Towards a bilingual lexicon of information technology multiword units

Radosław Moszczyński
Department of Formal Linguistics, University of Warsaw

The article presents a proposal of an electronic, English-Polish translation dictionary covering the language of computer science. The dictionary will focus on multiword units and phraseology typical for this domain. It is supposed to answer the needs of technical translators, who can easily access simple terminological databases, but lack good production dictionaries that would go beyond single terms. The proposed dictionary aims at filling this gap by focusing on multiword units and their modifications, as well as on individual terms' collocational patterns.

The dictionary will be based on the idea of 'extended phraseology' proposed by Müldner-Nieckowski. According to this idea, phraseology is not limited to idioms in the traditional sense of the word, but also covers phrasemes (i.e. units with conventionalized structure, but without figurative meaning), as well as phraseograms (syntactically incomplete units that carry some semantic value). Such a broad approach to phraseology in the planned dictionary will allow translators to create texts that sound natural to computer science experts and to maintain consistency on the stylistic level on top of terminological consistency.

The dictionary will be created in electronic form, with the aim to make it available free of charge on the Internet as part of the Freedict project.

1. Introduction

This paper outlines a project aimed at creating a bilingual lexicon of multiword units constrained to the domain of information technology. Section 2 defines the goals of the project and the applications for which the lexicon is intended. Section 3 describes the general design decisions and the development process. Section 4 outlines future plans and actions related to this project.

2. Goals of the project

The idea for this project was first formed by the needs of translators and localizers of IT materials. A former translator myself, I have always felt that generally available bilingual dictionaries and proprietary terminology resources were never going beyond the level of individual words or terms. With only such resources available, keeping grammatical and stylistic consistency on the level of phrases was difficult even when style guides were available[1], especially in large projects processed by several individual translators. In such projects preserving phrasal consistency required a very skilled and determined editor.

These concerns, which stemmed from pure practice, are also confirmed by researchers. Leroyer says that 'the language of written business communication is characterized by the extensive use of phraseology, not only in terms of collocations and idiomatic expressions, but also of standard phrases in prototypical business genres' (Leroyer 2006: 183), and quotes lexicography manuals which lament 'disastrous lack of phraseological information in most specialized dictionaries' (Leroyer 2006: 197). Although Leroyer is mainly concerned with business communication, I believe his remarks are also relevant to technical texts. Thus, the first goal of this project is to facilitate the work of translators and allow for greater

[1] Such concerns were also expressed by several linguists in an informal survey I carried out among employees of a large localization company.

overall consistency by providing them with a reference to the most common multiword units.

The second goal is to provide machine-readable input for computer-aided translation (CAT) and computer-aided review (CAR) tools. As far as CAT is concerned, a lexicon of multiword units could be a step towards subsentential segmenting[2], and could be integrated with terminology-lookup mechanisms to suggest translated words along with their collocational patterns, taking into account the context of the source word. In terms of CAR, it would allow to achieve greater precision, i.e. limit the number of false-positives reported during automatic review[3]. There are several other possible CAT and CAR applications to be explored, but these go beyond the scope of this article.

3. Lexicon design and development process

The collection of multiword units for the planned resource will follow the idea of 'extended phraseology' as defined by Müldner-Nieckowski (2007). The main principle of extended phraseology is that phraseology covers not only idioms in the traditional sense (i.e. multiword units with non-compositional semantics), but also phrasemes (i.e. units which exhibit a considerable degree of repeatability in language, and which have at least one constituent that is not freely substitutable, but which do not have the metaphorical quality of traditional idioms) and phraseograms (i.e. syntactically incomplete multiword units which nevertheless carry some semantic value[4]).

Some examples of phrasemes taken from IT materials would be *grant privileges*, *take up disk space*, *run a platform*, *reserved word*. Phraseograms include e.g. *persistent across (sessions)*, *remove for (users)*, *download to (directory)*[5].

I believe that taking such a broad approach to phraseology is valid in technical translation, where the main problem is not finding equivalents of individual terms (the number of dictionaries available both in print and online is huge; even if a particular term is not available in any dictionary, a skilled translator can easily find an equivalent by using search engines or exploring multilingual online encyclopedias), but rather building coherent phrases around those terms, which sound natural for professional users of the translated materials. I also believe that

[2] Without going into too much detail, CAT tools store previous translations in a translation memory, which is then used for populating new material with previous translations. Translation memory engines can populate not only texts for which the source matches exactly one of the segments stored in the translation memory, but can also provide fuzzy matches, i.e. translations that need some adjustment by human translators. The longer the source sentence, the smaller the chance of receiving an exact or a fuzzy match. Subsentential segmenting/matching could remedy this.

[3] For example, if the project glossary against which consistency is checked contains an entry for 'order' translated into Polish as 'rozkaz', a typical CAR tool will report false-positive issues if the English text contains a multiword unit such as 'out of order', which contains 'order' in English, but does not contain 'rozkaz' in the Polish translation, because in this context 'order' does not mean a 'request'. A lexicon with the multiword unit 'out of order' defined as an integral entity could eliminate this problem.

[4] Or, for my purposes, carry some value from the point of view of linguistic consistency in translation.

[5] The words in parentheses are not part of the sample phraseograms.

the planned resource will find use in CAT and CAR applications, whose development seems to be focused on collaboration, networking and usability functions, instead of exploring the numerous possibilities opened by modern natural language processing tools and techniques.

Since the planned lexicon is intended primarily for use in an electronic medium, I am not making any specific assumptions regarding the macrostructure of a potential human-readable dictionary derived from it. Based on the information available in the electronic source, the structure could be alphabetical, grouped by specific technical domains or syntactic properties of multiword units, etc.

The microstructure, in its most basic form, would contain a source-language headword, a list of phrasemes, phraseograms, idioms and conventional phrases formed with it, and a list of equivalents of these multiword units in the target language. The descriptions are planned to contain information about possible variations of the units and modifications they can undergo. The lexicon's formal representation will be TEI-conformant XML.

During subsequent development stages, more linguistic information will be added, to be used in CAT and CAR related applications. In particular, each multiword unit will be assigned a formalized representation that will constitute input for natural language processing tools (see Piotrowski (1999) for an example of such applications).

A general framework for such a representation was presented by Bański and Moszczyński (2008). Detailed description of the framework goes beyond the scope of this paper. In short, the framework is based on the Idioms As Regular Expressions (IDAREX) formalism developed by Xerox in 1990s (see e.g. Segond and Breidt (1995)). The framework follows the IDAREX approach to multiword units, but uses XML as the means of representing them, which has several benefits. Most notably it makes processing easier (as libraries for processing XML are available in most, if not all, modern programming languages), and it allows embedding the formalized multiword units in other XML documents by using several namespaces in a single document.

I will use Freedict[6] as the general framework for creating the lexicon and will be following the incremental development process described by Bański and Wójtowicz (2009). The goal is to publish a minimal version of the lexicon as soon as possible, suitable for use by human translators, then refine the design and content to allow CAT and CAR applications described above, as well as implement user feedback.

Linguistic data for the lexicon will come mainly from user interface and documentation materials of open source software, as they are freely available in open formats that facilitate processing. The data will be used to build corpora compatible with Poliqarp, a corpus query engine developed at the Polish Academy of Sciences, which features a powerful query syntax and allows to gather statistical data[7]. Where possible, data will be gathered from bilingual files used for localizing

[6] See http://freedict.org.

[7] See http://poliqarp.sourceforge.net for details.

software (in PO and XLIFF formats) and converted into the Poliqarp format using an automatic tool developed by a student from University of Warsaw.

4. Summary and further research

In the sections above I described the potential benefits and explored the possibilities of creating a specialized multiword unit lexicon for translation and localization applications.

The work outlined here will be followed by creating a corpus of texts, coming both from UI (user interfaces) of software, as well as from UA (user assistance) materials. The corpus will be used to identify a set of phrasemes and produce a basic version of the lexicon, which will be then made available on the Freedict website. In parallel, I plan to refine the design of the dictionary, as well as the formal representation of multiword units and its interface with TEI guidelines.

References

Bański, P.; Wójtowicz, B. (2009). 'A repository of free lexical resources for African languages: the project and the method.' In *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages - AfLaT 2009*. Athens, Greece. 89-95

Bański, P.; Moszczyński, R. (2008). 'Enhancing an English-Polish electronic dictionary for multiword expression research.' In *Proceedings of the Sixth International Language Resources and Evaluation Conference* (LREC '08). Marrakesh, Morocco: European Language Resources Association (ELRA).

Leroyer, P. (2006). 'Dealing with phraseology in business dictionaries: focus on functions – not phrases.' In *Linguistik Online* 27. No. 2/06. 183-194.

Müldner-Nieckowski, P. (2007). *Frazeologia poszerzona*. Warsaw: Oficyna Wydawnicza Volumen.

Piotrowski, T. (1999). 'Tagging and conversion of a bilingual dictionary for XeLDA, a Xerox computer-assisted translation system.' In *Papers in Computational Lexicography COMPLEX '99 Proceedings*. Budapest: Hungarian Academy of Sciences. 113-120.

Segond, F.; Breidt, E. (1995). *IDAREX. Formal description of German and French multi-word expressions with finite-state technology*. Technical Report MLTT-022. Grenoble: Rank Xerox Research Center.