop_abstractbook_05092023.pdf#page=253

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequency norms for 544 homonyms. *Behavioral Research Methods, 44*(4), 1015–1027. https://doi.org/10.3758/s13428-012-0199-8

Azuma, T. (1996). Familiarity and relatedness of word meanings: Ratings for 110 homographs. *Behavior Research Methods, Instruments, & Computers 28*, 109–124. https://doi.org/10.3758/BF03203645

Azuma, T., & Van Orden, G. C. (1997). Why SAFE Is Better Than FAST: The Relatedness of a Word's Meanings Affects Lexical Decision Times. *Journal of Memory and Language*, *36*(4), 484–504. https://doi.org/10.1006/jmla.1997.2502

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247).

Curtis, A. J., Mak, M. H. C., Chen, S., Rodd, J. M., & Gaskell, M. G. (2022). Word-meaning priming extends beyond homonyms. *Cognition, 226*, 105175. https://doi.org/10.1016/j.cognition.2022.105175

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review, 22*, 13–37. https://doi.org/10.3758/s13423-014-0665-7

Filipović Đurđević, D. (2019). Balance of meaning probabilities in processing of Serbian homonymy. *Primenjena Psihologija, 12*(3), 283–304. https://doi.org/10.19090/pp.2019.3.283-304

Filipović Đurđević, D., & Đurđević, Đ. (2021). Categor 1.1 – slobodan softver za pomoć u razvrstavanju odgovora u kategorije i analizi kategorija. In V. Džinović & T. Nikitović (Eds.), *Kvalitativna* *istraživanja kroz discipline i kontekste: osmišljavanje sličnosti i razlik. Institut za pedagoška istraživanja* (pp. 73–78). https://hdl.handle.net/21.15107/rcub_ipir_428

Filipović Đurđević, D., & Kostić, A. (2008). The effect of polysemy on processing of Serbian nouns. *Psihologija*, *41*(1), 69–86. https://doi.org/10.2298/PSI0801059F

Filipović Đurđević, D., Đurđević, Đ., & Kostić, A. (2009). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija*, *42*(1), 95–106. https://doi.org/10.2298/PSI0901095F

Filipović Đurđević, D., & Kostić, A. (2017). Number, Relative Frequency, Entropy, Redundancy, Familiarity, and Concreteness of Word Senses: Ratings for 150 Serbian Polysemous Nouns. In S. Halupka-Rešetar and S. Martínez-Ferreiro (Eds.), *Studies in Language and Mind* (pp.13–77). Faculty of Philosophy, University of Novi Sad. http://digitalna.ff.uns.ac.rs/sadrzaj/2017/978-86-6065-446-7

Filipović Đurđević, D., & Kostić, A. (2023). We probably sense sense probabilities. *Language, Cognition and Neuroscience, 38*(4). https://doi.org/10.1080/23273798.2021.1909083

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, *67*(4), 407–425. https://doi.org/10.1016/j.jml.2012.07.010

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass 3*, 111–127. https://doi.org/10.1111/j.1749-818X.2008.00104.x

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 25,* 1366–1383. https://doi.org/10.1037/0278-7393.25.6.1366

Gernsbacher, M. A. (1984). Resolving 20 Years of Inconsistent Interactions Between Lexical Familiarity and Orthography, Concreteness, and Polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. https://doi.org/10.1037/0096-3445.113.2.256

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*, 395–427. https://doi.org/10.3758/BF03201693

Gortan-Premk, D. (2004). *Polisemija i organizacija leksičkog sistema u srpskome jeziku*. [Polysemy and the organisation of the lexical system in Serbian language]. Zavod za udžbenike i nastavna sredstva.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Harris, Z. S. (1954). Distributional Structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22*(6), 1331–1356. https://doi.org/10.1037/0096-1523.22.6.1331

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hulme, R. C., Barsky, D., & Rodd, J. M. (2019). Incidental Learning and Long-Term Retention of New Word Meanings From Stories: The Effect of Number of Exposures. *Language Learning*, *69*(1), 18–43. https://doi.org/10.1111/lang.12313

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278–305. https://doi.org/10.1016/0010-0285(81)90011-6

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114, 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2), 115–124. https://doi.org/10.1037/a0026727

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182. https://doi.org/10.1037/0033-295x.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295x.85.5.363

Klein, D. E., & Murphy, G. L. (2001). The Representation of Polysemous Words. *Journal of Memory and Language*, *45*(2), 259–282. https://doi.org/10.1006/jmla.2001.2779

Klein D.E. & Murphy G.L. (2002). Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language, 47*(4), 548–570. https://doi.org/10.1016/S0749-596X(02)00020-7

Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language, 81*(1–3), 205–223. https://doi.org/10.1006/brln.2001.2518

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, *20*(1), 1–24. https://doi.org/10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making Sense of Word Senses: The Comprehension of Polysemy Depends on Sense Overlap. Journal of Experimental Psychology: *Learning Memory and Cognition*, *34*(6), 1534–1543. https://doi.org/10.1037/a0013012

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika* [Frequency dictionary of contemporary Serbian language]. Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

Krstev, C., Pavlović-Lažetić, G., & Obradović, I. (2004). Using textual and lexical resources in developing Serbian wordnet. *Romanian Journal of Information Science and Technology*, *7*(1-2), 147–161. https://rgf.bg.ac.rs/LicnePrezentacije/ivan_obradovic/Radovi/RJIS_2004.pdf

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037//0033-295x.104.2.211

Lin, C.-J. C., & Ahrens, K. (2005). How many meanings does a word have? Meaning estimation in Chinese and English. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language acquisition, change and emergence: Essays in evolutionary linguistics* (pp. 437–464). City University of Hong Kong Press.

Ljubešić, N., & Lauc, D. (2021). BERTić – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 37–42). Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.5/

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers, 28*(2), 203–208. https://doi.org/10.3758/BF03204766

Lopukhina, A., Laurinavichyute, A., Lopukhin, K., & Dragoy, O. (2018). The mental representation of polysemy across word classes. *Frontiers in Psychology, 9,* 192. https://doi.org/10.3389/fpsyg.2018.00192

Mak, M. H. C., Curtis, A. J., Rodd, J. M., & Gaskell, M. G. (2023). Episodic memory and sleep are involved in the maintenance of context-specific lexical information. *Journal of Experimental Psychology: General*, *152*(11), 3087–3115. https://doi.org/10.1037/xge0001435

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings* (pp. 1–12). https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Tomas. In *Advances in Neural Information Processing Systems* (pp. 3111–3119). MIT Press. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2019). Advances in pre-training distributed word representations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation* 1 (pp. 52–55). https://doi.org/10.48550/arXiv.1712.09405

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39–41.

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition, 17*(2), 141–147. https://doi.org/10.3758/BF03197064

Mišić, K., & Filipović Đurđević, D. (2022a). Redesigning the exploration of semantic dynamics: SSD account in light of regression design. *Quarterly Journal of Experimental Psychology, 75*(6), 1155–1170. https://doi.org/10.1177/17470218211048386

Mišić, K., & Đurđević, D. F. (2022b). The influence of the experimental context on lexical ambiguity effects. *Teme 46*(2), 577–596. https://doi.org/10.22190/TEME210418030M

OpenAI. (2023). GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774

Parker, A. J., Taylor, J. S. H., & Rodd, J. M. (2023). Readers use recent experiences with word meanings to support the processing of lexical ambiguity: Evidence from eye movements. *PsyArXiv.* https://doi.org/10.31234/osf.io/b49vk

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark. C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237).

Popović Stijačić, M. (2021). *Značaj perceptivne informacije u obradi reči –perspektiva utelovljene kognicije* [*The importance of perceptual information in word processing – embodied cognition perspective*]. Doctoral dissertation, University of Novi Sad.

Popović Stijačić, M., & Filipović Đurđević, D. (in preparation). Modality Specific Perceptual Strength, Concreteness, Imageability, Context Availability, Emotional Valence, Arousal and Age of Acquisition Norms for 2100 Serbian Nouns.

Rice, C. A., Tokowicz, N., Fraundorf, S. H., & Liburd, T. L. (2019). The complex interactions of context availability, polysemy, word frequency, and orthographic variables during lexical processing. *Memory & Cognition, 47*, 1297–1313. https://doi.org/10.3758/s13421-019-00934-4

Vujanić, M., Vujanić, M., Gortan-Premk, D., Gortan-Premk, D., Dešić, M., Dešić, M., Dragićević, R., Dragićević, R., Nikolić, M., Nikolić, M., Nogo, Lj., Nogo, L., Pavković, V., Pavković, V., Ramić, N., Ramić, N., Stijović, R., Stijović, R., Tešić, M., … Fekete, E. (2007). *Rečnik srpskoga jezika* [Dictionary of the Serbian Language]. Matica Srpska.

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science, 15*(2), 411–427. https://doi.org/10.1177/1745691619885860

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language, 46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9*(5), 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3

Schmalhofer, F., McDaniel, M. A., & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes, 33*(2), 105–132. https://doi.org/10.1207/S15326950DP3302_01

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97-123.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Soler A. G. & Apidianaki, M. (2021). Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, *9*, 825–844. https://doi.org/10.1162/tacl_a_00400

Twilley, L.C., Dixon, P., Taylor, D., Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*, 111–126. https://doi.org/10.3758/BF03202766

Ulcar, M., & Robnik-Sikonja, M. (2019). High Quality ELMo Embeddings for Seven Less-Resourced Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4731–4738). https://doi.org/10.48550/arXiv.1911.10049

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *ArXiv*, *abs/1909.10430*. https://doi.org/10.48550/arXiv.1909.10430

Yenicelik, D., Schmidt, F., & Kilcher, Y. (2020). How does BERT capture semantics? A closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP* (pp. 156–162). Association for Computational Linguistics.

Zipf, G. K. (1945). The Meaning-Frequency Relationship of Words. *The Journal of General Psychology*, *33*(2), 251–256, https://doi.org/10.1080/00221309.1945.10544509

6. Appendix A

Table 1. Descriptive statistics of all measures calculated from this classification. More detailed data is available in the supplementary materials

Measures	Correction	Coders	Categories	PoS	M	SD	Min	Max	n
NoS	No	Main	DC	Adjectives	4.51	2.03	1	11	108
NoS	No	Main	DC	Nouns	4.17	1.97	1	11	100
NoS	No	Main	DC	Verb	6.3	2.75	2	15	99
NoS	No	Main	DAC	Adjectives	5.5	2.25	1	13	108
NoS	No	Main	DAC	Nouns	5.01	2.28	1	11	100
NoS	No	Main	DAC	Verb	7.32	2.65	2	15	99
NoS	No	Control 1	DC	Adjectives	3.97	1.84	1	8	108
NoS	No	Control 1	DC	Nouns	3.52	1.73	1	7	100
NoS	No	Control 1	DC	Verb	5.72	2.56	2	12	99
NoS	No	Control 1	DAC	Adjectives	4.9	1.84	2	9	108
NoS	No	Control 1	DAC	Nouns	4.96	1.7	2	9	100
NoS	No	Control 1	DAC	Verb	7.21	3.09	2	15	99
NoS	No	Control 2	DC	Adjectives	4.34	1.86	1	8	108
NoS	No	Control 2	DC	Nouns	4.12	2.06	1	8	100
NoS	No	Control 2	DC	Verb	5.37	3.2	1	13	99
NoS	No	Control 2	DAC	Adjectives	5.24	1.81	1	8	108
NoS	No	Control 2	DAC	Nouns	5.61	1.99	2	11	100
NoS	No	Control 2	DAC	Verb	6.24	2.7	2	13	99
NoS	10%	Main	DC	Adjectives	3.74	1.62	1	9	108
NoS	10%	Main	DC	Nouns	3.41	1.63	1	10	100
NoS	10%	Main	DC	Verb	5.11	2.23	2	14	99
NoS	10%	Main	DAC	Adjectives	4.71	1.88	1	10	108
NoS	10%	Main	DAC	Nouns	4.22	1.8	1	10	100
NoS	10%	Main	DAC	Verb	6.08	2.17	2	13	99
NoS	10%	Control 1	DC	Adjectives	3.97	1.84	1	8	108
NoS	10%	Control 1	DC	Nouns	3.52	1.73	1	7	100
NoS	10%	Control 1	DC	Verb	5.72	2.56	2	12	99
NoS	10%	Control 1	DAC	Adjectives	4.9	1.84	2	9	108
NoS	10%	Control 1	DAC	Nouns	4.96	1.7	2	9	100
NoS	10%	Control 1	DAC	Verb	7.21	3.09	2	15	99
NoS	10%	Control 2	DC	Adjectives	3.48	1.53	1	7	108
NoS	10%	Control 2	DC	Nouns	3.34	1.86	1	8	100
NoS	10%	Control 2	DC	Verb	4.37	2.52	1	10	99
NoS	10%	Control 2	DAC	Adjectives	4.59	1.62	1	7	108
NoS	10%	Control 2	DAC	Nouns	4.87	1.86	1	10	100
NoS	10%	Control 2	DAC	Verb	5	2.15	2	10	99
Entropy (H)	No	Main	DC	Adjectives	0.51	0.18	0.18	0.95	108
Entropy (H)	No	Main	DC	Nouns	0.48	0.18	0.09	0.92	100

Entropy (H)	No	Main	DC	Verb	0.63	0.19	0.22	1.14	99
Entropy (H)	No	Main	DAC	Adjectives	0.59	0.18	0.18	1.03	108
Entropy (H)	No	Main	DAC	Nouns	0.56	0.18	0.1	0.92	100
Entropy (H)	No	Main	DAC	Verb	0.71	0.17	0.22	1.05	99
Entropy (H)	No	Control 1	DC	Adjectives	0.51	0.18	0.16	0.87	108
Entropy (H)	No	Control 1	DC	Nouns	0.45	0.19	0.15	0.79	100
Entropy (H)	No	Control 1	DC	Verb	0.63	0.2	0.2	1.03	99
Entropy (H)	No	Control 1	DAC	Adjectives	0.59	0.16	0.27	0.91	108
Entropy (H)	No	Control 1	DAC	Nouns	0.56	0.17	0.15	0.9	100
Entropy (H)	No	Control 1	DAC	Verb	0.72	0.19	0.29	1.13	99
Entropy (H)	No	Control 2	DC	Adjectives	0.5	0.17	0.1	0.82	108
Entropy (H)	No	Control 2	DC	Nouns	0.51	0.17	0.18	0.82	100
Entropy (H)	No	Control 2	DC	Verb	0.62	0.2	0.3	1.01	99
Entropy (H)	No	Control 2	DAC	Adjectives	0.6	0.16	0.1	0.82	108
Entropy (H)	No	Control 2	DAC	Nouns	0.6	0.16	0.18	0.93	100
Entropy (H)	No	Control 2	DAC	Verb	0.62	0.19	0.3	1.01	99
Entropy (H)	10%	Main	DC	Adjectives	0.47	0.18	0.15	0.89	108
Entropy (H)	10%	Main	DC	Nouns	0.46	0.16	0.18	0.9	100
Entropy (H)	10%	Main	DC	Verb	0.59	0.18	0.22	1.14	99
Entropy (H)	10%	Main	DAC	Adjectives	0.56	0.19	0.15	0.95	108
Entropy (H)	10%	Main	DAC	Nouns	0.53	0.16	0.19	0.9	100
Entropy (H)	10%	Main	DAC	Verb	0.67	0.16	0.22	1.05	99
Entropy (H)	10%	Control 1	DC	Adjectives	0.51	0.18	0.16	0.87	108
Entropy (H)	10%	Control 1	DC	Nouns	0.45	0.19	0.15	0.79	100
Entropy (H)	10%	Control 1	DC	Verb	0.63	0.2	0.2	1.03	99
Entropy (H)	10%	Control 1	DAC	Adjectives	0.59	0.16	0.27	0.91	108
Entropy (H)	10%	Control 1	DAC	Nouns	0.56	0.17	0.15	0.9	100
Entropy (H)	10%	Control 1	DAC	Verb	0.72	0.19	0.29	1.13	99
Entropy (H)	10%	Control 2	DC	Adjectives	0.47	0.16	0.19	0.79	108
Entropy (H)	10%	Control 2	DC	Nouns	0.48	0.17	0.21	0.82	100
Entropy (H)	10%	Control 2	DC	Verb	0.58	0.2	0.17	0.94	99
Entropy (H)	10%	Control 2	DAC	Adjectives	0.59	0.13	0.25	0.79	108
Entropy (H)	10%	Control 2	DAC	Nouns	0.59	0.14	0.26	0.91	100
Entropy (H)	10%	Control 2	DAC	Verb	0.57	0.2	0.17	0.94	99
Redundancy (T)	No	Main	DC	Adjectives	0.18	0.11	0	0.53	108
Redundancy (T)	No	Main	DC	Nouns	0.2	0.14	0	0.69	100
Redundancy (T)	No	Main	DC	Verb	0.16	0.08	0	0.39	99
Redundancy (T)	No	Main	DAC	Adjectives	0.17	0.1	0.04	0.51	108
Redundancy (T)	No	Main	DAC	Nouns	0.17	0.12	0	0.67	100
Redundancy (T)	No	Main	DAC	Verb	0.15	0.06	0	0.33	99
Redundancy (T)	No	Control 1	DC	Adjectives	0.11	0.09	0	0.46	108
Redundancy (T)	No	Control 1	DC	Nouns	0.18	0.13	0.02	0.51	100
Redundancy (T)	No	Control 1	DC	Verb	0.13	0.07	0	0.35	99

Redundancy (T)	No	Control 1	DAC	Adjectives	0.12	0.06	0.01	0.22	108
Redundancy (T)	No	Control 1	DAC	Nouns	0.17	0.11	0.02	0.5	100
Redundancy (T)	No	Control 1	DAC	Verb	0.12	0.05	0.04	0.22	99
Redundancy (T)	No	Control 2	DC	Adjectives	0.18	0.14	0	0.68	108
Redundancy (T)	No	Control 2	DC	Nouns	0.19	0.12	0.06	0.63	100
Redundancy (T)	No	Control 2	DC	Verb	0.18	0.1	0	0.49	99
Redundancy (T)	No	Control 2	DAC	Adjectives	0.16	0.08	0.04	0.34	108
Redundancy (T)	No	Control 2	DAC	Nouns	0.17	0.11	0.05	0.63	100
Redundancy (T)	No	Control 2	DAC	Verb	0.18	0.1	0	0.49	99
Redundancy (T)	10%	Main	DC	Adjectives	0.14	0.11	0	0.51	108
Redundancy (T)	10%	Main	DC	Nouns	0.12	0.08	0	0.39	100
Redundancy (T)	10%	Main	DC	Verb	0.11	0.07	0	0.35	99
Redundancy (T)	10%	Main	DAC	Adjectives	0.14	0.09	0.03	0.51	108
Redundancy (T)	10%	Main	DAC	Nouns	0.12	0.09	0	0.64	100
Redundancy (T)	10%	Main	DAC	Verb	0.12	0.06	0	0.29	99
Redundancy (T)	10%	Control 1	DC	Adjectives	0.11	0.09	0	0.46	108
Redundancy (T)	10%	Control 1	DC	Nouns	0.18	0.13	0.02	0.51	100
Redundancy (T)	10%	Control 1	DC	Verb	0.13	0.07	0	0.35	99
Redundancy (T)	10%	Control 1	DAC	Adjectives	0.12	0.06	0.01	0.22	108
Redundancy (T)	10%	Control 1	DAC	Nouns	0.17	0.11	0.02	0.5	100
Redundancy (T)	10%	Control 1	DAC	Verb	0.12	0.05	0.04	0.22	99
Redundancy (T)	10%	Control 2	DC	Adjectives	0.11	0.09	0	0.37	108
Redundancy (T)	10%	Control 2	DC	Nouns	0.11	0.07	0.01	0.3	100
Redundancy (T)	10%	Control 2	DC	Verb	0.12	0.1	0	0.43	99
Redundancy (T)	10%	Control 2	DAC	Adjectives	0.12	0.06	0	0.26	108
Redundancy (T)	10%	Control 2	DAC	Nouns	0.13	0.05	0.05	0.23	100
Redundancy (T)	10%	Control 2	DAC	Verb	0.13	0.1	0	0.43	99