Production of NLP-oriented bilingual language resources from human-oriented dictionaries

Vera Fluhr-Semenova*, Christian Fluhr[†], Stéphanie Brisson*

* SCIPER

46 rue du Moulin a Tan, 91150 Etampes, France fluhrsciper@aol.com, 101376.156@compuserve.com

[†]CEA/DIST

CE-Saclay, 91191 Gif-sur-Yvette Cedex, France

fluhr@tabarly.saclay.cea.fr

Abstract

In this paper, the main features of manually produced bilingual dictionaries, which have been originally designed for human use, are considered. The problem is to find the way to use such kind of dictionaries in order to produce bilingual language resources that could make a base for automate text processing, such as machine translation, cross-lingual interrogation in text retrieval, etc.

The transformation technology suggested hereby is based on XML-parsing of the file obtained from the source data by means of serial of special procedures. In order to produce well-formed XML-file, automatic procedures suffice. But in most cases, there are still semantic problems and inconveniencies that could be retired only in interactive way. However, the volume of this work can be minimized due to automatic pre-editing and suitable XML mark-up.

The paper presents the results of R&D project which was carried out in the framework of ELRA'1999 Call for proposals on Language resources Production.

The paper is based on the authors' experience with English-Russian and French-Russian dictionaries, but the technology can be applied to other pairs of languages.

1. Introduction

Bilingual language resources are indispensable for applications like crosslingual querying, alignment of sentences for translation memories (Fluhr et al., 2000), multilanguage information classification or filtering, machine translation, multilingual text mining, etc.

All these fields of application are considered as main subject both in the United States policy in Linguistic Engineering and in the 5th Framework Program of the European Union. It is also a major subject of cooperation between USA and European Union. Its importance is due to the fact that, in the globalization of the economy using electronic networks, more and more information for strategic, commercial, scientific and technology watch is in other languages than English, and also because E-commerce requires the use of the customers' mother tongue in order to be effective.

In order to perform the research and development included in the 5th Framework Program and develop the tools capable to manage multilingual information, bilingual language resources of great volume and high quality are needed. It concerns primarily the resources of general lexis.

The existing general bilingual language resources have about 30,000 (or less) entries. The experience shows that this volume is not sufficient for real use in NLP-applications (Fluhr et al., 1997). In particular, this volume has to be at least doubled if one wants to get rid of mismatches in cross-lingual interrogation and to guarantee the quality of text information retrieval.

However, the creation of new great language resources "from zero" is extremely expensive. That's why it seems very important to reduce this work and to find the way to use existing bilingual dictionaries, originally designed for human use. Among them, there are fundamental ones which accumulate a huge knowledge on bilingual word-to-word correspondences.

But the point is that in general it is not evident that these dictionaries can be transformed into wellstructured electronic data suitable a) to be managed directly by NLP software tools and b) to provide the accuracy of language processing.

These problems, as well as their solutions, are considered in the present paper.

The paper is based on the authors' recent experience of processing dictionaries edited in Russia, for Russianspeaking users. All the examples quoted here have been found in those dictionaries¹.

The purpose of the project was not just to produce bilingual language resources involving Russian language but also to prove that the existing Russian high-quality dictionaries can be used to facilitate the development of bilingual language resources for the languages of the European Union and other important countries.

Moreover, the methods that are being tested and validated during the project in constructing bilingual language resources are language-independent and can be reproduced (and the tools can be reused) for other language pairs.

2. Bilingual human-oriented dictionaries

A number of linguistic resources have been developed in the former Soviet Union (Semenova, 1998). Among those are dictionaries and especially bilingual dictionaries. They include many languages and are of a very high quality and good coverage.

¹ this concerns primarily the "New Large English-Russian Dictionary" edited by Yu. Apresyan and E. Mednikova (Moscow, 1993) and "Nouveau dictionnaire français-russe" edited by V. Gak and K. Ganshina (Moscow, 1994).

Generally, they have much more entries than those found in Western countries.

2.1. Source data: manual production

Most of these dictionaries exist only in paper form, i.e., as printed books (at least, till recently) Originally, they were created by means of manual lexicographical methods, i.e. without any computer assistance. While preparing the edition the authors stored their data on paper card indexes. The final version of the dictionary, just for the book edition, was produced by means of typewriting.

When the dictionary is being prepared manually and the data are provided by several authors, their manners of data description can not be strictly the same. Usually they differ, at least slightly. Therefore, the format of data presentation in such kind of dictionaries is inevitably heterogeneous.

2.2. From paper form to electronic form

During the last 5 years the main bilingual dictionaries which connect Russian language with those of Western Europe have been scanned (first of all, by the company MediaLingua² in Moscow), in order to produce electronic versions suitable for office applications.

These new dictionaries are now distributed in electronic form, mostly on CD, and are accessible from MS Windows applications. In particular they may be integrated into MS Word.

The use of these dictionaries still remains human, but now it becomes computer-aided. This permits to use space more freely than in printed format, and therefore to present the data in a way more comfortable and more convenient for human use:

- not to merge all the homonyms and meanings of the entry word into the same paragraph, but to present them as separate paragraphs;
- not to use tildes instead of entry word occurrences;
- to use phraseology samples more effectively because each sample is now accessible from different entries;
- to use hypertext links for cross-references;
- to search for words not only of source language but of the target language as well, etc.

These new opportunities and new solutions (Volovich and Zorky, 1997) are mostly concerned with data format rather than content, because the usage of the dictionary remains human.

2.3. Explanations instead of translations

The fact that a dictionary is designed for human use means, in particular, that in the right part of the dictionary line one can find not a direct translation equivalent to the left part, but an explanation of what can be used as such translation.

For example, there are following explanations for the noun "*ponderable*":

"something that could be weighted; something appreciable".

Accordingly, an explanation for the noun "impertinent" is "somebody who likes to intrude in other people's affairs".

It seems evident that such kind of explanation is useful for those who knows the target language well. Such user is able to understand the explanation and to use it in order to produce the most convenient lexical unit which is a) translation equivalent to the source word and b) goes well with all the context of the phrase. But NLP-tools are not capable to use the explanations in the same way.

The explanations, if they reside in the dictionary, will provoke processing errors – such as incorrect phrases in machine translation or noise in information retrieval, etc.

It seems very important to get rid of explanations in NLP-oriented language resource, especially if one desires to produce a pair of consistent dictionaries. If an explanation resides in the data, it will produce a false entry in the inverted dictionary. Explanations in the left part of the dictionary can not be admitted.

2.4. Separators. Multiple roles of comma

It must be added that in the last sample quoted in the previous subsection, the syntactical construction "*somebody who*" in Russian language needs a comma between the words "*somebody*" and "*who*". But commas are usually used in the dictionaries as separators between the translation equivalents.

Therefore, the list of translation equivalents of the word "*impertinent*" may be presented as "*jerk, rascl, somebody, who likes to intrude in other people's affairs*".

Every native speaker can easily understand that the first two commas in this list are separators while the third one is a "syntactic" comma. But it seems difficult to distinguish them automatically, i.e. - with a software tool.

The problem is even more complicate because the commas have also a third role in such kind of dictionaries: they are often used instead of the word "or".

For example, the translation equivalents of the word "*piping*" may be listed in the following way: "*hydraulic, hydromechanical mining*".

Sometimes all three types of commas may be found within the same line or within the same paragraph. For example, the translation equivalents for the noun "scathe" may be presented as "damage, something, that makes a scathe, the reason of sadness, of regret". For a native speaker it is clear that the first and the third commas are separators, the second one is "syntactical" and the fourth one replaces the word "or". But for a software tool, it seems impossible to distinguish all the cases.

Sometimes commas occur in the same line with the word "or". For example, an adjective "self-assumed" was translated as "taken or got voluntarily, without permission or right". If it will be left "as is" it will produce a wrong correspondence During manual corrections it was split into the following list of equivalents separated by comma: taken voluntarily, taken without permission, taken without right, got voluntarily, got without permission, got without right".

² http://www.medialingua.ru

This operation gets rid of two incorrect pairs of words (where the first one is more or less correct, but the second is not correct), and constructs 6 correct pairs at their place:

Left part	Right part
self-assumed	taken voluntarily,
self-assumed	taken without permission,
self-assumed	taken without right,
self-assumed	got voluntarily,
self-assumed	got without permission,
self-assumed	got without right

Table 1: Correct pairs of words

2.5. Manual pre-editing

The list of difficulties which can complicate the process of transforming the source data into bilingual language resources, is certainly far from being exhausted by the two previous subsections. But it is already evident that the correct selection of useful information from the source data is possible only if these are pre-edited by a native speaker of the target language. This phase can not be omitted but it can be reduced as much as possible due to automatic identification of all the doubtful cases. Consequently, manual pre-editing may be accomplished by randomly jumping from one such case to another.

Thus, in the work described in the next section, more than 3 000 doubtful lines were automatically identified and manually processed.

A lot of MS Word macros were developed, in order to make this work "as automatic as possible".

3. Bilingual language resources production

3.1. The idea of the method

The main idea of the method is to transform the source data into a well-formed XML text. However, it is not so simple to realise. For example, the first mark-up based on the list of styles, is "flat" and therefore doesn't correspond to the real structure of articles. An intelligent software tool was developed in order to rebuild the hierarchic structure and to deal with logical operators and cross-references.

3.2. Scheme of data processing

In the framework of the project, a technology was developed for the purposes of improvement and simplifying the structure of such kind of dictionaries, transforming them into XML format and then producing valid language resources for various purposes (synonyms, phonetics, orthography variants etc.).

This technology allows the production of pairs of bilingual language resources in XML format, where every language resource is consistent with its reverse counterpart (this feature is extremely important for the robustness of cross-lingual interrogation).

Such kind of language resources inversion is possible due to the tabular form of presentation.

The scheme of data processing is presented on the fig. 1.

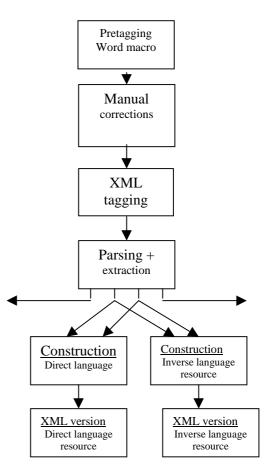


Figure 1. Scheme of data processing aimed to language resources production

3.3. Source data

The source data we started with was a collection of more than 240 MS Word files with a style sheet including more than 40 styles. These files have been produced by means of scanning of a dictionary edited as a book.

The content of these files has been explored using several MS Word macros and C++ programs in order to understand what is the use of styles, what are the values of part of speech or domain tags.

In particular, it was found that the styles application is not strictly the same in different files, and sometimes even within the same file. It is explained by the fact that the styles were applied by several people, without common rules.

Special tools were developed in order to correct human errors (made in the dictionary creation, its scanning, styles application, etc.). Then, each of the dictionary files was processed by the following treatments.

3.4. Pre-tagging

The first treatment of the file consists of a MS Word macro that transforms the word file into a tagged file. Each style is put between a beginning tag and an ending tag. At this time, the tagged file is not a well formed XML file because parenthesis and brackets had not a good style attribution.

3.5. Manual corrections

At this time the tagged file must be processed manually to tag in different ways some ambiguous separation character like the comma. It is also necessary to transform some expressions which have a common part to produce two complete expressions. At this step, translations that are definitions (and that cannot be found by program) are manually tagged.

3.6. XML tagging

The next step is a program that transforms the modified file into a well-formed source XML file. This file has the full extended content of the original file but is now XML tagged. At this time no DTD has been build for this source file. This source file can be used to build up other linguistic resources different from the ones expected in the first work (for example a monolingual English language resource with part of speech, phonetics, etc.).

This processing is made by a Perl program.

3.7. Parsing and extraction

The next step contains an XML parser (Expat). It ensures that the preceding file is well formed. Then it extracts the useful information for a bilingual dictionary. The results are given in a tabular format : one line for each pair "source word – target word". This step is a Perl program. The result is composed into several files :

- a file containing all entries having a part of speech;
- a file containing the compounds that have no part of speech and so are supposed to be noun phrases;
- a file of pairs of source words that are orthographic variants or synonyms;
- a file of eliminated pairs for various reasons;
- a file of translations that are explanations or definitions.

3.8. Construction of language resources

The preceding steps were performed on 241 MSWord files. The present step reads all "source word – target word" pairs from the two first types of file (entries with part of speech and compounds), sort them according to source word, part of speech, domain, identification of synonym list. The output the final language resource in a tabular form.

The file containing the compounds must be controlled manually because some compounds are not noun phrases but adverb phrases.

The same program reads the same files but inverses source and target file. After a sort, it produces an inverse language resource.

The produced language resources in tabular form contain more than 350 pairs of words.

The tabular form of the direct and inverse language resources are transformed into a XML format conforming to the DTD. The direct language resource contains more than 95,000 entries, the inverse one – more than 120,000.

4. Use of results

The language resources produced in our project will be used in a metacrawler (meta-browser) developed in the CEA (French Commissariat for Atomic Energy) for exploring the "invisible" Web (i.e., the information on the Web that can only be accessed through a search engine).

The results will also be used in the SPIRIT crosslingual system (Fluhr et al, 1998).

Even the first experience to use SPIRIT crosslingual text information retrieval for the English-Russian pair with an English-Russian dictionary of 70,000 entries (Brisson, 1998) showed that the mismatches were much less frequent with the dictionary of this size, than for the use of SPIRIT for the French-English and English-French pairs (about 32,000 entries in each dictionary).

One more important result of the project is the above-mentioned well-formed XML file. It contains the same information as the source data but it may be managed and processed by XML-parsers. It may allow to produce other language resources from the same source.

5. References

- Brisson, S., 1998. *Mise en forme d'un dictionnaire permettant une interrogation bilingue de bases textuelles russes.* DEA thesis. Denis Diderot University (Paris VII).
- Fluhr, C., D. Schmit, F. Elkateb and K.Gurtner, 1997. Multilingual database and crosslingual interrogation in a real internet application. In *AAAI 1997 Spring Symposium Series*, workshop "Cross-language Text and Speech retrieval", Stanford University, Caligornie.
- Fluhr, C., D. Schmit, Ph. Ortet, F. Elkateb, K. Gurtner and Kh. Radwan, 1998. Distributed cross-lingual information retrieval. In G.Greffenstette (ed.), *Crosslanguage information retrieval*. Kluwer Academic Publisher.
- Fluhr, C., F. Bisson and F. Elkateb, 2000. Mutual benefit of sentence/word alignment and crosslingual information retrieval. In Jean Veronis (ed.), *Parallel Text Processing*. To be published in Text, Speech and Language Technologies series.
- Semenova, Vera, 1998. On Russian Resources for Language Engineering. *The ELRA Newsletter*. 2:6-7.
- Volovich, M., and Zorky, K., 1997. Vocabulary in books and on displays. In A. Narin'yani (ed.), Dialogue'97. Computational Linguistics and its Applications. Yasnaya Polyana – Moscow, Russia.