

Rapid Ukrainian-English Dictionary Creation

Using Post-Edited Corpus Data

Marek Blahuš¹, Michal Cukr¹, Ondřej Herman^{1,2}, Miloš Jakubíček^{1,2}, Vojtěch Kovář^{1,2}, Jan Kraus¹, Marek Medved^{1,2}, Vlasta Ohlídalová^{1,2}, Vít Suchomel^{1,2}

¹Lexical Computing, Brno, Czechia

² Faculty of Informatics, Masaryk University, Brno, Czechia
E-mail: firstname.lastname@sketchengine.eu

Abstract

This paper describes the development of a new corpus-based Ukrainian-English dictionary. The dictionary was built from scratch, we used no pre-existing dictionary data. A rapid dictionary development method was used which consists of generating dictionary parts directly from a large corpus, and of post-editing the automatically generated data by native speakers of Ukrainian (not professional lexicographers). The method builds on [Baisa et al. \(2019\)](#) which was improved and updated, and we used a different data management model. As the data source, a 3-billion-word Ukrainian web corpus from the TenTen series ([Jakubíček et al., 2013](#)) was used.

The paper briefly describes the corpus, then we thoroughly explain the individual steps of the *automatic generation—post-editing* workflow, including the volume of the manual work needed for the particular phases in terms of person-days. We also present details about the newly created dictionary and discuss directions for its further development.

Keywords: Ukrainian; post-editing; dictionary; lexicography

1. Introduction

For decades, language corpora have served as source data for dictionary building. In the last years, corpora were also used for automatic generation of various dictionary parts ([Rundell & Kilgariff, 2011](#); [Kosem et al., 2018](#); [Gantar et al., 2016](#); [Kallas et al., 2015](#)). These automatic outputs were then post-edited by professional lexicographers to ensure the data quality in the resulting dictionary.

With the advancement of technology, it is now possible to create whole dictionaries using this scenario of automatic generation and post-editing by native speakers (not necessarily professional lexicographers). The methodology was used before ([Baisa et al., 2019](#)); we have improved the process and used it in a new project aimed at creating a Ukrainian–English dictionary using a 3-billion-word Ukrainian corpus.

This paper covers all our work on this particular project. We describe building, cleaning and tagging the new multi-billion web corpus of Ukrainian. Then, we discuss the rapid dictionary creation method and our particular implementation which is different from ([Baisa et al., 2019](#)) especially in the data management approach.

In the last part, we describe the resulting dictionary that contains more than 55,000 verified headwords but due to time and budget constraints, we were able to fully complete only 10,000 entries, so there is still large space for improvements.

2. New Ukrainian Web Corpus

We were able to identify three Ukrainian corpora the new dictionary could be based on:

- General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova, 2020; Starko, 2021)
- UberText Ukrainian corpus by Lang-uk¹, a web corpus of 665 million tokens
- ukTenTen14 web corpus from 2014, consisting of 2.73 billion tokens

Of these corpora, the first one is not available for download. The second one is a rather small, topic-specific corpus (mostly news). It is distributed in the form of shuffled sentences, which prevents the selection of headwords by document frequency. For our dictionary work, we took the third one, enlarged it and updated it into a new Ukrainian web corpus. In this stage we followed the methodology of the TenTen corpora family (Jakubíček et al., 2013).

The crawler (Suchomel & Pomikálek, 2012) was instructed to download from Ukrainian top-level domains .ua and .укр and generic domains such as .com, .org, or .net. A character trigram based model trained on a 200 kB sample of manually checked Ukrainian plaintext was used to stop crawling websites that did not contain text in Ukrainian.

The crawl was initialized by nearly 6 million unique seed URLs:

- 194 manually identified news sites
- 94,000 websites from web directories
- 336,000 URLs of web pages found by search engine Bing by searching Ukrainian words
- 5,410,000 URLs found in ukTenTen14

Table 1: Number of documents by TLD in the final merged and cleaned data from 2014 and 2022

TLD	documents	tokens	% corpus tokens
ua	4 640 585	2 122 675 553	65
com	1 099 646	591 327 114	18
org	1 089 027	397 328 162	12
net	318 197	143 994 060	4.4
eu	16 046	8 759 810	0.27

Data obtained by the crawler were converted to UTF-8 with the help of the Chared tool (Pomikálek & Suchomel, 2011) and cleaned by jusText (Pomikálek, 2011). The

¹ <https://lang.org.ua/en/corpora/>, accessed in April 2023.

Table 2: Websites contributing the most tokens to the final merged and cleaned data from 2014 and 2022

Website	description	documents	tokens	% corpus tokens
uk.wikipedia.org	encyclopedia	791 134	243 194 981	7.4 %
uapatents.com	government, patents	36 829	36 339 611	1.1 %
pulib.if.ua	tech encyclopedia	11 669	26 054 618	0.79 %
techtrend.com.ua	tech encyclopedia	18 746	22 706 445	0.69 %
litopys.org.ua	text library	4 592	22 501 121	0.69 %
ligazakon.ua	legal	17 622	22 334 382	0.68 %
uad.exdat.com	(site down in 2023)	8 220	15 928 398	0.49 %
alls.in.ua	(site down in 2023)	14 022	15 292 017	0.47 %
maidan.org.ua	politics, news	11 791	14 826 687	0.45 %
ua.textreferat.com	essays, schoolwork	18 536	14 614 928	0.45 %
economy.nayka.com.ua	economic news	6 025	14 418 873	0.44 %
uatxt.ensayoes.com	(site down in 2023)	7 401	13 575 562	0.41 %
gazeta.dt.ua	news	9 306	13 047 965	0.40 %
uadocs.exdat.com	(site down in 2023)	7 221	12 810 126	0.39 %
zakon-ua.com	(legal, down in 2023)	6 385	12 249 178	0.37 %

result was merged with the old ukTenTen14 and with 1,040,000 articles from Ukrainian Wikipedia downloaded by the Wiki2corpus tool.² Duplicate paragraphs were removed by Onion (Pomikálek, 2011) and manual cleaning was performed according to Suchomel & Kraus (2021).

The final size of the merged Ukrainian corpus is 3,280 million tokens and 2,593 million words in 7.2 million documents with 52% texts downloaded in 2014 and 48% texts downloaded in 2020. Sizes of parts of the corpus coming from selected TLDs and websites are in Table 1 and Table 2, respectively. As can be seen there, the most contributing sites are encyclopedias, technology sites, news sites and legal related sites. Distribution of genres and topics assigned using the method described in Suchomel & Kraus (2022) can be found in Table 3 and in Table 4, respectively.

The corpus was then tagged using RFTagger (Schmid & Laws, 2008) and lemmatized using CST lemmatiser (Jongejan & Dalianis, 2009). The RFTagger model was trained on the Universal Dependencies corpus for Ukrainian³ and the Brown corpus of the Ukrainian language (Starko & Rysin, 2023). Training was also supplemented by an additional morphological database generated from the Ukrainian Brown dictionary (Starko & Rysin, 2020). The model for the CST lemmatiser was trained on Ukrainian Brown dictionary using Affixtrain.⁴ As the last step, heuristic postprocessing of the tagging and lemmatization was applied based on manual inspection of the corpus data.

3. Rapid Dictionary Development by Post-editing

The post-editing methodology we are building on assumes that all lexicographic content is automatically generated from an annotated corpus, and step-by-step post-edited, re-

² <https://corpus.tools/wiki/wiki2corpus>

³ https://github.com/UniversalDependencies/UD_Ukrainian-IU

⁴ <https://github.com/kuhumcst/affixtrain>

Table 4: Subcorpus sizes by topic

Topic	documents	tokens	% corpus
society	484 425	244 264 763	7.4 %
business	100 807	76 791 974	2.3 %
science	76 214	61 546 682	1.9 %
arts	86 876	51 562 726	1.6 %
health	60 184	39 859 674	1.2 %
home	81 442	39 208 234	1.2 %
recreation	23 241	14 499 974	0.44 %
games	18 548	11 291 503	0.34 %
sports	23 357	7 331 632	0.22 %
technology	5 561	1 622 874	0.049 %

Table 3: Subcorpus sizes by genre

Genre	documents	tokens	% corpus
news	1 507 101	584 037 607	18 %
encyclopedia	1 080 862	510 102 047	16 %
legal	165 224	87 684 930	2.7 %
blog	57 846	36 407 663	1.1 %
discussion	24 547	17 370 881	0.53 %

informing the corpus to maximize the mutual completion between the data and the editors, thereby minimizing the editorial effort. Central to this process are two databases: the corpus and the dictionary draft which get mutually updated. The entry components are generated separately according to their dependencies, as illustrated in Figure 1.

After an entry component is generated and post-edited by human editors, the edited data are incorporated into the corpus annotation and used for generating further entry components. For example, having word sense post-edited leads to the introduction of sense identifiers in the corpus, which in turn yields sense-based analysis for a distributional thesaurus or example sentences (which would not be very reliable otherwise).

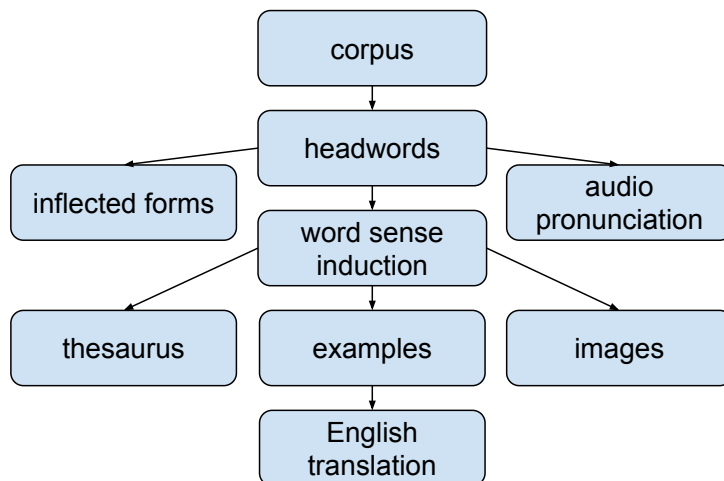


Figure 1: A high-level workflow overview of the post-editing process

In the next sections we explain in detail how we developed a large-scale dictionary with a fraction of human effort compared to the standard setting in which the lexicographers themselves interrogate the corpus. We show the method can rely on existing (imperfect) NLP tools but requires a radical change to the typical lexicographic workflow and a robust data management process between the corpus, the dictionary and the editors.

3.1 Training the Native Speakers

Annotators should be native speakers of the source language, but they are not expected to have any previous lexicographic training. For tasks that involve translation, written capacity in the target language (English) is required. English was also the prevailing language of instruction.

Good training helps annotators understand their tasks well and leads to high-quality output. Each step in the dictionary creation process needs to be clearly explained—containing all relevant information; giving illustrative examples; describing potential conflicts or marginal cases; mentioning the recommended amount of time per entry in each particular task.

Therefore, the training for each task consists of three parts:

1. e-learning describing the task in general, providing English examples, explaining the underlying linguistic concepts, including test questions to verify that the annotator understood the essence of the task
2. half-day face-to-face training where we explain the whole task with real Ukrainian examples and language-specific issues
3. manual of 2-3 pages with the necessary instructions

Most of the time, annotators work using the Lexonomy on-line dictionary building tool (Měchura, 2017; Jakubiček et al., 2018). We have developed a dedicated user interface (customized entry editor) in Lexonomy for each task.

3.2 Headwords

The annotator sees a list of headword candidates (i.e. combinations of lemma and part of speech) and their task is to assign a flag to each according to its perceived correctness. Flagging can be performed with the mouse, but using keyboard shortcuts is preferred. Available flags are given short English names and color codes. The key to attributing flags to headword candidates, reproduced here as Figure 2, is shown to the editor all the time.

After familiarizing themselves with the concepts of *lemma* and *part of speech* and having learned about specifics of handling them in Ukrainian and in the applied tagger, annotators train by using the key to flag headword candidates.

In this project, a total of 119,615 headword candidates were evaluated, 87% of which received at least two annotations and 24% were annotated at least three times. Multiple annotations are taken to create a margin for detecting errors and conflicts of opinion. Eight annotators took part in the annotation effort, the work was split into 289 batches and in total 285,177 annotations were made.

The most frequently assigned flag was “ok” (38.4%), followed by “not a lemma” (25.9%) and “wrong POS” (21.2%), then came “proper name” (5.1%) and “I don’t know” (5.0%), later “non-standard (register or spelling)” (2.7%), and at last “not Ukrainian” (1.6%).

Total time annotators spent on this task was 2114 hours, i.e. one annotation took on average 27 seconds. Speed varied greatly between annotators, ranging from 12 seconds

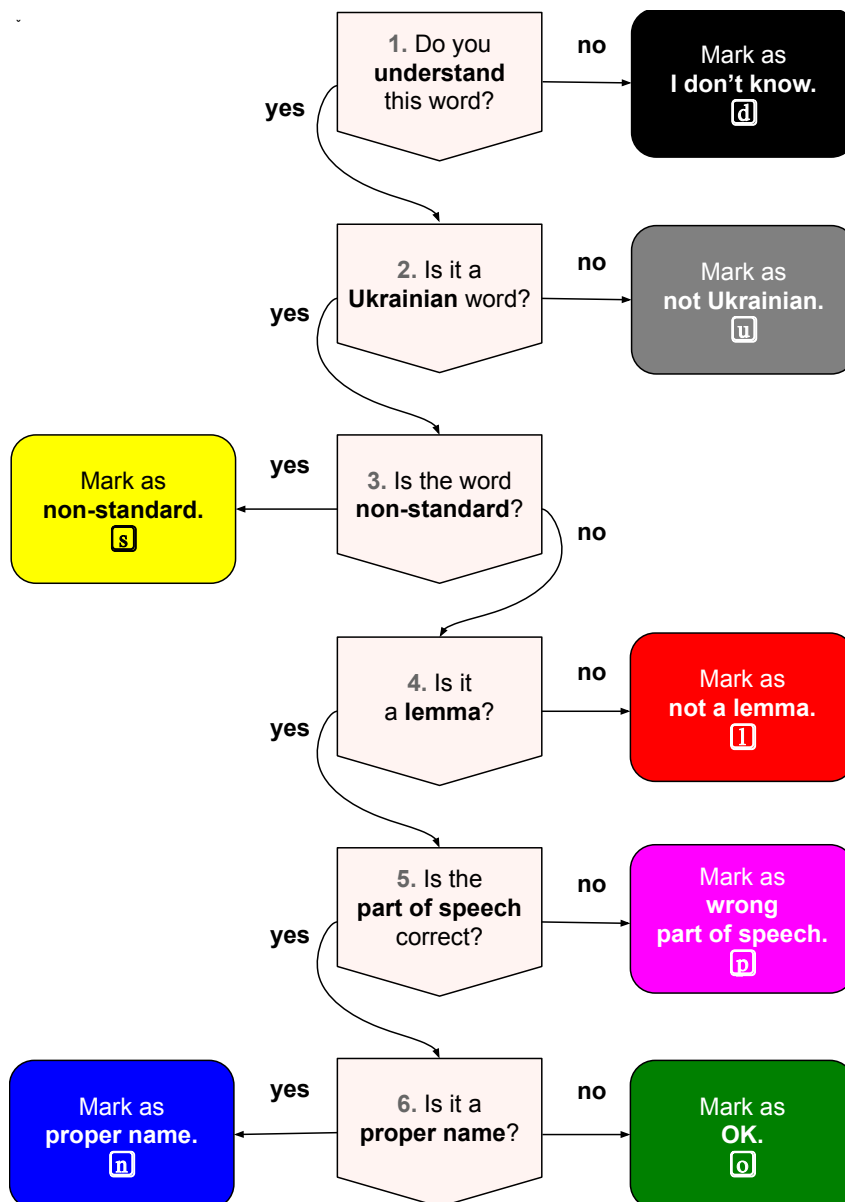


Figure 2: Key to attributing flags to headword candidates, color-coded and with keyboard shortcuts

to 64 seconds per annotation, influenced by factors such as annotator’s self-confidence, computer skills (use of clicking vs. pressing keys), reliance on external resources, work habits or tiredness.

Out of the presented headword candidates, 49,131 (41%) were eventually accepted as correct headwords into the final dictionary. Major contributor of noise in the input data was inter-POS homonymy, produced by early versions of the used tagger from before we managed to reduce it by integrating a larger morphological database. If only lemmas are counted, 66% of the candidates made it into the dictionary. The lempos to lemma ratio has decreased from 1.45 among the headword candidates to only 1.02 among the accepted headwords. Low homonymy between parts of speech was expected since it is a strong property of Slavic languages.

3.3 Headword Revision


In Headword Revision, annotators get the chance to review headword candidates that were rejected in the Headwords task but could be turned into correct headwords. For each such rejected headword, Lexonomy displays a form in the right-hand pane (see Figure 3), whose exact content varies depending on what is signalled to be an issue with the headword (e.g. not a lemma, wrong part of speech, non-standard spelling).

For instance, if only part of speech is believed to be wrong, then the lemma field is pre-filled and the annotator is asked to select a different part of speech from a dropdown. However, they still have the option to modify the lemma as well, at their own discretion. For cases of ambiguity, it is possible to enter multiple revisions for a headword. The annotator can also decide that the headword be ignored (without revision), or accepted as is (call it correct). Due to the decisive character of this task, it should be commissioned with priority to annotators with high proficiency in the language and good performance in the Headwords task.

ПОЛІКЛІНІК PoS: noun

CORRECT HEADWORD SHOULD BE:

lemma: PoS: proper name?



I DO NOT UNDERSTAND THIS WORD

THIS IS NOT A UKRAINIAN WORD

THIS HEADWORD IS CORRECT

EXAMPLES

1. Хоча керівництво **поліклініки** й надалі переконує – внески добровільні.
2. Зусиллями міської влади і депутатів басейн повернули на баланс **поліклініки** .
3. Його буде встановлено на першому поверсі хірургічного корпусу **поліклініки** .
4. Один випадок був зафіксований й на території однієї з **поліклінік** міста .
5. Проведення медоглядів у **поліклініках** у присутності батьків є логічним.

Figure 3: Interface for the Headword Revision task

In this project, 54,503 headword candidates were sent for revision. Some of them eventually underwent revision more than once (in order to explore inter-annotator agreement), what resulted in 5,820 duplicate entries (though with possibly differing annotations). Four annotators contributed to this task, which was split into 66 batches.


To make an annotation, the user clicks a radio button. If the headword is to be corrected, then they also enter the correct lemma, pick the correct part of speech and indicate whether it is a proper name.

In 94.9% of cases, a revision was resolved by providing an alternative headword. In 3.2%, annotators said that the displayed headword was in fact correct. The remaining 1.9% were cases of unrecognized words or words considered non-Ukrainian. In the typical situation when correct headwords were provided to replace an incorrect headword, in 91.6% there was just one replacement headword, in 7.4% there were two and in 1.0% three or more (up to six).

Total time annotators spent on this task was 722 hours, i.e. annotating one entry took on average 43 seconds. Speed fluctuated a lot across annotators in this task too, with the fastest person taking just 28 seconds per entry and the slowest one needing 77 seconds.

3.4 Word Forms

The Forms task is concerned with inflection. Ukrainian is an inflected language and we want to collect as many inflected forms of each headword (lemma) in the corpus as possible. Annotators are first trained to distinguish inflection from derivation. Then, in Lexonomy, their task is to tell apart correct and incorrect items in a list of possible inflected forms for each headword. A link to concordance is available for case of doubt, but in practice, most items are resolved swiftly. “Correct” is the default, so the annotator needs to act only in case of incorrect forms. This task has a threshold only slightly higher than the Headwords task, it can be introduced quite early in the process and no other later tasks depend on it, which makes it a universal task for times of delay etc.

держати (verb)  I DON'T KNOW

Inflected forms:























	form	correct?	
1.	держати	=headword	
2.	держити	 	
3.	держало	 	
4.	держ	 	
5.	держимо	 	
6.	держатиме	 	
7.	держатимуть	 	
8.	Держать	 	

Figure 4: Interface for the Forms task

In this project, word forms were sought for 42,694 headwords, for which there was a total of 578,327 form candidates (i.e. an average of 13.5 form candidates per headword). Among the form candidates, nearly all (99.2%) only appeared with a single headword. This means that the task was not as much about checking the form-lemma relationship, rather than about checking the correctness and acceptability of the form itself (the used tagger is permissive and accepts even some archaic and corrupted word forms). All seven annotators available at the time were made to work on this task. The work was divided into 43 batches.

The observed ratio of reported incorrect forms was 21.6%. In almost four out of five such cases (79.4%), the rejected form candidates started with a capital letter – and, for lemmas which start with a small letter themselves, such word forms differing in letter case are highly unlikely in Ukrainian. In fact, 77.4% of all capitalized word forms ended up marked incorrect.

Annotators spent 1269 hours checking the word forms, which means that they took on average 107 seconds per headword, or 8 seconds per word form. The fastest annotator needed only 2.5 seconds per word form (or 35 seconds per headword), while the slowest required 22 seconds per word form (or 328 seconds per headword). Explanation of these inter-annotator differences must be looked for in the same factors as mentioned with the Headwords task.

We did not make any automatic judgments on the correctness of words forms, but we benefited from the large corpus to extract a rather satisfying list of them – both in terms of precision (we have shown above that majority of the presented candidates were correct) and recall (although we did not attempted to quantify it, as we explicitly did not aim at acquiring a “full” word form list, whatever it should mean). The average number of unique word forms per lemma was 18.0 for verbs, 13.1 for adjectives, 9.4 for pronouns, 7.3 for nouns, 5.4 for numerals, and below 1.3 for other parts of speech (uninflected). Depending on details of the used processing pipeline, orthographic or phonetically motivated variants of words may have been represented either as “inflected forms” or as separate headwords.

3.5 Audio Recordings

Instead of relying on phonetic transcriptions to indicate pronunciation or on the traditional stress marks to indicate word stress, we make an audio recording of the headword’s pronunciation by a native speaker and store it as a part of the dictionary entry. This is the only part of the entry creation process that is done fully manually, since we want to be in control of the quality of the result and automatic text-to-speech output could not be post-edited. However, apart from having to face a few challenges such as preserving a steady loudness or maintaining a low noise level, it turned out to be also one of the simplest tasks. This is also the only step that does not use Lexonomy, but a specially developed audio-recording software, and the only step which necessitates physical presence of the annotator in dedicated premises (a soundproof audio cabin with high-quality recording hardware) during the whole work time.

In this project, we recorded audio pronunciation for all the 55,632 headwords in the final dictionary. Some of the headwords were recorded multiple times and, due to the recording occurring in parallel with the rest of the dictionary building, we also made recordings of

some headwords which eventually did not make it into the final dictionary. In total, 57,800 audio files were created (i.e. 3.9% overhead). The work was divided into 60 batches, 59 of which were assigned to the same annotator so that same voice is used throughout the dictionary. Only the last batch (1.3% of headwords) was assigned to a different person because the original speaker was not available anymore.

The recording station in the audio booth was controlled with a special small 6-key keyboard (the available keys were marked with pictograms meaning YES, NO, SKIP, DOWN, UP, QUIT, respectively). This was done to save desk space else occupied by a regular keyboard, concentrate all controls in a single location, reduce the chance of typos, limit noise generated by keystrokes and improve user comfort for the annotator. The processing of each headword consists of seeing it displayed on the screen, recording its pronunciation, then listening to the recording to check its quality, and possible re-recording if the quality is not sufficient.

It took the annotators 553 hours, or about 14 weeks (of 40 work hours each), to make the recordings. That means an average of 36 seconds per headword. This time, however, includes regular break time, because it is demanding if not impossible for a non-trained person to stay concentrated in a small booth and keep speaking using a fresh voice of stable strength for the whole day. In fact, in most of the cases when a headword had to be recorded repeatedly, the reason for this was the software stepping in with an automatic low-volume alert.

3.6 Word Senses

Identification of word senses for each headword is an important step in the dictionary building process, because all subsequent tasks are performed on sense level instead of headword level, and therefore dependent on the word-sense distinctions made here. After annotators learn that there is not a single perfect solution for the problem (Kilgarriff, 1997), reaching common ground with regard to granularity of sense distinctions is attempted by means of joint practice and discussions on each other's proposed solutions.

Annotators' invention is effectively limited to automatically induced word sense data (read more in 4.2.5), represented in Lexonomy as *example usages* (i.e. collocations, each including a longest-commonest match (Kilgarriff et al., 2015)) and grouped into clusters, each of which could be considered a word sense candidate. Having reviewed this data, however, the annotator has the freedom to establish a number of senses of their choice, to distribute the collocations among them freely, not to assign a collocation to any particular sense (by marking it either as "mixed sense" or "error") and even come up with a sense not linked to any of the collocations (the latter is allowed so that no important word senses are lost due to possible deficiencies of the word sense induction algorithm). Each sense is also given a disambiguating gloss (in the language of the dictionary), one or more English translations, and a mark saying whether it is offensive in meaning.

The Word Senses task might be the most difficult task to be properly trained, and the quality of its outcome directly influences the quality of data in all upcoming tasks. In this project, 10,098 post-edited word sense disambiguations were performed in this way, for a total of 10,016 distinct headwords. In each processed entry, there were on average 43 collocations, divided into 9 clusters. Four annotators were chosen and trained for this task and the work was divided into 55 batches.

In terms of part of speech, 45.2% of the annotated headwords were nouns, 24.7% adjectives, 21.6% verbs and 5.8% adverbs; the remaining 2.7% were other parts of speech, for which word sense disambiguation is not always applicable. Figure 5 shows an example of one cluster, with three collocations. Annotators assign collocations to senses by clicking numbered buttons. The available buttons multiply as soon as more senses are declared – which is done by providing a disambiguating gloss, English translation(s) and possibly switching a toggle to mark offensiveness. To reflect real-world conditions, English translations can be shared by multiple senses, again by means of numbered buttons, thus reducing the need for typing. And when a collocate is not self-explaining, the annotator has the option to view a corresponding concordance in the corpus.

Group 1				
Mark all: <input type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>				
example usage	actions	collocate	relation to headword	concordance
<i>бродіння відбуватися</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	бродіння NOUN	"відбуватися" молочнокисль ...	🔗
<i>відбувається масаж внутрішніх органів</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	орган NOUN	"відбуватися" масаж ...	🔗
<i>заміщення відбуватися</i>	<input checked="" type="button" value="1"/> <input type="button" value="NEW"/> <input type="button" value="MIXED"/> <input type="button" value="ERROR"/>	заміщення NOUN	"відбуватися" поступовий ...	🔗

Figure 5: Part of the interface for the Word Senses interface

For 60.1% of headwords, a single sense was identified; 18.5% was split into two senses; 10.5% into three senses; 5.0% into four senses; the remaining 5.8% into five senses or more. Overall average number of senses identified for the processed headwords was 1.84. Among annotators, the average number of identified senses reached from 1.38 to 2.31. The highest number of senses (not necessarily an ideal) was routinely found by an annotator who happened to have some formal education in the field of linguistics. The same annotator was also the only one who would, exceptionally, go to great lengths by establishing more than 10 senses for a headword.

Annotators spent a total of 1203 hours on the Word Senses task, i.e. about 7 minutes per headword. Three of the annotators had very close averages (from 7.3 to 9.1 minutes per headword), only the fourth annotator (the one with linguistic education) differed substantially when she took a much lower average of 4.5 minutes per headword.

Of the listed collocations, only 1.8% were declared incorrect or incomprehensible, and 1.5% could not be conclusively attributed to particular sense (when there were more senses to choose from). Remaining collocations were either all attributed to a single sense, or distributed among two senses (in the average ratio of 79:21) or three senses (67:23:10). Even with four senses, the least-frequent one still corresponded to approximately three collocations, which indicates that even in highly competitive situations, all senses were solidly backed up by corpus data (in contrast with senses defined without any corpus evidence, which are disregarded in this computation).

Annotators entered a total of 26,715 English translations (usually single words, but sometimes multi-word expressions, and exceptionally even descriptions of concepts that lack a direct English translation), which means an average of 2.65 translations per headword. This is close to the average number of pre-generated machine translations of the headword into English, which was 2.45.

Only a tiny fraction (25, 0.1%) of the identified senses has been marked offensive, although the annotators were aware of this possibility and each of them used it at least once. We believe that more of the headwords could be used in an offensive or derogatory way and suspect that the annotators may have under-annotated them under the influence of the previous tasks, in which we had to repeatedly stress that also bad words are to be included in the dictionary and that they should be treated *as any other word*.

3.7 Thesaurus

In the Thesaurus task, annotators are trained to evaluate thesaurus candidates (i.e. selected related headwords, read more in 4.2.6) for a given headword *in one of its senses* (this subdivision into senses is maintained across the rest of the dictionary building). Each thesaurus item can be put into one of three categories: Synonym, Antonym and Similar word (i.e. not a synonym or antonym, but still somehow related). A fourth option, named Other, is the default choice and results in the candidate being discarded.

океан NOUN I DON'T KNOW

translations: **ocean**

thesaurus candidates:

	candidate	type			
1.	море <small>NOUN</small>	synonym	antonym	similar	other
2.	затока <small>NOUN</small>	synonym	antonym	similar	other
3.	озеро <small>NOUN</small>	synonym	antonym	similar	other
4.	річка <small>NOUN</small>	synonym	antonym	similar	other
5.	пустеля <small>NOUN</small>	synonym	antonym	similar	other
6.	ріка <small>NOUN</small>	synonym	antonym	similar	other
7.	гора <small>NOUN</small>	synonym	antonym	similar	other

Figure 6: Interface for the Thesaurus task

In this project, two annotators were assigned to the Thesaurus task and, at the time of writing, they had processed in total 10,377 entries (headwords in individual senses), divided into 12 batches. Each entry contained exactly 20 thesaurus candidates.

Out of all thesaurus candidates, three fourths (75.5%) were discarded (marked as Other), while 15.0% were accepted as Similar, 8.2% were classified as Synonyms and 1.3% as Antonyms. In the training phase, we realized that one annotator had developed a preference towards marking many *related* words as Similar, while the other preferred Synonym in these cases. During data inspection, we found out that Similar:Synonym ratio was 83:17

for the first annotator and 61:39 for the second one. We could, however, not find solid grounds on which we could convince one or the other to change their preference. The percentage of identified antonyms was consistently low with both annotators.

Work on the Thesaurus took 364 hours, with one of the annotators being significantly slower (527 seconds per entry) than the other (74 seconds).⁵

On average, 4.9 thesaurus candidates were accepted for each headword. Since the candidates were scored by Sketch Engine and shown in that order, we would expect that items higher in the list have a higher chance of being accepted as thesaurus items. And indeed, the probability that a candidate item had been accepted was found to be inversely proportional to the item's rank; it was 48.9% for the first item in the list, 38.1% for the second one, 32.3% for the third one; 19.3% for position ten, 15.6% (minimum) for position fifteen. Positions 16–20 were exceptions, because they had been reserved for top-scored thesaurus candidates for the headword, regardless of sense. These items had a higher chance of being accepted (21.2–27.3%), comparable to that of the (sense-specific) positions 5–9.

3.8 Usage Examples

Choosing a good, easy to understand, illustrative dictionary example for a headword (in one of its senses) is a challenging task. So although GDEX (Rychlý et al., 2008) is used to pre-select candidate sentences (read more in 4.2.7), annotators need to be well trained to choose the best one of the five pre-selected sentences and redact them when necessary (shorten them or remove controversial information). In rare cases, annotators may even be forced to come up with an example sentence of their own (for this purpose, they have on hand a link to the first one hundred GDEX-scored collocation lines from the corpus as source of inspiration), although writing example sentences anew is strongly discouraged for reasons of time expense and authenticity.

In the user interface, the annotator selects their preferred sentence by clicking on a button next to it. Clicking directly on the sentence activates a text input field in which its text can be modified as needed. After an example sentence is selected, it changes color from red to green and another text field opens below it, pre-filled with machine translation of the original sentence into English. It is the annotator's responsibility to check and fix the English translation as needed and to make sure that the sentences in the two languages stay in sync.

Four annotators were trained in this activity and 20 batches were finished at the time of writing this paper. In those batches, the annotators processed a total of 14,474 entries, each containing five pre-selected and pre-translated example sentences. The work took them 693 hours, which averages to 2.9 minutes per entry. The average time spent on an entry varied greatly across the annotators (0.8 minutes, 4.3 minutes, 6.0 minutes, 14.7 minutes). The differences are likely to have been caused by each annotator's differently strong criteria for a good example. Slower annotators edited their chosen examples more heavily, often fully rewriting them because they thought it necessary.

It seems that the position of the five offered sentences in the list (they were order by decreasing GDEX score) correctly reflected their quality, or at least that sentences closer

⁵ Due to the charitable dimension of the project, the work with annotators had defects which would not be tolerated in a fully commercial setup.

examples:

1.

Термін дії проміжних нарядів не повинен перевищувати терміну дії загального наряду.

NO
 YES
2.

Це гарантує високу міцність і тривалий термін служби.

NO
 YES

translation This guarantees high durability and a long service period.
3.

Термін дії візи буде точно відповідати тривалості навчання.

NO
 YES

Figure 7: Interface for the Examples task

to the top attracted annotators' attention more and would be more probably chosen in case if multiple comparably good candidates were present. The chance of the sentence in position one to be chosen was 34.7%; position two 18.9%; position three 15.0%; position four 12.6%; position five 12.3%.

The average length of the chosen example in its original (from the corpus) and accepted (possibly modified) form was 63.1 and 56.5 characters (8.7 and 7.8 words), respectively, which suggests a welcome tendency of the annotators to produce shorter examples. The same tendency was found also with regard to the length of the sentence's English translation (decrease from 67.6 to 60.8 characters; from 11.3 to 10.2 words). Evaluation of Levenshtein distance (minimum number of insertions, deletions, and substitutions) between the generated and post-edited Ukrainian sentences reveals that 67.3% of the 13,449 studied sentences did not need any modifications at all, and the average edit distance was 12.6 on the whole set (and 38.5 just on those sentences which needed modification).

The pre-generated machine translations of the original Ukrainian sentences into English and their final forms (often updated both for language and for linguistic deficiencies in the Ukrainian originals) differed more, as expected, but not substantially: the edit distance was 15.9 (and 34.6 on just the modified sentences, which is even a decrease). Also, surprisingly, 54.1% of the machine-translated sentences were considered good enough by the annotators to be left intact. This seems to suggest that the machine translation is reliable and saves time during annotation. Indeed, in cases when the Ukrainian sentence was left unmodified, 76.0% machine translations were also not modified; and other 7.7% only required up to 5 edits (insertions, deletions or substitutions) to be performed in order to fix the English sentence. The average edit distance of English sentences in these cases of unmodified Ukrainian sentences was 2.6 (or 10.6 just on the modified sentences).

3.9 Images

The Images task was not yet administrated at the time of writing this paper, but we foresee using an interface similar to the one depicted in (Baisa et al., 2019). Freely licensed images relevant to the headwords will be identified and a top list will be offered to annotators to choose from.

3.10 Final Review

Final Review is the last phase of the dictionary building process. In it, a complete dictionary entry is composed out of the collected components (see entry structure in 5.1) and visualized for the first time. The annotator’s task is to fix any typos and mistakes and to check the overall coherence of the entry.

For instance, senses (however well-defined) are perceived differently across annotators, who may produce translations, usage examples and images that are not fully compatible with each other. In Final Review, a skilled annotator has the last say and can modify or delete entry components to achieve coherence. Addition of information, however, is discouraged at this step. Final-reviewed entries have got their definite form (in terms of data management, not visualization), in which they will appear in the final dictionary.

4. Data Management

Baisa et al. (2019) reported on issues with data management. Although the paper itself is not very specific about this issue, we have learned from the authors that the issues were connected to the fact that the XML annotations from all the phases described above were exported from and imported into one centralized database. Once an annotation was imported into the database, it could not be easily changed and re-imported, because the entries could have been changed by following imports. The approach would be probably working fine if all the annotations and import/export processes were perfect and consistent, however, that was not the case. Every inconsistency in annotation and all the small bugs in the automatic import/export procedures, propagated and resulted in a decent amount of entries containing inconsistent information which must have been manually corrected, generating delays and additional costs. Moreover, as new versions of the source corpus were produced (e.g. due to improvements of lemmatization and tagging), some parts of the data became inconsistent with the corpus.

Therefore, our approach to the data management is different. We take the source corpus and the native speaker annotations as source data for fully automated procedure that creates respective dictionary parts, merges them into the complete dictionary and generates new data for annotation. The procedure is implemented as a Makefile which makes it easy to define dependencies among the individual components, and is illustrated in Figure 8. In case of any change (new annotations available, new version of the source corpus, ...) all the data are re-processed, new versions of partial dictionaries and derived corpora are created, and a new version of the dictionary is automatically generated. Also, new data for annotation, if needed, are created.

This approach gives us the flexibility to fix any problem or inconsistency in the source data, or in the manual annotations, that previously passed unnoticed, and re-generate

the whole dictionary easily. The fully automated procedure therefore enforces consistency across all the pieces of data involved in the process. Also, it can be used instantly for a new corpus and a new language.

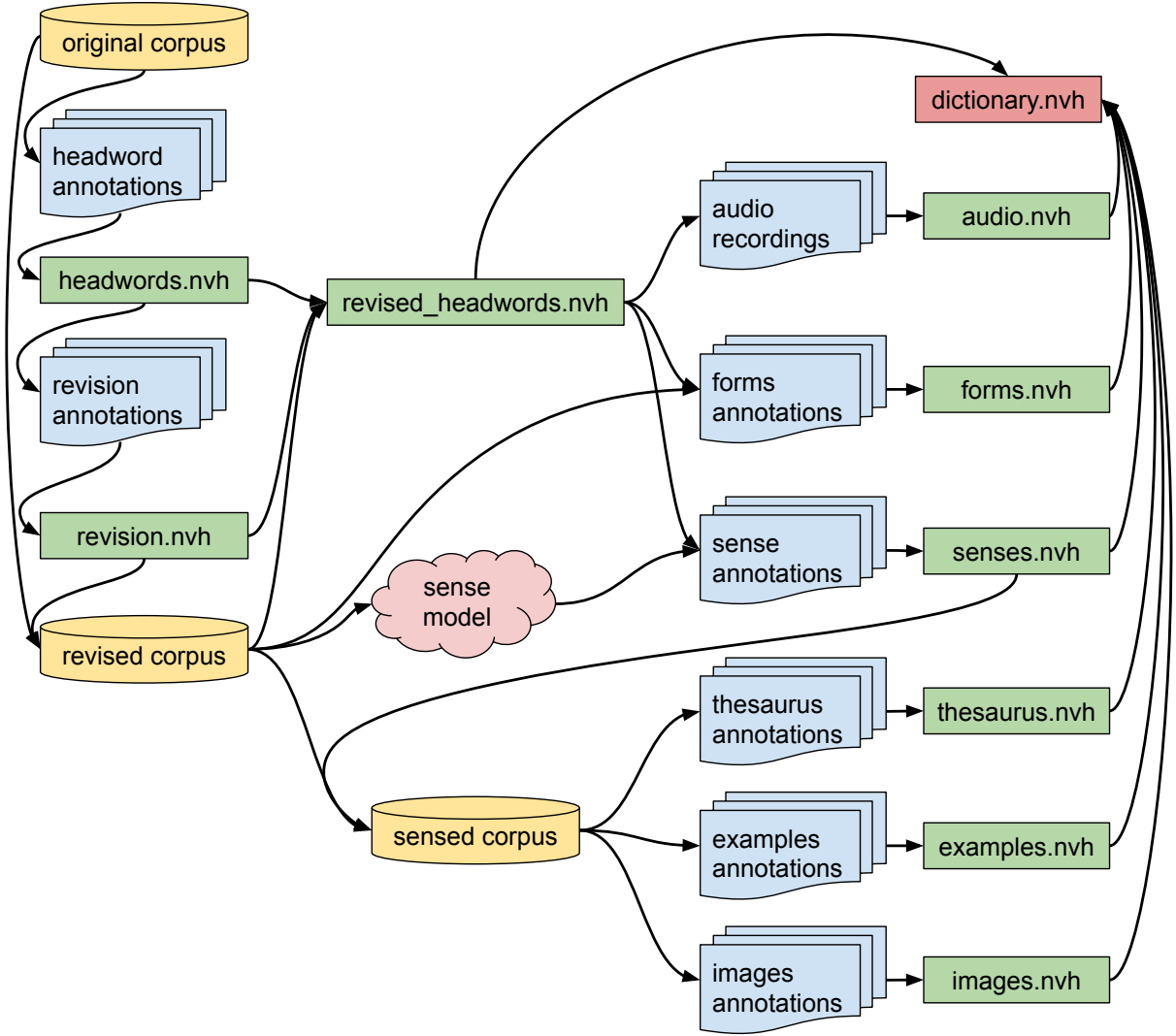


Figure 8: Illustration of the data management process

4.1 Formats

For the partial dictionaries (green rectangles in Figure 8) as well as for the resulting dictionary, we used the NVH – name-value hierarchy – format (Jakubíček et al., 2022), a text format easily readable for both humans and simple automatic text processing tools, which is suitable for dictionary data and significantly less complex than XML.

For the manual annotations (blue “documents”), XML was used as the internal format of the Lexonomy software where the annotators worked.

4.2 Generating the Dictionary

In this part we describe the automatic procedure more in detail. In Figure 8, every shape represents a target in a Makefile, and the arrows represent the dependencies among the particular targets. Typically, there is one Python script (or a few calls of the standard UNIX tools) for each of the targets, which generates the target contents from its dependencies.⁶

For clarity, we have split the description into parts, but please bear in mind that all the content of this section is one fully automatic process that runs as a whole over partial data, and can be repeated as many times as needed.

4.2.1 Headwords

At the very beginning, there is a source corpus, tagged and lemmatized automatically, using available software tools. The first step of the procedure takes the word list of lemmoses (lemmas with a one-letter part-of-speech suffix) from the corpus and generates annotation batches (“headword annotations” in Figure 8) for the N most frequent words (by document frequency). In this project, a total of 102,323 lemmoses received 2 annotations from different annotators, and if they were not in agreement, further annotations were collected until there was at least 50% agreement.⁷

From the headword annotations, a partial dictionary `headwords.nvh` is generated, containing lemmos, its annotations, final decision and the percentage of agreement, for each of the headwords.

4.2.2 Headword Revision

If the final decision about a headword was *wrong_lemma*, *wrong_pos* or *non_standard*, a *revision* annotation was generated – a next step whose purpose was to fix mistakes of the automatic lemmatizer and tagger and find correct (or standard, respectively) lemmas and parts of speech of the words. Most of the items sent to revision were revised to a word that we already had in the dictionary, but we also obtained 6,177 words that we had not seen before and the dictionary would miss them if the *revision* step was not incorporated.⁸

The outputs of the revision annotations are merged into a partial dictionary `revision.nvh` which records the corrections. This dictionary is then used in two ways:

- Using the recorded revisions and the original corpus, we create a *revised corpus* that contains correct lemmas and parts of speech, and is used as a base for further

⁶ The Figure 8 is slightly simplified. In the real Makefile, there are few more targets of rather technical nature that would split some of the arrows into two. However, they are not important for understanding the principle of the procedure, so we don't discuss them here for the sake of clarity.

⁷ However, this is something we may want to change in the future because collecting too many annotations slightly complicated the task and led to a delay. For the next projects, we would recommend collecting 2 annotations only, and continuing directly to the *revision* step in case of disagreement.

⁸ There is still a theoretical possibility of missing an important word: if the lemma of a frequent word form was ambiguous and the tagger always returned one of the options and never the other; however, we did not encounter such a situation in the project. Also, this problem would be present with any approach based on a list of lemmas from a corpus.

processing, namely word senses. (If we did not take this step, the word sense model would not contain the 6,177 new words at all, and the data for other words would be incomplete.)

- We merge it with `headwords.nvh` and create `revised_headwords.nvh` which contains a final list of headwords for the dictionary together with frequencies and frequency ranks generated from the revised corpus. The next phases do not add more words into the dictionary, they just add more information to words that are already present.

4.2.3 Word Forms

For each of the valid words in `revised_headwords.nvh`, we generate a list of word forms present in the revised corpus into the word form annotation batches. The annotators mark them as correct or wrong and the correct word forms are then exported into `forms.nvh` which is later merged into the final dictionary.

4.2.4 Audio Recordings

Audio batches for recording are generated for all the valid words in `revised_headwords.nvh`. After recording, the audio files are kept separately and the metadata containing information about the location of the particular audio file, are compiled into `audio.nvh` which is then merged into the final dictionary.

4.2.5 Word Senses

From the revised corpus, we generate an automatic model of word senses for all words in the corpus. At first, we used traditional collocation-based approach described in [Herman et al. \(2019\)](#), but the result would frequently miss high frequency senses. The overall quality of the result was not sufficient and significant post-editing effort was necessary to extract useful information. For this reason, we switched to a word sense induction model based on [Bartunov et al. \(2016\)](#), which represents the senses of a word as word embeddings. Then, we map the senses from the model onto (some of) the collocations from word sketch, clustering the collocations. Each cluster of collocations is then considered a candidate sense. From these clusters of collocations, we generate sense annotation batches and ask the annotators to name, fix and translate the automatically identified senses, as discussed in [3.6](#).

These annotations are then processed into another partial dictionary `senses.nvh` that records the division of each word into senses, the collocations assigned to the particular senses, and the names and translations of the senses. Apart from being an input for the final dictionary, this partial dictionary is used to generate a *sensed corpus* from the revised corpus, where the basic unit of analysis is not a lemma (lempos) anymore, but a *sense*. In our particular implementation, a sense is a lempos concatenated with the sense name, e.g. *bank-n#river* vs. *bank-n#money*, but the exact string is not important, it could as well be *bank-n#1* and *bank-n#2*. The important moment is that now we can work with separate senses instead of lemmas (lemposes)—namely compile word sketches and thesaurus for senses so that word sketch and thesaurus for *bank-n#river* is different from word sketch and thesaurus for *bank-n#money*.

4.2.6 Thesaurus

For each *sense* recorded in `senses.nvh`, a list of similar words (and similar *senses*) is pulled from the *sensed corpus* using Sketch Engine’s thesaurus function (Rychlý & Kilgarriff, 2007) and put into thesaurus annotation batches. Because not all of the occurrences are clustered into senses, we merge thesaurus for the sense with the thesaurus of the (more general) lemma to get more quality data. The results of the annotation are again compiled into a partial dictionary `thesaurus.nvh`.

4.2.7 Usage Examples

For each sense recorded in `senses.nvh`, we generate a set of 5 best candidate example sentences from the corpus with the GDEX tool (Rychlý et al., 2008). For this purpose, a new Ukrainian-specific GDEX configuration was created. The candidate sentences are then automatically translated into English by the DeepL API⁹ and annotation batches are created from the extracted sentences and their automatic translations. The annotators are then asked to read all the sentences, select one best example, edit it (but only if needed) and check and edit (again, only if needed) its automatic translation into English.

The annotations are then processed into a partial dictionary `examples.nvh` which is then merged into the final dictionary.

4.2.8 Images

The images phase of the project is not yet finished at the time of writing this paper, but we intend to implement it in the same frame as the previous phases: automatically search for copyright-free images in several databases, based on English translations for each sense, let the annotators select one best image out of 10, and record the selections in a partial dictionary `images.nvh`.

5. About the Dictionary

So far we discussed the process of compiling the Ukrainian dictionary. This section summarizes some basic information about the resulting dictionary itself.

5.1 Entry structure

The entry structure of the dictionary may be clear from the description of the methods above—however, the following description shows it explicitly:

- **Headword (lemma + part of speech)** is the basic identification of every entry. It is also the primary key of the dictionary in the database sense—we don’t allow multiple entries with the same lemma and part of speech.
- **Flag** specifies the type of the entry: in the final dictionary, we have only *ok*, *name* and *non_standard* but we also keep all the rejected words with the other flags in

⁹ [deepl.com](https://www.deepl.com)

a separate database. *Non_standard* and *name* entries do not contain senses, and *non_standard* also contains a link to the standard form of the headword.

- **Frequency** of the word retrieved automatically from the document frequency in the corpus, i.e. number of the documents the headword occurred in.
- **Rank** of the headword according to the frequency (computed automatically from the frequency).
- **Pronunciation**, or precisely the location of the audio recording with the pronunciation, the output of the audio recording phase.
- **List of word forms**, the output of the word forms post-editing phase.
- **List of senses** identified in the sense annotation phase. Only words marked *ok* have senses and translations. Next, every sense contains:
 - **Sense identifier** or disambiguator which tells the senses apart and may explain them to an extent (but it is neither definition or explanation of the sense). It may be empty if the word is found monosemous (has only a single sense recorded in the dictionary).
 - **One or more translations to English**, as recorded in the sense annotation phase.
 - **Collocations** sorted by grammatical relations, as recorded in the sense annotation phase. Each collocation also contains a short example (typically 3-5 words) automatically extracted from the corpus.
 - **List of synonyms, antonyms and similar words**, as identified in the thesaurus annotation phase.
 - **One usage example** and its translation to English, both results of the example annotation phase.
 - **Image**, if appropriate, selected in the images selection step (not implemented yet). Every image consists of its location, source and license.

The structure is rather shallow, but we believe it contains the most important elements for a decent dictionary entry. Also, the modular nature of the process makes it possible to add further steps easily, such as definitions/explanations or translation into more languages.

In this dictionary, we did not take multi-words into account—but there are already tools available to identify multi-words from corpus n-grams and collocations that would make it relatively easy to enrich the dictionary in this direction.

5.2 Basic statistics

For organizational and budget reasons, we did not complete all the entries all the way through, some of them are “more complete” than others. A relatively long list of valid frequent headwords and word forms is a valuable multi-purpose resource, so we aimed at having a really long list of headwords first, and then continued with the other phases step by step, always starting with the most frequent headwords.

By the time of writing this paper, the project is still not finished, but mainly for budget reasons we slowed it down and now the work continues with only one remaining Ukrainian editor. This means the numbers below are not final but they reflect the state after less than 1 year of intensive work during which **6,918 hours** of manual post-editing work (or approximately 3.5 full-time person-years) were consumed. See Figure [10](#) for a breakdown by task.

зуб

зуб

зубець

зубка

зубний

зубожілий

зубожіння

зубожіти

зубок

зубопротезування

зубочистка

зубр

зубчастий

зубчатий

зубчик

зуб NOUN ★☆☆☆

rank: 3 776

Inflected forms

зубів, зуби, зубами, зуба, зуб, зубах, зубом, зубам, зубі, зубу, зубові

1 анатомічний

In English

tooth

Synonyms

ікло, моляр

Similar

коронка, протез

Examples

Існує один дуже хороший народний метод відбілювання зубів.
There is one very good folk method of teeth whitening.

2 механічний

In English

tine

Examples

Основні елементи циліндричного зубчастого колеса з прямим зубом.
The main elements of a spur gear with a straight tooth.

3 озброєний до зубів

In English

armed to teeth

Examples

Сюди ми прийшли на катамарані, озброєні до зубів.
We came here on a catamaran armed to the teeth.

Figure 9: Dictionary entry example for the word зуб (tooth). There are 7 senses of the word in total, here we show only the first three of them.

Overall, our database contains 123,574 annotated headword candidates (i.e. all the headwords from the corpus seen by at least one annotator). This figure includes the revised headwords that were originally not present in the corpus—without them, it is 117,397. Of these, 14,141 were only seen by one annotator (better than nothing but not reliable

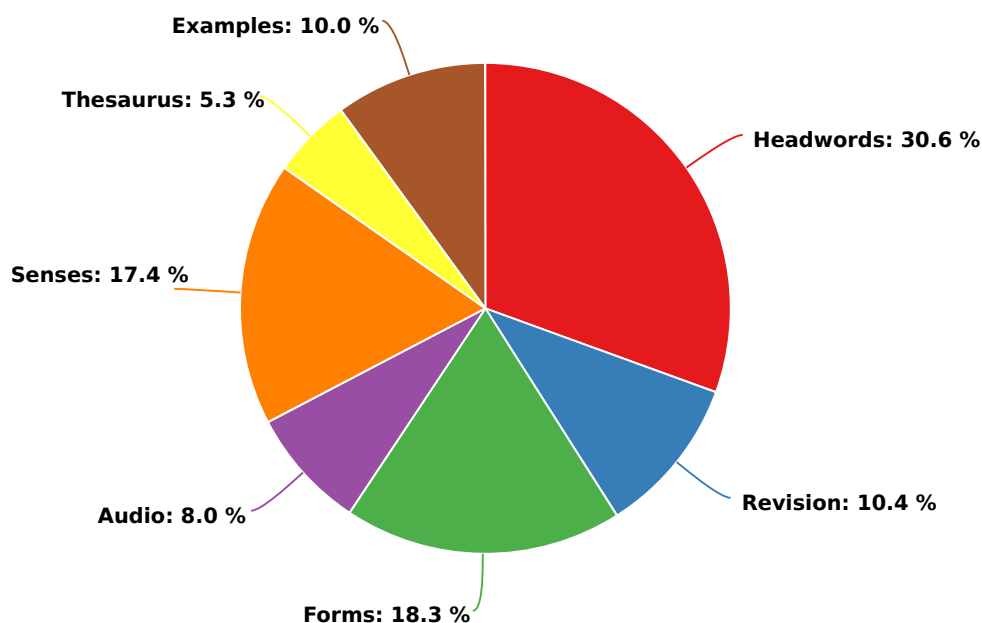


Figure 10: Workload by task (100% = 6,918 hours)

enough for the dictionary) which leaves us with 109,433 headword candidates with reliable annotation.

Of these, 55,632 ended with flags suitable for the final dictionary, namely:

- 46,987 common words (marked *ok*)
- 8,252 proper names
- 393 non-standard words

So we can say that the size of our dictionary is **55,632 entries**. All of these entries contain an audio recording of the pronunciation, as well as frequency and rank derived automatically from the corpus.

42,639 of these headwords contain list of their word forms which is in total 453,010 validated word forms.

The size of the dictionary in terms of complete entries, i.e. entries with verified senses, translations, thesaurus and usage example, is **9,785**. (We still plan to add images in the near future.) Of these, 3,901 entries are polysemous and 5,884 are monosemous. 1,057 words have more than three senses. Total number of senses in the dictionary is 17,973.

In all the process phases, we always proceeded according to document frequency. In other words, we went through the 109,433 most frequent words in the corpus, the dictionary contains the 55,632 most frequent Ukrainian words (according to the corpus) and we have complete entries with senses for the 9,785 most frequent words.

6. Conclusions

We have reported on a rapid corpus-based development of a new Ukrainian-English dictionary using a new process of automatic generating and step-by-step post-editing of

the dictionary. We described building the source corpus, then we went through all phases of the process in detail and explained our approach to dictionary data management during the process.

The resulting dictionary contains ca. 10,000 finished entries; another 45,000 entries for less frequent headwords are partly finished. Overall the process consumed less than 7,000 hours of paid editor's time which is a fraction of both time and money needed to build a similar dictionary in a traditional way with professional lexicographers.

In the future, we will continue working on the dictionary (2–5,000 more finished entries, adding images), and since we made the workflow setup really easy within this project, we are looking forward to running similar projects with new languages soon.

7. Acknowledgements

We cordially thank the Institute for Ukrainian (<https://mova.institute>) for permission to use their manually annotated corpus available through the Universal Dependencies project (https://github.com/UniversalDependencies/UD_Ukrainian-IU). We cordially thank Andriy Rysin, Vasyl Starko and the BrUK team for permission to use the Ukrainian morphological database they developed and made available at https://github.com/brown-uk/dict_uk. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic.

8. References

- Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medved, 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.
- Bartunov, S., Kondrashkin, D., Osokin, A. & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*. PMLR, pp. 130–138.
- Gantar, P., Kosem, I. & Krek, S. (2016). Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2), pp. 200–225. URL <https://doi.org/10.1093/ijl/ecw014>. <https://academic.oup.com/ijl/article-pdf/29/2/200/7199846/ecw014.pdf>.
- Herman, O., Jakubíček, M., Rychlý, P. & Kovář, V. (2019). Word Sense Induction Using Word Sketches. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings 7*. Springer, pp. 83–91.
- Jakubíček, M., Kilgarrieff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*. Lancaster, pp. 125–127. URL <http://ucrel.lancs.ac.uk/cl2013/>.
- Jakubíček, M., Kovář, V., Měchura, M. & Rambousek, A. (2022). Using NVH as a Backbone Format in the Lexonomy Dictionary Editor. In A.R. Aleš Horák Pavel Rychlý

- (ed.) *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022*. Brno: Tribun EU, pp. 55–61. URL <https://raslan2022.nlp-consulting.net/>.
- Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-editing Lexicography with Lexonomy and Sketch Engine. In *The XVIII EURALEX International Congress*. p. 65.
- Jongejan, B. & Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore: Association for Computational Linguistics, pp. 145–153. URL <https://aclanthology.org/P09-1017>.
- Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocations: Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Trojina, Institute for Applied Slovene Studies, pp. 1–20.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31, pp. 91–113.
- Kilgarriff, A., Baisa, V., Rychlý, P. & Jakubíček, M. (2015). Longest–commonest Match. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd ..., pp. 11–13.
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2018). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), pp. 119–137. URL <https://doi.org/10.1093/ijl/ecv014>. <https://academic.oup.com/ijl/article-pdf/32/2/119/28858872/ecy014.pdf>.
- Měchura, M.B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. pp. 19–21.
- Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masaryk University.
- Pomikálek, J. & Suchomel, V. (2011). chared: Character Encoding Detection with a Known Language. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011*. pp. 125–129.
- Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end. *A Taste for Corpora. In Honour of Sylviane Granger*, pp. 257–282.
- Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M. & McAdam, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada, pp. 425–432.
- Rychlý, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. pp. 41–44.
- Schmid, H. & Laws, F. (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK:

- Coling 2008 Organizing Committee, pp. 777–784. URL <https://aclanthology.org/C08-1098>.
- Shvedova, M. (2020). The General Regionally Annotated Corpus of Ukrainian (GRAC, uacorporus.org): Architecture and Functionality. In V. Lytvyn, V. Vysotska, T. Hamon, N. Grabar, N. Sharonova, O. Cherednichenko & O. Kanishcheva (eds.) *Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*. Lviv, Ukraine, April 23-24, 2020, volume 2604 of *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 489–506. URL <https://ceur-ws.org/Vol-2604/paper36.pdf>.
- Starko, V. (2021). Implementing Semantic Annotation in a Ukrainian Corpus. In N. Sharonova et al. (eds.) *Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2021)*. Kharkiv, Ukraine, April 22-23, 2021, volume 2870 of *CEUR Workshop Proceedings*. CEUR-WS.org, pp. 435–447. URL <https://ceur-ws.org/Vol-2870/paper32.pdf>.
- Starko, V. & Rysin, A. (2020). *Velykij elektronnyj slovnyk ukrayins'koyi movy (VESUM) yak zasib NLP dlya ukrayins'koyi movy*. Seriya "Ne vse splyva rikoyu chasu...". Vydavnychyj dim Dmytra Buraho. URL https://www.researchgate.net/profile/Vasyl-Starko/publication/344842033_Velikij_elektronnij_slovník_ukrainiskoi_movi_VESUM_ak_zasib_NLP_dla_ukrainiskoi_movi_Galaktika_Slova_Galini_Makarivni_Gnatuk/links/5fa110cd458515b7cfb5cc97/Velikij-elektronnij-slovník-ukrainiskoi-movi-VESUM-ak-zasib-NLP-dla-ukrainiskoi-movi-Galaktika-Slova-Galini-Makarivni-Gnatuk.pdf.
- Starko, V. & Rysin, A. (2023). Creating a POS Gold Standard Corpus of Modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 91–95. URL <https://aclanthology.org/2023.unlp-1.11>.
- Suchomel, V. & Kraus, J. (2021). Website Properties in Relation to the Quality of Text Extracted for Web Corpora. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021*. pp. 167–175. URL <https://nlp.fi.muni.cz/raslan/2021/paper19.pdf>.
- Suchomel, V. & Kraus, J. (2022). Semi-Manual Annotation of Topics and Genres in Web Corpora, The Cheap and Fast Way. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*. pp. 141–148. URL <https://nlp.fi.muni.cz/raslan/2021/paper22.pdf>.
- Suchomel, V. & Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In S.S. Adam Kilgarriff (ed.) *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Lyon, pp. 39–43. URL <http://sigwac.org.uk/raw-attachment/wiki/WAC7/wac7-proc.pdf>.