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

Natalia Kotsyba[1] and Igor Shevchenko[2]

[1] Institute of Slavic Studies, Polish Academy of Sciences

[2] Ukrainian Linguistic-Informational Centre, National Academy of Sciences of Ukraine

Abstract. The paper describes the process of digitalization and further processing of a Polish-Ukrainian electronic dictionary, its technical and linguistic preparation for future lexicographic works, mainly: 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. The approach presented here aims at producing an updated dictionary as well as a lexicographic editing environment and a tool set for further expansion and modification of the bilingual dictionary.

1 Introduction

Polish-Ukrainian lexicography, both paper and electronic, is represented nowadays by numerous small- or average-size dictionaries created on the basis of earlier paper editions with the addition of the most frequently used, essential new terminology covering the spheres of business, economy and tourism. An extensive review of existing Polish-Ukrainian lexicographic resources with their quality analysis – the macrostructure (choice of entries) and microstructure (entry content and design) – is presented in [1]. During the four years since the appearance of that publication, several new sources that deserve our attention became available. ABBYY Lingvo included a Polish↔Ukrainian dictionary in its version x.3 (2008) [5]. It is based on a modern paper edition and counts ca. 42000 words.[1] Trident Software Electronic Dictionary and Translator [3] includes the Polish↔Ukrainian language pair. Unfortunately no information about the sources and size of the dictionary is provided, and the project is commercial. Considerable progress, as compared to its state in 2005, can be seen in the development of the Multilingual Dictionary by Valentyn Solomko (updated in 2008), which is generated automatically from bitexts [6]. Dictionaries for each language pair in the MS Excel file format are available for download under GNU General Public License. The Polish-Ukrainian file contains 65000 words or word combinations with one-to-one correspondence of translation equivalents. This dictionary can be helpful for machine processing, but it is not particularly human-friendly. Summing up, as far as the size and the quality of entry description is concerned, there is still a need for a large modern electronic and freely available Polish↔Ukrainian dictionary suitable for both public use and linguistic research.

2 From paper to digital version, preparing dictionary background

A large electronic Polish-Ukrainian dictionary was developed by a 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. The basic core of the existing version of the Polish-Ukrainian electronic dictionary comes from the paper Polish-Ukrainian dictionary in two (three physical) volumes edited by Lukiya Humetska and published in Kyiv in 1958. This is the most comprehensive existing bilingual dictionary of very high lexicographic quality for Polish and Ukrainian. It contains about 100000 headwords. Since it was created half a century ago, its entry list and, sometimes, entry content are considerably outdated and do not fully reflect the modern state of both languages. Some domains (computers, finance) are not represented at all, while others (e.g., agriculture)

* The study and preparation of these results have received partial funding from the EC's 7th Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

[1] Information about the size comes from ABBYY developers and concerns the electronic version of the dictionary.

are described in excessive detail. The dictionary is too biased ideologically, which is not surprising taking into the consideration the time and political circumstances of its appearance. Nevertheless, it is a good ground for further lexicographic works.

2.1 Technical editing

The paper dictionary was scanned and processed through the FineReader optical text recognition program in order to receive a text out of the scanned images. The resulting text was saved in the MS Word format. Its quality left much to be desired. The first edition of the dictionary file was the most tedious one and included correction of errors generated by the poor physical quality of the original paper edition and failures of the optical character recognition (OCR) proper. Some mistakes were systematic, which allowed us to apply multiple automatic replacement both in content and formatting. OCR mistakes were more numerous than in ordinary text due to the bilingual character of the dictionary using two different alphabets – Latin and Cyrillic – with several similar-looking letters; 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.

Grammatical and stylistic mark-up is crucial in the digitalizing process as it helps define the structure of the dictionary (see Sections 4 and 6). It is also important to preserve its original formatting (italic or boldface), as it is crucial for successful parsing. It is often impossible to visually determine whether a letter belongs to the Cyrillic or Latin alphabet, cf. “c” and “с”, “k” and “к”, “p” and “р”, as well as “a, e, i, o, y”, or Cyrillic „т” that looks like Latin „m” (*m*) in italic. Therefore, a series of heuristics was used to unify chains of letters delimited by a space to a single alphabet. For one- and two-letter abbreviations, the automatic replacement function of MS Word was used to check the consistency of alphabets and formatting. Some misreadings had a regular character and were corrected automatically as well, either in a supervised (one after another) or unsupervised way (all at once).

Examples of typical automatic substitutions (taking into account adjacent spaces as well):

v) → 1) (number of meaning)

om. → orn. (stylistic label “ornithology”)

Spelling errors were also detected by preparing a frequency list of space-delimited chains and checking the ones that contain up to five symbols and have the lowest frequency.[2] According to Zipf’s law, these are candidates for misspellings. Even though such automatization facilitated the editing work considerably, much labour remained to be done by hand.

2.2 Preliminary edition of the content

While editing the technical side of the dictionary it was impossible to ignore its content either. The two peculiarities of this dictionary are that it was overloaded with Soviet ideology and contained an unforgivable number of Russisms (Polonisms were met more rarely). These were removed from the file and replaced with more neutral and literary correspondents respectively. All the changes were recorded into a separate file. Below are some examples of ideologically biased entries.

“Party” words[3]:

partyjny (“belonging to the party”). It is supplied with excessive examples of use and the party is understood as the Communist Party of the USSR in all usages: *aktyw* ~ партійний актив, -ву (партактйв); *grupa ~na* партійна група (партгрупа); *komitet* ~ партійний комітет, -ту (парт-ком, парткомітет); *konferencja ~na* партійна конференція (партконференція); *1 є g i t y m a-c i a ~na* партійний квиток,

[2] Another option, suggested by Janusz Bień, could be the use of the programme Kolokacje („Collocations”) by Aleksander Buczyński that can help detect unusual word combinations and in this way find words with wrong spelling. We did not experiment with it, though.

[3] We also leave here the original after-OCR format to give the idea what the dictionary text looked like after scanning and text recognition.

(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”).

The 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 антисоціаліСТЙЧНИЙ (“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-демократйчної революції в соціалістичну (“transformation of the bourgeois-democratic revolution into the socialistic”); *~dy burżuazyjne* буржуазні передсуди, -дів (“bourgeois prejudicies”); etc.

Russisms were used not only as translation equivalents, there were many of them in additional explanations of use, etc. Below are examples in the following format: *Russism → literary_Ukrainian_word (Russian_literary_equivalents) “English_translation”.

*нуждатися → мати потребу/потребувати (нуждаться) “have a need”; *могучість → могутність/міць (могущество) “power”; *віръовка → 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”; *скучний → нудний (скучный) “boring”; *плеск → плескіт (плеск) “splashing”; *покрасити → пофарбувати (покрасить) “paint, v”; *командировочні → добові/відрядні (командировочные) “travel allowance”; *полуботинок → півчобіток (полуботинок) “(kind of) shoes”; *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”.

3 Conversion to a database format

Working with the dictionary text in a text editor such as MS Word is very inconvenient, as it is impossible to directly access particular structural units of word entries, and the pace of processing large text files is very slow. This is why the dictionary was converted into a database where its structure is reflected in separate tables and their columns and rows. This was done in several steps. 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 (see Section 6 for details).

4 Automated detection of structural elements boundaries of the dictionary

Information about the entry word limits, defined in the original by bold font and restored in the post-OCR MS Word file, made it possible to mark the border between the headword and its explanation in the database by placing them in separate columns. The borders between lexical entries were marked by line breaks. The grammatical and stylistic information, highlighted by italics within the dictionary entry, was

marked up accordingly but retained in the same column for easier edition before the final, most detailed, parsing.

To mark the boundaries of structural elements in a semi-automatic mode we used a variety of complex context-dependent substitutions which took into account punctuation, the alphabet used (Latin or Cyrillic), text formatting: regular, italic or boldface font, and the content of the word entry. In cases where the context and the printing style were insufficient to clearly identify an element, the correction was made manually.

Upon analysing the word entry structure and formal signs of structural elements, we can see the following general picture:

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], [.])

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]

Here are examples of contextual replacements to identify structural elements of the word entry.

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
</Реє> space, [Latin, italic]	</Реє> space <ГрП> [Latin, italic]
{</Реє>, </Нзн>}, space, [Cyrillic]	{</Реє>,</Нзн>},space,<Екв>[Cyrillic]
</Реє>[,] space [-] [Latin bold]	</Реє> [,] space <ПСз> [-] [Latin bold]
[Cyrillic], space, [-][Cyrillic]	[Cyrillic],</Екв> <Усз>-[Cyrillic]
</НЗн> space, [(][Cyrillic italic]	</НЗн>space,<Уточ>[(][Cyrillic italic]
[Cyrillic italic], [)], space, [Cyrillic]	[Cyrillic italic], [)], space, </Уточ> <Екв> [Cyrillic regular]
</НЗн> space, [(],[Latin italic]	</НЗн>space,<ПКер>,[(],[Latin italic]
space, [див.] space, [Latin bold]	space,<Пос>[див.]</Пос>space,<Адр> [Latin bold]

Tab. 1. Examples of context replacements in the dictionary text for identification of structural elements

During the conversion some data were lost; in cases where entries were split between columns or pages this was systematic, although not too frequent. During the second edition the loose ends were added manually and further errors resulting from oversight during the first edition and parsing errors were corrected.

5 Defining the core vocabulary

Already in this simple format, the dictionary database has more functions than a simple text file, namely, we can work with the entry list of the dictionary. As the actual database resulting from the paper edition appeared too large for experimenting with lexicographic methods and producing preliminary ready-for-use results, it was decided to select a core vocabulary of ca. 30 thousand lexical entries for the pilot version of the dictionary. This selection is also the first part of the dictionary that is intended for public release for use through a web interface. The frequency parameter was chosen as the criterion of selection. A frequency list was generated from the IPI PAS corpus of the Polish language[4] with the help of the program Poliqarp 1.2, which allows for statistic reports on corpora. Since Poliqarp has restrictions on the length of query reports, a query for each part-of-speech (or a flexeme in IPIPAN Corpus tagset presentation) was run, which gave the additional advantage of supplying the frequency list with part-of-speech (POS) information.

In order to avoid proper names, or rather to separate them from common nouns, adjectives and nouns starting with a capital letter were excluded from the search. A typical query looks as follows:
[orth="[qwertyuiopasdfghjklzxcvbnmżźćńłóęąś].*" & pos="subst"] group by base sort by freq count all.

The table below shows the distribution of types generated for a given flexeme.

Flexeme	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

[4] Available at http://korpus.pl.

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

Tab. 2. Distribution of flexeme types

Gerunds, or so-called *-nie* forms, are treated in the IPI PAS corpus in a special way. They are included to both 'verb' and 'noun' categories, and their lemma is identical with the infinitive of the corresponding verb. Polish gerunds are an important part of the vocabulary; they are used more widely than their formal Ukrainian correspondents. However, their formation is not completely regular: they are often homonymous with abstract nouns. Their list was extracted from the corpus on the basis of the ending *nie. This list had to be manually cleaned afterwards.

In general, the procedure of extracting the lexicon basing on the frequency criterion gave us the following advantages: 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. This information gives valuable information for further manipulation with the lexicon. For example, 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. This is the case with Polish *obuwać*, *obuć*[5], *rozzuwać się*, *prześpiewanie*, *zakipieć*, etc.

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.

6 Parsing the lexical entry and recording it in a lexicographic database

The next step of the work is a proper lexical entry parsing that enables creating a lexicographic editing tool. The selection of the structural elements of the dictionary is carried out according to the original lexical entry design. Polygraphic formatting peculiarities can be used for automatic identification of text structure. In order to convert the primitive table into a lexicographic database, special labels are defined to mark the beginning and the end of entries' structural parts. The following formal boundaries of structural elements have been detected from the analysis of text entries.

STRUCTURAL ELEMENTS	LABEL
Polish register unit (word or phrase)	Рєє
Grammatical and semantic properties of a word equivalent	ГрПа
Homonym number	Ом
Meaning number	НЗн
Ukrainian equivalent word	Екв
Polish inflectional element	ПСз

[5] 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.

Ukrainian inflectional element	Усз
Grammatical and semantic properties of a word equivalent	ГЕк
Phrase (collocation)	Кол
Polish prepositional agreement element	Пкер
Ukrainian prepositional agreement element	Укер
Phraseology label	Фрз
Reference label	Пос
Comparison label	Пор
Reference address	Адр
Specification of meaning	Уточ
Additional form (phonetic variant or verb aspect match elements)	Дод

Tab. 3. Structural elements of words, and their labels.

In comparison with monolingual dictionaries, the bilingual dictionary has more a complex and specific structure. The main difference is that the explanatory dictionary in its left-hand part describes formal elements of the lexical unit and in its right-hand part deals with the content, its semantic elements. Therefore the left-hand and right-hand parts of the word entry are clearly separated one from another in (almost) all cases. The bilingual dictionary is characterised by a slightly different situation: the left-hand side of the word entry describes grammatical characteristics and semantic features of the source-language units, while the right-hand one describes the content represented by equivalents of words and phrases in another language (in our case Ukrainian). Moreover, elements of the left-hand and right-hand parts are given in a mixed order, creating a complex, intertwined structure.

6.1 Parsing steps

Let us consider a relatively simple bilingual dictionary entry:

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!

We can see in the entry the Polish headword „dobry”. Its three meanings are rendered by different Ukrainian equivalents: „добрий” („good”), „підхожий” („suitable”), „придатний” („fit”). Further we have Polish phrases (collocations) as examples of word usage, and their Ukrainian equivalents. We can notice Polish words in a truncated form in the entry, where the initial part of the word is marked with a tilde. When used independently (space- or punctuation-separated mode) the tilde indicates the register word as a whole. Besides, in the above example there are tags for prepositional agreement with appropriate values, both of the Polish entry word and its Ukrainian equivalents, phraseological label ◊, stylistic tags like *розм.* and so on.

Having replaced polygraphic formatting marks with explicit labels – HTML tags for boldface and/or italic fonts – we can get the entry to look as shown below. The dictionary text that was marked up in this way became the ground for further automatic entry parsing and additional tagging of the structural elements:

<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!

After the rearrangement of the labels by means of complex contextual replacements we receive the following structural elements in a linear form with explicit marking of the limits (beginning and end) of all structural elements of the entry:

<Реє><B>dobry</B></Реє> <НЗн>1)</НЗн> <Екв>добрий</Екв>; <Кол><B>~</B>re słowo</Кол> <Екв>добре (ласкаве) слово</Екв>; <Кол>ludzie <B>~rej</B> woli</Кол> <Екв>люди доброї волі</Екв>; <Кол>z <B>~rej</B> woli</Кол> <Екв>з доброї волі, добровільно</Екв>; <НЗн>2)</НЗн> <ПКер>(do czego) </ПКер> <Екв>підхожий</Екв> (для чого); <Кол><B>~</B> 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 <B>~ra</B> na płaszcz <Екв>матерія придатна на плащ</Екв>; <Фрз>◊</Фрз> <ГрП><I>розм.</I></ГрП> <Кол>a to <B>~re!</B></Кол> <Екв>от тобі й маєш! от тобі й на!</Екв> <ГрП><I>розм.</I></ГрП> <Кол><B>~ra</B> nasza!</Кол> <Екв>наша бере!</Екв>

The linear format can be further split into a hierarchical tree on the basis of links between entry elements. The figure below shows that the first meaning of the Polish headword corresponds to one Ukrainian equivalent. Additionally, three examples of collocations with the headword are given together with their Ukrainian equivalents. The phraseology zone includes two Polish phrases marked as colloquialisms, the former corresponding to two Ukrainian equivalents, and the latter only to one.

6.2 Tree-structured entry record

<<Реє><B>dobry</B></Реє>
<НЗн>1)</НЗн>
<Екв>добрий</Екв>;
<Кол><B>~</B>re słowo</Кол>
<Екв>добре (ласкаве) слово</Екв>;
<Кол>ludzie <B>~rej</B> woli</Кол>
<Екв>люди доброї волі</Екв>;
<Кол>z <B>~rej</B> woli</Кол>
<Екв>з доброї волі, добровільно</Екв>;
<НЗн>2)</НЗн>
<ПКер>(do czego) </ПКер>
<Екв>підхожий</Екв> (для чого);
<Кол><B>~</B> 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 <B>~ra</B> na płaszcz</Кол>
<Екв>матерія придатна на плащ</Екв>;
<Фрз>◊</Фрз>
<ГрП><I>розм.</I></ГрП>
<Кол>a to <B>~re!</B></Кол>
<Екв>от тобі й маєш! от тобі й на!</Екв>

<ГрП><I>розм.</I></ГрП>

<Кол><B>~ra</B> nasza!</Кол>

<Екв>наша бере!</Екв>

Fig. 1. The entry „dobry” as a tree structure.

Another example of a word entry with more structural elements:

ale 1) але; та *(рідше)*; 2) *(після заперечної частини речення)* a; nie tutaj, **~** tam не тут, а там; ◊ **~**i tak певна річ, звичайно; *прик.* nikt nie jest bez **~** немає людини без вади.

We can see here, *inter alia*, a clarification of the meaning, in this case through providing the context of usage: *після заперечної частини речення* “after the negative part of a sentence”; additional information about the frequency of use for one of the equivalents: *рідше* “more rarely”; not fully synonymous equivalents separated with a semicolon; a mark indicating a set expression, *прик.* “saying”.

Upon the replacement of the formatting tags with explicit labels this entry looks as follows:

<B>ale</B> 1) але; та <I>(рідше)</I>; 2) <I>(після заперечної частини речення)</I> a; nie tutaj, <B>~</B> tam не тут, а там; ◊ <B>~</B>i tak певна річ, звичайно; <I>прик.</I> nikt nie jest bez <B>~</B> немає людини без вади.

Upon contextual replacements inserting structural labels:

<Реє><B>ale</B></Реє> <НЗн>1)</НЗн> <Екв>але; та</Екв> <I>(рідше)</I>; <НЗн>2)</НЗн> <Уточ><I>(після заперечної частини речення)</I></Уточ> <Екв>a</Екв>; <Кол>nie tutaj, <B>~</B> tam</Кол> не тут, а там; <Фрз>◊</Фрз> <B>~</B>i tak <Екв>певна річ, звичайно</Екв>; <Прк><I>прик.</I></Прк> <Кол>nikt nie jest bez <B>~</B></Кол> <Екв>немає людини без вади</Екв>.

6.3 Generalized structure of the word entry

Thus, a generalized structure of the word entry for the Polish-Ukrainian dictionary can be presented with certain simplification in the following way. Elements of the right-hand side of the dictionary, i.e. Ukrainian equivalents with their appropriate labels, are in italics.

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

Phrase (recurring)

Phrase equivalents(recurring)

Grammatical parameters (stylistic labels)

Set expression (recurring

Set expression (recurring)

Grammatical parameters (stylistic labels)

Verbal forms with reflexive **się**

Inflectional elements (recurring)

Variants or parallel forms (recurring)

Headword variant (phonetic variant or verb aspect counterpart)

Inflectional elements (recurring)

Variants or parallel forms (recurring)

Linguistic characteristics (labels of grammatical categories, style, terminology)

Inflectional elements (recurring)

Labels of style, terminology (recurring)

Number of meaning

Linguistic characteristics (labels of grammatical categories, style, terminology)

Valency frame (labels)

Specification

Inflectional elements (recurring)

Variants or parallel forms

Inflectional elements

Set expression (recurring)

Phrase equivalents(recurring)

Linguistic parameters (stylistic labels)

Set expression (recurring)

Set expression (recurring)

Linguistic parameters (stylistic labels)

Fig. 2. Generalized tree structure of the word entry in the Polish-Ukrainian dictionary.

7 Reversing the language direction in a bilingual dictionary

It is desirable in a bilingual lexicographic system to be able to access this system not only through the source-language entry list (the left-hand part of a bilingual dictionary) but from the target-language units (the right-hand part) as well. Thus, the reversal of the bilingual dictionary so that the left-hand and the right-hand parts of the entries change places becomes another important task. The objective actually is to

transform the $Language_1 \rightarrow Language_2$ dictionary into a $Language_2 \rightarrow Language_1$one. This task is far from being trivial because, as we can see, the information about the correspondence between words and word combinations of the two languages is recorded according to lexicographical tradition in a laconic, compressed form, most economic and convenient for the user. This problem is solved through "unfolding" the word entry into a set of basic equivalents, i.e., separating rows of original words or phrases and their respective equivalents in the other language, along with the corresponding grammatical, stylistic and thematic information.

The conversion of a word entry of the initial dictionary into a set of elementary equivalents requires several operations. First of all, abbreviated words with tildes are to be replaced with their full versions, i.e., „~ra", „~re", „~rej" are restored to „dobra", „dobre", „dobrej". This is done automatically by searching the first letter (after the tilde) of the shortened word in the full-form word; the search is carried out from right to left. The part from the entry word on the left of this letter gives us the string to be inserted instead of the tilde. The next step is to detect the limits of the equivalents together with their source-language counterparts. The boundary is defined due to obligatory occurrence of the equivalent expression from the target language after any source-language word or phrase. One word is often translated as several words and/or phrases. Equivalents are often presented by short synonymic rows, where synonyms are separated by commas. A comma inside an equivalent expression often, although not always, means a limit between synonymous equivalents. Therefore, it can be used for dividing an entry into basic sets of equivalents automatically.

Here is a fragment of our sample entry **dobry**:

z **~rej** woli з доброї волі, добровільно;

with the first step it turns into the line:

z **dobrej** woli з доброї волі, добровільно,

with the second step the line is split into two more basic sets of equivalents:

z **dobrej** woli з доброї волі 1;

z **dobrej** woli добровільно 2.

The equivalent rank, taken from the order of the equivalent expression in the entry, is assigned automatically. It usually indicates a kind of priority, a higher frequency or higher standard of the translation equivalent of the entry in question. This information can be useful for further stages of work with the reverse dictionary. In our example we receive information about the priority of the translation equivalent „з доброї волі" (lit. „of one's free will") for the Polish phrase „z dobrej woli", although in general another translation equivalent, „добровільно" „voluntarily", is equally common.

Sometimes a comma inside the equivalent zone is not a sign to separate two different (synonymous) values, but is a part of an equivalent phrase, as in:

~ (ten), który to powiedział „той, який (що) це сказав"

In this case, „той, який (що) це сказав" (lit. „the person who (that) said this") is an integral equivalent. At the same time, brackets are another indicator of variability of the translation equivalent and point to a compressed translation. Thus, we have two elementary equivalents here:

ten, który to powiedział „той, який це сказав"

ten, który to powiedział „той, що це сказав"

Apart from a pair of equivalent words or phrases with the same meaning, an elementary equivalent set, as we define it, should also include various labels available for this pair. For this particular dictionary these are: grammatical category, peculiarities of morphological forms, stylistic and terminological tags, as well as an extended valency frame that also includes information about prepositional agreement. Although prepositional agreement is also a kind of valency information, a significant difference in rendering information about proper valency frames is that the former ones are given in italics, and the latter ones in

regular type and, normally, in brackets. Clearly all phraseology, proverbs, etc., found in the original dictionary, preserve their status in the reverse dictionary as well.

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!

The next step is to swap the elementary equivalents, which is a trivial operation of replacement of the left-hand side of the line with the respective right-hand side:

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!

However, the result of this reversing operation for basic equivalents is still quite distant from a genuine reverse bilingual dictionary formed according to lexicographic rules. This is why the further stage of work requires a number of compression operations, folding the entry back into a different combination of units. First, a list of words and word combinations available in the initial dictionary $Language_1 \rightarrow Language_2$ in the alphabetic order of $Language_2$ is created. In our case, basic equivalents extracted from the dictionary become the basis for the Ukrainian word list. The next step is the formation of word entries of the reverse dictionary. The equivalents extracted from the „dobry" entry, will appear in the entries containing relevant Ukrainian equivalent expressions: „добрий" („good"), „ласкавий" („kind"), „добровільно" („voluntarily"), „підхожий" („suitable"), „придатний" („fit"), „мати" („have"), „на" („on"), „брати" („take") and others. Clearly equivalents, for example for „мати", used either as a frequent functional verb or a noun ("have" or "mother"), will be gathered from various Polish headwords. To receive the basic (so-called dictionary) forms of words, the lemmatization procedure will obviously have to be used. The Ukrainian Grammatical Dictionary together with its supporting software developed at the ULIF NASU

can serve for this purpose. Besides, it should be noted that main words of collocations should be determined during the compilation. These words will be the input to collocations in the reverse dictionary. If this choice is made and a system of grammatical identification of lexical units (lemmatization and paradigmatization) is available, the further creation of the inverse dictionary can be carried out automatically. Of course, some post-processing manual check and edition will be necessary anyway.

8 Database and an editing tool

After all basic cleaning and parsing stages the dictionary database is ready for further lexicographic work. A special editing environment is highly desirable for the more convenient work of the lexicographers, enabling them to introduce systematic changes into the dictionary. The lexicographical database of the explanatory dictionary of the Ukrainian language ("Словник української мови") developed at the ULIF NASU can be used as a model. In particular this system allows the user to view entries, directly access individual structural elements, as well as modify entries, replace elements, change the sequence of homogeneous structural elements, remove entries and add new ones to the dictionary. Thus, the lexicographical system is both a reference system for the user (an electronic dictionary) and an operating tool for lexicographers who compile or edit a dictionary. It should be noted that the structure of a bilingual dictionary differs significantly from a monolingual explanatory one, which turns the creation of a bilingual lexicographical database into a special independent task for which new solutions have to be found. An essential property of bilingual lexicographic systems is enabling users to enter the dictionary through either of the two languages' word list, which requires a reverse dictionary creation technology.

The approach presented here can produce an updated dictionary, as well as a lexicographic system as a computer tool set for further expansion and modification of the bilingual dictionary.

9 Future work

Lemmatization and paradigmatization allows us to conduct further interesting experiments. The word list of the Ukrainian part of the dictionary, with a frequency index, can be mapped against the word list of the explanatory Ukrainian dictionary. This can help us detect more outdated words, Russisms and Polonisms in an automatic way. It would also be interesting to see whether there are words of high frequency in the explanatory dictionary that are not used in the bilingual one and analyse this group.

On the other hand, we need to complete the bilingual dictionary with new terminology, e.g., of computer science, business, law, technology. Preliminary word lists for these fields to work with have already been extracted from the explanatory dictionary. Since bilingual terminology is usually presented by one-to-one correspondents, and our system allows for the reverse language direction to work with lexical entries, the source language of terms is no longer so important. Further work on existing lexical entries from the point of view of consistency of the grammatical description and presentation of semantic correlation of meanings within lexemes must be done as well.

Another practical task, important for language didactics, is extraction of automatic interlingual homonymy, or so-called translator's false friends.

We also plan to use Polish-Ukrainian corpus (PolUKR)[6] for acquisition of more translation equivalents, either automatically or manually.

Bibliography

[1] Kotsyba, Natalia and Magdalena Turska (2006). Polsko-ukraińska leksykografia – współczesny stan i perspektywy. In *Semantyka i konfrontacja językowa*, t. 3, red. V. Koseskiej-Toszewej, SOW,

[6] www.corpus.domeczek.pl

Warszawa.

[2] Słownik polsko-ukraiński we dwóch tomach (1958). Kolegium redakcyjne: A. I. Gęsiorski. T. Ł. Humecka (redaktor naczelny), M. Kiernycki, M.J. Onyszkiewicz, M.I. Rudnycki. Kijów.

[3] Trident Software. Polish-Ukrainian electronic dictionary and translator. http://www.slownik.ukraincow.net/

[4] Turska, Magdalena and Natalia Kotsyba (2007). Polish-Ukrainian Parallel Corpus and its Possible Applications. In *Proceedings of the International Conference 'Practical Applications in Language and Computers', 7–9 April 2005, Łódź*, Peter Lang GmbH.

[5] Universal (Pl-Ua) within ABBYY Lingvo x3 version (2008). Electronic version is based on: Polish-Ukrainian and Ukrainian-Polish dictionary by Anna Malecka and Zbigniew Landowski, edited by Vyacheslav Busel, ITF "Perun", 2007.

[6] Соломко, Валентин. Багатомовний електронний словник: http://slovnyk.org

[7] Шевченко И.В., Широков В.А., Рабулець А.Г. (2005). Электронный грамматический словарь украинского языка. // Труды международной конференции „Megaling'2005. Прикладная лингвистика в поиске новых путей". 27 июня–2 июля 2005 года. Меганом, Крым, Украина., p. 124–129.

[8] Широков В.А. (2005). Елементи лексикографії. Київ: Довіра.

[9] Широков В.А., Рабулець О.Г., Шевченко І.В., Костишин О.М., Якименко К.М. (2007). Інтегрована лексикографічна система „Словники України", версія 3.1. Київ. CD-видання.