Extracting terms and terminological collocations from the ELAN Slovene-English parallel corpus

Špela Vintar
Dept. of Translation and Interpreting
Faculty of Arts
University of Ljubljana
spela.vintar@guest.arnes.si

Abstract
In many scientific, technological or political fields terminology and the production of up-to-date reference works is lagging behind, which causes problems to translators and results in inconsistent translations. Experience gained in various projects involving parallel corpora show that automatic extraction of terms and terminological collocations is an achievable goal, however methods and techniques may vary according to the language pair selected. The paper describes how a methodology for multi-word term extraction and bilingual conceptual mapping was developed for a Slovene-English corpus. We used word-to-word alignment to extract a bilingual glossary of single-word terms, and for multi-word terms two methods were tested and compared. The statistical method is broadly applicable but gives results of very limited use, while the method of syntactic patterns extracts highly useful terminological phrases, however only from a tagged corpus. A vision of further development is given and how these methods might be incorporated into existing translation tools.

1 Introduction

Through the past decade Central and Eastern Europe has been undergoing a process of rapid economic, cultural and ideological transition, which invariably affects all aspects of communication, particularly translation. In Slovenia, new international contacts, foreign investments as well as preparations for full membership in the EU have caused an enormous increase in the volume of translations produced and required. The translation of EU-legislation alone is an overwhelming task for such a small translation market as over 160,000 pages are due within the next two years, and other fields such as localization and web technologies are facing similar demands for translations, mostly from or into English.

In order to accomplish these tasks with maximum efficiency and quality, two aspects should be particularly stressed. The implementation of state-of-the-art translation technologies such as translation memories and terminology management systems can prove very useful both in supporting translators' work and in providing reusable resources for the future, however a thorough analysis of the text and translation environment is essential before these tools are introduced.

The second aspect is related to the first one and involves terminology work in general, which tends to be even less well-organized under such circumstances. Continuous overload in translation work and massive production of translated texts inevitably means that terminologists cannot keep up with the linguistic, technological and terminological developments within the field, and terminographical reference works are compiled – if at all – with a backlog that sometimes completely annihilates their usefulness. Lack of reliable terminological resources in turn results in even greater inconsistency.

A possible solution for these problems is the creation of a domain-specific parallel corpus, a kind of archive of past translations and their originals, which serves as reference as well as terminology source for further translations within the domain. A parallel corpus is however not equivalent to a terminology database, and the way leading from former to the latter – if done manually – involves laborious text scanning and comparison of different sources.

The aim of the work presented in this paper was to develop a methodology that would support the work of terminologists and translators by automatic term extraction and bilingual conceptual mapping. First, a method for automatic extraction of single-word terms and their translation equivalents from the parallel corpus is described, then we focus on the problem of extracting multi-word terms. Two approaches were tested, evaluated and compared, first the statistical method based on extracting recurring word combinations and then the so-called linguistic method of extracting terms on the basis of their syntactic structure.

The results obtained so far still leave a lot of room for improvement, nevertheless they show that it is possible to automatically retrieve bilingual terminological collocations from a parallel corpus and use them either as a translation resource, a basis for new terminological databases or a supplement to existing ones.

2 The corpus

All experiments described here were performed on the Slovene-English parallel corpus of 1 million words that was compiled in 1999 within the framework of the EU ELAN (European Language Activity Network) project (Erjavec 1999, Vintar 1999). The corpus consists of 15 source texts and their translations covering a wide range of domains, the largest subcorpora being EU-related texts and computer manuals.

The first level of corpus annotation involved **sentence alignment**, **tokenization** encoding in TEI-conformant SGML. The next level was necessary especially for the Slovene part of the corpus, because almost any kind of computational processing of Slovene requires **lemmatization**. Slovene is a highly inflectional language with a morphological richness that is – in corpus linguistics' terms – reflected in a basic tagset of over 1600 tags, i.e. different combinations of grammatical categories and paradigms. Thus if we wish to retrieve items of language on the basis of frequency counts, the results will only be representative if wordforms are converted into lemmas.

The lemmatization was kindly performed by the company Amebis, a leading developer of language technologies in Slovenia.

The second part of the experiment involved term extraction on the basis of syntactic patterns, and for this another step of corpus annotation was needed – the POS-tagging. Although this is considered a routine operation for English and several taggers are freely available on the web, for smaller languages like Slovene this might not be as easy. After several tools had been tested (Džeroski et al 1999), the Slovene part of the corpus was tagged using the TnT tagger developed by Thorsten Brants in Saarbrücken (Brants 2000).

The corpus is freely available and can be accessed at http://nl.ijs.si/elan/.

3 Automatic extraction of single-word terms and their translation equivalents

To extract a bilingual glossary of terms from a parallel corpus we must first identify single-word terms for each language. The use of the word **term** might be slightly confusing, because terms as entities in the specialised vocabulary of a certain domain are only rarely single words but mostly comprise two or more words. The average lengths of terms in Slovene and English will be discussed in more detail below. Here, a more appropriate description would be words of potential terminological relevance, thus if *operating system* is a term, we extract both *operating* and *system* as terminologically relevant words.

The basis for deciding whether a word is terminologically relevant is its relative frequency in the specific text as compared to its frequency in a general language corpus. By extracting **keywords**, i.e. words that occur with a higher relative frequency than would be expected, we obtain lists of words characteristic of the domain that the text belongs to. Such a list can however contain only words that occur several times. To make up for this lack the list was supplemented with the words that were labelled unknown by the lemmatizer, because these would also turn out to be mostly terms.

Now that candidate terms were identified for each half of the fifteen parallel texts the next step would be to find their translation equivalents. For this task the Twente word-alignment software was used, a Unix-based freely available tool that extracts a bilingual lexicon from a bitext by calculating statistical probability for each word being the translation of another word in an aligned sentence pair (Hiemstra 1998). The lexicon extraction was run on a pre-processed corpus from which all **stopwords** – highly frequent functional words such as articles and prepositions - had been previously removed and the remaining wordforms were converted into lemmas.

The result is a bilingual lexicon containing all words that occur in the corpus and their suggested translation equivalents, together with a probability rate. After some analysing it was seen that the results are only reliable for those items that occur more than 4 times in the corpus and are at the same time matched together as translation equivalents with a probability of over 0.50. We therefore devised Perl filters according to these rules and the initial term list and extracted only the word pairs that matched all the criteria. The final outcome is a bilingual text-specific glossary of single-word terms (see Figure 1).

Frekv.	Slovensko	Angleško	Ver.
45	agencija	Agency	0.58
9	agraren	agrarian	0.78
9	akt	regulations	0.79
18	AKTRP	AAMRD	0.84
5	analiza	analysis	0.69
5	C	C	1.00
6	carinski	Tariff	0.83
8	celovit	integrated	1.00
5	cena	prices	0.96
15	center	Centre	0.57
10	časoven	Timescale	0.70
26	članica	Member	0.94
8	človek	persons	0.63
8	človeški	Staff	1.00
14	dejavnost	activities	0.80

Figure 1: A bilingual glossary of single-word terms

Using this method, around 17% of words of the entire lexical inventory of the text are extracted as terminologically relevant and matched with a translation equivalent, with a precision rate of over 98%. The main drawback of this method is the fact that it retrieves only single words that occur at least four times in the corpus. The value of such a bilingual glossary for translation purposes is thus very limited, however we should not underestimate its usefulness for identifying multi-word terms, as we shall see below.

4 Automatic extraction of multi-word terms

The basic idea underlying the tests described below was to identify multi-word terms first in a monolingual context and then try to link them to their respective translation equivalents or related terms in the target language.

Before a methodology for extracting multi-word units from a parallel corpus could be developed, some characteristics of terms in English and Slovene had to be established. We were especially interested in the following: How many words do terms usually have in Slovene and in English? What is the structure of multi-word units in terms of syntax, morphology and orthography? Which kinds of terms can be successfully retrieved by computational methods?

To find answers to the above questions, an existing terminology database of EU-related terms was thoroughly analysed. This MultiTerm database was created by terminologists at the Slovenian Government's Office of European Affairs and currently contains over 16,000 entries in Slovene and English, partly also other European languages. The database is regularly updated and can be accessed at http://www.gov.si/evroterm.

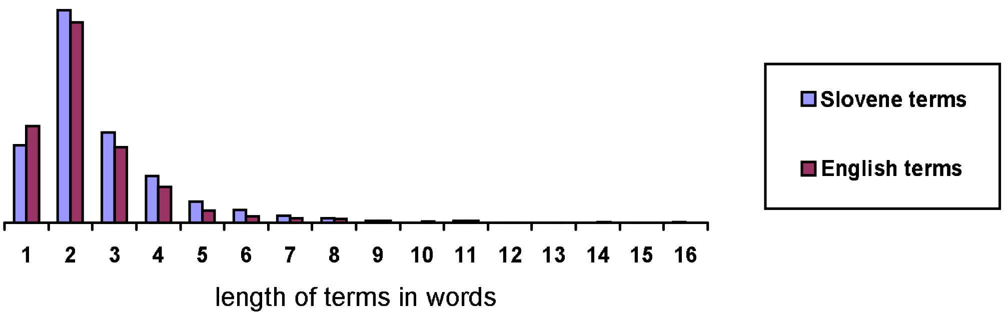


Figure 2: Length of Slovene and English terms

An analysis of a random sample of 2,000 entries in both languages showed that the length of terms can vary from 1 to over 26 words, with the majority of entries in both languages being two-word items, usually an adjective+noun sequence. The second most frequent type is a single-word term. As less than 5% of all entries exceed 6 words, we decided that automatic extraction should be limited to sequences of 2-6 words. Terms of length over 6 words are less fixed and show such variability that automatic extraction becomes a very complex task. Figure 2 shows the length of Slovene and English terms.

As the next step we manually analysed the syntactic patterns of terms for both Slovene and English and ordered them according to frequency. These patterns were needed for the second part of the experiment, the "linguistically motivated" method, as we shall see below.

4.1 Statistical method

Statistical methods in computational and corpus linguistics generally share the fundamental approach to language viewed as a string of characters, tokens or other units, where patterns are discovered on the basis of their recurrence. Accordingly, when we approach the extraction of multi-word units from a statistical point of view we initially retrieve recurring word sequences.

For the purposes of this study the SENTA system was used (Dias et al 1999), a software that extracts word sequences from 2 to 6 words on the basis of two statistical measures, the Mutual Expectation and the LocalMax. The system takes raw text with minimal or no markup as input and returns lists of both contiguous and non-contiguous phrases, which occurred in the corpus with a frequency higher than could be expected by coincidence.

The system was run on both parts of the Slovene-English corpus, whereby the raw version of the corpus was used. An examination of the results first showed a significant difference in recall for Slovene and English: around three times more phrases had been extracted from English than from Slovene. Given the morphological characteristics of Slovene with its rich inflectional inventory, this is not surprising. While it might be expected that lemmatization would diminish this discrepancy, it would undoubtedly cause other problems. Multi-word terms *(Direktive Sveta Evropske unije)* are fixed units that include inflected words, and converting them to their base form *(direktiva svet evropski unija)* alters the structure of the term, sometimes beyond recognition.

Moreover, the system extracts all multi-word units regardless of their type or form, so that we also find sequences like *so as to, is a system that, the* ____ *of,* etc., for which we have no use in terminology. Clearly the output must be thoroughly filtered before the results can be used in any productive way. But how do we know whether a certain sequence is a term?

Before answering the above question, the notion of "term" should receive some redefining. In terminology science, a term in its narrow sense designates a standardised name for a defined entity or a concept, and in its broader sense any fixed linguistic unit with a defined meaning within a special language. The scope of our study however is translation-oriented terminology, and we argue that for translation purposes the traditional understanding of terminology and terminography should be extended to include more information about term usage, collocations, variants and related concepts. Thus, some of the phrases we present here as "useful" might not be regarded as terms in their traditional sense. For the sake of clarity and precision it is therefore more appropriate to speak of terminological collocations.

Returning to the problem of filtering, it was decided to use two criteria to determine whether a sequence was terminologically relevant or not. The first was using a stopword filter and the rule that a term can never begin or end in a stopword. This would filter out things like *of the Prime Minister* and *the National Assembly is* and leave only *Prime Minister* and *National Assembly*. Of course, stopwords may very well occur within the phrase itself, as in *Ministry of Foreign Affairs.*

The second filter was based on the assumption that a multi-word term is likely to contain single-word terms (Maynard/Ananiadou 1999), so we used the lists of single-word terms devised in the first part of the research and filtered the phrases according to the number of terms they contained. A three-word phrase for example had to contain at least two single-word terms to be selected.

This two-step filtering proved to be a suitable method for selecting terminological collocations, however it considerably reduced the initial amount of phrases. Around 17% of the phrases were selected, finally amounting to 897 Slovene and 1,703 English terms. A sample section of the list can be seen in Figure 3.

external debt	global competitiveness
external trade	goods trade
favourable than	gods were
fell substantially	government expenditures
financial results	graph below
food manufacturing	harbour transport
foreign currency	high ____ growth
freight transport	higher _____ rate
fuel oil	household electricity
funds allocated	housing construciton
general government	income tax
giro accounts	increased their

Figure 3: Multi-word terms extracted with a statistical method

Considering the size of the corpus and the terminological richness of the texts, this result is not very encouraging. After the phrases extracted had been annotated in the corpus, an attempt was made to find pairs of translation equivalents. It turned out however that there were only few cases where the same term would be selected both in the Slovene and the English sentence of the aligned pair, so it was mostly not possible to identify bilingual correspondences between terms.

On the whole, the somewhat disappointing outcome of the statistical method as applied to Slovene and English provoked us to rethink the methodology and try to include more linguistic information in the extraction of terminology.

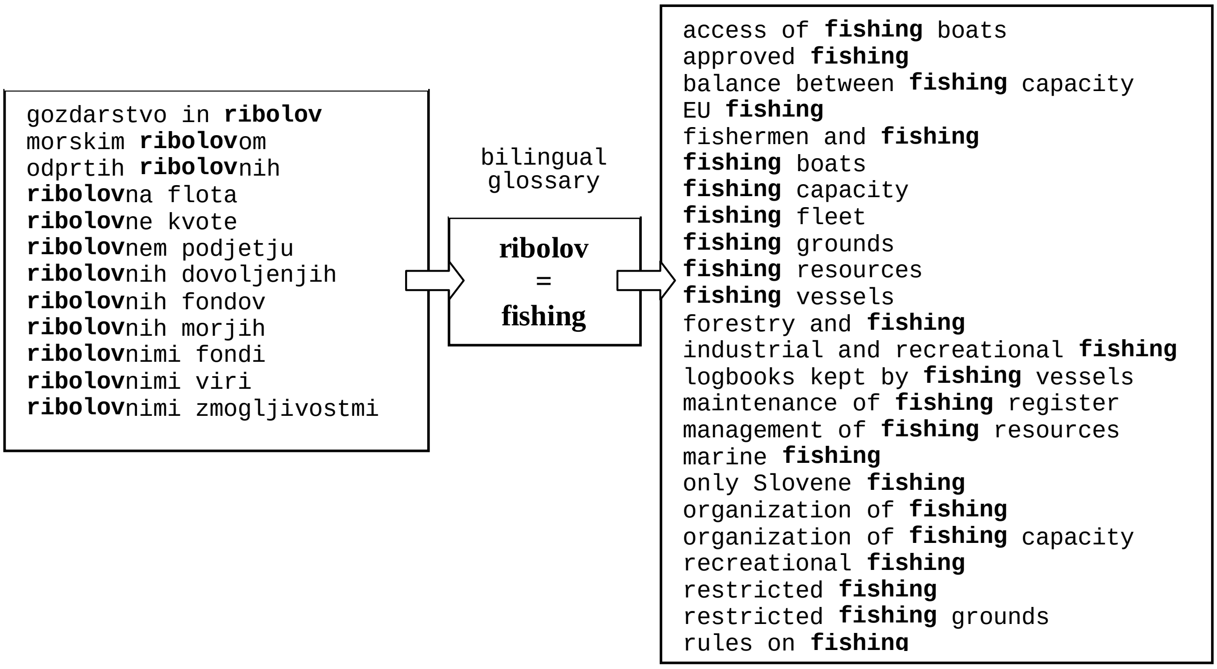
4.2 Syntactic method

The idea underlying this method is that multi-word terms are constructed according to more or less fixed syntactic patterns, and if such patterns are identified for each language it is possible to extract them from a morpho-syntactically tagged corpus (Heid 1999). Thus if we know that the pattern “noun + preposition *(of)* + noun + noun” is a valid four-word syntactic pattern for English terms we may use this information to extract phrases like the following: *acceleration of GDP growth, accession of candidate states, Adaptation of EU Legislation, amount of insurance coverage, calculation of pedigree value, crime of money laundering, Department of Plant Protection, Eradication of Potato Ring, fluctuation of labour force* etc.

The syntactic patterns for Slovene and English were identified on the basis of the existing termbank of EU terminology as described above. After that we used Perl scripts to extract all word sequences matching the patterns from the tagged corpus, for each language separately. The results of the extraction were very good, requiring almost no filtering or post-processing. Because this method is not based on frequency and extracts even phrases that occur only once, recall was considerably higher than with the statistical method, retrieving almost ten times as many phrases than the first experiment. Although not all of the phrases would be considered terms in the traditional sense of the word, most of them either contain terms or show illustratively how terms collocate with intensifying or specifying adjectives or names.

In view of the fact that the needs of translators often remain unanswered by what traditional terminological dictionaries or databases have to offer, we believe that such corpus-extracted collocations can be helpful to translators in forming adequate technical texts.

In order to establish a link between the two separate piles of terminological phrases and enable bilingual querying, the bilingual lexicon of single-word terms described above can be used. This conceptual mapping is then conducted in three steps (Figure 4).



gozdarstvo in **ribolov**
morskim **ribolov**om
odprtih **ribolov**nih
ribolovna flota
ribolovne kvote
ribolovnem podjetju
ribolovnih dovoljenjih
ribolovnih fondov
ribolovnih morjih
ribolovnimi fondi
ribolovnimi viri
ribolovnimi zmogljivostmi

access of **fishing** boats
approved **fishing**
balance between **fishing** capacity
EU **fishing**
fishermen and **fishing**
fishing boats
fishing capacity
fishing fleet
fishing grounds
fishing resources
fishing vessels
forestry and **fishing**
industrial and recreational **fishing**
logbooks kept by **fishing** vessels
maintenance of **fishing** register
management of **fishing** resources
marine **fishing**
only Slovene **fishing**
organization of **fishing**
organization of **fishing** capacity
recreational **fishing**
restricted **fishing**
restricted **fishing** grounds
rules on **fishing**

Figure 4: Bilingual conceptual mapping

In the first step, the translator would enter the search word in one language *(ribolov)* and retrieve all terminological collocations containing the word. In the second step, the search term is translated using the bilingual lexicon *(ribolov = fishing)*, and thirdly, the list of terminological phrases in the target language is consulted.

5 Conclusions and future work

The paper presents a methodology for the extraction of terminological collocations from a parallel corpus for translation purposes. It shows that statistical methods are useful especially because they are generally applied to raw texts, however the results they produce are not always equally satisfactory. Working with linguistically annotated texts will improve the results and minimize the efforts of post-processing, but the annotation itself tends to be a very time-consuming phase in the process of corpus building, especially for minor languages.

Nevertheless we believe that terminological collocations and phrases extracted in this way, coupled with the bilingual conceptual mapping described above, could be used effectively either to supplement existing terminological collections or to be used in addition to traditional reference works.

In future we envisage the development of techniques for the alignment of exact translation equivalents of multi-word terms in Slovene and English, and one way of doing so is by finding correspondences between syntactic patterns in both languages. Probably though this will not be a sufficient criterion and will have to be combined with other methods such as cognate recognition or NP-chunking.

Even if at this point the trouble of compiling the corpus, installing various tools and going through complex phases of processing and post-processing seems unrealistic to be used in a real terminological or translation environment, in reality such methods might only be a step away. Translation memory systems already store translations in a format similar to a parallel corpus, and terminology tools already involve functions such as Autotranslate that statistically calculate the most probable translation equivalent. By refining these functions and making them language specific, we could soon be facing a new generation of tools for terminologists and translators. It remains to be seen, however, whether they can really be implemented into translation environments on a broad scale.

Acknowledgments

The author would like to thank Tomaž Erjavec for his help and support. Thanks go also to Gael Dias for his generosity in sharing his experience, knowledge and resources.

References

1. Brants, Thorsten (2000) TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference, ANLP-2000, Seattle, WA.
2. Dias, Gael et al (1999) Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: Proceedings of the 9th Conference on Artifcial Intelligence. Berlin: Springer.
3. Džeroski, Sašo; Erjavec, Tomaž in Zavrel, Jakub (1999): Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. Jožef Stefan Institute Research Report IJS-DP 8018, 1999. http://nl.ijs.si/lll/
4. Erjavec, Tomaž (1999) A TEI Encoding of Aligned Corpora as Translation Memories. In: Proceedings of the EACL Workshop on Linguistically Interpreted Corpora. Bergen: ACL.
5. Heid, Ulrich (1999) Extracting Terminologically Relevant Collocations from German Technical Texts. In: Sandrini, Peter (ed.): Terminology and Knowledge Engineering (TKE '99), Innsbruck. Wien: TermNet, 241-255.
6. Hiemstra, Djoerd (1998) Multilingual Domain Modelling in Twenty-One: Automatic Creation of a Bi-directional Translation Lexicon from a Parallel Corpus. In: Coppen,

Peter-Arno et al, ur.: Proceedings of the 8th CLIN meeting, 41-58. http://wwwhome.cs.utwente.nl/~hiemstra/

7. Maynard, Diana; Ananiadou, Sophia (1999) Identifying Contextual Information for Multi-Word Term Extraction In: Sandrini, Peter (ed.): Terminology and Knowledge Engineering (TKE '99), Innsbruck. Wien: TermNet, 212-221.
8. Vintar, Špela (1999) A Lexical Analysis of the IJS-ELAN Slovene-English Parallel Corpus. In: Vintar, Š. (ed.) Proceedings of the workshop Language Technologies – Multilingual Aspects. Ljubljana: Faculty of Arts.