Applying Oxford-PWN English-Polish Dictionary to Machine Translation

Jassem Krzysztof

Adam Mickiewicz University, Faculty of Mathematics and Computer Science, Poznań
[jassem@amu.edu.pl]

The paper reports on the process of converting a large human-readable English-Polish dictionary further abbreviated as OPEP (Oxford-PWN English-Polish), to the format applicable for machine translation. The report concludes with an evaluation of the *Translatica* system which uses the converted data in transfer-based translation.

1. Translatica system

The *Translatica* system originates from the *PolEng* project developed in 1996-2002 at Adam Mickiewicz University (AMU) in Poznań. *PolEng* developed a transfer-based system translating texts from Polish into English, with the lexicon based on Internet texts. *Translatica* expands *PolEng* capabilities by translation in the reverse direction and the usage of broader lexicon obtained from the contents of OPEP.

The main features of *PolEng* inherited by *Translatica* are:

- bottom-up parsing based on the CYK algorithm
- phrasal structure representation
- Perl-like formalism of transfer and synthesis rules.

It is worth noting that the same formalism for description is used for both directions

The formalism for the description of grammars is a kind of CFG. The authors have tried to use available resources to describe the English grammar, e.g. AGFL (2002)but it turned out that they would hardly comply with the elaborated translation engine (see Graliński (2002) for details on the engine). In order to use the engine it was necessary for us to compose grammar rules ourselves.

According to the agreement between the authors of *PolEng* and the authors of OPEP, the lexicon for the English-to-Polish direction should be based on the OPEP contents. The paper reports the work that was done in 2003 in order to adopt OPEP contents to *Translatica* – the *PolEng* successor.

2. Building a lexicon for an MT system with the Polish language

Two approaches for building an MT lexicon are mainly discussed in the literature. In Pinkham and Smets (2002) the authors distinguish between "HanC systems" – based on "Hand-crafted Dictionary" and the "Lead systems" – based on "Learned Dictionary". Systems of the first type use traditional bilingual dictionaries to determine the transfer between word senses, whereas in "Lead systems" the transfer part of the dictionary is trained on bilingual text corpora. The authors of the publication show the advantages of the "Lead approach" – particularly for new pairs of languages and limited time for development.

By contrast, the experiments of Ilaraza, Mayor, and Sarasola (2001) demonstrate the advantages of building an MT lexicon on the basis of large traditional dictionaries. The authors compare translation produced by the system that uses raw bilingual dictionary to tha given by the system whose dictionary is a merge of a Basque lexical database and the Morris bilingual English-Basque dictionary. The authors state that the output of the latter system is distinctly better.

Another dichotomy is mentioned by Baldwin, Hutchinson and Bond (1999). On the basis of English-Japanese translation the authors compare the systems that store entries as source/target language pairs to those which consider both languages separately. In the source/target approach "a word has as many senses as it has translation equivalents". In the latter approach sense distinctions are specific for each language, which in the authors' opinion is "more cognitively justifiable". The main argument for the "monolingual" approach in MT is that the decision on the sense disambiguation may be postponed until the process of generation whereas in the "bilingual approach" the semantic constraints on the source side are used for disambiguation in both source analysis and transfer. Another argument against "bilingual" dictionaries is their uni-directionality.

One of the main criteria for choosing the type of lexicon for an MT application is the availability of resources. Because of the scarcity of aligned Polish-English bi-texts the "Lead approach" has lost its main benefit for our purposes: rapid deployment – in order to use the approach it would be necessary to build aligned corpora (the same reason has determined the choice of the transfer method used for the translation, rather than the corpus-based one). On the other hand, our group have had to free disposal quite large traditional dictionaries: OPEP English-Polish dictionary and a few Polish dictionaries mentioned in section 4. The situation called for the "HanC approach".

The decision whether the entries should be bilingual or monolingual was to large extend determined by the conditions of the agreement between the PolEng group and PWN. The dictionary publishers (as well as the authors of the system) liked the system to use the linguistic knowledge included in the dictionary material to the maximum extend. The system dictionary should mirror the OPEP material as closely as possible. This called for the bilingual description. The uni-directionality was not a counterargument either as the Polish-to-English part had already been developed.

3. Main goals

The main goals posed to the conversion process were:

- to lose as little information as possible from OPEP
- to extract and formalize syntactic and semantic information given in OPEP
- to supplement data with all information necessary for transfer-based machine translation – in accordance with the *Translatica* algorithm.

4. Resources

Before the start of the work, the group consolidated the following resources:

1. Lexical resources at free disposal:
   - OPEP in the electronic form, XML format
   - PolEng lexicon (Polish-to-English)
   - Polish lexicon of inflected forms delivered by PWN – the lexicon developed by the organization of Polish Scrabble players, further referred to as the *scrabble dictionary*
   - other dictionaries of Polish published by PWN, e.g. Bańko (2000)
   - lists of entries (e.g. proper nouns) from PWN encyclopedias.

The PolEng lexicon comprises some syntactical information that could prove useful for the other direction; the scrabble dictionary handles Polish inflection exhaustively; Bańko (2000) includes some information on syntactical features of entries, in the form well suited for computer processing.

2. Lexical resources at limited disposal (via Internet), e.g.
   - Meriam-Webster Dictionary (www.m-w.com)
   - Internet English-Polish dictionary (www.dict.pl)

These and other dictionaries available on-line helped lexicographers understand the meaning of some entries or suggest alternative equivalents

3. Text corpora concordancers:
   - British National Corpus (http://sara.natcorp.ox.ac.uk)
   - Collins Cobuild Corpus (http://www.cobuild.collins.co.uk/form.html)
   - WordCorp (http://www.webcorp.org.uk/index.html)
   - PolEng Internet corpus – the corpus of Polish Internet texts collected while working with the Polish-English translation.

Consulting such tools helped to verify syntactic features of words – such information is not given exhaustively in OPEP.

WordNet has proved to play a key role in assigning semantic values. The semantic hierarchy used in *Translatica* is a subtree of the *WordNet* lattice.

4. Translation tools:
   - grammar description
   - syntactic-semantic parser
   - tools for transfer and synthesis.

Translation tools impose specific constraints on the type of information that should be stored in the dictionary.

5. *Translatica* dictionary formalism

The dictionary formalism of the *Translatica* system assumes one lexicon for each direction (source/target approach). For example, the description of an entry in the English-to-Polish direction gives the constraints under which a word (or a phrase) is translated into appropriate equivalents.

5. Basic problem – time limitations

As is shown in section 6, full and detailed conversion of a single entry consumes a lot of man-work (apart from computer work). The group could afford 27 man-months to accomplish the task of conversion (3 lexicographers, 9 months). The available time was not sufficient to manually elaborate each entry of OPEP (even after automatic pre-processing). The group assumed the following approach:

- function words should be described almost from scratch
- out of 55 900 entries in OPEP, ca 20 000 most frequent ones (according to BNC) should be conversed automatically and then elaborated manually
- the rest of the dictionary should be conversed only automatically
- errors resulting from manual and automatic conversion should be corrected semi-automatically.

6. Processing of the dictionary

The process of dictionary conversion involved the following stages:

- automatic conversion of dictionary information
- automatic morphological description
- manual description/verification of 20 000 most frequent lexemes
- semi-automatic correction of errors
- manual correction of errors found while testing the translation system.

6.1. Automatic conversion of dictionary information

In Mayfield and McNamee (2002) the authors present an interesting idea that aims at simplifying the conversion of bilingual dictionaries from human-readable to computer-readable form. They have created a language called ABET (APL Bidict Extraction Tool) that allows for automating "the processes that are the same across most extraction tasks". At the time the paper was written the authors had converted 50 MB of on-line "bidicts" of varying formats and the longest ABET script they needed consisted of mere twenty-four lines.

Although we think that a language like ABET may prove beneficial in dealing with more than one dictionary of a simple format we do not think that the idea would work for traditional off-line dictionaries that include deep linguistic knowledge, rarely given in a systematic way. In our attempt to convert the dictionary we have come across so many specific "little problems" that it is hard to imagine for us that any generalizing language could be of much help.

The script used for the conversion job was written in Perl. The section lists the major problems in the conversion (and does not mention those "little nuisances" which are particularly not amenable to description in a language of a higher level).

The task consisted in the following steps:

Separating entries

In the approach suggested in Mayfield and McNamee (2002) the macro HEADER-FIND is responsible for separating entries. The macro identifies headers according to the specification of the dictionary. Such approach would not solve the problems that we encountered in separating entries in OPEP:

- Some entries have references to other entries. The entry *upon* is described in OPEP only with the reference to the entry *on* (i.e. *upon= on*). In such a situation the reference was forwarded (the idea of reference is used in the *Translatica* dicitionary as well). Quite a few entries have references only in one of the senses. An example may be *brainstorm*, whose informal meaning is described as equal to *brainwave* (whereas other senses are described separately). Such situation requires different treatment ("copy and paste inside"). Another type of reference concerns entries which are word forms of other entries. The entry *bidden* has a reference to *bid* as its whole description, whereas the entry *built* has a reference (to *build*) in one sense as well as its own set of equivalents for other senses. Our decision was to discard the entries like *bidden* from the dictionary and discard only referenced senses in the entries like *built*.

- Entries may have graphical variants. We decided to separate such variants into multiple entries (tha other variants received references to the first variant). There are two ways of denoting graphical variants in OPEP: one with the usage of braces inside words, e.g. *bias(s)ed*, the other by means of a comma, e.g. *baldachin, baldaquin* (the comma is not used consistently, however, e.g. in the entry *births, marriages and deaths* the comma obviously does not separates variants – to solve that disambiguation automatically we assumed that variants of the same word must share the first letter). On the other hand graphical variants that differ only with the first letter capitalization (e.g. *balkanization, Balkanization*) should be merged to one entry only. The entry *backwards* has a "graphical alternative" *backward* but *backward* itself constitutes a separate entry in OPEC (giving a reference to *backwards*!). In cases like this the entries should not be duplicated during conversion
- Some words are indexed and treated as separated entries in OPEP (e.g. *billet* as *billet*^{1} /*an order*/ and *billet*^{2} /*of wood*/) although they are the same parts of speech. We merge such entries into one (with separate equivalents).
- There are entries that lack direct equivalents (e.g. *behalf*) – the equivalents are given only for phrases that include them. Since we demand each entry to have at least one non-empty equivalent such entries must be automatically tagged and individually decided: either the entry should be removed (only lexical phrases should remain) or an artificial equivalent should be found for the entry.
- Phrasal verbs need special treatment. We decided to separate entries that are described in OPEP as phrasal verbs from simple verbs. This approach requires considerable caution. For example, the phrasal verb *get out of* is in OPEP described as a separate phrasal, but some important uses of the multiword *get out of* are mentioned also in the description of the phrasal verb *get out*, and in the description of the simple verb *get*. To make things worse, these uses partially overlap.

As it will be seen in the sections to follow, the process of separating entries takes place also in the next phases of the conversion.

Automatic acquisition of attribute values

This phase consists in extracting from OPEP the values of attributes needed by the *Translatica* translation algorithm.

- The automatic procedure converts OPEP flexional description into *Translatica* encoding. Some verbs in English have different inflected forms for different senses (e.g *speed, sped, sped* as contrasted to *speed, speeded, speeded*). In such a case we decided to divide an entry into two.
- The syntactical information on the complements should be concluded from various parts of the OPEP description. Transitive verbs (decoded as *vt* in OPEP) are treated in our approach as having an equivalent with a noun complement in accusative in Polish. The complementation may partially be derived from `PREP` elements which in the OPEP XML describes prepositional complements. Direct complements are listed in OPEP as `OBJ` elements for verbs and `INDIC` elements for nouns. Complementation may and should also be concluded from human-readable descriptions, e.g. from strings like *on or about smth*.
- Assigning semantic attributes from various tags in OPEP has been partially done automatically. For example the values given in the element *`COLL`*, which describes the sense, (e.g. for the word *speech*, the senses are: *oration, faculty, language, subject*) are automatically converted into *Translatica* semantic hierarchy by means of a WordNet query.
- The attribute of *context*, which comprises "domain", "style" and "dialect" in *Translatica* should be concluded from various tags in OPEP, mainly from qualifiers..

Merging senses

This phase aims at diminishing numbers of equivalents for entries in order to simplify computations in the translation process.

- Some "second-best" equivalents are replaced: the sense they represent is discarded and the equivalent is inserted as a synonym of a "better" equivalent. A sense is converted into a synonym only if strict conditions are fulfilled: all attributes must have same value (this automatic procedure may be undone during manual verification).
- It often happens that different senses have the same equivalents. For example, all

first five senses the verb *get* in OPEP include the Polish equivalent *dostać*. In such a case we would like to merge the senses into one (according to the paradigm that in *Translatica* dictionary a word has as many senses as it has equivalents). In order to make the merging it was necessary to cautiously process the values of attributes, e.g. the value of complementation attribute for the merged entry should include complementation patterns of all merged senses.

Automatic conversion of lexical phrases

For most word-senses the OPEP dictionary gives examples of usage as well as idioms that include the word-sense. Both types of phrases were automatically copied into the list of idioms in *Translatica* database. It was up to the lexicogrpahers to distingush between the two types and make appropriate decisions. Idioms in the *Translatica* dictionary are described with the same set of attributes as single words. Some syntactical and semantic values could be obtained automatically from OPEP in the same way as for single words (e.g. the human semantic value for an object denoted as *sb*). Still, there was more to do for lexicographers with idioms than with single words.

Some idioms and their equivalents are denoted in OPEP by slash marks (e.g. *a banking/an educational ~ system bankowy/ edukacji*). Such idioms needed to be separated automatically.

Cleaning up and adjusting

OPEP uses its own metalanguage which should be parsed during conversion. For example, the word *or* serves to indicate either alternative uses, alternative translations, or alternative complements. We have decided to treat *or* in the following way in the conversion process: alternative uses should be divided into separate idioms, from the alternative equivalents all except the first one should be discarded, alternative complements should be merged. Similar treatment is used for the metaword *also*. The word *beacon*, in its third sense, has the following form in OPEP: *(also: radio ~)*. This should be conversed into a new entry: *radio beacon*.

The usage of words in braces should be disambiguated in this phase also. Usually the braces denote optional occurrence, like in the entry *bulgur (wheat)*. If the braces appear at the source side, two entries are generated (e.g. *bulgur* and *bulgur wheat*); if they appear in the target side, the second alternative is omitted.

6.2. Automatic morphological description

The morphological information for English words is given in OPEP. The main resources that contributed to determining the Polish inflection were the *PolEng* Polish-to-English dictionary and the PWN scrabble lexicon of inflected forms.

6.3. Manual verification/modification of data

The data produced by automatic conversion needed manual verification and modification. Verification showed errors in the conversion process – which could be corrected by mere adjusting conversion procedures. Some linguistic aspects required human linguistic knowledge as well as consulting other sources than OPEP.

An important factor was the labor organization. As the lexicographic group consisted of three (occasionally four) lexicographers who were controlled by a co-ordinator it was necessary to find the way in which linguistic data could be modified more than once and the labor could be distributed between persons. We decided to convert the data into a classical SQL database. This idea made it possible to keep the lexical database in order even if some of the work overlapped.

The main aspect (in view of transfer translation) which is not well handled by automatic conversion is complementation. OPEP delivers almost none explicit information on left-hand complements (specifiers) like prepositions (and their translations) that tend to precede specific nouns. This information should have been encoded manually on the basis of other resources (e.g. *PolEng* dictionary). The right-hand complements are treated in OPEP more exhaustively but many of them are given implicitly in examples of usage, e.g. *we'll never get by without him* is an example given in OPEP that should, for the translation needs, be treated as a phrasal verb *get by* complemented by a PP *without sb*.

The lexicographers were asked to query English corpora in order to check for complements that are not listed in OPEP. This was a hard task that required good knowledge of English and could not be executed (semi)automatically.

Word senses in the *Translatica* dictionary are described semantically by means of a subtree from the WordNet semantic hierarchy. OPEP delivers some hints on semantic values of words and semantic values of their complements but it was up to the lexicographer to choose the most appropriate ones.

The lexicographers found processing idioms and phrases the most challenging. It was up to lexicographers:

- to convert idioms and phrases into a canonical form
- to distinguish phrases from word usage examples so that only the former are included in the dictionary
- to determine complementation of idioms (basing on lexicographers' intuition and on-line corpora)
- to describe admissible gaps in phrases
- to determine the syntactic role of a phrase (e.g. to determine that *this morning* should be treated as an adverb rather than a noun).

6.4. Semi-automatic correction

Before that stage several types of errors were present in the lexicon. The errors might have resulted from erroneous automatic conversion, mistakes of lexicographers or the development of description formalism during the project. Semi-automatic correction consisted in automatic search for errors (spelling and syntactical – inconsistent with a formalism) and manual correction.

7. The dictionary status

The status of the dictionary after the conversion process is the following:

- English to Polish:

73294 lexemes; 101878 inflected forms; 120805 Polish equivalents; 79971 lexical phrases

- Polish to English:

49989 lexemes, 974989 inflected forms, 54952 equivalents, 42596 lexical phrases.

It is worth noting that the OPEP dictionary was automatically reversed to enrich the Polish-to-English lexicon with one-to-one equivalents.

8. Evaluation

The evaluation of the Polish-to-English translation (*PolEng*) was executed in the Allied Irish Bank (Dublin) in November 2003. At that time the English-into-Polish translation was not available yet. The lexicon included only entries developed manually from various traditional dictionaries (before OPEP enrichment).

As the system provides add-ins to MS-Office application as well as a plug-in to the Internet Explorer, the evaluation dealt with types of documents specific to those applications. For each type the tester selected 7-15 documents from the banking domain and tried to evaluate the speed, coverage of words/phrases and accuracy. The testing was a "black-box" type – the tester was not capable of estimating which component of the translation was responsible for erroneous translations. The tester used both relative and absolute evaluation method. *PolEng* was compared to *PolTrans* – a translation system available on-line. Not surprisingly, *PolTrans* passed the test better as far as the speed is concerned, *PolEng* had higher numbers in accuracy. The attempt to find the absolute estimation was made in the following way: for each document the number of translated words was divided by the number of all words in the text giving *the word completeness*. The tester then divided texts into "phrases": whole sentences or parts of sentences. The number of phrases translated completely (e.g. all word translated) was divided by the number of all phrases, giving the phrase completeness coefficient. For each phrase the accuracy of translation was estimated in terms of "accurate", "good", "moderate", "illegible". The number of accurate and good translation divided by the number of all translated phrase resulted in the accuracy coefficient.

An extract of the document is presented below. The range goes from worst to best translated documents.

Type of document	Word completeness%	Phrases completeness%	Accuracy%
Word 97	95,87 – 99,07	72,45 – 98,84	68,37 – 93,02
Word 2000	95,05 – 99,63	72,45 – 98,84	59,56 – 93,02
Internet Explorer	73,78 – 98,09	46,15 – 94,23	33,85 – 81,69
PowerPoint	94,71 – 99,7	62,64 – 97,65	54,9 – 81,95

Similar tests have not yet been done for English-to-Polish translation. At the time this report is written, the quality of the translation in this direction "looks" distinctly lower. The main reason is worse elaboration of transfer rules (less time) in that direction. Other possible reasons are discussed in *Conclusions*.

9. Conclusions

The *Translatica* group has reached the following conclusions:

1. Only a small part (smaller than expected) of the conversion of a human-readable dictionary into a machine-readable lexicon could be properly carried out fully automatically.
2. The quality of translation is not proportional to the completeness of the description in the dictionary. A larger number of equivalents (if they are not constrained strongly) often results in decreasing rather then increasing the standard of translation.
3. The dictionary description should be limited: the more information – the slower translation
4. WordNet as an entry point for semantic description proved helpful but it has two drawbacks: 1) the structure of lattice assumed in WordNet is hard to deal with computationally (as opposed to the structure of a tree), 2) Semantic hierarchy needed for disambiguation between English and Polish rarely subsumes the WordNet hierarchy
5. It proved very beneficial for the organization of labor to store the lexicon in the form of a standard SQL database.

Another conclusion concerns the translation algorithm itself. The transfer translation in both direction differs strongly in one specific aspect: problem of homography. In Polish homography is quite rare because of rich inflexion. An exemplary homographic sentence: *Przesłał (sent) mi (me) długi (long, debts) list (letter)* which should be properly translated into *He sent me a long letter* (the subject is not obligatory in a Polish sentence) might as well be interpreted as *A letter sent me debts* (according to the free order of components in a Polish sentence). This problem may be easily handled by semantic disambiguation: a letter is more likely to be an object than a subject of a *send* action. In translating from English, homography should be disambiguated as soon as possible (e.g. by statistical POS-tagging) in order to achieve good and robust translation.

References

AGFL (2002): http://www.usenix.org/events/usenix02/tech/freenix/full_papers/koster/koster_html/node4.html

Baldwin, T., Hutchinson, B. and Bond, F. (1999): A Valency Dictionary Architecture for Machine Translation. In: *Eighth International Conference on Theroretical and Methodological Issues in Machine Translation*: TMI-99, Chester, UK, pp. 207-217

Bańko, M. (2000): *Inny słownik języka polskiego (A Different Dictionary of Polish),* Wydawnictwo Naukowe PWN

Graliński, F. (2002): Wstępujący parser języka polskiego na potrzeby systemu POLENG (Bottom-up Parser of Polish designed for the POLENG System) In: *Speech and Language Technology*. Volume 6, Poznań 2002, [http://www.ceti.pl/~poleng/zasoby/publikacje/index.html]

Ilaraza, A. Diaz de, Mayor, A. and Sarasola, K. (2001): Building a lexicon for an English-Basque MT system from heterogeneous wide-coverage dictionaries. In: *Proceedings of MT 2000, Machine Translation and Multilingual Applications in the New Millenium*, University of Exeter, United Kingdom, 20-22 November 2000

Mayfield, J. and McNamee, P. (2002): Converting on-line bilingual dictionaries from human-readable to machine-readable form. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland

OPEP (2002): *Wielki Słownik Angielsko-Polski* (*English-Polish Dictionary),* Wydawnictwo Naukowe PWN

Pinkham, J. and Smets, M. (2002): Modular MT with a learned bilingual dictionary: rapid deployment of a new language pair. In: *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 800-806.