BUILDING BILINGUAL TERMINOLOGY DATABANKS FOR GREEK FROM PARALLEL TEXTS

António Ribeiro, Gabriel Lopes and João Mexia

In this paper we present a method of building bilingual terminology databanks for Greek from parallel corpora. Parallel texts are good sources of information to obtain the translations of bilingual terminology. They provide the usual ways terms are used in different languages. This work uses parallel corpora from the Official Journal of the European Communities in English, Greek and Portuguese in order to extract translations of terminology. The method starts by aligning the parallel texts and extracting terminology found in those texts. Then, translations of terms are identified by analysing the similarity of the distribution of the extracted terms in the aligned corpora. We show some results given by this promising way of accomplishing this task.

1. Introduction

Bilingual terminology databanks are useful resources either for machine translation, cross-language information retrieval or even for human translators themselves. They provide domain specific terms which have typical translations in other languages. However, terminology databanks are not available for every domain. Furthermore, their manual compilation is quite time consuming and requires much human effort.

Parallel texts are good sources of information to obtain translations of terminology and are becoming ever more widely available. Parallel texts are texts which are mutual translations. For example, the European Commission produces daily hundreds of pages of legislative texts in the eleven official languages of the European Union[1].

Still, before it may be possible to identify translations of words or terms automatically, it is necessary to make correspondences between the pieces of text in the different languages. This process is called *alignment*. Since translations of terms are not necessarily done word by word, multiword terms should be extracted from the parallel corpora. It is possible to do so using statistical techniques.

In this paper, we present a method to extract translations of terms in order to compile a bilingual terminology databank from parallel texts in English, Greek and Portuguese. Some previous work has already been done on alignment of parallel texts and extraction of translation equivalents, specially using English as one of the languages but not much with Greek. Greek presents a further challenge compared to other Western European languages:

[1] Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

it has a different alphabet. Even if one takes advantage of a possible conversion of the character set (e.g. '*α*' to '*a*', 'β' to 'b', and so forth), it still poses some interesting problems for bilingual research because it is a highly inflectional language. Thus, it is a good language that makes possible to assess how language independent language methodologies can be.

This paper is structured as follows: we begin by giving a general overview over previous work in the next section. In section 3, we present the parallel corpus that was used. Section 4 describes briefly the technique used to extract terms. Sections 5 and 6 discuss the aligner and how translations of terms are identified. We present some results in section 7. Finally, we draw some conclusions in section 8 and discuss future work in section 9.

2. Previous Work

Early work on parallel texts alignment was performed at sentence level counting words or characters (see [1] and [4]). The algorithms grouped sequences of sentences till they had proportional sizes. In [5], two sentences were aligned if the number of correspondence points associating them was greater than a threshold. The sentence aligner used a bilingual dictionary derived from previous alignments which is progressively refined as the alignment proceeds. In [12], the use of cognates enhanced the alignment results. Cognates, which are similar words like *Parliament* and *Parlement* in English and French respectively, provide more and better clues for alignment models. [7] also recurred to cognates in order to define correspondence points. These correspondence points were subsequently filtered if they laid outside an empirically defined search space.

The requirement for clear sentence boundaries was dropped in [3] for alignment of English–Chinese parallel texts. For the English texts, terms were extracted by matching pre-selected syntactic regular expressions, typical of noun phrases, to tagged text. Translations of terms were then identified comparing vectors that stored distances between consecutive occurrences of terms (DK-vec's) using dynamic time warping.

[6] presents a methodology to align Greek–English parallel texts at sentence, lexical and word level using shallow linguistic processing and statistical models. Translations of noun phrase terms are extracted from sentence aligned parallel texts using syntactic patterns. However, they also acknowledge problems concerning sentence identification and sentence delimiters ([6], p. 123). This was also one of the early problems in [1] and [4].

In the following sections we present a methodology to identify translations of terms based on statistical approaches. It is a language independent methodology that does not require *a priori* language knowledge and which does not recur to heuristics.

3. Source Parallel Texts

We worked with a parallel corpus selected at random from the Official Journal of the European Communities [2] and from The Court of Justice of the European Communities[2] in English, Greek and Portuguese. For each language, we included:

- five texts with Written Questions asked by members of the European Parliament to the European Commission and their corresponding answers (average: about 60k words or 100 pages / text);
- five texts with records of the Debates in the European Parliament (average: about 400k words or more than 600 pages / text);
- five texts with judgements of The Court of Justice of the European Communities (average: about 3k words or 5 pages / text).

The table below shows the number of words per sub-corpus. The average number of tokens per text is inside brackets[3]. 'el' stands for Greek, 'en' for English and 'pt' for Portuguese.

Language	Written Questions	Debates	Judgements	Total
el	272k (54k)	1,9M (387k)	16k (3k)	2222k
en	263k (53k)	2,1M (417k)	16k (3k)	2364k
pt	284k (57k)	2,1M (416k)	17k (3k)	2381k
Total	819k (55k)	6,1M (407k)	49k (3k)	6967k

4. Extraction of Terms

In order to identify multiword terms in the several languages, we have used a methodology proposed in [11]. This methodology is based on the idea that the more cohesive a group of *n* words is, the higher its cohesiveness score is. The algorithm assumes that the score of a good multiword unit must be a local maximum, i.e. the cohesion of the set of *n* words is higher than any subset of *n*–1 words contained in it and higher than the cohesion of any superset of *n*+1 words which contains it. Thus, the algorithm is able to select, for example, 'common rules and standards' as a relevant multiword term instead of 'common rules and' or of 'common rules and standards for' because the scores of these multiword units are lower.

The methodology has proved to be quite adequate to be used across several languages. In this way, we are able to capture multiword terms for each language.

[2] http://curia.eu.int.

[3] html markups were discarded since they would provide extra clues to the aligner, biasing the results.

5. Alignment of Parallel Texts

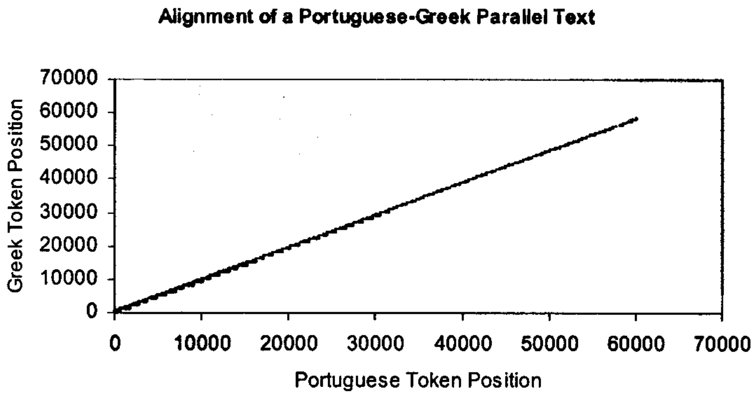
To make correspondences between the texts in English, Greek and Portuguese, we took advantage of *clues* in the parallel texts that help identify what piece of text in one language should correspond to in the other language.

The text aligner looks at words which are identical for a pair of languages, at numbers, punctuation and even at the outline of documents (paragraph structure, lists of items). Although we could have made the conversion of the Greek letters to the Latin character set, we decided not to do so at this stage in order to check how far our methodology could go using simple strategies. For average size texts (e.g. the Written Questions), the number of common words accounts for about 4% of the total number of words (about 2k words / text). These words end up being mainly numbers and names.

Here are a few examples from a Greek–English parallel text: *1998* (numbers, dates), *Eureka* (acronyms), *Greenpeace* (names of organisations, proper names) and *Poitiers* (names of cities). A sample of an alignment of a Greek–Portuguese parallel text is shown below:

Greek	**Portuguese**
Όσον αφορά το καθεστώς του ευρωπαίου πολίτη , το άρθρο 8 , παράγραφος	No que respeita ao estatuto do cidadão europeu , o n º
1	**1**
, του σχεδίου της Συνθήκης για την Ευρωπαϊκή Ένωση προβλέπει ότι	do artigo 8 º do Projecto de Tratado da União Europeia prevê que
«	**«**
πολίτης της Ένωσης είναι κάθε πρόσωπο που έχει την υπηκοότητα ενός κράτους μέλους	é cidadão da União qualquer pessoa que tenha a nacionalidade de um Estado – membro
» . ¶ (1)	**» . ¶ (1)**
Η Επιτροπή λυπάται για την καθυστερημένη απάντησή της	A Comissão lamenta o atraso com que esta resposta é dada
. (2) SEC (91) 1855	**. (2) SEC (91) 1855**
τελικό	final

The characters in bold highlight the tokens used as correspondence points for the alignment of the texts. The positions of these tokens in the texts are used as the co-ordinates of correspondence points, i.e. the aligner uses the byte number of the tokens in the text as co-ordinates. The byte number is the offset of the token from the beginning of the file measured in number of bytes. If we plot the correspondence points on a graph, we tend to get a wellbehaved set of points along a diagonal line. The figure below shows an example from a Portuguese–Greek parallel text:



n

With these points we can build a linear regression equation which helps to decide which are the good points and which are the bad points for alignment. This filtering is needed because it may be the case that the position of a token in a particular text in one language was wrongly paired with the position of the identical token in the parallel text of the other language. This leads to noisy points which usually lay far away from the diagonal line. The filtering of those noisy points is based on the confidence bands of linear regression lines and on the histogram of distances between the actual value of the y co-ordinate of the correspondence point and its expected value computed with the linear regression equation. That is, for each point (x,y), the aligner computes the expected $\hat{y}$ and checks how far the actual value y is from it. If it lays outside the confidence interval, than that point is filtered out. An overview of the algorithm is given below is (see [8], [9], for a more detailed account):

1. Take two parallel texts A and B;
2. For each text, build a table with the positions of each word. The positions are given by the offset from the beginning of the file measured in number of bytes;
3. Define the points (0,0) and (length of text A, length of text B) as the extremes of the initial segment where more correspondence points will be searched;
4. Build a set of candidate segments using the co-ordinates of identical sequences of characters which occur with the same frequency within the segment; use the co-ordinates of previously identified translations should an extracted bilingual lexicon be already available;

Since not all points defined using this rule are good points, we build a linear regression with all points and use histograms and confidence bands to filter the noisy points:

5. Filtering out bad points.
   5.1. Build a linear regression line using the co-ordinates of the segments;
   5.2. Build a table with the distances between the expected and the real positions of *y* at each point. Use the linear regression equation to compute the value of the expected $\hat{y}$ value given a co-ordinate *x*;
   5.3. Compute the confidence bands of the linear regression line and remove all points outside the band;
6. Re-apply steps 4 to 6 (recursive algorithm) for each piece of parallel text between two consecutive segments in order to find more correspondence points.

6. Extraction of Translations of Terms

The key issue for the extraction of translation of terms is to find a correlation between the co-occurrences of terms in the aligned texts segments. In general, the more often two terms appear together in aligned segments, the greater the chance they are translations.

The aligner splits the parallel texts into aligned segments. These aligned segments can be used to track the *distribution similarity* of translations. In order to measure how similar two terms are, i.e. to measure whether a term in a particular language is a translation of a term in another language, we use a similarity measure. We start by building a table which counts the number of co-occurrences of two specific terms in the aligned segments. This is called a *contingency table*. The contingency table for the pair 'Επιτροπή των Ευρωπαϊκών Κοινοτήτων' – 'Comissão das Comunidades Europeias' ('*Commission of the European Communities*') is shown below:

N: 162347	(595) 'Επιτροπή των Ευρωπαϊκών Κοινοτήτων'	× 'Επιτροπή των Ευρωπαϊκών Κοινοτήτων'
(601) 'Comissão das Comunidades Europeias'	(a) 499	(b) 102
× 'Comissão das Comunidades Europeias'	(c) 96	(d) 161650

N is the total number of aligned segments. The Greek term occurs in 595 aligned segments whereas the Portuguese term occurs in 601 segments. The table stores the *number of aligned segments* that contain (a) both terms ('Επιτροπή των Ευρωπαϊκών Κοινοτήτων' and 'Comissão das Comunidades Europeias'), (b) the Portuguese term but not the Greek term, (c) the Greek term but not the Portuguese term and (d) neither term.

The difference between the number of occurrences of both terms may result from different translations or from some occasional misalignment. Different translations may be due to syntactic constraints or to alternative translations made by the human translator as shown below:

Greek	Portuguese
Αντιπροσωπεία *της* Επιτροπής των Ευρωπαϊκών Κοινοτήτων	Delegação *da* Comissão das Comunidades Europeias
Τα έξοδα στα οποία υποβλήθηκαν η Γαλλική Κυβέρνηση και η Κυβέρνηση του Ηνωμένου Βασιλείου καθώς και η Επιτροπή	As despesas efectuadas pelos Governos francês e do Reino Unido e pela Comissão das Comunidades Europeias

By using the score provided by the Average Mutual Information, it is possible to identify correct translations for terms across the several languages. A comprehensive analysis of similarity measures was carried out in [10] where, as a conclusion, this similarity measure proved to be appropriate for the task of identifying translation equivalents. The Average Mutual Information is computed as follows:

$$I(X;Y) = \sum_{x=\{0,1\}} \sum_{y=\{0,1\}} p(X=x, Y=y) \log_2\left(\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}\right)$$

where *X* and *Y* are the two terms to be tested as translations. In this formula, p(x=1, y=0) is the probability that term *X* occurs but term *Y* does not.

7. Results

We used the corpus described in section 3 to test our methodology. The texts were aligned and terms were extracted. The table below presents some translations of terms:

English	Greek	Portuguese
JUDGMENT OF THE COURT	ΑΠΟΦΑΣΗ ΤΟΥ ΔΙΚΑΣΤΗΡΙΟΥ	ACÓRDÃO DO TRIBUNAL DE JUSTIÇA
Advocate General	γενικός εισαγγελέας	advogado – geral
On those grounds	Για τους λόγους αυτούς	Pelos fundamentos expostos
Language of the case	Γλώσσα διαδικασίας	Língua do processo
Furthermore	Επιπλέον	Além disso
Commission of the European Communities	Επιτροπή των Ευρωπαϊκών Κοινοτήτων	Comissão das Comunidades Europeias
Member States	κρατών μελών	Estados – membros
Act of Accession	Πράξεως Προσχωρήσεως	Acto de adesão
President of the Chamber	πρόεδρο τμήματος	presidente de secção
First Chamber	πρώτο τμήμα	Primeira Secção

We can see some terms quite characteristic of the domain like ‘Commission of the European Communities’ or ‘Advocate General’. However, in our list, we also got other terms which are not characteristic but which reflect general patterns of language usage like ‘Furthermore’, ‘Επιπλέον’ and ‘Além disso’, or ‘United Kingdom’, ‘Ηνωμένο Βασίλειο’ and ‘Reino Unido’, in English, Greek and Portuguese, respectively. The fact is that the extractor of terms, described in section 4, tends to capture typical sequences of tokens. It is not wise enough to distinguish what terms are specific to the domain and which are of general usage.

Furthermore, we should stress that we also got some problems due to the inflectional nature of the languages. ‘Member States’ had some alternative translations in Greek as ‘κρατών μελών’ or ‘κράτη μέλη’ depending on the case. As a result, their scores were low.

8. Conclusions

In this paper we have presented a methodology to compile a bilingual terminology databank for Greek–English and Greek–Portuguese by extracting translations of terms from parallel texts. It is a language independent methodology which does not recur to heuristics and does not require *a priori* language knowledge. Although the aligner introduced some errors with occasional misalignments of parallel text segments due to different word orders, it was possible to extract reliable translations. Furthermore, we believe that better results could be obtained by lemmatising the texts and/or extracting typical sequences of characters using a methodology similar to the one used to extract the terms. All in all, the results obtained look rather promising.

9. Future Work

We intend to take advantage of the conversion of the Greek character set to the Latin character set in order to have more correspondence points between two texts. By identifying similar words after the conversion, like ‘Demokratía’ (‘Δημοκρατία’) and ‘Democracia’, in Greek and Portuguese, or ‘Noembríou’ (‘Νοεμβρίου’) and ‘November’ in Greek and English, it will be possible to find more correspondence points. This will further improve our results and lead to more accurate translations of terms. There are also some other issues to be analysed more carefully like word inflection and term extraction. Although it is not as hard in English as in Greek or Portuguese, word inflection results in terms which have lower frequency in the texts and which may not be extracted. Moreover, when they are extracted, they may lead to several alternative translations since the correct translation depends on the syntactic constraints. This is clearly a problem that must be analysed and which we wish to look into more detail in the near future.

10. Bibliography

[1] Brown, P., Lai, J. and Mercer, R., Aligning Sentences in Parallel Corpora, 1991, in *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, U.S.A., 169–176.

[2] ELRA (European Language Resources Association), *Multilingual Corpora for Co-operation*, 1997, Disk 2 of 2, Paris, France.

[3] Fung, P. and McKeown, K., Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping, 1994, in *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, U.S.A., 81–88.

[4] Gale, W. and Church, K., Identifying Word Correspondences in Parallel Texts, 1991, in *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, Pacific Grove, California, U.S.A., Morgan Kaufmann, 152–157.

[5] Kay, M. and Röscheisen, M., Text-Translation Alignment, 1993, in *Computational Linguistics*, volume 19, number 1, 121–142.

[6] Piperidis, S., Papageorgiou, H. and Boutsis, S., From Sentences to Words and Clauses, 2000, in Véronis, J. (ed.), *Parallel Text Processing*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 117–138.

[7] Melamed, I., Bitext Maps and Alignment via Pattern Recognition, 1999, in *Computational Linguistics*, volume 25, number 1, 107–130.

[8] Ribeiro A., Lopes G. and Mexia J., Using Confidence Bands for Parallel Texts Alignment, 2000, in *Proceedings of the 38th Conference of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, China, 432–439.

[9] Ribeiro, A., Lopes, G. and Mexia, J., A Self-Learning Method of Parallel Texts Alignment, 2000, in White, J. (ed.), *Envisioning Machine Translation in the Information Future – Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000* – Lecture Notes in Artificial Intelligence, Cuernavaca, Mexico, volume 1934, Springer-Verlag, Berlin, Germany, 30–39.

[10] Ribeiro, A., Lopes, G. and Mexia, J., Extracting Equivalents from Aligned Parallel Texts – Comparison of Measures of Similarity, 2000, in Monard, M. and Sichman, J. (eds.), *Advances in Artificial Intelligence – Proceedings of the International Joint Conference IBERAMIA 2000 / SBIA 2000 – 7th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2000) and the 15th Conference of the Brazilian Society of Artificial*

Intelligence (SBIA 2000) – Lecture Notes in Artificial Intelligence, São Paulo, Brazil, volume 1952, Springer-Verlag, Berlin, Germany, 340–349.

[11] da Silva, J., Dias, G., Guilloré, S. and Lopes, J., Using Localmaxs Algorithms for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, 1999, in Barahona, P. and Alferes, J. (eds.), *Progress in Artificial Intelligence* – Lecture Notes in Artificial Intelligence, volume 1695, Springer-Verlag, Berlin, Germany, 113–132.

[12] Simard, M. and Plamondon, P., Bilingual Sentence Alignment: Balancing Robustness and Accuracy, 1998, in *Machine Translation*, volume 13, number 1, 59–80.

António Ribeiro
PhD Student
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática
Quinta da Torre, P-2829-516 Caparica
Portugal
ambar@di.fct.unl.pt

Gabriel Lopes
Principal Researcher
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática
Quinta da Torre, P-2829-516 Caparica
Portugal
gpl@di.fct.unl.pt

João Mexia
Full Professor
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Matemática
Quinta da Torre, P-2829-516 Caparica
Portugal