# Dictionary Express: First Phases
# Rapid dictionary-making method for European, Asian and other languages

Marek Blahuš[1], Michal Cukr,[1] Miloš Jakubíček[1,2], Vojtěch Kovář[1,2] and 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-mail: firstname.lastname@sketchengine.eu

**Abstract**
Dictionary Express (DE) is a new methodology combining automatic tools for lexicography and manual checking (annotation) of words, their forms, usage etc. The main goal of the project is to accelerate dictionary making faster and less demanding by separating the process into simple tasks, as opposed to the traditional dictionaries made entry-by-entry. This means the non-automatic work can be done by a small team of native speakers who are not professional linguists, supervised by a smaller team of developers and lexicographers. The data is acquired from big corpora of current web language usage, which helps the dictionary to be more accurate and up to date with the current language trends. In the past, several "rapid dictionaries" have been created using this method. The time needed to complete a DE project depends on the quality of the tagging of the corpus and the amount of the weekly workload. A DE project for Czech is now in the making, and apart from creating a new Czech dictionary, it focuses on analysing the rapid dictionary-making process and the input/output data. In this paper, we present the main annotation tasks of the DE methodology, the data preparation, and some interesting phenomena that occurred during the first phases of the Czech Dictionary Express.

**Keywords:** corpus annotation, semi-automatic lexicography, Dictionary Express, dictionary drafting, post-editing lexicography

## 1. Introduction

Dictionary Express (DE) is a group of projects focused on accelerating and simplifying dictionary-making processes. The aim of these projects is to create mono- and bilingual dictionaries faster by using automatic tools and large language corpora. The projects operate with a complex methodology toolkit.
The process of creating an express dictionary is semi-automatic. Some tools and phases are fully automated (e.g. generating a list of possible entries from a corpus), while others require manual annotation by native speakers (e.g. post-editing the list of possible entries to create a lexicon of existing words).
In the recent past, four dictionaries have been made using the DE methodology: a dictionary of Urdu, Lao, Tagalog [2] and Ukrainian [4]. This paper presents the first tools and phases of the DE methodology on a new project for Czech (Czech Dictionary Express, or CDE): preparation, headword annotation [7] and revision, word form checking, as well as the future stages of the project. It points out the differences and updates of the methodology, tools and workflow of CDE in comparison to the previous projects. Most DE tools can be applied on every language (including European and Asian languages), although the dictionary-making process may look slightly different.

## 2. What is Dictionary Express

The aim of the DE methodology is making dictionaries of different languages as fast as possible, using automatic tools and a team of native speakers. These native speakers (called *annotators*) are not professional linguists and are required only to have a secondary education (with some level of linguistic knowledge). The projects are coordinated by a team of coordinators, who supervise the project, develop automatic tools and provide annotators with editing interfaces.
The data is acquired from large web corpora (preferably with multiple billions of words; for Tagalog however, a 230-million-token corpus was used [2]). The size is crucial not only for the quality of headwords including mid- or low-frequency words, but also of multi-word expressions, collocations, and phraseology.[1] To collect such amounts of data, DE uses corpora made of texts on the web. The fact that the data comes from common-use texts and that it is manually checked by native speakers without higher linguistic education means the final dictionary contains language of current direct use, supported by transparent and traceable corpus and annotation data. It is, however, not made and checked by professional lexicographers. Moreover, some might argue the "common-use texts" are only of the common use on the internet – that DE projects don't concentrate on the spoken form or other written forms of language (which would be more complicated given the corpora of these forms of language would be much smaller).
The aim of the Czech Dictionary Express project is not only to create a new dictionary, but also to analyse the creative process and the dictionary's final form, and to compare it with other existing Czech dictionaries created in more traditional ways. This is possible, because unlike in the previous projects, the coordinators of CDE understand the given language and are native speakers themselves.
A dictionary created by the DE methodology describes a lexicon of correct headwords (lemma and POS tag pairs) of the given language (acquired from the automatically lemmatized and tagged corpus, manually annotated and revised; see Section 2). The entry of each headword contains:

- a set of used forms (acquired from the corpus and manually checked; see Section~\ref{forms}),
- an audio recording of the word pronunciation,
- a set of word senses (see Section~\ref{sense}) containing
    - thesaurus – synonyms, antonyms, cohyponyms etc.,
    - dictionary examples,
    - images (photos etc.),
    - translations to other languages (e.g. English, Korean).

The DE methodology describes all phases of creating such a dictionary. Most tasks consist of an automatic part and a manual part. The automatic part comprises of acquisition of a word list or adding new data to the words in this list. The manual part requires the annotators to go through the word list and to check the state of the data (for example whether a headword belongs to the selected language, whether a form or a sense belongs to the headword etc.).
Figure 1 shows a simplified diagram of the tasks.
First, a corpus is selected. Different tasks may use different data from the same corpus or from other corpora. After some tasks, the corpus is improved (e.g. is again lemmatized using data from the Headword Annotation task).
From the selected corpus, a headword list is automatically acquired. Headwords, i.e. lemma-POS pairs (e.g. *cat-noun*), are ordered by document frequency. From the ordered list, a number of the most frequent words is used (100,000 in CDE) and prepared for Headword Annotation (see Section 3), i.e. checking if a word belongs to the selected language, if it is a correct lemma etc.
Headword Annotation, the headword list is revised (see Subsection 3.2) and used for improving the corpus and acquiring data for the following tasks. The Forms task (see Section 4) consists of checking word forms of each headword. The Senses task (see Section 5) handles word senses of each headword. The audio pronunciation task consists of recording audio of pronunciation of every lemma; it is the only task done completely manually.

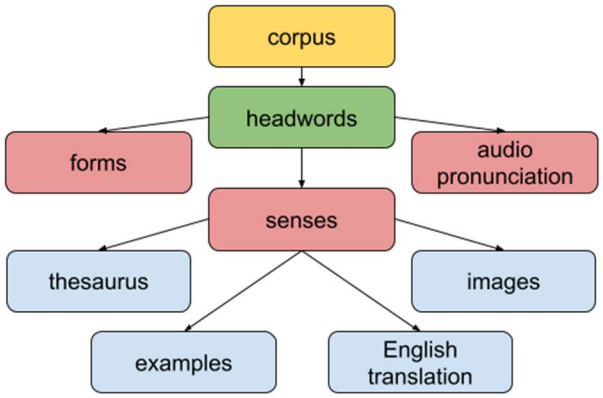


Figure 1. Simplified workflow process of Dictionary Express manual post-editing. [4]

The key feature of Dictionary Express projects is dividing the annotator's work into as simple tasks as possible. The dictionary is not created entry-by-entry, but takes the component-by-component approach.[1] This speeds up the dictionary-making process significantly (annotators don't divide concentration to several tasks, as professional linguists usually do) and makes it easier to train annotators (training concentrates only on one simple task).
Before each task, the annotators go through a short training. They are presented with as little linguistic terminology as possible, and go through a short sandbox batch of data in the interface. They can ask and discuss any technical or linguistic phenomena. The training is sometimes preceded with short online exercises or theoretical materials. During the training, a manual for the given task is provided, both a written and an online version. The training is preferred to be face-to-face, not online, so the discussion is more natural and the coordinators get to know the annotators and their behaviour better.
While a task is being completed, the coordinators check the annotation data regularly and discuss the annotation choices with the team, in an online discussion where anyone can raise a question or make a comment about linguistic phenomena.
The data acquired in the manual processes may also serve other purposes than only creating the dictionary. For example, the data from the Headword Annotation task may improve the tools for lemmatization (e.g. lemmatization databases or lemmatizers) as mentioned above.

## 3. Headword annotation and revision

### 3.1 Annotation

Preparing the manual Headword Annotation task means deriving a wordlist from the corpus. The wordlist consists of a number of most frequent headwords. In CDE, the document frequency was used. This means frequency is not counted

from every usage in the corpus, but from the number of documents the word is used at least once. The document frequency cannot be distorted by the random high usage of some words in a small number of documents, and represents the words better in the widest range of genres and themes.
The headwords are sorted alphabetically and distributed among the annotators in batches. The alphabetical order groups together similar words, so the annotator can see and choose which ones are correct. (E.g. the annotator sees *starý-adj*, *starý-noun*, *Starý-adj* and *stará-adj*, and can choose that only *starý-adj* is the dictionary form of a Czech headword.)
The annotators never work with code. Instead, they use a simple interface in the Lexonomy tool.[8] The annotator goes trough a list of headwords and chooses a *flag* for each of them. There are several flags they can choose from, so the annotation produces more complex data than "right" or "wrong" (see [7]). This data not only helps the wordlist to be improved in the next phases, but can also provide some interesting insights into the dictionary and the language use itself.
The annotators don't see the context of the headword, which would slow the process down significantly, and are advised against searching it online or in existing dictionaries – this prevents copying data from external sources.

### 3.2 Revision

Each headword is annotated by at least two different annotators. After some amount of annotations have been done, a small group of annotators is chosen to revise the annotations which have been found problematic. This group is called the inspectors. For the revision task, a new interface is set up.
Revised is only a fraction of headwords that are not fully correct or not fully incorrect. These are controversial (have been assigned different flags) or have been annotated with one of the flags that mean "something is off": non-standard words, words with typos, words that are not lemmas, words with parts of speech.
An inspector goes through this word list and corrects the headwords. They are advised to look at the headwords more carefully, more objectively. Not only do they see the headword, but also the flags the words have been annotated with, and they see the context of the use of the word (in the form of Good Dictionary Examples, GDEX [9].)
Whenever an inspector sees a headword, they can choose one of the four following actions:

- modify the headword (change the lemma or the POS tag; they can also create another headword);
- state that they don't know the headword or the headword is incorrect ("I do not understand this headword");
- state that the headword doesn't belong to the language of the dictionary;
- state that the headword is correct (with the option to mark it a proper name).

Any revised headword is connected to its use in the corpus. This affects the data used in the next annotation task – the Forms task.

## 4. Forms

The Forms task follows the headword annotations and concentrates on the possible forms of words. As Czech is an inflected language, this phase concentrates on the inflection – particularly on the forms of nouns, adjectives, numerals, verbs and adverbs.
The Forms task can run simultaneously with headword annotation and revision. However, each headword used in the Forms task has to be already annotated and revised.
Figure 2 shows the Forms interface. A Forms annotator goes through a list of accepted headwords and simply marks each of its possible forms as correct or incorrect. The option to mark a form correct yet *non-standard* is also available, as well as the option to look at the context of each inflected form.

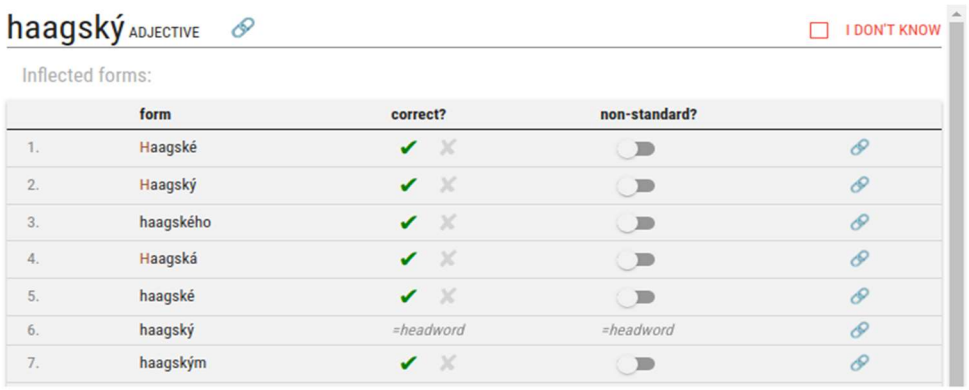


Figure 2. An example of Forms task interface of the headword *haagský-adjective*.

In the past, the letter case was a difficult issue for the task forms. (See Subsection 3.1.) For better orientation in the list of forms, the uppercase letters are highlighted in red.
If not sure, the annotator can access the context of the form in the corpus. The button on the very left of each form works

as a link to the context of the headword presented using the Sketch Engine tool.[5] Since this slows down the annotation process significantly, the annotator should use the button seldom.
The list of possible forms was generated from the combination of three large corpora of the Czech Web Corpus family [10], which uses automatic tools for Czech, such as Majka morphological analyser [12] and desamb [11]. The manual form annotations and the manual headword annotations are used to create a rapid Czech dictionary but also serve as valuable data for enhancing the databases that improve these automatic tools.
The list of possible forms depends on annotator/inspector choices made in previous tasks. If a headword was manually corrected in the revisions, all of its forms from the context automatically connect to the new headword. For example, if the headword *čtyry-numeral* was revised as the non-standard form of *čtyři-numeral*, all of the forms of *čtyry-numeral* now show under the headword *čtyři-numeral*.

### 4.1 Form removal

To make the form annotation more time effective, a simplification was carried out: Forms with non-matching letter case have been removed.
The idea behind keeping all forms was that some words might be used in interesting ways combining letter cases. For example, the "Czech" form *MUDRuje* comes from the form *mudruje*, the headword being *mudrovat-verb* (*to talk smart*, *to philosophize*) in combination with *MUDR* – a short of *medicinae universae doctor*, the doctor of medicine title. However, these forms were found very sparse.
On the other hand, the number of forms with different letter case than the lemma was huge. An average batch of 1,000 headwords had 32,295 forms. This means one headword had on average more than 32 forms, often reaching more than 100 forms. (For comparison, in the Ukrainian project, a batch of 1,000 headwords had an average of 13,299 forms.)
From a list of the first form-annotated 15,000 headwords, only 2 lemmas have only uppercase letters and 57 lemmas have an uppercase letter in the middle or at the end.
At the point where 415,519 forms have been annotated, 104,946 forms with uppercase letters in the middle or at the end (leaving out lowercase forms and forms with only an uppercase initial letter) have been **rejected** and only 813 of such forms have been accepted. Of these 813 accepted forms (377 fully uppercase), 558 have been rejected by the coordinators and a big number of the rest belongs to the lemmas with uppercase letters in the middle or at the end.
This means only a tiny fraction of forms with uppercase letters in the middle or at the end are or should be accepted. On the other hand, these forms make up more than 25 % of all forms. Lemmas with uppercase letters in the middle or at the end are so sparse they can be left with all their supposed forms (this makes the implementation of form removal considerably simpler).
These findings led to a major change in the list of forms of lowercase lemmas and lemmas with only an uppercase initial letter: Forms with an uppercase letter that didn't have a lowercase variant were replaced by their lowercase variant, with the initial letter of the form matching the one of the lemma (e.g. *baráku* instead of *BaRáKU*, *Pavlovi* instead of *PAvLovi*). Other forms with uppercase letters in the middle or at the end were removed.

Figure 3. The context of the same forms with various letter cases are now mashed together.

Despite the fact that the number of forms drops by more than 25 %, the changes to the lexicon are minimal. Most of the removed forms would be marked incorrect anyway.

## 5. Senses

For the CDE project, the word sense induction is being prepared. The Senses task is one of the most time-consuming tasks.
It is preferred, although not absolutely necessary for the Headword Revision task to be done before the Senses task data

generation. The automatic word senses generation is a very complex task and requires a lot of computation time, often reaching weeks of computation. Therefore, it is preferred to run another task (forms, rest of revisions etc.) while the data for senses are being processed.
In the Senses task, the annotators go through a list of revised headwords. For each headword, examples[1] of typical usage are generated using the Word Sketch method [6], and automatically clustered into groups based on a word sense model [3]. The task of the annotator is to name one or more patterns of collocations of the headword and connect each of them

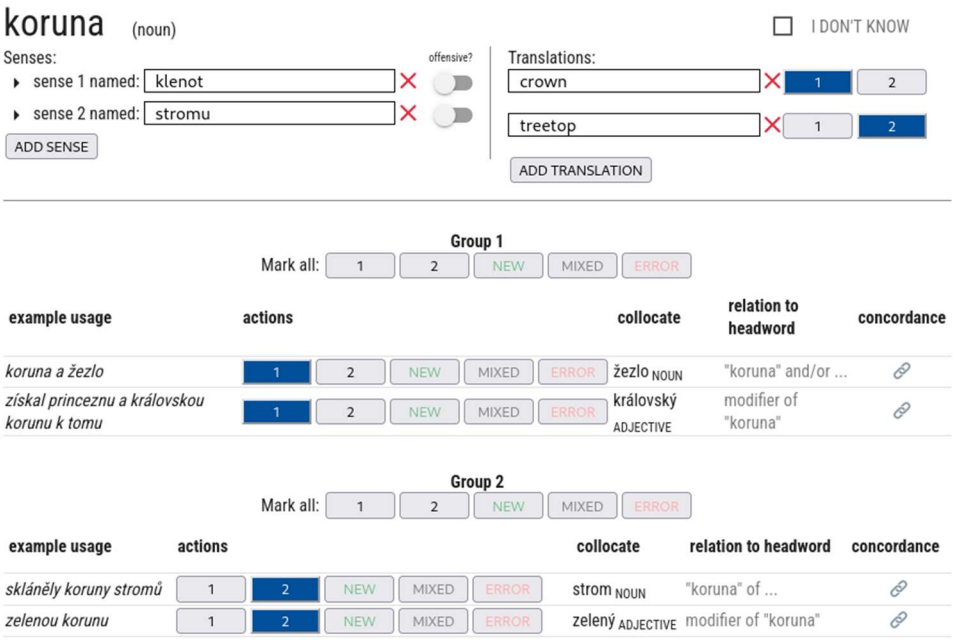


Figure 4. An example of Senses task interface of the headword *koruna-noun*.

to the listed examples. The annotator sees the interface of the Senses task, as presented in Figure 4 on the Czech headword *koruna-noun* (which has several meanings in Czech, including a "crown" and a "treetop").
The annotator has added a second sense and named both senses by typical collocates (there are more options of naming a sense).

- The first English translation, "crown", is automatically recognized; the second one, "treetop", has been manually added. Both were linked each to one of the senses (the translations can also be linked to more or all senses) -- hence the blue buttons "1" for the first sense and "2" for the second sense.
- Each of the automatically found examples from the Word Sketches listed below has been linked to one of the senses: All of the first group to the first sense and all of the second group to the second sense. This can be done by selecting a number separately for each example, or by selecting a number common for the whole group.

When the "New" button is selected, a new, unnamed sense is created, to which the example (group of examples) is linked. The "Mixed" button is used if the word in an example (group of examples) can have two or more of the senses. The "Error" button is reserved for the rare situations when the annotator does not understand the example or there is an issue with the example.
Similar to the Forms tasks (Section 3), the annotator can access the context of the example using the link button at the very right.
Word sense induction requires deeper understanding of linguistic structures. The annotators have to understand that meanings of words are not always (if ever) exact and matching each example by 100 %. On the Sense annotation training, the problem of semantics and meanings is discussed deeper, while still not leaving the area of expertise of a common native speaker.
For CDE, it is being discussed whether the Senses and Translation are to be unified (as in Figure 4). The unification might make the dictionary making quicker, but goes against the idea that the dictionary-making process is divided into tasks as simple as possible. Also, for the Translation task, the annotators have to have an advanced knowledge of the English language.

## 6. Conclusion

This paper provides an overview of the Dictionary Express semi-automatic methodology. It focuses on the four main tasks of the project: the headword annotation and revision, the form annotation and the word sense induction. The

methodology has been and is being used for several languages of different origins, including Asian and European languages, and is constantly being improved and localised. The paper shows the tools of DE on the example of the Czech Dictionary Express project, which is currently in the making and whose goal is to analyse the methodology, its input/output data, its strengths and its weaknesses compared to similar dictionaries.

**Note**

[1] In this section, we use the word "examples" not to address the dictionary examples, but typical collocations of a headword, e.g. in the context with its collocates.

## References


Dictionary Express – automated dictionary generation, https://dictionary.express/

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

Bartunov, S., Kondrashkin, D., Osokin, A., Vetrov, D.: Breaking sticks and ambiguities with adaptive skip-gram. artificial intelligence and statistics pp. 130–138 (02 2015)

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: 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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1 (2014). https://doi.org/http://dx.doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. Information Technology 105 (01 2004)

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

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

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

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

Šmerk, P.: K morfologické desambiguaci češtiny [online] (2008 [cit 2023-11-07]), https://is.muni.cz/th/wteg5/

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