Semi-automatic Dictionary Creation for Czech

Using Automatisation to Create a Rapid Czech Dictionary

František Kovařík

Faculty of Arts, Masaryk University
Arna Nováka 1, 602 00 Brno, Czech Republic
frkov@mail.muni.cz

Abstract. There are many ways to make lexicographer's work faster and more accurate by using automatised and semi-automatic tools. In our project, we create a Czech dictionary using corpora and automatic tools, as well as subsequent manual annotations. We examine the semi-automatic method used in previous projects on different languages – its efficiency, accuracy and speed. This paper is to introduce the project, its preparations, its initial phases, as well as the goals of its research.

Keywords: Dictionary, manual, automatic, semi-automatic, headwords, annotation, revision, tag, lemma, POS, Czech

1 Introduction

Corpus-based and computational tools help lexicographers create dictionaries rapidly and accurately in more areas of expertise than ever before. This leads to easier dictionary creation and also helps the lexicographer skip the parts of creation which can be automatised, so they can focus on more difficult or interesting tasks. It also allows native speakers who are not professional linguists to help maintain accuracy and objectivity of dictionary entries while also boosting creation speed.

As described in previous papers [1,2], lexicographers of Lexical Computing have created dictionaries using a unique semi-automatic methodology. The methodology consists of multiple tools. Some are fully automatic. Some require manual inspection. Manual annotations were done by native speakers (we will refer to them as *annotators*). These were (with one exception) not professional linguists. The lexicographers themselves (we will refer to them as *coordinators*) didn't speak the languages of the created dictionary, and only had a limited knowledge of the language they were examining. In the course of multiple projects, the methodology continued to evolve and four dictionaries have been created: Lao, Tagalog, Urdu and Ukrainian. These consist of translations to English and/or Korean, as of sense distinction using thesaurus and/or pictures and the morphosyntactical behaviour.

The aim of this paper is to introduce a new rapid dictionary project called *Czech Dictionary Express*. We use the existing methodology to create a dictionary

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 93–100, 2023. © Tribun EU 2023

of the Czech language and explore the possibilities, just as the downsides of this semi-automatic approach. We examine the first phases of the semi-automatic dictionary creation. We describe the main questions and problems that can arise within these phases.

2 Project phases and overlapping

We split the project into multiple phases.

The phases follow each other according to their number, but they can also overlap. For example, we can generate more headwords to the lexicon (a tool from Phase 0 – see section 3) while the annotators already annotate the headwords generated earlier (Phase 1 – see section 4) and even earlier generated and annotated headwords are being revised (Phase 2 – see section 5.1) and so on.

3 Preparation phase

In the preparation phase (which we call Phase 0), two objectives have been met:

3.1 Objective 1: Generating headword batches

First a list had to be generated of the lexicon that was going to be used in the dictionary. The list consists of headwords, i.e. lemma-POS couples (for example *místnost-noun*). These were taken from a corpus combining three of the largest Czech corpora, the csTenTen web corpora: csTenTen12, csTenTen19 and large part of csTenTen17. [3] The lexicon of the corpus is thus derived from relatively present-day Czech used on the web.

The lexicon was split into separate word batches, most containing 1 000 words.

Firstly, we only produced 30 batches, containing in total 15 000 most frequent unique words from the corpus. The batches 1–15 were identical to the batches 16–30 so inter-annotator agreement could be generated easily. In the process, we discovered the annotation went faster than expected, so we enlarged the lexicon to a final 80 000 words, producing over a hundred more batches.

One of the batches has been seven times multiplied and given to all the annotators so we could compare their annotations all together. All the other batches were only duplicated (e.g. batch 2 is identical to batch 17) and given to two different annotators. This was done to investigate the inter-annotator agreement and also to prevent errors and recognise difficult words.

In the following research, we want to make our method even more accurate by duplicating the batches once again, so every thousand words has been annotated by three different annotators. This could help us further explore the inter-annotator agreement and compare the two methods - two annotations vs. three of them.

3.2 Objective 2: Annotator recruitment

The annotation team consists of eight annotators, all of which are Czech native speakers and have finished their secondary education. They didn't receive full linguistic university education yet are relatively educated in the language area. This helps provide the sort of annotation data for later research use: the annotators don't assess the language too complexly, yet they do understand the subject enough so they can judge Czech headwords by their intuition.

Each annotator was asked about their local and social background – where they and their relatives live and lived and what schools and languages did they study. These information could be used later when examining the annotations separately.

4 Headword annotation phase

After preparations have been met, the project could step into the headword annotation phase which we called Phase 1.

4.1 Headword annotation

The headword annotation consists of a simple task of assigning a single *flag* to a potential headword. The annotator goes through a list of potential headwords (lemma-POS couples) and assigns a *flag* to each of them as follow:

1. If they don't understand the lemma, don't know it from the use of language or think it is not a proper word, the annotator is to choose the flag ***I don't know***.

2. If they know the lemma from another language or assume it is used in another language, but don't know it from the use of Czech, the annotator is to choose the flag ***not Czech***.
(Note: The flags *I don't know* and *not Czech* are handled very similarly in the proceding phases.)

3. If the given lemma is a word in Czech (including non-lemma forms), but there is another word in standard contemporary Czech that is used much more often, the annotator is to choose the flag ***non-standard***. Here, intuition of a common user of contemporary Czech should be preferred to the knowledge acquired in schools. Non-standard forms include the past, literary, dialectical, non-written and other word forms.

4. If the given lemma is a word form in standard contemporary Czech but it is not the lemma form, the annotator is to choose the flag ***not a lemma***.

5. If the suggested lemma is a correct lemma form in standard contemporary Czech but the POS tag cannot be considered corresponding to the lemma, the annotator is to choose the flag ***wrong part of speech***.

6. If the suggested headword contains a correct lemma form in standard contemporary Czech and the POS tag can be considered corresponding to the lemma, the annotator should choose whether the lemma is a ***proper name*** (flag) or not. For the proper nouns the flag ***OK*** is to be chosen.

4.2 Annotator training

Some additional training was needed so the annotators could understand their task. This required a work manual, a short introduction and presentation of the project in a workshop and also a discussion (brainstorming) about the language-related problems that can arise. Before, our limited knowledge about such problems in Czech was based on our language intuition and experience of the preceding dictionary-creation (Lao, Tagalog, Urdu and Ukrainian – see section 1). For annotators to understand the basic linguistic, language-neutral terminology used in our project, an interactive online course was provided. (For each phase, a course is needed. The course for headwords contained information about how to approach foreign words, non-standards words, proper names etc.)

A significant difference from the preceding projects is that the coordinators are newly also native speakers of the examined language (Czech) and can thus better comprehend the subject and anticipate difficulties.

4.3 Language-related annotation problems

Here are some of the language-related problems and solutions discussed on the training and during the annotation:

- **Only single words**: The batches contain only single words in combination of POS tags. This should be considered when we come across words of which their dictionary form usually includes another word. This in Czech mostly concerns the reflexive verbs (reflexiva tantum). For example the verb ″bát se″ doesn't have an equivalent without the reflexive pronoun ″se″. Yet the batches would only contain the headword ″bát-verb″ – this form should be accepted in spite of not having the obligatory pronoun.

- **Presumption of correctness**: POS tagging can be in some cases very complex. We encourage the annotators to accept the POS tag provided by the automatic tagger of the corpus. Only if the POS tag should be considered objectively wrong for certain, POS tag is not to be accepted. (E.g. the word ″prostřednictvím-preposition″ is to be considered OK, because it can behave like a preposition in this norm, even though it comes from the noun ″prostřednictví″. On the other hand, ″hajný-adjective″ should be considered having a wrong POS – in spite of being derived from an adjective, it behaves only as a noun in modern Czech.) We also advise not to depend fully on the information learned in previous education but to follow the intuition of a native speaker and the knowledge of the language behaviour in general.

- **Abbreviations** have been decided to be handled as usual (single) words. This means their POS tag should correspond to their sentence usage. For example the abbreviation ″dr.″ (doktor, doctor) is a noun, the abbreviation ″např.″ (například, for example) is an adverb. The lemma of the abbreviations needs to have or lack a dot according to the used standard to be accepted (″cca″ for circa without a dot, ″např.″ for například with a dot).

- **Single letters** which do not stand alone as words (e.g. ″ě″ which is not a word in Czech or ″A″ for which the lemma ″a″ should be used) or standard used abbreviations (e.g. ″r″ – the proper form is ″r.″ for rok) are not to be considered proper lemmas.

- **Vulgar** and otherwise taboo words should be looked upon as normal part of the lexicon and annotated as such.

- **Negation** of words: When should it and when should it not be accepted in the lemma? We decided not to accept negation in a lemma if the word is not considered negative tantum (doesn't have a non-negated form; e.g. ″nenávidět″) or secondary negative tantum (the negated form has a distinct meaning from a simple semantic negation of the non-negated variant). The annotators should always think about if the non-negated form is used (E.g. ″neodmyslitelně″ is used very often in Czech. The word ″odmyslitelně″ on the other hand is practically never used.)

- **Interjections**: Which forms should be accepted? We decided to only accept the most transparent forms (e.g. ″kikirikí″ or ″kykyryký″, but not ″kykyrykýhyhý″). While this could be considered a very subjective decision, we predict that in most cases there will be more and less transparent forms. As in the POS disambiguation annotation, we encourage the annotators to consider the lemma right if they don't consider the form strongly non-standard.

- Other wanted properties of a lemma were discussed, such as preserving the gender in the noun lemmas (i.e. ″stolař″ and ″stolařka″ should be considered two separate lemmas).

4.4 Findings

Before the annotations have begun, our vision was to firstly generate and annotate 15 000 potential words twice (approximately 10 000 future dictionary entries), possibly extending this number to 50 000 in the future. This estimation has been based on the speed of the previous projects. However, the annotations of Czech headwords were faster than previous annotations. One batch took a single person approximately 2 hours to annotate (meaning the double annotation took 4 hours), whereas a similar batch in Ukrainian took a single person 6 hours (12 for double annotation). One of the expected reasons is that Czech has a significantly better tools for Corpus creation and management (e.g. Majka morphological analyser [5] and desamb [4]) and bigger corpora than the other languages. This also means more headwords get the flag *OK* (they contain the right lemma and POS tag) than in the previous projects. (For example, only 38.4 % of the Ukrainian headwords have been annotated as *OK*. [2] The same flag got 65.7 % of the Czech headwords in 149 batches completed to the day of writing this paper.)

We finally decided to extend the number of twice annotated headwords to 80 000.

We have chosen one batch that every annotator should annotate that could provide us some interesting data. This data have been used to recognise the annotation style of each annotator and recognise some interesting linguistic problems.

As mentioned before, other batches have been annotated by two different annotators. We are considering annotating these for a third time since annotations since the speed of annotations is higher than expected. Before this, experimental third annotation of one or two batches will be made and we will examine the statistics provided by these experiments.

5 Subsequent phases

In this section, two of the nearest subsequent phases are described. The tasks follow on from the annotation data provided in Phase 1 and the headwords lists generated in Phase 0. Both phases are going to be launched simultaneously, but they could also follow each other if needed in other projects.

5.1 Revision phase

Revisions of the headwords annotations (also called Phase 2) are done by experienced annotators who proved capable in Phase 1. We have chosen 4 annotators who worked the longest and we have examined their annotation data. In the data of every annotator, we spotted recurring difficulties. These will be discussed on an upcoming training for revisions.

The task of revisions is to go through headwords at least once annotated *non-standard*, *not a lemma* or *wrong POS*. Headwords with the *I don't know* or *not Czech* flag in combination with the *proper name* or *OK* flag are also to be revised. The same goes for the headwords with the combination of the *proper name* and *OK* flag. The revising annotator sees a headword and its annotation flag and is supposed to select one of these options: Either they can enter the correct headword the annotated headword corresponds to. Or they can state they don't understand the word or the word is not Czech. The last option is to state the annotated headword is correct.

The headwords annotated only as *OK* or only as *proper name* are not revised. The same goes for headwords annotated only in any combination of the *I don't know* and *not Czech* flags. This focuses Phase 2 only on a fraction of the headword list with the need for revisions. As mentioned in subsection 4.4, the number of the headwords annotated with the *OK* flag is greater than in the previous projects since the Czech tagger is more precise and the used corpora are bigger than for the languages before. The speed of revisions can thus be also expected to increase.

5.2 Forms

Another aspect of the dictionary we want to create, besides the lemmas and POS tags, are the inflected word forms (we call this task Phase 3). In Czech, nouns, adjectives, pronouns, numerals and verbs can be inflected and adverbs can be comparative. The form annotators will go through a list of headwords who have been decided to be standard Czech lemmas with corresponding POS tags in Phase 1 and later Phase 2 (section 4 and subsection 5.1). For each headword, a list of possible word forms will be generated from the corpus. The task is to mark which of them are correct standard forms of the given headword.

6 Conclusion

This paper introduces the project of creating an express Czech dictionary using a semi-automatic method. In the first section, we describe the goals and priorities of the project (mainly the speed of creation) and the preparations needed to set up the creation. First, a list of headwords (lemma-POS pairs) are created from large corpora using automatic tools. In the following sections, first three phases of the project are introduced. First phase focuses on headwords: whether each lemma is correct and used in standard Czech and whether the POS tag corresponds with it. Second and third phase follow the first phase. The second phase focuses on revising the headwords that in the first phase have not been annotated as completely correct or completely incorrect. In the third phase, annotators decide which automatically found words are correct inflected forms of a headword.

After the first three phases, more automatic and manual tasks are going to follow. The most demanding, time-heavy phases are estimated to be the ones focused on meaning distinction: sense-recognising, thesaurus words and examples. [2]

References

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

2. Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medveď, M., Ohlídalová, V.: Rapid Ukrainian-English Dictionary Creation Using Post-edited Corpus Data. In: Medveď, M., Měchura, M., Tiberius, C., Kosem, I., Kallas, J., Jakubíček, M., Krek, S. (eds.) Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Lexical Computing CZ s.r.o., Brno, Czech Republic (2023), https://elex.link/elex2023/wp-content/uploads/114.pdf
3. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family (2013), https://api.semanticscholar.org/CorpusID:107998183
4. Šmerk, P.: K morfologické desambiguaci češtiny [online] (2008 [cit 2023-11-07]), https://is.muni.cz/th/wteg5/
5. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of the Raslan Workshop 2009. Masarykova univerzita, Brno (2009), https://nlp.fi.muni.cz/raslan/2009/papers/13.pdf