Vocabulary Size of Czech Native Speakers: A Statistical Approach

Marek Blahuš [1], Miloš Jakubíček [1,2], Vojtěch Kovář[1,3], František Kovařík[1,3]

[1]Lexical Computing, Brno, Czech Republic
[2]Faculty of Informatics, Masaryk University, Brno, Czech Republic
[3]Faculty of Arts, Masaryk University, Brno, Czech Republic
E-mails: firstname.lastname@sketchengine.eu

Abstract

This paper explores the theory of measuring vocabulary size, including the various methods that can be used and the parameters that have to be set. We have examined the experiments carried out on English and Dutch. Goulden et al. (1990) claims the average native speaker knows about 17,000 English base words (non-derived words). Keuleers et al. (2015) and Brysbaert et al. (2016) claim the average native speaker with secondary education knows about 42,000 headwords (lemmas). We have conducted an experiment similar to that of Keuleers and Brysbaert on Czech, with the input of 100,000 letter sequences from the wordlists of large web corpora. We assume the vocabulary size of Czech native speakers (as well as the vocabulary size of native speakers of any language) could be bigger, exceeding 57,000 (Czech) headwords, should we provide the participants with more inputs (150,000 sequences, or even more) or should we count the specialized terminology of their fields of interest.

Keywords: vocabulary size; native speaker; manual annotation; semi-automatic dictionary drafting; Dictionary Express

1. Introduction

Measuring the vocabulary size of speakers of a language has always been a difficult task. Before we even begin to measure one's vocabulary size, many questions arise: What does it mean to know a word? Which words belong to the language and which don't? What method should we use to measure one's vocabulary size? And even – what is a word? There have been several attempts to measure the vocabulary size of native English speakers. Goulden et al. (1990) stated that the results of these studies range from thousands to hundreds of thousands of words, up to 216,000 words. Goulden then carried out an experiment using *Webster's Third New International Dictionary* as a base for the general vocabulary and concluded "that the average educated native speaker [of English] has a vocabulary of around 17,000 base words". This was a rather complex experiment with a lot of factors in play, so it's highly insufficient to say that "a native speaker of a language simply knows around 20,000 words" without adding a lot of context.

In this paper, we are looking into the basic theory of vocabulary size. We examine how different aspects of vocabulary can strongly manipulate the results. We set a baseline knowledge of vocabulary measurement for examining the vocabulary of native speakers

of specific languages. We then apply this knowledge to test the basic quantitative and qualitative aspects of the vocabulary size of Czech.

We present a practical experiment done on Czech native speakers as a part of the Czech Dictionary Express (CDE) project.(Kovařík et al., 2024b) This project is part of a group of dictionary-making projects done on several smaller languages.(Baisa et al., 2019; Blahuš et al., 2023)

2. Measuring vocabulary size

In linguistic theory, the (passive) vocabulary of a language user is the sum of words they recognise and understand, but don't necessarily productively use in their everyday life. When measuring the vocabulary size, we work with the receptive knowledge of a word, i.e. if the user recognises the word in some of its forms, meanings, uses, etc.(Pignot-Shahov, 2012) The hypothetical size of a user's vocabulary depends on the categories we set to measure. This is why some researchers come to small numbers, such as thousands of words, while others talk about vocabulary sizes a hundred times bigger.

If we measure only the number of headword lemmas the user knows, we inevitably neglect the depth of the knowledge of these headwords, e.g. if the user knows the word in its usual senses, or even one meaning of the headword (since we often recognize some words by their phonology without understanding their semantics).

Another important question is *what* we measure, and how the measured language deals with the measured item. We can measure the number of word forms (e.g. "governors", "using") a speaker understands, or the number of lemmas (the dictionary form, e.g. "governor", "reuse"), base words (e.g. "govern", "use"), or word families (e.g. all the words like "govern", "governor", "governing" together).

After their experiment, Goulden et al. (1990) estimate the vocabulary size of average educated native speakers of English to be around 17,000 base words. This excludes derivatives, abbreviations, alternative spellings, inflected words, proper words and compound words. From the 267,000 headwords of *Webster's Third New International Dictionary* with a full entry, only around 54,000 can be considered base words.

Pignot-Shahov (2012) states word families "are similar to lemmas but include all words related to the headword regardless of their word class", and sets an example of *teach*, *taught*, *teaches*, *teacher* and *teachable* as belonging to the same word family. Goulden et al., however, add the word *misgovern* to the family of *govern*, *governor*, *government* and *ungovernable*. For some languages like Czech, which heavily use prefixation as part of their derivation systems, and where prefixes do not always simply add the quality of meaning they represent, this could be problematic.

As a conclusion of a newer experiment, Brysbaert et al. (2016) estimate the median vocabulary size of 20-year-old Americans is about 42,000 lemmas and 11,100 base words (their categories for base words are stricter than those of Goulden et al.). Brysbaert's test was similar to a previous one, by Keuleers et al. (2015), carried out on approximately 279,000 Dutch native speakers from Belgium and the Netherlands in an online survey. The test showed each participant 100 different letter sequences, around 70 of which were words randomly chosen from a 53,000-word Dutch vocabulary, the rest being non-words. Each

participant was asked to state whether they recognised the sequence as a Dutch word or not. They were informed that the whole test takes about 4 minutes, and they can repeat it with new sequences as many times as they want. The results of the Dutch test showed similar results to the next one about American English (a vocabulary size of about 42,000 lemmas), and that the vocabulary size increases slightly with the age of the speakers.

There are not many studies on what the vocabulary size of a Czech native speaker is. Těšitelová (1987) estimates the active vocabulary of Czech native speakers ranges from 3,000 to 10,000 words (5,000 words on average) and that the passive vocabulary is about 3–6 times bigger, corresponding to the Brysbaert studies of Dutch and American English. The numbers Těšitelová mentions are, however, not backed by an experimental study.

3. Manual annotation of Czech headwords

For our experiment, we use data from the first phases of Czech Dictionary Express, a rapid dictionary-making project. The Dictionary Express projects aim to create dictionaries using a semi-automatic method by automatically collecting dictionary data from large corpora and manually annotating them. The process is divided into small tasks (e.g. headword selection, word form selection), which are done simultaneously for all the words before moving on to another task. Each DE project uses a large language corpus with preferably billions of tokens; CDE uses a combination of multi-billion corpora from the Czech web family, known also as csTenTen.(Suchomel, 2018; Kovařík, 2023)

3.1 Headwords annotation

The first task of the DE projects is to create the vocabulary of the dictionary, i.e. to select the desired size and to pick the right headwords. A document frequency wordlist of lemma-POS (part-of-speech) pairs (i.e. headwords) is gathered from the corpus, sorted alphabetically so similar headwords are close to each other, and separated into headword "batches". Batches are then annotated using a customized interface within the Lexonomy web tool (Měchura, 2017), as represented in Figure 1: On the left a set of headwords, all but one (the last one with the "no flag" statement) already annotated, on the right a diagram displayed to help the annotator.

Each headword is annotated at least two times by two different annotators (often more) – native speakers with secondary education, but not academically trained linguists. Afterwards, the more controversial headwords went through another process – revision – described in section 3.2.

The annotators of CDE went through 100,000 automatically lemmatised and POS-tagged letter sequences from the top of the document frequency wordlist. Based on their native speaker intuition, they marked each headword with one of the "flags":

65. "[I] don't know the word",
66. "not Czech",
67. "non-standard" (based on the Czech "spisovnost" standard),
68. "not a lemma" (including non-lemma words and lemmas with typos),
69. "wrong part of speech",

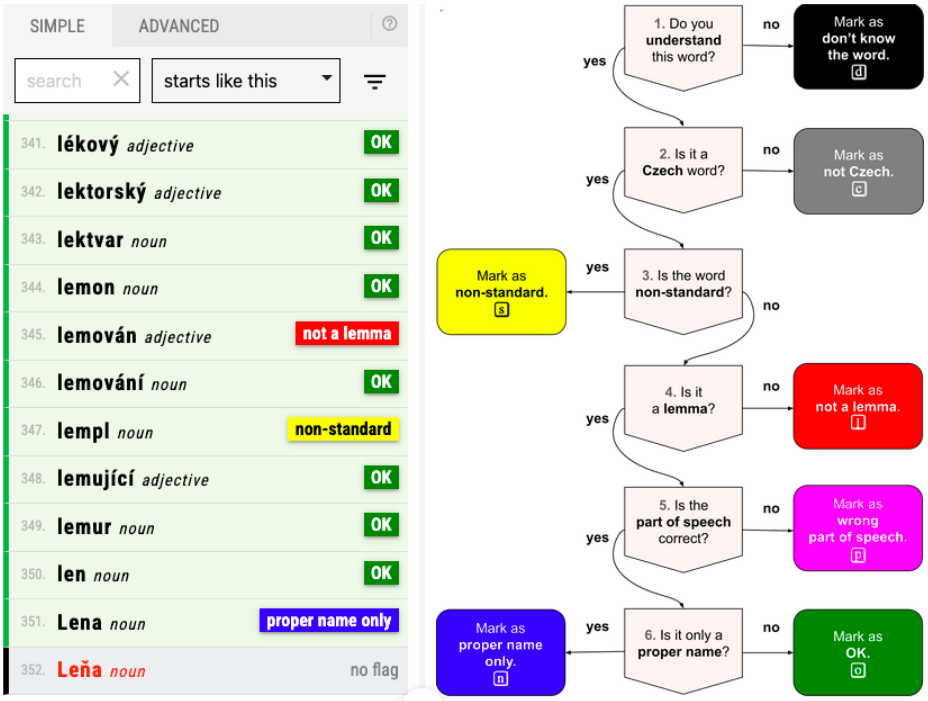

Figure 1: Example of a headword annotation

70. "proper name only" (for words that are correct but only proper names),
71. or "ok" (for words that are correct and common).

This order represents the "weight" of the flag – since only one flag can be assigned to each headword, a word that is "non-standard" and "not a lemma" at the same time should be only marked as "non-standard" (non-standard has a higher priority, as we can see in Figure 1).(Kovařík et al., 2024b)

Before the annotations, the native speakers were given workshops and annotation manuals for each task (headwords annotation, revision, word forms, ...) to synchronise the thought processes behind the annotations. They could also discuss difficulties, many of which are addressed in Kovařík (2023) and Kovařík et al. (2024b), in an online group chat with trained linguists – the dictionary coordinators.

During the annotation, the annotators didn't see the context of the headword and were asked not to search for the headwords on the internet or in other dictionaries.

3.2 Headwords revision

After the headwords have been annotated, another step is the revision phase. For this, a group of experienced annotators, called the inspectors, from the former group has been chosen. The inspectors went through the headwords with insufficient annotation agreement and the ones with the majority of "non-standard", "not a lemma" and "wrong POS" flags. As opposed to the annotators, the inspectors saw the flags given to the words by other annotators, and the context in which the headword can be found in a corpus. They could therefore consider the words more objectively.

Figure 2 shows the interface of the revisions (from top: the headword + the anticipated flag + the annotations; the revision choice window; up to ten examples of usage from the corpus). For each headword, the inspectors could decide if the headword was partially correct or non-standard (in which case they could add the correct or standard variants), completely incorrect, not Czech or a correct Czech word.

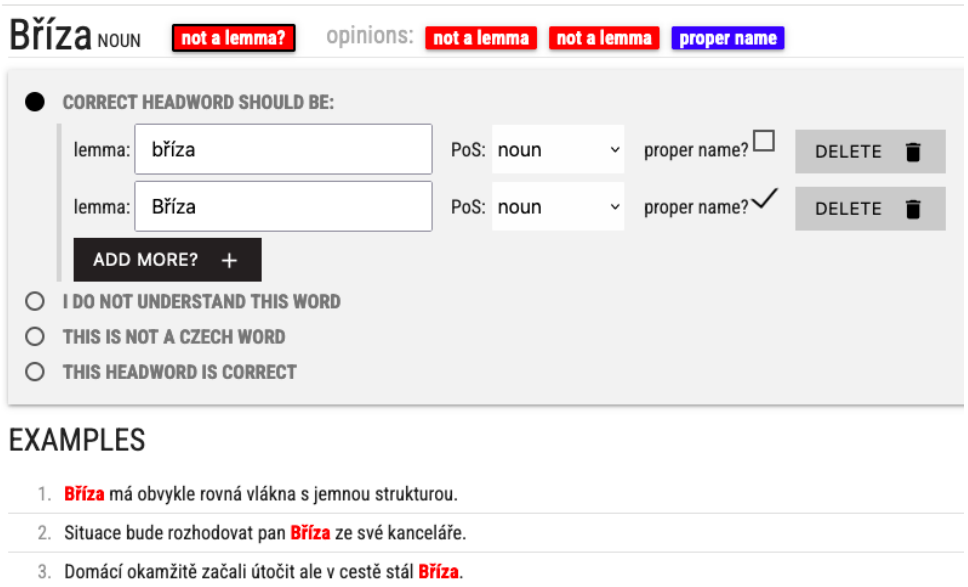

Figure 2: Example of a headword revision; taken from Kovařík et al. (2024a)

4. Czech vocabulary statistics

After the headwords and annotations had been revised, two separate databases of headwords were created – a database of the headword annotations (without revision) and a database of the headwords with the final flag, accepted into the dictionary or rejected. The revised headwords are being used as a vocabulary of the dictionary and for gathering data for the next stages (word forms selection, word sense induction, etc.). The non-revised headwords can be used as a database for inter-annotator agreement (and later intra-annotator agreement as well).

4.1 Current statistics of the vocabulary of CDE dictionary (revised headwords)

The revised headwords, which are outside of the experiment for measuring a native speaker's vocabulary since they represent more of a dictionary vocabulary than the one of the native speakers (the inspectors looked at other people's annotations and context), show the following statistics: From the 100,000 currently annotated letter sequences, 72,278 have been marked "correct" ("ok" or "propen name only") or "partially correct" (i.e. "non-standard", "not a lemma" or "wrong POS") by the most annotators. After the revision (subsection 3.2), 60,723 have been decided to be "ok" (common words) and 12,557 as "proper names only", and new headwords have been added in the task of revision, so the vocabulary now holds 79,756 headwords (63,690 common words and 16,066 proper names).

4.2 Headwords vs word families

According to Brysbaert et al. (2016), Goulden found 54,000 base words, which could relate to some 28,000 word families. However, Webb (2020) states in the Introduction (page 9) that Goulden found the average native speaker knows 15,000–20,000 *word families* (despite that Goulden wrote about 15,000–20,000 of the 54,000 *base words*), implying the number of word families is similar or same to the number of base words. There are more educated opinions on this, and we are not going to get into the business of word families and base words, since this would require a lot of manual work of experienced linguists and the criteria for word families are complicated even in English where they are most researched (for Czech, this could be even worse). Instead, we are looking into the number of headwords, as Keuleers et al. (2015) and Brysbaert et al. (2016) did.

The problem with homographs Goulden et al. (1990) mentioned doesn't concern our experiment, since the annotation process is focused on lemma-POS pairs. We don't focus on word meaning. However, the annotators had to be sure they knew that the word truly existed within the language's vocabulary, implying they knew how this word could be used.

4.3 Statistics of the headword annotations

If we want to learn more about the vocabulary size of the Czech native speakers, we need to look into the non-revised annotation data. We have found the annotation process closely resembles the experiment of Keuleers et al. (2015) and Brysbaert et al. (2016), where native speakers marked letter sequences as "known" or "unknown" in the Dutch/English language. The headword annotation process is a little bit more complicated (a reason for the annotators to get a complex training) with more options to be chosen from, but it can serve a similar purpose.

There are many ways to look at the knowledge of a word.Pignot-Shahov (2012) In this research, we are looking solely at the receptive knowledge of the formal aspect of a headword – what a word *looks* like, not what it *means*. We ignore the categories such as shallow vs deep or precise vs partial knowledge (Henriksen, 1999), word fluency, recognition speed, etc. We also ignore all the categories and problems related to the L2 speakers, since we are testing only the language knowledge of native speakers.

For the Dutch experiment, Keuleers et al. (2015) used "73,500 stimuli (52,847 words and 20,653 non-words)" and found "an average 20-year-old student [...] knows 42,000 lemmas and 4,200 multiword expressions". For the English experiment, Brysbaert et al. (2016) used 61,800 headwords and showed the participant 67 of them, combined with 33 non-words. They found that among English native speakers "[t]he median score of 20-year-olds is 68.0% or 42,000 lemmas; that of 60-year-olds 78.0% or 48,200 lemmas".

Our experiment is different from the two headword experiments in that we don't have a dictionary base of "real Czech headwords" combined with a set of non-word letter sequences. Instead, we use 100,000 most frequent lemma-POS pairs from the document frequency wordlist of the corpus and do not pre-decide which are words in Czech and which are not. If we were to use some database of "correct Czech words", such as revised headwords or an existing dictionary, the results could be more precise because they wouldn't be manipulated by subjective factors such as the task interpretation of the annotator (such as English

words that have only recently been introduced to the Czech language) or accidental errors made throughout the annotation. This would, however, disqualify words that only some annotators know and use because of their geographical, cultural, professional or social setting.

The base for the annotation was the 100,000 most frequent words from the corpus, using the document frequency wordlist. The vocabulary was made gradually. First, the 15,000 headwords were annotated (document frequency rank 1–15,000), then the next 35,000 (rank 15,001–50,000), then another 30,000 (rank 50,001–80,000) and finally another 20,000 headwords (rank 80,001–100,000). We call these frequency groups the "scopes".

From each scope, three or four batches of 1,000 headwords have been selected to represent the frequency scope. Each of these batches has been annotated by 7 different annotators. We can see the statistics in Table 1. The "ok" and "proper name only" columns present the percentage of words that were annotated "ok" or "proper name only" in each scope (regardless of whether these words made it into the dictionary). The "false 'ok'" and "false 'proper name only'" columns present the percentage of words that were annotated "ok" or "proper name only" but didn't make it past the revision phase – these words were either:

- not added to the final vocabulary:
  - in the same form (some examples:
    - the typo "doměnka-n" was added as the correct "domněnka-n";
    - the non-standard "doznět-v" was added as "doznít-n";
    - the headwords with the wrong POS were revised to the correct POS;
    - the proper name "Times-n" was corrected to the full multiword expression "New York Times-n", "Financial Times-n" and "Times Square-n"; etc.),
  - or to the same category (e.g. "Facebook-n" marked by an annotator as "ok" was added as a proper name);
- or were not added to the dictionary at all.

While there might be some words from the false "ok" and false "proper name only" categories that are actual Czech words that only some of the annotators know, efforts were made to accept as many correct Czech words as possible into the dictionary.

Scope	"ok"	"proper name only"	false "ok"	false "proper name only"
1-15,000	86.84%	4.05%	1.00%	0.23%
15,001–50,000	64.57%	12.83%	2.51%	1.25%
50,001–80,000	53.42%	10.65%	2.67%	2.40%
80,001–100,000	32.66%	13.43%	2.68%	3.91%
Average	59.37%	10.24%	2.22%	1.95%

Table 1: Percentage of words marked "ok" or "proper name only" in each scope and in all the scopes combined.

In this experiment, we only take the flag "ok" as the equivalent to the "yes" in the experiment of Keuleers et al. (2015). The "non-standard", "wrong POS" and "wrong

lemma" flags can create homophones ("wrong lemma" was among other things used to mark typos, and a word marked "non-standard" can be the non-lemma or non-standard variant of the word already marked "ok" because of its high annotation priority).

The average participant marked 57.15% (59.37% minus 2.22%) of the headwords later accepted into the dictionary as "ok". The estimated average Czech native speaker's vocabulary size could be at least 57,150 Czech headwords. We expect the number to be even higher because:

72. The percentage of common words marked "ok" in the last rank was bigger than 30%, implying the next scope (words with document frequency rank 100,001–150,000 or even more) still holds a lot of headwords that could be part of the participants' vocabulary. (This also applies to the studies we base our research on – how can we be sure we have the maximal numbers if we are only showing the participant 53,000 lemmas?)

73. There is always a problem with measuring the vocabulary of multiple people, because this doesn't include the specialised vocabularies defined by the user. Most of the participants were university students of linguistics in their 20s – this means some of their sociolect and professional vocabulary is going to be accepted, while the vocabulary of others may not. More on this in subsection 4.4.

74. The "non-standard", "not a lemma" and "wrong POS" could hold some words which the participants know but which weren't counted in the final statistics. (This might add up to the 6,476 words that have been added in the revision, but most likely less than that.)

4.4 The peripheries of users' vocabularies

We mentioned above that Goulden et al. concluded the average vocabulary size of an English user is around 17,000 from some 54,000 base words of English. This means the average native speaker doesn't know more than half of the base words. In fact, as Goulden mentions, some authors like Diack (1975) divide the vocabulary into frequency levels, the first including basic words almost all English speakers know (like *asphalt* or *density*) and the last including English words almost nobody knows or uses (some of which you might have never heard of like *alburnum* and *radula*, but also, to the linguist's satisfaction, *enclitic* too). According to Diack, only a few people ever reach the highest level of knowledge and learn to know almost all the basic words of English.

But language acquisition is rarely done like this – that the speaker should only learn the words from a more difficult group after they have learned all the words from the previous one. The vocabulary of a speaker depends on many factors – work, hobbies, social circle, geographical setting and others. For all these activities, a specialised vocabulary has to be learned, and this can be achieved even though the speaker doesn't know other difficult words in other specialised word-spaces. While the word *house* is known by almost every speaker of English, the *f-hole* is known only to a small number of English speakers, specifically the ones interested in bowed stringed instruments.(Merriam-Webster, n.d.)

When we measure the vocabulary size based on only a general language dictionary or corpus, we measure the general language, i.e. the intersection between vocabularies of

different users. For example, we imagine some of the Czech mycologists have a vocabulary thousands of words bigger than the numbers we have found because they add a long list of mushroom names (e.g. *penízovka*, *kotrč*, *závojenka*, *čirůvka*) and special adjectives (e.g. *žlutomasý*, *polovolný*, *mlženka*), as well as the activities associated with mushroom hunting (and we are not even counting the Latin names). To truly measure a person's vocabulary size, we also need to include a measurement of their specialised vocabulary; otherwise, we are only measuring the size of their general vocabulary and could be missing a lot of data.

5. Conclusion

In this paper, we have looked into the vocabulary size experiments done in the past on English and Dutch. The experiments conclude that a usual native speaker knows approximately 42,000 words in their native language. We have conducted an experiment on the 100,000 most frequent headwords from the document frequency wordlist of a Czech web corpus. We have found native speakers in their 20s with secondary education know at least 57,000 headwords on average. This number could possibly be even bigger if we used 150,000 or even 200,000 headwords as the base of the experiment. Even then, the wordlist may not include many words the participants know in their fields of interest (e.g. their hobbies or professions). However, more experimental data is needed before we make any confident estimations on the vocabulary size of Czech native speakers. This will be the subject of our future research.

Software

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P. & Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. In *Proceedings of the 6th Biennial Conference on Electronic Lexicography*. Brno, Czech Republic: Lexical Computing CZ s.r.o., pp. 805–818. URL https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019_Proceedings.pdf.

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medveď, M., Ohlídalová, V. & Suchomel, V. (2023). Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data. In C.T.I.K.J.K.M.J.S.K. Marek Medveď Michal Měchura (ed.) *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. Brno, Czech Republic: Lexical Computing CZ s.r.o., pp. 613–637. URL https://elex.link/elex2023/wp-content/uploads/114.pdf.

Brysbaert, M., Stevens, M., Mandera, P. & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Front. Psychol.*, 7, p. 1116.

Diack, H. (1975). *Wordpower. Your Vocabulary and Its Measurement.* Paladin. St Albans.

Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11, pp. 341–363. URL https://api.semanticscholar.org/CorpusID:145483070.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Stud. Second Lang. Acquis.*, 21(2), pp. 303–317.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68(8), pp. 1665–1692. PMID: 25715025.

Kovařík, F. (2023). Semi-automatic Dictionary Creation for Czech. *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)*, 17. URL https://nlp.fi.muni.cz/raslan/raslan23.pdf.

Kovařík, F., Blahuš, M., Cukr, M., Jakubíček, M. & Kovář, V. (2024a). Dictionary Express : First Phases Rapid dictionary-making method for European, Asian and other languages. In N.S.M. Inoue Ai; Kawamoto (ed.) *AsiaLex 2024 Proceedings : Asian Lexicography - Merging cutting-edge and established approaches*. Tokyo: Toyo University, pp. 84–89. URL https://www.asialex.org/pdf/Asialex-Proceedings-2024.pdf.

Kovařík, F., Kovář, V. & Blahuš, M. (2024b). On Rapid Annotation of Czech Headwords : Analysing the First Tasks of Czech Dictionary Express. In A.O.B.I. Despot Kristina Š.; Anić (ed.) *Lexicography and Semantics : Proceedings of the XXI EURALEX International Congress*. Cavtat: Institut za hrvatski jezik, pp. 336–344. URL https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf.

Merriam-Webster (n.d.). F-HOLE. In *Merriam-Webster.com dictionary*. URL https://www.merriam-webster.com/dictionary/f-hole. Accessed: 2025-6-28.

Měchura, M. (2017). Introducing Lexonomy: an Open-source Dictionary Writings and Publishing System. In M.J.J.K.S.K.V.B. I. Kosem C. Tiberius (ed.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing.

Pignot-Shahov, V. (2012). Measuring L2 Receptive and Productive Vocabulary Knowledge. *Language Studie Working Papers*, 4(II), pp. 37–45.

R.L.G. (2013). Lexical facts. URL https://www.economist.com/johnson/2013/05/29/lexical-facts.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In P.R. Aleš Horák & A. Rambousek (eds.) *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018*. Brno: Tribun EU, pp. 111–123. URL https://nlp.fi.muni.cz/raslan/2018/paper10-Suchomel.pdf.

Těšitelová, M. (1987). *O češtině v číslech*. Academia, malá jazyková knižnice edition.

Webb, S. (2020). *The Routledge Handbook of Vocabulary Studies*. Routledge Handbooks. New York, NY.