
František Kovařík, Vojtěch Kovář, and Marek Blahuš

ON RAPID ANNOTATION OF CZECH HEADWORDS

Analysing the First Tasks of Czech Dictionary Express

Abstract Czech Dictionary Express has been introduced as a project of a semi-automatically made dictionary of the Czech language. (Kovařík, 2023) The Dictionary Express method (formerly known as *rapid dictionaries*) has been used for several different languages (Baisa et al., 2019; Blahuš et al., 2023). In this paper, we analyse the automatic and manual tools used in Czech Dictionary Express and inspect the statistical and qualitative data such tools provide. As the first task of the project – the headword annotation – comes to an end, we examine some opportunities and difficulties of the method used, as well as the data acquired in the process.

Keywords dictionary drafting; post-editing lexicography; corpus annotation; semi-automatic dictionary making; Dictionary Express; Czech

1. Introduction

Czech Dictionary Express (hereinafter referred to as CDE) is a project for creating a Czech dictionary from scratch. It uses a list of headwords generated automatically from large web corpora. This list is subsequently manually checked and revised by a team of Czech native speakers (the *editors*), who also inspect word forms and help disambiguate word senses in the later tasks. The process is controlled by supervisors (the *coordinators*), trained linguists and lexicographers.

One of the goals of CDE is to analyse the rapid dictionary-making method Dictionary Express (DE) used for several different languages. (Baisa et al., 2019; Blahuš et al., 2023) In the last 50 years, there has been a great effort in automating the dictionary-making process, so lexicographers can focus on more challenging and interesting tasks and linguistic phenomena. (Rundell & Kilgarriff, 2011) Dictionary Express projects incorporate tools that help simplify dictionary making. Their aim is the possibility to create a dictionary of any natural language with a reasonably big corpus in a relatively short time span (ideally within a year) with a team of native speakers (not professional linguists, with secondary education), supervised by professional linguists and lexicographers who don't need to speak the described language (in the case of CDE, they do). (Further information on the DE methodology can be found on the <https://dictionary.express/> website. (DE, 2023))

This paper is to show the process of the headword annotation (the first manual task of the method) and interesting data gathered in the process of acquiring Czech headwords to make conclusions about the advantages and/or challenges of the semi-automatic approach.

2. Headword Annotation Task

The lexicon of the dictionary data was acquired from *Czech Web* (also known as *csTenTen*), a large Czech corpus consisting of texts from three large Czech web corpora: *Czech Web 2012, 2017* and *2019*. (Suchomel, 2018) The texts were deduplicated and parts of the corpora were discarded for containing spam and automatically translated nonsensical texts. (Further information on *Czech Web* corpora family can be found on the web page. (cst, 2024)) The corpus was automatically lemmatized using the tools *Majka* and *desamb* (both recognized tools for the Czech language working with the Czech attributive tagset described by Jakubiček & Kovář, 2011). (Šmerk, 2008; 2009)

After the automatic acquisition, the project moved to the headword annotation task. In this section, we look closely at the annotation process and discuss interesting linguistic phenomena that have been discovered during the annotation process, mostly associated with a subjective view of word form, lemma form, and language standard.

2.1 Headword Annotation

The annotator team consists of 8 editors. 100,000 most frequent headword candidates (supposed lemmas with POS tags, e.g., *místnost-noun*) are annotated with a flag from a list of possible flags (see diagram in Figure 1): The word can be a non-word (the **I don't know** flag), a proper but **not Czech** word, a Czech but **non-standard** word, a standard Czech word but **not a lemma**, a standard Czech lemma but incorrectly POS-tagged (**wrong POS**), or it can be accepted as a **proper name** or a common word (the **OK** flag).

Before the task had started, each annotator was presented with an online course on lemmata and POS-tags, presented on the English language. They also attended a workshop where they learned about the task and some specific situations that might occur. With the help of an annotation manual, they then trained for the task on a small batch of word.

While working, the annotator sees a diagram (Figure 1) similar to the one used in the Ukrainian dictionary project. (Blahuš et al., 2023) The flag-assignment can be done using a mouse, but using a keyboard is faster and is preferred. The diagram shows the flags' colour coding and a button for the keyboard assignment of each flag.

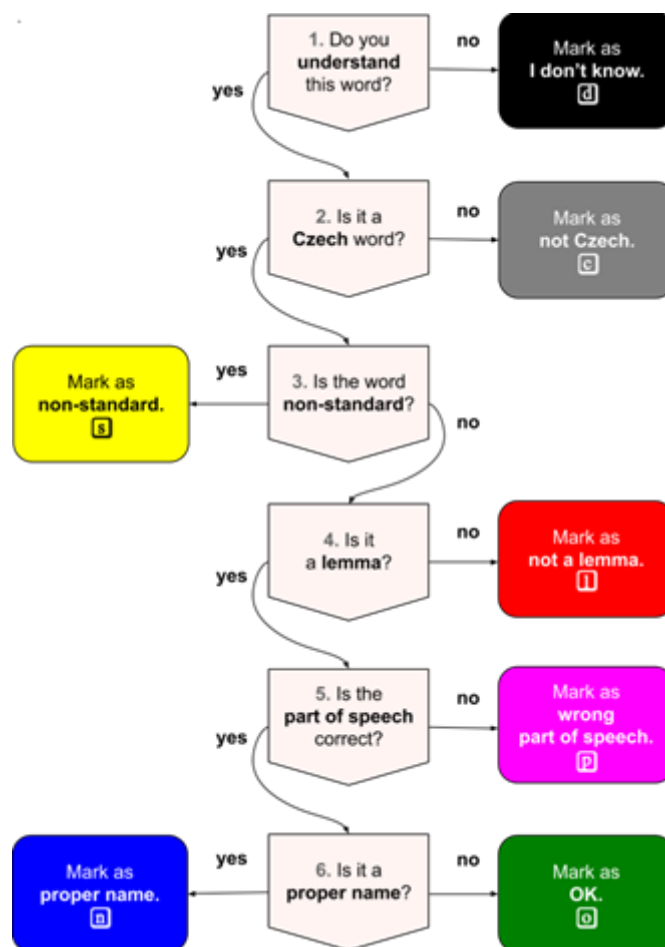


Fig. 1: Key to attributing flags to headword candidates.

The annotators didn't know the context of any particular headword and were advised against searching the words on the internet or in existing dictionaries. This made the process faster and prevented copying data from existing dictionaries.

For the annotation, the headword list is ordered alphabetically. The reason for this is that the annotator sees similar words close together, and can easily identify which words from the group are correct/incorrect.

2.2 Disagreement Patterns

Headwords are annotated at least twice. The 15,000 most frequent words have been annotated three times and 1,000 words have been annotated eight times, once by each annotator.

After analysing the data from the first 15,000 annotated headwords, it is possible to notice some rough patterns. Especially disagreement between two or more different annotators gives insight into what problems we should concentrate on.

Some level of disagreement is suppressed by applying **the presumption of correctness** (i.e., “if the headword suggested by the lemmatizer/tagger can be

considered correct, it should”) and pointing out expected problems (as described in Kovařík, 2023). The presumption of correctness is an important principle. The CDE is centred around swift dictionary making and language phenomena that make sense for the average educated native speaker, not around “linguistic purity”. The task of the annotator is to correct the assumption of the lemmatizer/tagger only if it is transparently wrong, in contradiction with the target language.

However, not all difficulties can be solved like this and after some time we could see new recurring problems:

- Some annotators marked the **diminutives** with the *not a lemma* flag. Diminutives usually go as separate lemmas in Czech and often have a different meaning than just a simple diminution of the non-diminished word. Diminutives should thus be treated as separate lemmas and not forms of their non-diminished relatives.
- The manual emphasises that **short forms of adjectives** (e.g., *zdráv*) should be marked with the *not a lemma* flag. Only the long (the so-called “složený”, composed) form of adjectives (e.g., *zdravý*) should be accepted as a lemma. Despite this, annotators continued to mark the short forms of adjectives *OK*.
- As expected, controversy arises over the **view of the standard** (a classification of a word as “standard” or “non-standard”). This includes disagreement over the **dialectisms** (e.g., *tož*), **past forms** (mostly the difference between the suffix *-ismus* and more recently accepted *-izmus*; some words ending with *-ismus* are viewed more historical than others by different people) and more (e.g., the difference in *o/ó*, like in the word *salon/salón*). Most of these words are marked *non-standard* by one annotator and *OK* by another.
- For some words, it’s difficult to **decide the POS**. For example, the word *zima* is a noun but can also in some sentences be considered an adverb: *Je mi zima*. Another problem is the nominalised adjectives (nouns derived from adjectives) that, with no context, can be mistaken for pure adjectives. Examples of these are *zlatá-noun*, *zelená-noun*, *výborná-noun*, *známý-noun*, *ženská-noun*, *zraněný-noun*. All of these should be accepted (because of the presumption of correctness). Controversy also arises over words that are considered pronouns or numerals but are used as other parts of speech in a sentence (e.g., *víckrát* could be a numeral or an adverb).
- Does the **non-negated form** exist? Is it used in Czech? This question arises frequently, e.g., for the words *překonatelný*, *přetržitý*, *vyhnutelný*, *vyzpytatelný*, *zbytný* or *zvykle*.
- In many cases, one annotator marked a headword *OK* and another *not Czech*. This disagreement arises most often in the cases of **words coming from English** that are only partially accepted by the Czech native speakers (*link*, *market*, *teenager* etc.), or in the cases of the only partially accepted

proper names (*Juraj, Trump, Times* etc.) or name forms (*Jozef* which is more a Slovak form than Czech etc.).

- Deciding whether a word coming from other languages is or isn't part of target language is a difficult linguistic task which many times doesn't have a clear answer. However, the Dictionary Express method can at least provide useful data on how the native speakers approach a given (foreign) word.

2.3 Inter-Annotator Agreement and Annotation Statistics

One of the aims of the CDE project is to study the view of the language from the perspectives of annotators. This means not only defining the language-specific problems but also observing the annotation statistics.

Statistics provided by the frequency-acceptance relationship show that the less frequent a word is, the less likely it is going to be a proper Czech word (Figure 2). Hence, there is more or less a direct proportion between word frequency and chance of accepting the word.

Figure 2 also shows how the results of each annotator differ. Less frequent words show a greater difference between some annotation statistics. The outstanding results (the two lowest) were checked manually by the coordinators of the project, and are caused mostly by the alphabetical order of annotated headwords. They are connected to the quality of the particular annotations.

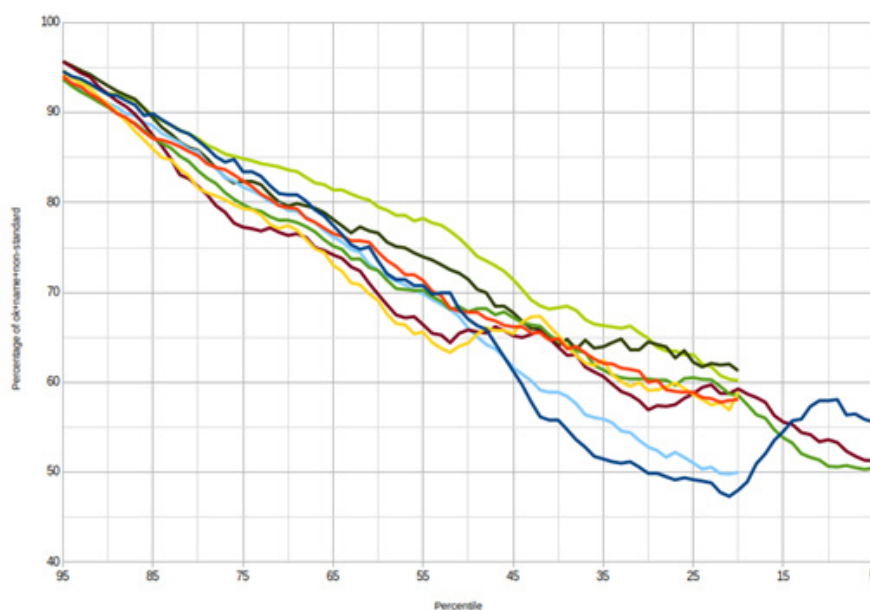


Fig. 2: The less frequent the headword is in the corpus, the less likely it is going to be a proper Czech headword. Vertical axis presents how many words have been marked as *OK*, *proper name* or *non-standard*. Horizontal axis represents frequency (most frequent words on the left) divided by percentile. Each coloured line represents an annotator. Calculated from 95,000 annotated headwords.

Especially useful is the number of cases when the annotators' opinions are the same and cases when they are different – the so-called inter-annotator agreement. Tables 1 and 2 show statistics of 205,926 headword annotations.

- **Table 1** demonstrates, how many times annotators agreed upon a flag marked to a headword.
 - In the left column: The first number is the maximum times the annotators agreed upon a flag for a headword. The second number is the times the headwords have been annotated. Hence, “4/8” means “headwords for which a maximum of 4 annotators agreed on a flag and which have been annotated 8 times”.
 - In the right column are the counts of these headwords.
- **Table 2** shows the statistics of flags marked to words with an inter-annotator agreement greater than 50 %. E.g., 47,593 headwords were mostly marked *OK*.

Maximum agreements / times annotated	Number of headwords
1/1:	15,000
1/2:	12,784
2/2:	42,290
1/3:	782
2/3:	3,813
3/3:	18,331
2/4:	18
3/4:	78
4/4:	904
3/8:	14
4/8:	57
5/8:	102
6/8:	112
7/8:	194
8/8:	521

<i>OK</i>	47,593
<i>name</i>	5,996
<i>not a lemma</i>	4,510
<i>non-standard</i>	527
<i>wrong POS</i>	690
<i>I don't understand</i>	3,474
<i>not Czech</i>	1870

Table 1. Inter-annotator agreements

Table 2. Flag statistics (more than 50 % agreement)

3. Annotation Revisions

In the annotation task of the project, more than 100,000 headwords have been annotated with two or more flags. Some of the headwords still need to be revised: those marked with two or more different flags and those marked *non-standard*, *not a lemma* and *wrong POS*. A group of experienced editors is selected to examine the annotations and decide, as objectively as possible, which flag is appropriate and how to approach partially unaccepted headwords and non-standard forms. We call this process *the revision(s)* and the editors charged with the task *the inspectors*.

The inspector goes through a list of words annotated in the headword annotation task. This list consists of words mostly annotated with the *non-standard*, *not a lemma* or *wrong POS* flags and of words with a combination of different flags. Words annotated

only with the *I don't know* and/or *not Czech* flags, only with the *proper name* flag or only with the *OK* flag are not included. This means only a fraction of the words (approximately 25,000 of the 100,000 in CDE) need to be revised.

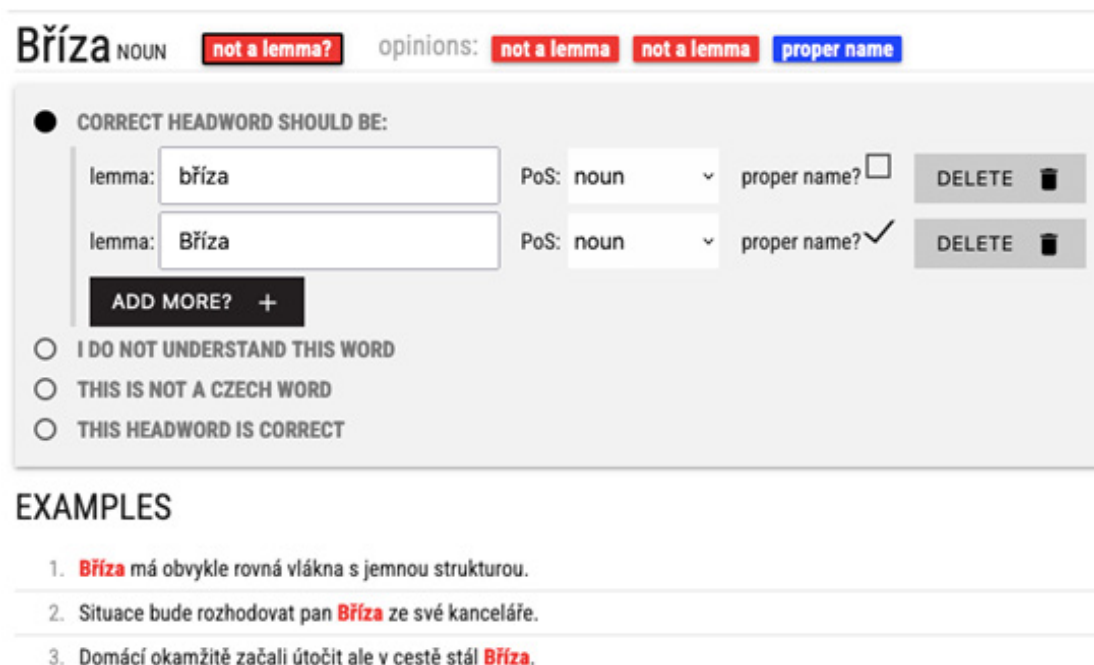


Fig. 3: Example of a revision entry.

Figure 3 shows the Lexonomy interface (Měchura, 2017), where the revisions get done. The inspector looks at a headword and its previous annotations. The annotations direct them to a more objective approach to the given headword than in the previous task. At the inspector's disposal is also the context typical for the headword (demonstrated as a list of Good Dictionary Examples (Rychlý et al., 2008).

Based on the provided data and knowledge of Czech as their native language, the inspector decides that:

1. the headword is slightly incorrect or has two or more different correct forms, and enters the correct form of the headword with the option of marking it a proper name;
2. OR the headword's lemma is not a proper Czech lemma;
3. OR they don't know the headword, in which case the headword is marked as a non-word;
4. OR the headword is correct.

If the headword is a non-standard form of a formally similar standard Czech headword (e.g., *čtyry-numeral*, it should be corrected to the standard form (e.g., *čtyři-numeral*).

If the headword is an incorrect or non-standard union of two separate words (e.g.,

zachvilka-noun in Czech), it should be corrected to the correctly used phrase of two or more words (e.g., *za chvílku – adverb* in Czech).

The revision is a significantly more complicated task than the headword annotation because it expects the inspector to view the headwords more objectively. Thus, the inspectors were chosen from the editors who had more experience with annotation and who proved more consistent while annotating headwords.

4. Subsequent Tasks

The creation of an annotated and revised headword list is only the first task of the methodology of creating a full dictionary of a language. Subsequent tasks of CDE are being prepared and executed. These include:

- annotation of inflected forms of nouns, adjectives, verbs, adverbs and other, language-specific categories;
- word sense disambiguation, translation to English, choosing the right dictionary examples etc.;
- recording of word pronunciation (the only manual-only task of DE).

5. Conclusion

In this paper, we described the annotation and revision tasks of the Dictionary Express method for dictionary-making and demonstrated it on the ongoing Czech Dictionary Express project. A headword list is created using large, automatically lemmatized corpora. In the annotation task, the annotators go through this list and assign flags such as *not Czech*, *non-standard*, *not a lemma* or *OK*. In the revision task, the inspectors go through the list again, look at the assigned flags and decide more objectively which headwords should remain in the list and in what form.

By the time of writing, almost 100,000 headwords have been assigned at least two flags. This provides interesting data for an insight into the language and its perception. Some of the data have been discussed in the paper.

A list of headwords has been created for the dictionary. The Czech Dictionary Express project continues with tasks concerning inflected forms. Later on, word senses will be assigned to headwords through a combination of automated tools and manual work, as well as translation to another language (English), dictionary examples and other parts of the dictionary.

References

DE, 2023. (2023). *Dictionary Express – automated dictionary generation*. Retrieved April 17, 2024, from <https://dictionary.express/>. Accessed:

cst, 2024. (2024). *csTenTen – Czech corpus from the web*. Retrieved April 17, 2024, from <https://www.sketchengine.eu/cstenten-czech-corpus/>.

Baisa, V., Blahus, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medved', M., Mechura, 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* (pp. 805–818). Brno, Czech Republic. Lexical Computing CZ s.r.o.

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medved', M., & Ohlidalová, V. (2023). Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data. In M. Medved', & M. Mechura (Eds.), *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.

Jakubíček, M., Kovář, V., & Šmerk, P. (2011). Czech morphological tagset revisited. In R. Horák (Ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing 2011* (pp. 29–42). Brno. Tribun EU.

Kovařík, F. (2023). Semi-automatic dictionary creation for czech. *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)*, 17.

Měchura, M. (2017). Introducing Lexonomy: An Open-source Dictionary Writings and Publishing System. In I. Kosem, C. Tiberius, et al. (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing.

Rundell, M., & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A Taste for Corpora: In honour of Sylviane Granger. Studies in Corpus Linguistics*, 45 (pp. 257–282).

Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M., & McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal, & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425–432). Barcelona. Institut Universitari de Lingüística Aplicada.

Suchomel, V. (2018). *csTenTen17, a Recent Czech Web Corpus*. In A. Hořák, & A. Rambousek (Eds.), *Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018* (pp. 111–123). Brno. Tribun EU.

Šmerk, P. (2008). *Towards Morphological Disambiguation of Czech*. Ph. D. thesis proposals, Faculty of Informatics, Masaryk University.

Šmerk, P. (2009). Fast Morphological Analysis of Czech. In *Proceedings of the Raslan Workshop 2009*, Brno. Masarykova univerzita