General architecture and lexical entry structure of the Polish-Ukrainian electronic dictionary

Natalia Kotsyba

Igor Shevchenko

Plan

- the process of digitalization and further processing of a Polish-Ukrainian electronic dictionary
- its technical and linguistic preparation for future lexicographic works
- post-OCR problems and ways of their automatic correction,
- conversion of the dictionary file into a database
- defining the core set of lexical entries with the help of frequency lists
- lexical entry parsing procedure
- automatic dictionary direction reversal
- further expansion and modification of the bilingual dictionary

From paper to digital version, preparing dictionary background

- joint group of linguists of the Institute of Slavic Studies of the Polish Academy of Sciences and the Ukrainian Linguistic-Informational Foundation of the National Academy of Sciences of Ukraine during 2005–2009
- paper Polish-Ukrainian dictionary in two (three physical) volumes edited by Lukiya Humetska and published in Kyiv in 1958
- 100000 headwords

The paper version

- does not fully reflect the modern state of both languages
- entry list and, sometimes, entry content are considerably outdated
- some domains (computers, finance) are not represented at all
- while others (e.g., agriculture) are described in excessive detail
- too biased ideologically
- nevertheless, it is a good ground for further lexicographic works

Technical editing

- scanned, OCRed with FineReader and saved as MS Word doc format
- OCR mistakes more numerous than in ordinary text due to the bilingual character of the dictionary:
- two different alphabets – Latin and Cyrillic – with several similar-looking letters, cf. “c” and “с”, “k” and “κ”, “p” and “р”, as well as “a, e, i, o, y”, or Cyrillic „т” that looks like Latin „m” (*m*) in italic
- omnipresent stylistic and grammatical mark-up in an abbreviated form that is not found in standard OCR dictionaries
- shortened forms with the common part replaced by the special character ~ (tilde), etc.
- systematic mistakes allowed automatic replacement both in content and formatting

Preliminary edition of the content

- overloaded with Soviet ideology
- contained large number of Russisms
- Polonisms were met more rarely
- obvious translation mistakes were corrected (few)
- now widely used neologisms, e.g. *komputer*, *mysz 2* „computer”, „mouse 2” were added (few)

“Party” words

- *partyjny* (“belonging to the party”), excessive examples of use and the party is understood as the Communist Party of the USSR in all usages:
- *aktyw* ~ партійний актив, -ву (партак-тйв); *grupa ~na* партійна група (партгрупа); *komitet* ~ партійний комітет, -ту (парт-ком, парткомітет); *konferencja ~па* партійна конференція (партконференція); *1 є g i t y m a-c i a ~па* партійний квиток, (партквиток); партійний працівник, -ка (парт-працівнйк); *praca ~na* партійна робота (партробота); *staż* ~ партійний стаж, -жу (партстаж); *szkolą ~na* партійна школа (партшкола); *zebranir л:е парт,пні ..* к>ри, -рів (партзбори); *zjazd* ~ партійний з'їзд, -ду (партз'їзд): (“activists, group, committee, conference, membership card, worker, work, experience, school, meeting, congress”)
- derivation for *partia* (“party”) in its political sense is also overrepresented: *partyjność* (“the state of belonging to the Party”), *POP (Partyjna Organizacja Podstawowa) skr.* первинна партійна організація (“primary party organization”), etc.

“Anti” words

- *przeciwsocjalistyczny* антисоціалістичний (“antisocialistic”)
- *przeciwreligijny* антирелігійний (“antireligious”)
- *przeciwrepublikański* антиреспубліканський (“antirepublican”)
- *przeciwżydowski* антиєврейський (“anti-Jewish”)
- *przedkołchozowy* доколгоспний (“pre-kolkhoz”)
- *okres ~ od socjalizmu do komunizmu* перехідний період від соціалізму до комунізму (“the transferring period from socialism to communism”)
- *~ rewolucji burżuazyj-no-demokratycznej w socjalistyczną* переростання буржуазно-демократичної революції в соціалістичну (“transformation of the bourgeois-democratic revolution into the socialistic”)
- *~dy burżuazyjne* буржуазні передсуди, -дів (“bourgeois prejudicies”)

*Russism → literary_Ukrainian_word (Russian_literary_equivalents) “English_translation”

used not only as translation equivalents but also parts of additional explanations of use

- *жарений *кофе → смажена кава (жарений кофе) “roasted coffee (beans)”
- *нуждатися → мати потребу/потребувати (нуждаться) “have a need”
- *могучість → могутність/міць (могущество) “power”
- *вірьовка → мотузка/шнур (веревка) “rope”
- *лагер → табір (лагерь) “camp”
- міліцейський *участок → дільниця (участок) “police station; lot”
- *похожий → подібний (похожий) “similar”
- *сахарний → цукровий (сахарный) “sugar, adj”
- *жарке → печеня (жаркое) “stowed meat”
- *гравіровка *печатей → гравірування печаток (гравировка печатей) “engraving seals”;
- *покрасити → пофарбувати (покрасить) “paint, v”;
- *командировочні → добові/відрядні (командировочные) “travel allowance”
- *флажок → прапорець (флажок) “flag”
- *пересахарити → перецукрувати (пересахарить) “put too much sugar”
- *прощитатися → прорахуватися (просчитаться) “miscalculate”
- *передаточний → передавальний (передаточный) “transformational”
- *снотворний → снодійний (снотворный) “soporific”; *напиток → напій (напиток) “drink, n”
- *приємного апетиту! → Смачного! (приятного аппетита) “Bon appétit!”
- *італьянське → італійське (итальянское) “Italian”; *ізумруд → смарагд (изумруд) “emerald”
- *шокірувати → шокувати (шокировать) “shock, v”
- *готовитися → готуватися (готовиться) “prepare”

Conversion to a database format

- converted into a database where its structure is reflected in separate tables and their columns and rows
- First, dictionary text was split into entries with the most primitive structure: the headword and the rest
- This format enabled relatively convenient check and further edition of the dictionary, already as a database.
- After the second edition the larger part of the dictionary entry was further parsed and recorded into a more complex database

Defining the core vocabulary

Flexeme (distribution of flexemes in IPI PAN corpus)	Tag	Types
Adjective (starting with lowercase letters only)	adj	7157
Adjective (including those starting with a capital letter)	adj	7283
Adverb	adv	2762
Conjunction	conj	67
Punctuation	interp	43
Predicative	pred	19
Preposition	prep	66
Particle	qub	448
Substantive (including those starting with a capital letter)	subst	19957
Substantive (starting with lowercase letters only)	subst	16798
Verb	verb	12411
Verb (together with gerunds)	verb	12546
Sum (without proper name candidates and gerunds)		39771

Advantages of extracting the lexicon basing on the frequency criterion

- singling out words of low frequency that were included into the original dictionary version;
- receiving a list of words of high frequency that was not included into the original dictionary version
- Inter-POS homonymy was accounted for due to POS limitation of the search, while intra-POS homonymy had to be ignored—the same frequency value was assigned for all homonyms within the same part of speech.
- Polish words that were not found in the IPI PAS corpus at all (or received a minimal frequency rank) but whose Ukrainian equivalents receive high frequency rank in the Ukrainian corpus call for revision as suspects for archaisms (Polish *obuwać*, *obuć*, *rozzuwać się*, *prześpiewanie*, *zakipieć*, etc.)
- There are 21 uses of forms lemmatized *obuć* "put on shoes" in the IPI PAS corpus, 19 of them are participles form *obuty*, still in wide use, and only two are finite past verb forms *obuł*, both from a novel written in 1985. No occurrence of its aspectual counterpart *obuwać* has been found at all.

Automated detection of structural elements boundaries of the dictionary

- Left-hand part
- Headword (bold, new line)
- * opt. homonym ([(I, II, III, IV)], [space])
- * optional (additional forms, e.g., perfect aspect forms of verbs, phonetic variations, etc.)
- grammatical forms ((* opt. [hyphen], [form], [comma]), * opt. hyphen [form], space)
- mark grammatical categories [sort of] for declensions ((italic, [form], * opt. (dot, comma)), italic, [form], * opt. dot)
- tags of style
- tags of topics and terminology
- * opt. valency frame ([(], ((* opt. prepositions), forms) [)], space)
- clarification / definition (__italic__: [(], [content] [)], [space])
- interpretation: the basic form (Cyrillic, * opt. [[(], option ,[)], [space]], END :{[,], [;], [.]}, space)
- * opt. phrases (bold: [1st part], [space], [2nd part] (* opt. [space], [3rd part]) sign [:])
- * opt. verbal form "się" ([;], [space], [/ / ~ się], [space], [right side], [.])

Automated detection of structural elements boundaries of the dictionary

- Right-hand part
- * opt. meaning number (integer, symbol [)], space)
- tag style / theme and terms (italics, * opt. [* opt. (point, point)], [dot] [space])
- * opt. option value ([Cyrillic: (a, b, in)] [)], [space])
- *opt. valency frame ([(], ((* opt. prepositions), forms) [)], space)
- clarification / definition (__italic__: [(], [content] [)], [space])
- interpretation: the basic form (Cyrillic, * opt. [[(], option,[)], [space]], END :{[,], [;], [.]}, space)
- *opt. grammatical forms ((* opt. [hyphen], [form], [comma]), * opt. hyphen [form], comma)
- *opt. collocation examples ([;], * opt .[~], [variable part], [space], * opt. [the rest of the collocation], [space], [construction], {[;], [.]})
- *opt. phraseological ([;], [space], [<*>], [space], * opt. [tag style]) [newline]

Examples of contextual replacements

CONTEXT	REPLACEMENT PATTERN
[new line] [Latin, bold]	[new line] <Рєє> [Latin, bold]
[Latin, bold], *opt.[,] space, [non-bold]	[Latin, bold] </Рєє>, *opt.[,] space, [non-bold]
space, [integer], [closed bracket], space	space, <НЗн> [integer], [closed bracket], </НЗн> space
[Latin, bold], space {[I], [II], [III], [IV]} space	[Latin, bold], space <Ом>{[I], [II], [III], [IV]} </Ом> space

Parsing steps

- **dobry** 1) добрий; **~re** słowo добре (ласкаве) слово; ludzie **~rej** woli люди доброї волі; z **~rej** woli з доброї волі, добровільно; 2) (do czego) підхожий (для чого); **~** do tej roboty підхожий для цієї роботи; 3) (na co) придатний (на що); materia **~ra** na płaszcz матерія придатна на плащ; ◊ *розм.* a to **~re!** от тобі й маєш! от тобі й на! *розм.* **~ra** nasza! наша бере!
- <B>dobry</B> 1) добрий; <B>~</B>re słowo добре (ласкаве) слово; ludzie <B>~rej</B> woli люди доброї волі; z <B>~rej</B> woli з доброї волі, добровільно; 2) (do czego) підхожий (для чого); <B>~</B> do tej roboty підхожий для цієї роботи; 3) (na co) придатний (на що); materia <B>~ra</B> na płaszcz матерія придатна на плащ; ◊ <I>розм.</I> a to <B>~re!</B> от тобі й маєш! от тобі й на! <I>розм.</I> <B>~ra</B> nasza! наша бере!

Explicit marking of the limits of all structural elements of the entry

- <Реє><В>dobry</В></Реє> <НЗн>1)</НЗн> <Екв>добрий</Екв>; <Кол><В>~</В>re słowo</Кол> <Екв>добре (ласкаве) слово</Екв>; <Кол>ludzie <В>~rej</В> woli</Кол> <Екв>люди доброї волі</Екв>; <Кол>z <В>~rej</В> woli</Кол> <Екв>з доброї волі, добровільно</Екв>; <НЗн>2)</НЗн> <ПКер>(do czego) </ПКер> <Екв>підхожий</Екв> (для чого); <Кол><В>~</В> do tej roboty</Кол> <Екв>підхожий для цієї роботи</Екв>; <НЗн>3) </НЗн> <ПКер> (na co) </ПКер> <Екв>придатний</Екв> <УКер> (на що) </УКер>; materia <В>~ra</В> na płaszcz <Екв>матерія придатна на плащ</Екв>; <Фрз> ◊ </Фрз> <ГрП><І>розм.</І></ГрП> <Кол>a to <В>~re!</В></Кол> <Екв>от тобі й маєш! от тобі й на!</Екв> <ГрП><І>розм.</І></ГрП> <Кол><В>~ra</В> nasza!</Кол> <Екв>наша бере!</Екв>

Tree-structured entry record

<Рее><В>dobry</В></Рее>
<НЗн>1)</НЗн>
<Екв>добрий</Екв>;
<Кол><В>~</В>re słowo</Кол>
<Екв>добре (ласкаве) слово</Екв>;
<Кол>ludzie <В>~rej</В> woli</Кол>
<Екв>люди доброї волі</Екв>;
<Кол>z <В>~rej</В> woli</Кол>
<Екв>з доброї волі, добровільно</Екв>;
<НЗн>2)</НЗн>
<ПКер>(do czego) </ПКер>
<Екв>підхожий</Екв> (для чого);
<Кол><В>~</В> do tej roboty</Кол>
<Екв>підхожий для цієї роботи</Екв>;
<НЗн>3) </НЗн>
<ПКер> (na co) </ПКер>
<Екв>придатний</Екв>
<УКер> (на що) </УКер>;
<Кол>materia <В>~ra</В> na płaszcz</Кол>
<Екв>матерія придатна на плащ</Екв>;
<Фрз>◊</Фрз>
<ГрП><І>розм.</І></ГрП>
<Кол>a to <В>~re!</В></Кол>

Generalized structure of the word entry

- Headword
  Homonym number
  - Inflectional elements (can recur)
    - Variants or parallel forms (recurring)
  - Headword variant (phonetic variant or verb aspect counterpart)
    - Inflectional elements (recurring)
    - Variants or parallel forms (recurring)
    - Linguistic characteristics (labels for grammatical categories, style, terminology)
  - Inflectional elements (recurring)
  - Labels of style and/or terminology (recurring)
  - Number of meaning
- Linguistic characteristics (labels of grammatical categories, style, terminology)
- Valency frame (agreement labels)
- *Specification*
- *Word equivalent*
  Inflectional elements (recurring)
- *Specification of meaning*
- *Variants or parallel forms*
- *Inflectional elements (recurring)*
  Specification of meaning
- Collocation (recurring)
- *translation equivalents of the collocation (recurring)*
  - *Grammatical parameters (stylistic labels)*
- Set expression (recurring)
- *Set expression (recurring)*
  - *Grammatical parameters (stylistic labels)*

Structural mark-up inconsistencies found with the help of parsing

- Additional explanatory information added to the description of meanings in brackets, sharing the general function of meaning specification, has to be further divided according to its meaning and function, e.g.

pociągać

- (до відповідальності) lexical combination (in Ukrainian)
- (про хворого) lexical domain coverage
- (збуджувати потяг до себе) synonim or synonimic expression
- (лише докон.) grammatical restriction of the word itself
- (kogo do czego) керування – syntactic valency
- (za sobą) lexical combination (in Polish)

Reversing the language direction in a bilingual dictionary

- **dobry** 1) добрий;
- **dobre** słowo добре слово 1;
- **dobre** słowo ласкаве слово 2;
- ludzie **dob**rej woli люди доброї волі;
- z **dobrej** woli з доброї волі 1;
- z **dobrej** woli добровільно 2;
- **dobry** 2) (do czego) підхожий (для чого);
- **dobry** do tej roboty підхожий для цієї роботи;
- **dobry** 3) (na co) придатний (на що);
- materia **dobra** na płaszcz матерія придатна на плащ;
- ◊ *розм.* a to **dobre!** от тобі й маєш! 1
- ◊ *розм.* a to **dobre!** от тобі й на! 2
- ◊ *розм.* **dobra** nasza! наша бере!

Reversing the language direction in a bilingual dictionary

- добрий; **dobry** 1)
- добре слово 1; **dobre** słowo
- ласкаве слово 2; **dobre** słowo
- люди доброї волі; ludzie **dob**rej woli
- з доброї волі 1; z **dobrej** woli
- добровільно 2; z **dobrej** woli
- підхожий (для чого); **dobry** 2) (do czego)
- підхожий для цієї роботи; **dobry** do tej roboty
- 3) (na co) придатний (на що); **dobry**
- materia **dobra** na płaszcz матерія придатна на плащ;
- ◊ от тобі й маєш! 1 *розм.* a to **dobre!**
- ◊ от тобі й на! 2 *розм.* a to **dobre!**
- ◊ наша бере! *розм.* **dobra** nasza!

Conclusions and future work

- complete the bilingual dictionary with new terminology, e.g., of computer science, business, law, technology
- reverse language direction
- consistency of the grammatical description and presentation of semantic correlation of meanings within lexemes
- extraction of automatic interlingual homonymy, or so-called translator's false friends
- using Polish-Ukrainian corpus (PolUKR) for acquisition of more translation equivalents, either automatically or manually www.corpus.domeczek.pl