

František Kovařík, Marek Blahuš, Miloš Jakubíček &
Vojtěch Kovář

AGREE TO DISAGRE

Intra-Annotator Agreement in Semi-Automatic Vocabulary Building

Abstract Intra-annotator agreement is the rate at which an annotator makes the same choices in the same situation. This paper examines the annotation process in a semi-automatic dictionary-making project, Czech Dictionary Express. In the vocabulary-building process, its annotators went through 100,000 headwords from a corpus frequency wordlist, marking each as incorrect or (partially) correct. The aim of the Dictionary Express projects is a high dictionary-making speed, so the annotators weren't given the context of the headwords.

In this paper, we examine the shift in annotators' opinions on headwords before and after context was provided. We argue against providing no context of the headwords, because, unlike the (partially) accepted words, which can still be cut from the vocabulary at later stages, the rejected words cannot be put back in, and this is the case of some correct words that have not been recognised by the annotators without context. We show that the context-based annotation process doesn't take significantly longer. We also take a brief look at a study about using LLMs for the same annotation process.

Keywords Inter-annotator agreement; intra-annotator agreement; dictionary drafting; post-editing lexicography; corpus annotation; semi-automatic dictionary making; dictionary express o Czech

1 Introduction

The process of creating any dictionary at some point incorporates vocabulary selection – a stage of thousands of macro- and micro-level choices about the composition of the dictionary wordlist and the specific words that get to be in the dictionary and receive the often resource-expensive lexicographic treatment. The semi-automatic approach to this problem, described in this paper using the Dictionary Express methodology (formerly “rapid dictionaries from scratch” (Blahuš et al., 2023)), consists of automatically assembling frequency wordlists and subsequently annotating them by native speakers. This annotation is as simple as possible, requiring only an answer on whether the word is suitable for recording in a lexicographic resource; the rest of the dictionary entry for the accepted words is created in later tasks (more in Section 2).

One problem with manual annotation is whether a human annotator's introspection can be trusted for long-term consistency – especially since projects like Dictionary Express use native speakers who are not fully trained linguists (although problems with word acceptance can arise as frequently inside as outside the linguistic community). The rate by which an annotator agrees with their own past decisions, with reference to the inter-annotator agreement (IAA; “agreement among annotators”), is sometimes called the **intra-annotator agreement** (IaAA).

In this paper, we examine data from two separate annotation processes in the Czech Dictionary Express project (see Section 2), carried out by a group of annotators twice, with a one-year gap between the two experiments. We ask whether and how certain techniques and interface changes can influence annotators' micro-choices about accepting a word in their native language, and whether these techniques and changes could improve or harm the selection quality of the

final vocabulary (see Section 3). In the final part, we apply statistical measures to the intra-annotator data and mention a similar experiment conducted on LLMs (see Section 4).

2 Semi-automatic Vocabulary Selection

Dictionary Express is a group of projects that aims to create general-purpose, descriptive general-language dictionaries for different (often small) languages using a semi-automatic approach. DE dictionaries often include forms of both proper names (place names, people names, abbreviations, organisation names, etc.) and forms, senses, thesaurus synonyms and English translations of common words, and more. The methodology uses data gathered from large web corpora, manually annotated by a group of native speakers. The projects are divided into smaller, simple tasks, each creating chunks of dictionary entries for the whole vocabulary, with the first task – the headword selection – centred around creating the vocabulary itself (Kovařík, 2023).

The annotation team usually consists of 10 or fewer annotators, most of whom are not university-level linguists. This is often due to the lack of expert linguists or resources in small-language native-speaker communities. Although Czech lacks neither, the Czech Dictionary Express (CDE) project simulates this environment, allowing researchers to experiment with the tools (Kovařík et al., 2024).

At the beginning of each project, a list of general-language words (single words only; no multiwords are included at this stage) is gathered and alphabetised (so similar words are close to each other) for different frequency ranges. For CDE, the wordlist was gathered from the Czech Web 2012+17+19 corpus, which contains over 11,7 billion words (*CsTenTen – Czech Corpus from the Web*, 2024).

After the initial annotation workshops, the DE annotators are given the task of *headword selection* – they go through batches of the general-language wordlist and assign each *headword* (usually lemma-POS pairs) one of seven flags: *I don't know the word, not Czech* (not in the language of the dictionary), *non-standard, not a lemma* (also for wrong lemmas, typos etc.), *wrong POS*, and finally *proper name* and *ok* (for correct common words).

Until recently, this has been done without providing the typical context of the headwords. The annotators would only see the wordlist and a simple diagram of the flag priority (e.g. if a word can be marked *non-standard* and *not a lemma*, it should be marked *non-standard* since this flag has a higher priority). The context from corpus concordances would be provided at a later stage, in the revision task, in which another group of annotators would go through a smaller list of headwords that have been marked with **different** flags, or flags **of a specific problem** (*non-standard, not a lemma, wrong POS*), but **not the “rejection” flags** (*I don't know the word, not Czech*).

The project coordinators assumed that combining a context-free annotation with a context-based revision would lead to a quicker vocabulary-building process (we argue against this in Subsection 3.4). It was not considered desirable to provide the initial annotators with context, since each word has to be marked by at least two different annotators, whereas revisions are done on a much smaller part of the candidate vocabulary by a single annotator. There have been concerns, however, about the desired objectivity of the vocabulary, and especially about the proper Czech words unanimously rejected by the annotators working under time pressure and without context because they simply didn't recognise a lemma-POS form of a commonly used word. Where the human factor comes into play, errors are always to be expected.

Headwords rejected in the annotation phase were not taken to the revision nor the vocabulary, thus “lost forever” to the project. The problem of incorrectly rejected headwords is a type II (false negative) error. It leads to correct words omitted in the dictionary, and presents a much bigger problem in the semi-automatic methodology than the opposing type I (false positive) error: Where accepted incorrect words can be discarded in the later stages of the projects, the rejected correct words can never be put back into play (except on rare occasions in which the lexicographers/coordinators notice the missing word themselves).

To assess this problem, another interface feature has been added to the headword annotation task recently, and serves as a part of our experiment: If an annotator would assign the headword a different flag than *ok* (common word) or *proper name*, they would get a message with an additional question on whether they really want to do that, accompanied by the common context of the word in given language (from concordances of the word in Czech Web corpus), as seen in Figure 1. In this paper, we refer to this feature simply as *The Message*.

Applying this approach may yield more vocabulary headwords, but it may also slow down the annotation process. In the next section, we discuss whether the headwords not omitted thanks to *The Message* are worth the delay, how many there could be, and whether they could serve as proper candidates to the recorded vocabulary.

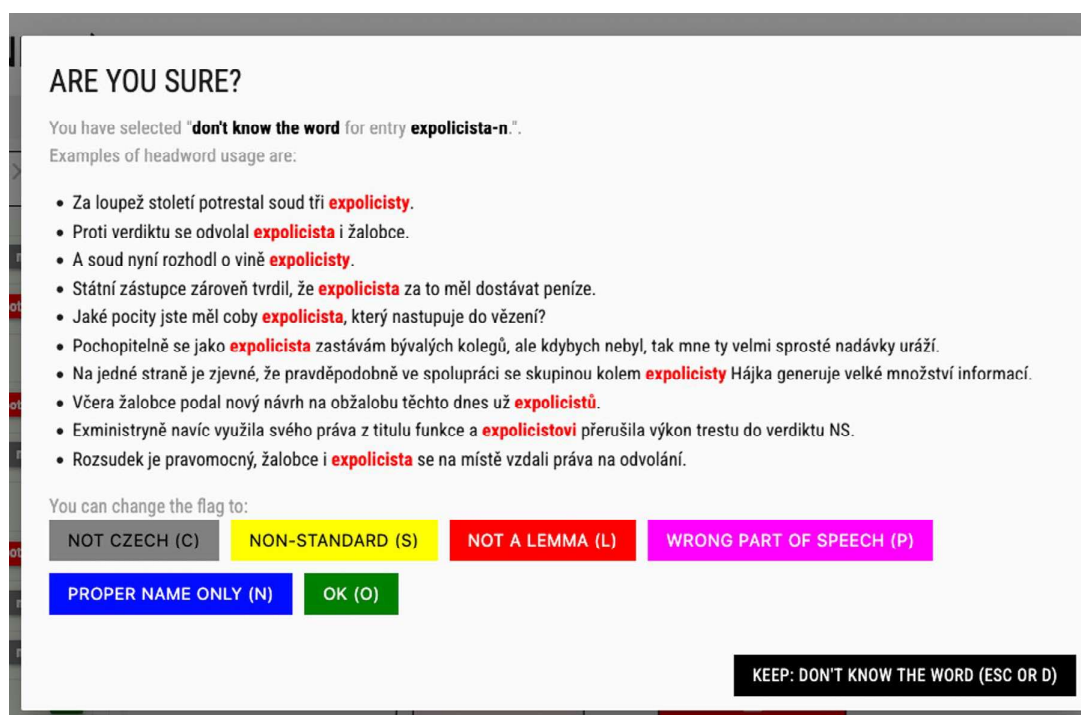


FIGURE 1 The Message appearing after marking the noun “expolicista” as I don’t know in the Lexonomy (Měchura, 2017) interface

3 IaAA

In this experiment, three Czech native speakers have each been given 14 batches of 1,000 headwords (14,000 headwords in total) from different document frequency ranks: 3 batches from 15,000 most frequent headwords, 4 batches from the next 35,000 (ranks 15,001–50,000), 3 batches from the next 30,000 (ranks 50,001–80,000), 3 batches from the next 20,000 (ranks 80,000–100,000), and finally 1 batch from the very low frequency ranks

(ranks 100,001–150,000). They annotated each batch twice, with a one-year gap between annotations. The first annotation didn't include *The Message* with context (Figure 1). In the second annotation, two of the three annotators (annotator 1 and 2) have been warned by *The Message* when selecting a flag other than *ok* or *proper name*; the third annotator (annotator 3) has never been given *The Message* except for the last batch. (Since contributing to a separate experiment, the last batch of the ranks 100,001–150,000 received a special treatment, and has been annotated twice with a gap of several weeks, first time without and second time with *The Message*.)

3.1 Quantitative analysis: First vs second annotation

In this subsection, we present the data gathered in the experiment, with the focus on the difference between the numbers of “rejection” flags (*I don't know* and *not Czech*), which do not lead to putting the word into dictionary and can possibly disqualify the word from the subsequent revision process with context, and (partial) “acceptance” flags (*non-standard*, *not a lemma*, *wrong POS*, *proper name* and *ok*).

Table 1 presents the intra-annotator statistics of each annotator. *Words annotated* is the number of words annotated in the experiment (the third annotator saw slightly fewer words because of a software error). *Acc (1st annotation)* and *acc (2nd annotation)* are the amounts of words accepted; same goes with *rej (1st annotation)* and *rej (2nd annotation)*, which have been rejected the first or second time.

The columns correspond to the frequency ranks of the words.

It is important to understand that not every annotator works the same way. First off, we see that the words in the most frequent ranks get accepted more often in the first annotation, and that annotators 1 and 3 tend to accept fewer words than annotator 2 (all the annotators have been given the same words). We can also see that the annotator 3, who has not been given any headword context for the first or second time (except the last batch), tended to accept words more often in the second run, and that this shift is pretty close to that of annotator 2, who has been given *The Message* in the second run.

Annotator 1, however, given the context of the initially rejected headwords (*2nd annotation*), has accepted almost every headword candidate. By giving the context, the *ok*- or *proper name*-flag rate has risen by almost 500 out of 3,000 in the third rank, and by more than 900 out of 3,000 in the fourth rank. The rate of the other (partially) accepted flags has also increased. If most of these headwords are proper words suitable for dictionary recording, this would be a great help for the semi-automatic drafting. We look into some of them in the next subsection.

The acceptance rate has increased for all annotators in the second run, i.e., for those given *The Message* as well as for those who were not. We should thus remember that *The Message* is not the only factor at play, and that the annotators have a year of annotation experience between the two experiments. However, we argue that it can be a powerful tool for the annotators in the next subsection, because it is not always simple to recognise the word or lemmatisation/tagging error by looking at a lemma-POS pair without any context.

3.2 More on the acceptance shift

A word that presented a problem for the annotators in the early stages of the project was the adjective *rybařený*. Although the word is known for some Czech speakers as a rare slangism

TABLE 1 Word counts of accepted and rejected words in the first and second annotation

frequency rank	1-15,000	15,000-50,000	50,000-80,000	80,000-100,000	100,000-150,000
annotator 1					
words annotated	3,000	4,000	3,000	3,000	1,000
acc (1st annotation)	2,907	3,478	2,251	1,587	803
acc (2nd annotation)	2,982	3,984	2,966	2,945	977
rej (1st annotation)	93	522	749	1,413	197
rej (2nd annotation)	18	16	34	55	23
annotator 2					
words annotated	3,000	4,000	3,000	3,000	1,000
acc (1st annotation)	2,908	3,733	2,541	2,349	591
acc (2nd annotation)	2,947	3,796	2,668	2,430	759
rej (1st annotation)	92	267	459	651	409
rej (2nd annotation)	53	204	332	570	241
annotator 3					
words annotated	2,999	3,998	2,994	2,998	1,000
acc (1st annotation)	2,918	3,382	2,226	1,820	479
acc (2nd annotation)	2,927	3,588	2,401	1,978	616
rej (1st annotation)	81	616	768	1,178	521
rej (2nd annotation)	72	410	593	1,020	384

(in some forms of the climber slang, e.g. *rybařená hora* as “fished mountain” – a mountain that someone is trying to climb), it should be regarded as more or less an occasionalism, an infrequent word with unclear sense. And yet, it appeared amongst the most frequently used headwords.

While looking at the context, it was clear that it was a mistake made by the lemmatiser. The correct lemma should be the noun *rybaření* (fishing), which indeed is a common and frequently used Czech word. Given the word without its context, the annotators rejected the word, marking it *I don't know*.

The adjective *rybařený* belongs to the *rej*→*acc* words, i.e. words that have been rejected in the first run (and lost forever), but (partially) accepted in the second run (have the potential to be saved by their context). While looking into the words that the annotators 1 and 2 rejected in the first run, but accepted after looking at the context, we mostly see:

- abbreviations and foreign company names, e.g. *GSM*, *ICT*, or even *International* (supposedly as a part of a longer company/organisation name);
- incorrectly lemmatised words (e.g. the nouns *plavka*, or city name *teplce* without a capital letter) – these can still occur in the dictionary if their “correct” counterpart (*plavky* – “swimsuit”, *Teplce*) is frequent enough, which, however, is not the case for many (e.g. mentioned noun *rybaření*);
- words hard to recognise without context, such as *transparent* (noun) – “banner”, *pozměňovací* (adjective) – “amendatory”, *disperze* (noun) and *disperzní* (adjective) – “dispersion” and “dispersive”, *disponibilní* (adjective) – “disposable”, *dohrávka* (noun) – “postponed

match”, *vleknout* (verb) – “to drag (on)”, *vylhany* (adjective) – “lied out”, *brejkový* (adjective; connected to break dance), *brzdovka* (noun; sometimes instead of *brzdová kapalina* – “break fluid”), *buchar* (noun) – “power hammer”, *bující* (adjective) – “proliferating”, *leitmotiv* (noun), *pučista* (noun) – “putschist”, *psychedelie* and *psyché* (nouns) – “psychedelia” and “psyche”, *pudit* (verb) – “to force out”, etc.

TABLE 2 Intra-annotator agreement (how many times did the annotator select **the exact same flag** as before) and Cohen’s kappa (weighted on “rejected”, “partially accepted” and “accepted”).

frequency rank	1-15,000	15,000-50,000	50,000-80,000	80,000-100,000	100,000-150,000	All
annotator 1						
agreement	2808	3260	2085	1422	679	10254
disagreement	192	740	915	1578	321	3746
Cohen’s Kappa	0.572	0.37	0.324	0.177	0.479	0.349
annotator 2						
agreement	2774	3381	2348	2132	662	11297
disagreement	226	619	652	868	338	2703
Cohen’s Kappa	0.75	0.683	0.708	0.656	0.656	0.715
annotator 3						
agreement	2827	3195	2330	2150	696	11198
disagreement	172	803	664	848	304	2791
Cohen’s Kappa	0.758	0.635	0.737	0.708	0.669	0.742

Note that the words from the last group are not only terms used in a specific specialist text types (*disperze* or *pozměňovací*), but also common words (*vylhany*) or their variants (*vleknout* as a less used variant of *vléci*).

3.3 Agreement and Cohen’s kappa

For evaluating the data, we cannot use the final vocabulary as reference “correct words”, since there might have been words omitted because of a false negative error (e.g. the noun *rybaření* “disguised” as adjective *rybařený*). We use Cohen’s Kappa to evaluate the intra-annotator agreement.

Table 2 presents the kappa and agreement statistics of the experiment for each frequency rank, and for all words together. For each annotator, the first line presents the words for which the annotators selected **the exact same flag**, the second the words where the two flags were different. For the kappa statistics, we are using Cohen’s Kappa quadratically weighted for 3 values:

1. Rejection flags: *I don’t know* and *not Czech*.
2. Partial acceptance flags: *non-standard*, *not a lemma* and *wrong word*.
3. Acceptance flags: *ok* and *proper name*.

In the overall score, annotators 2 and 3 get a kappa of over 0,7 (substantial agreement (Viera & Garrett, 2005)), whereas annotator 1 gets a kappa of 0,349, which isn’t as high as we could expect (possibly because of *The Message*).

We can also see that, even though agreement on the exact same flag decreases with frequency, it is not the case with the weighted kappa between “rejected”, “partially accepted”, and “accepted”.

3.4 The Demands of Context

One of the advantages of semi-automatic dictionary making is its speed. The Dictionary Express projects were previously called “rapid dictionary-making projects”. One of the concerns of providing the headword annotators with context was that it could slow down the vocabulary-building process.

Thanks to good language resources for Czech (corpus, lemmatiser and POS-tagger), the usual headword annotation is faster compared to other “medium-sized” languages. A batch of 1,000 headwords takes about 1,5–2 hours on average for one annotator to annotate, without looking into the context.

Our results show that annotators 1 and 2, given the headwords without context, finished a batch in approximately 1,63 and 1,80 hours. While working with *The Message* (i.e. given word context if they would reject a word), the first annotator finished a batch in approximately 2,58 hours on average, whereas the second one even faster than before, in 1,52 hours.

While keeping in mind that both annotators had a year’s worth of experience more at the second run and could have gotten faster at their work, the work didn’t slow down significantly while providing context.

On the other hand, adding context to the annotations might slow down the revision process, because it yields more words with the partial acceptance flags. However, based on semi-automatic dictionary coordinators’ experience, the revision task takes up only a small amount of time compared to other tasks (headword annotation, sense disambiguation, etc.).

4 Large Language Models as annotators

In her bachelor’s thesis, Černá¹ (2025) has tested whether ChatGPT-4o or other large language models could substitute or supplement the work of the annotators. Part of her experiment included intra-model testing. ChatGPT-4o has been given words from the 100,000 document frequency wordlist 3–6 times. Consensus on a final flag was reached if a flag was assigned two more times than any other (i.e., in 3:1, 3:1:1, 4:2, 4:1:1, or 3:1:1:1 ratio). If no flag had the most “votes” in the 6th run, no flag was assigned.

Table 3 was taken from Černá’s thesis. The second column is the number of occurrences, i.e. the number of annotated headwords. The fourth column shows the percentage of these words in the experiment.

Most of the words (77.40 %) reached a consensus on a flag in the first three runs or four runs. On the same dataset (100,000 most frequent headwords), annotator 3, who didn’t receive *The Message*, had an agreement of 80.85 %; the other annotators had agreements of 81.81 and 73.65 %.

¹ While not being a direct part of the DE team, Černá was collaborating with the project and has made several important contributions, particularly about the use of LLMs.

Although we don't expect human annotators not to reach consensus after the 5 runs as LLMs did, the “intra-annotator agreement” of GPT is apparently close to the IaAA of native speakers. Keep in mind, however, that this experiment was done on a fairly resourced language – Czech – and the consensus of the large language models might be less accurate for smaller languages.

TABLE 3 Consensus in different runs of ChatGPT-4o annotation. (Taken from Černá, 2025.)

	Number of occurrences	Percentage
3 runs (3:0)	377	75.40 %
4 runs (3:1)	10	2.00 %
5 runs (3:1:1)	26	5.20 %
6 runs (4:2, 3:1:1:1)	9	1.80 %
6 runs, weak consensus (3:2:1, 2:1:1:1:1)	15	3.00 %

5 Conclusion

This paper has presented the statistics of agreement in the Czech wordlist annotation.

The vocabulary creation for Dictionary Express resources does not end with a single annotation of each headword by a single annotator, without context. Instead, it is backed up by several opinions from different annotators and the work of an “inspector”, another, more skilled annotator who sees the context of use of each word, as well as the previous annotations, and thus has a better objective overview of the headword. This leads to greater objectivity in the language vocabulary selection, done with minimal resources (number of annotators, time, etc.). Human error is expected at some rate in manual annotation.

However, given that only the headwords accepted in the annotation get to see the revision process, we argue that the annotation process should not go without context. Another approach would be to include fully rejected headwords in the revision process, for a “third opinion”. We have shown that, given the context, annotators changed their minds about rejecting some words, incorrectly lemmatised, POS-tagged, or just unrecognised in their lexicographic form. We have also shown that the time spent annotating doesn't significantly change when the annotators are given the context. The demands on time in the annotation and revision process would rise slightly, but provide more correct headwords.

References

- Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medved, M., Ohlidalová, V. & Suchomel, V. (2023). Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data. In C. Tiberius, M. Medved & M. Měchura (Eds.), *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference* (pp. 613–637). Lexical Computing CZ s.r.o. <https://elex.link/elex2023/wp-content/uploads/114.pdf>
- Černá, S. (2025). *Czech Headword Annotation: Native Speakers Vs Large Language Models* [Bachelor's thesis, Masarykova univerzita, Filozofická fakulta, Brno]. <https://is.muni.cz/th/jzl31/>
- CsTenTen: *Czech Corpus from the Web* (2024). <https://www.sketchengine.eu/cstenten-czech-corpus/>
- Kovařík, F. (2023). Semi-Automatic Dictionary Creation for Czech. In A. Horák, P. Rychlý & A. Rambousek (Eds.), *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)* (pp. 93–100). NLP Consulting, s.r.o. <https://nlp.fi.muni.cz/raslan/raslan23.pdf>

- Kovařík, F., Kovář, V. & Blahuš, M. (2024). On Rapid Annotation of Czech Headwords: Analysing the First Tasks of Czech Dictionary Express. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress* (pp. 357–366). Institut za hrvatski jezik. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf
- Měchura, M. (2017). Introducing Lexonomy: An Open-Source Dictionary Writings and Publishing System. In M. Jakubíček, I. Kosem & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference* (pp. 662–679). Lexical Computing.
- Viera, A. J. & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, 37(5), 360–363.

Contact Information

František Kovařík

Lexical Computing/Faculty of Arts, Masaryk University, Brno, Czech Republic
frantisek.kovarik@sketchengine.eu

Marek Blahuš

Lexical Computing
marek.blahus@sketchengine.eu

Miloš Jakubíček

Lexical Computing/Faculty of Informatics, Masaryk University, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Vojtěch Kovář

Lexical Computing/Faculty of Informatics, Masaryk University, Brno, Czech Republic
vojtech.kovar@sketchengine.eu