Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data

Marek Blahuš, Michal Cukr, Ondřej Herman, Miloš Jakubíček, Vojtěch Kovář, Jan Kraus, Marek Medveď, Vlasta Ohlídalová, Vít Suchomel

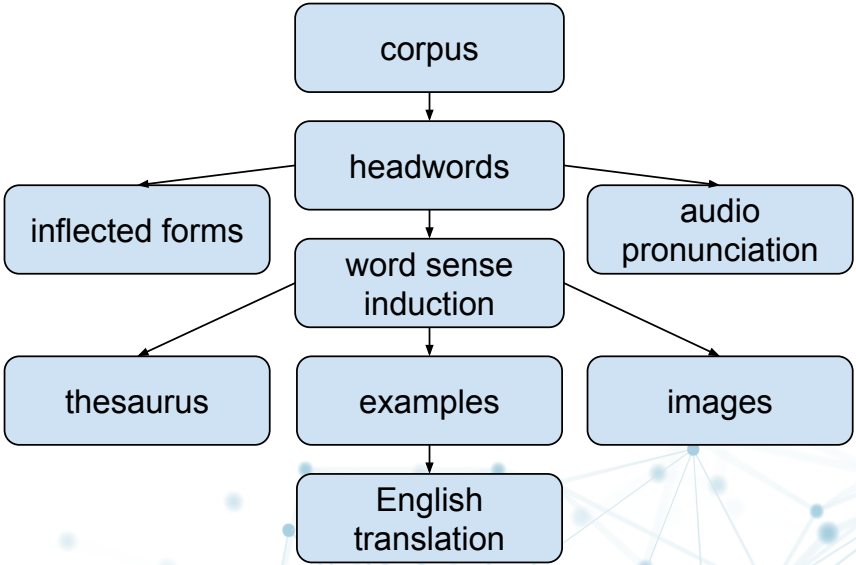
Lexical Computing Masaryk University

eLex 2023

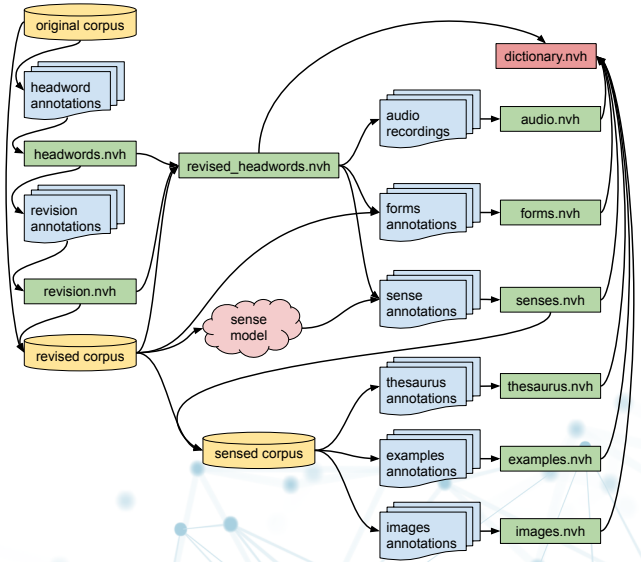
Automating dictionary production

- Parts of the dictionary can be generated automatically
  - collocations, examples, list of word forms, word sense induction, ...
- Let's generate dictionary entries automatically
  - and let native speakers post-edit them
  - „one-click dictionary"
- Problems
  - mistakes propagate to higher levels
  - dictionary entry is a complex structure
- Our solution
  - create entry parts separately
  - ask editors simple questions within intuitive interfaces
  - back-propagate the annotations into the corpus
  - `http://dictionary.express`

Basic workflow



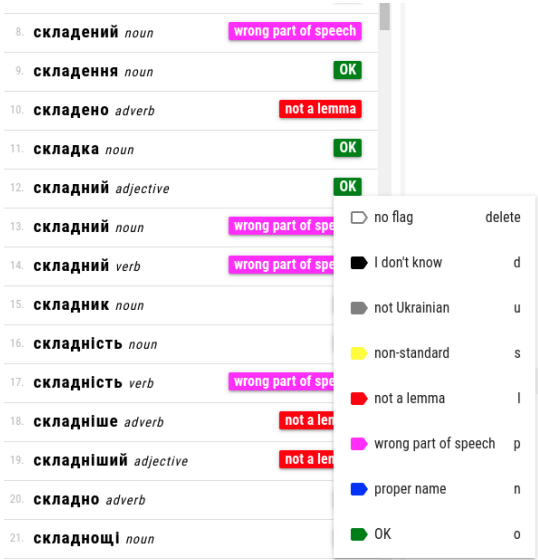
Detailed workflow



New Ukrainian web corpus

- Based on ukTenTen14
  - 52% downloaded in 2014
  - 48% downloaded in 2020
- Tagged by RFTagger
  - trained on Universal Dependencies corpus
  - supplemented by additional morphological database
- Lemmatized using CST lemmatizer
  - trained on Ukrainian Brown dictionary using Affixtrain

Headwords verification



Revision of incorrect headwords

поліклінік PoS: noun

- ● CORRECT HEADWORD SHOULD BE:
  - lemma: поліклініка PoS: noun proper name? ☐ DELETE
  - ADD MORE? +
- ○ 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.

Word forms

держати (verb)

☐ I DON'T KNOW

Inflected forms:

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

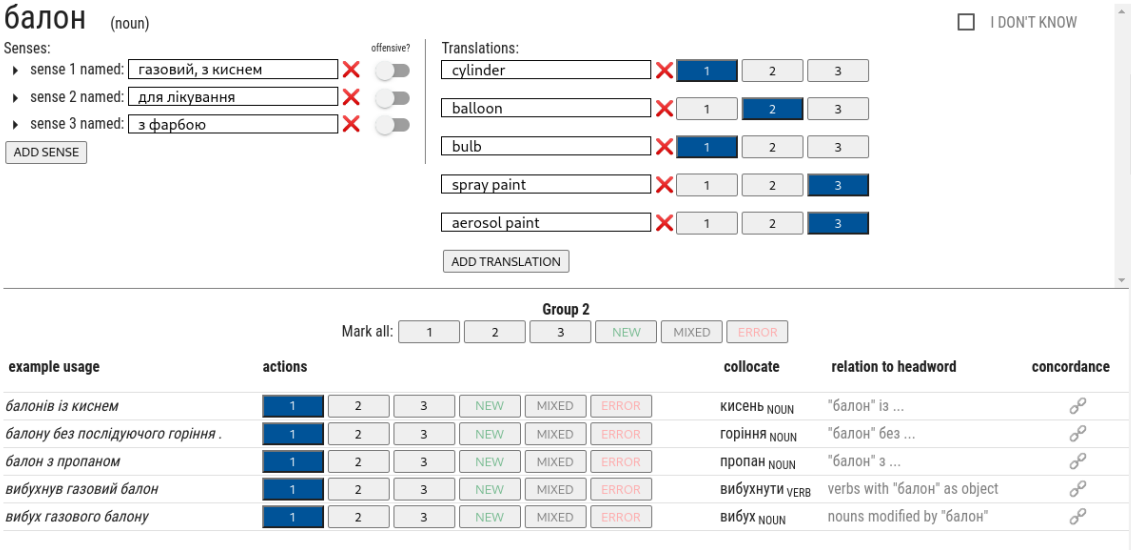
Word senses + translations

Group 1

Mark all: 1 NEW MIXED ERROR

example usage	actions	collocate	relation to headword	concordance
бродіння відбуватися	1 NEW MIXED ERROR	бродіння NOUN	"відбуватися" молочнокисль ...	🔗
відбувається масаж внутрішніх органів	1 NEW MIXED ERROR	орган NOUN	"відбуватися" масаж ...	🔗
заміщення відбуватися	1 NEW MIXED ERROR	заміщення NOUN	"відбуватися" поступовий ...	🔗

Word senses + translations



Thesaurus

океан NOUN

I DON'T KNOW

translations: ocean

thesaurus candidates:

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

Examples

термін NOUN період but not: ~~термінологія~~ ☐ COMPLETELY WRONG

translations: term · deadline · period · date · duration

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

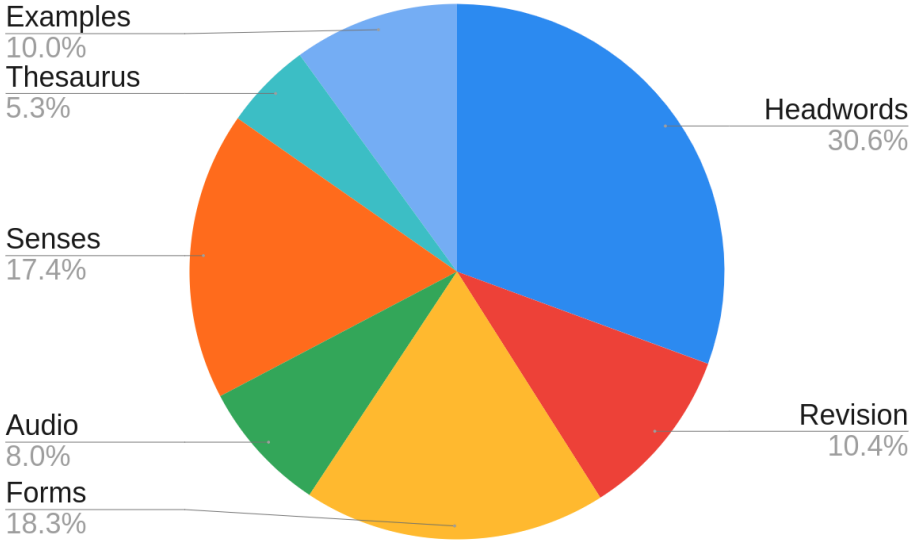
Circumstances

me working on my everyday batch from Kyiv underground hide during air alarm. Station Klovska on Pechersk in Kyiv. On closer photo you can even read: "Lexonomy Ukrainian senses". My daughter is standing near. My cat is sleeping under my coat. 26 Jan 2023, 10:18

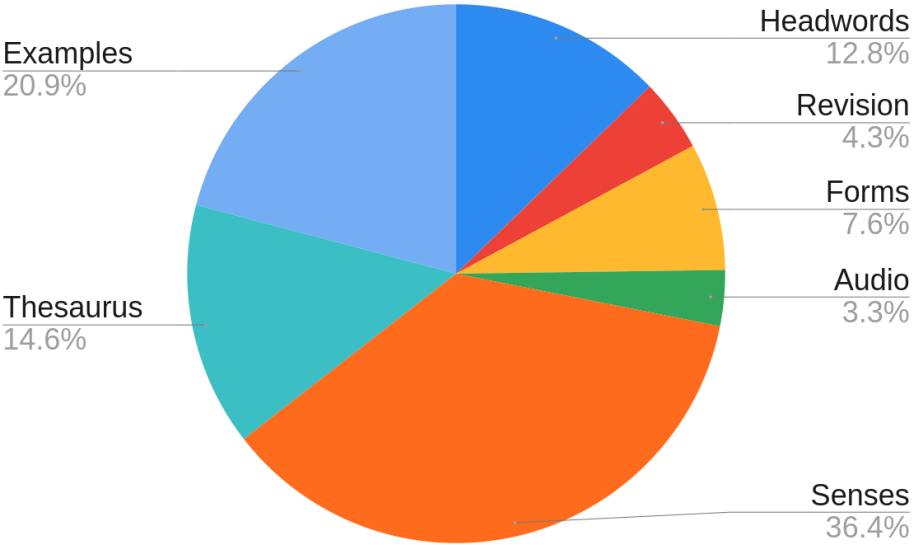
Some statistics

- 109,433 headwords annotated
  - 55,632 included into the dictionary
- 453,010 validated word forms
- 9,785 completed entries, with word senses, thesaurus and examples
  - 17,973 word senses in total
  - 60.1% single sense
  - 18.5% two senses
  - 10.5% three senses
  - 5.0% four senses
  - 4.9 thesaurus candidates on average
- Total annotation time
  - 6,918 hours (3.5 person-years)
- Total number of clicks
  - about 2 million clicks (half of a mouse lifetime)

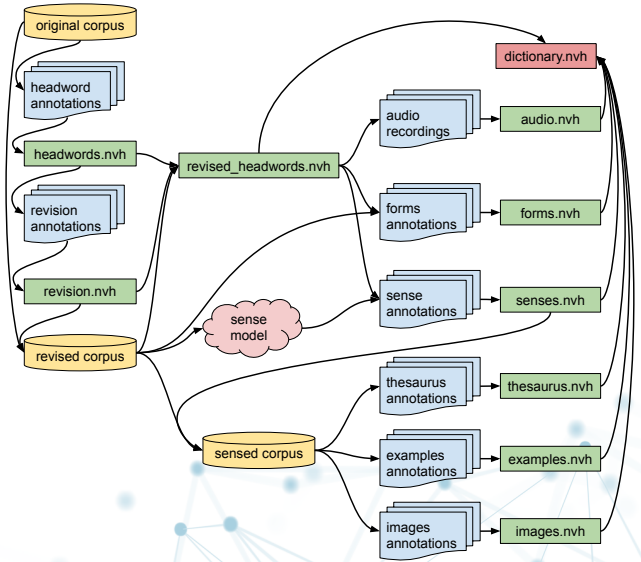
Annotation time



Annotation time normalized



Data management



Example entry

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.

Example entry

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.

Show collocations

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.

Example entry

4 зуб за зуб

In English

a tooth for a tooth

Examples

Це вихід за межі закону помсти, “око за око, і зуб за зуб”.

This is going beyond the law of revenge, "an eye for an eye and a tooth for a tooth."

Show collocations

5 тримати язик за бубами

In English

keep one's mouth shut

Examples

Медичні працівники повинні тримати язик за зубами.

Medical professionals should keep their mouths shut.

Conclusions

- New Ukrainian dictionary
  - 10k full entries, 55k partial entries
  - public sample
    `https://lexicography.sketchengine.eu/ukrainian_sample/dictionary.html`
- New Ukrainian web corpus
  - 2.6 billion words
- Dictionary Express workflow
  - ready to use software
  - straightforward setup for other languages

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`.