DEVELOPING AN ENGLISH-GREEK COMPARABLE CORPUS USING WEB TEXTS

George Mikros, Villy Tsakona, Maria Drakopoulou, Alexandra Koutra,
Evangelia Triantafylli and Sofia Trypanagnostopoulou
University of Athens

1. Introduction

The goal of the paper is to present a project involving the compilation of comparable corpora including web texts in English and Greek. The project has been developed as part of a course in "Introduction to Bilingual Lexicography", in the Interfaculty M.A. Programme "Lexicography: Theory and Applications" of the Faculty of English Studies.[1] The development of the corpus aimed both at training prospective lexicographers in creating and using such resources in their work and at assisting them with projects assigned to them in the course of their post-graduate studies.

In what follows, we will first provide the characteristics of comparable corpora and the advantages of their use in lexicography (section 2) and then present the details of the *English-Greek Comparable Corpus* (henceforth EGCC): the corpus size and content, the compilation procedure, and the metadata gathered for the texts included. Part of the corpus has been used to develop a quantitative method for judging content comparability between English and Greek texts (section 3). In particular, the medical subcorpus was used as a test bed in order to evaluate the suitability of the corpus as a linguistic resource for the extraction of bilingual terminology for lexicographical and educational uses (section 4). Section 5 summarizes the main findings of the study and discusses future prospects.

2. Comparable corpora and lexicography

Lexicography has entered the era of electronic corpora since the beginning of the eighties. Not only do corpora provide lexicographers with evidence on recurring linguistic phenomena, but they also allow them to obtain information on the frequency and the social context of their occurrence (see, among others, Sinclair 2003a, 2003b, Atkins and Rundell 2008, Hanks 2008, Krishamurthy 2008). Bilingual lexicography, in particular, relies on the use of corpora consisting of texts coming from two languages, thus *comparable* corpora come into play:

> A comparable corpus is one which selects similar texts in more than one language or variety. There is as yet no agreement on the nature of the similarity, because there are very few examples of comparable corpora (EAGLES 1996, in Maia 2003).

As a result, the term *comparable corpora* is often used for both corpora consisting of original texts in one language and their translations in one or more languages (see,

[1] The student lexicographers who participated in the project were Giannis Anagnostopoulos, Maria Drakopoulou, Despina Eleftheriou, Alexandra Koutra, Efthalia Krommyda, Maria Marin, Evangelia Triantafylli, Sofia Trypanagnostopoulou, and Georgios Fokas. The project was coordinated by Prof. Eleni Antonopoulou.

among others, Olohan 2002a, 2002b); and corpora consisting of original texts in two or more languages, matched by criteria such as the time of composition, text category, intended audience, etc. Following Granger (2003: 19) and Atkins and Rundell (2008: 476-479), we will refer to the first category as *translation corpora* and to the second one as *comparable corpora*.

Due to their design features, comparable corpora appear to be more suitable for lexicographical use than translation ones.[2] More specifically:

- Comparable corpora include original texts in two or more languages; interference is thus avoided.
- It is much easier to find original texts on a particular subject than to find a pair of texts consisting of the original and a translation. As a result, a larger corpus can be compiled and a greater degree of variety within the corpus can be achieved. Hence, the corpus becomes more reliable.
- It is not always possible to find translations of all genres, either because some of them are not usually translated (e.g. e-mails, chat) or because there are usually more translations in one direction (e.g. from English to Greek) than in another (e.g. from Greek to English).
- Comparable corpora are versatile: besides lexicography, they can be used in a wide range of other research areas, such as discourse analysis, pragmatics, translation and contrastive studies. They also offer wider possibilities for terminology extraction, information retrieval and knowledge engineering than translation corpora (see, among others, Zanettin 1998, Granger 2003, Maia 2003, Bekavak et al. 2004.)

More specifically, comparable corpora can be used by lexicographers for the creation of bilingual terminological databases to assist terminology translation and for the retrieval of information on the collocations and the use of terminology in context. Therefore, the compilation of bilingual specialised dictionaries or even general ones can be based on this kind of corpora (especially if the size and representability of the corpus allows it). Given that the present project was part of a post-graduate programme aiming at training prospective lexicographers, what seems to be equally important is the fact that comparable corpora allow students to improve their skills in creating their own resources, retrieving terminology and collocations typical of specific genres and registers, and producing their own bilingual lemmas.

3. The development of the corpus

3.1 Corpus structure design

EGCC is the first comparable corpus involving Greek texts. It is based on web texts in a variety of topics and genres. More specifically EGCC contains:

- 14 different topics;
- 2 genres (academic, informative).

[2] However, it is interesting to note that translation corpora can be (and have been) used in lexicography. For example, a Greek-English dictionary of collocations has been based on a translation corpus (Sidiropoulou 2008).

The texts collected are written in the time span 2000-2008 and collected in pairs (i.e. a Greek text for every English one, or vice versa). As to the corpus size, our initial aim was to gather 25,000 words for each topic subcorpus, so as to achieve a balanced design. However, at this point the total corpus size is 517,799 words (290,627 English ~ 227,172 Greek) and its quantitative description can be seen in Table 1:

Table 1: EGCC in numbers

	English		**Greek**			
Topic	**Texts**	**Size (in words)**	**Texts**	**Size (in words)**	**Total Texts**	**Total Size (in words)**
Book reviews	15	15403	15	11351	30	26754
Culture	18	15154	18	13318	36	28472
Environment	9	4771	9	7377	18	12148
Hobbies	10	10141	10	12683	20	22824
Internal affairs	10	36779	10	32668	20	69447
Horoscopes	14	1963	14	1706	28	3669
Humanities	3	22042	3	11510	6	33552
International	56	42218	56	28852	112	71070
Medical	20	11102	20	9494	40	20596
Movie reviews	30	25070	30	14455	60	39525
Science	21	27637	21	20862	42	48499
Social	53	38853	53	28442	106	67295
Sports	35	19935	35	17945	70	37880
Technology	15	19559	15	16509	30	36068
Total	**309**	**290627**	**309**	**227172**	**618**	**517799**

The text files are stored in txt format. In order to achieve a systematic recording of the corpus content, specific guidelines are followed. For each text, the following metadata are stored in an excel file:

a. Filename consisting of the following: text topic (3 letter abbreviation) _ genre (3 letter abbreviation) _ counter (1…n) _ text language (2 letter abbreviation) (e.g. spo_inf_12_en.txt, med_aca_4_gr.txt)
b. Language: ‘En’ for English and ‘Gr’ for Greek
c. Title: the text title
d. E-address
e. Publisher
f. Author(s)
g. Date: the date the text was written
h. Topic: ‘boo’ for Book reviews, ‘cul’ for Culture, ‘env’ for Environment, ‘hob’ for Hobbies, ‘hom’ for Internal Affairs, ‘hor’ for Horoscopes, ‘hum’ for Humanities, ‘int’ for International, ‘med’ for Medical, ‘mov’ for Movie reviews, ‘sci’ for Science, ‘soc’ for Social, ‘spo’ for Sports and ‘tec’ for Technology
i. Genre: ‘inf’ for Informative and ‘aca’ for Academic
j. Medium: ‘dig’ for Digital

Given that one of our main aims was to create a corpus to be used as an educational tool for students with no previous experience in natural language processing and corpus linguistics, we decided to give them the opportunity to learn how to work with raw text files which can be explored via the use of concordancers, thus avoiding complicated (and complicating) text preprocessing routines, such as stripping out xml metadata headers. At the same time, the metadata stored (topic, genre, and language) can be exploited to produce various subcorpora, since this kind of information is encoded in the filename of each text. Furthermore, all texts will be easily encoded in TEI using XML using simple scripts, when this corpus becomes available online.

3.2 Testing content similarity

In order to assess the comparability of EGCC from a quantitative perspective, we compared the text size of the members of each text pair. Our research hypothesis was that comparable texts in both languages would have similar size and it was tested by performing a two way ANOVA test with the text size (measured in words) as dependent variable and the language and topic as independent variables. Chart 1 shows the distribution of text size among the two independent variables:

Chart 1: Text size dispersion per topic

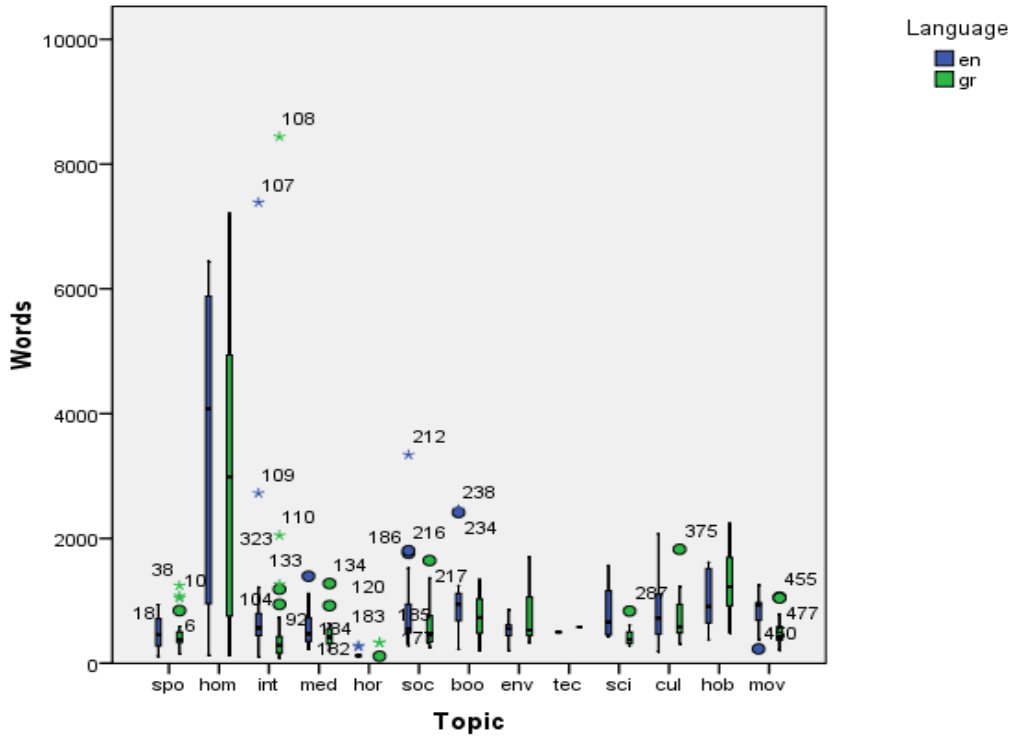


Chart 1 shows considerable variation in text size between topics, but low variation between language pairs. The visual impression is supported by the ANOVA results. The ANOVA is overall significant ($F = 2.63$, $p < 0.05$). The main effect of topic is also found to be statistically significant (4.11, $p < 0.05$), indicating that text size

varies systematically with topic in both languages. However, the main effect of language does not reach statistical significance ($F = 0.73$, $p > 0.05$), which means that the size in English-Greek text pairs is similar. The interaction effect for the language ~ topic variables is also non significant; in other words, *English-Greek text pairs are similar in size within topic subcorpora*.

In order to further assess the quality of the corpus, we set up a small scale experiment testing the *subjectivity* involved in the decisions made by the team of corpus compilers. Using the medical subcorpus of EGCC (see table 1), two raters were asked to grade each comparable text pair for content similarity using a scale from 1 to 10. The distribution of the ratings appears in Chart 2:

Chart 2: Content similarity of the EGCC medical subcorpus as graded by two raters

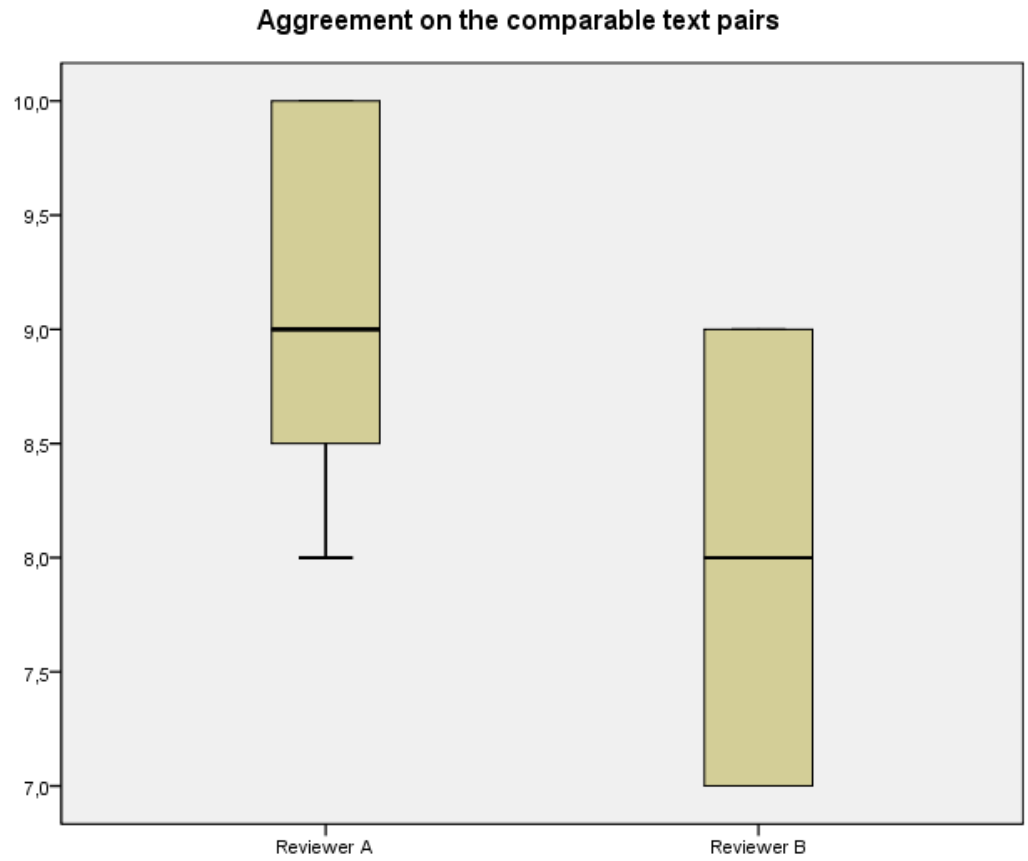


Chart 2 displays the average grade and the distribution of each rater's grades on the quality of the text content comparability. Rater A evaluated the content similarity with 9.2, while Rater B with 8.05. The homogeneity of the raters' grading was further analyzed using Cohen's Kappa which is 0.07. This score is considered a low one and shows that the two raters perceived content comparability differently. This result does not undermine the quality of the present corpus, since each topic has so far been compiled by the same person in both languages. However, we should bear in mind that there is a need for systematic training of corpus compilers before they start compiling texts for a larger comparable corpus, where many compilers would contribute to the *same* topic.

4. Using EGCC for automatic keyword extraction

4.1 Methodology

A comparable corpus such as EGCC can support a wide range of research activities in both language systems (see section 2). For the purposes of the present paper, we decided to use it as a linguistic resource for bilingual keyword extraction using the EGCC medical subcorpus, which consists of 40 files (20 for each language) with a total size of 20,596 words (see table 1).

Our main research hypothesis was that comparable corpora such as the EGCC can support the study of bilingual terminology via the use of simple automated methods, without state of the art keyword detection algorithms. However, this hypothesis assumes that the selected method of automatic terminology extraction provides us with reliable results in monolingual and bilingual term extraction.

Automatic terminology extraction from the EGCC was attained by using the "Keywords" function of *WordSmith Tools*, a well-known and widely used corpus analysis software. This procedure allows us to extract only *single word* candidate terms. Keywords are detected by comparing the frequency wordlist of a text with the frequency wordlist of a large general language corpus (Scott and Tribble 2006). The comparison is evaluated using the Log-Likelihood criterion (henceforth LL method), namely a robust statistical measure widely used in word frequency comparisons (Kilgarriff 1997, Kilgarriff and Rose 1998). The reference wordlist for Greek is generated from the *Hellenic National Corpus* (HNC),[3] while the reference list for English comes from the *British National Corpus* (BNC).

4.2 Evaluation

In order to check whether the automatic extraction of terminology via the "Keywords" function of *WordSmith Tools* is reliable, we also extracted terminology *manually*. We then evaluated:

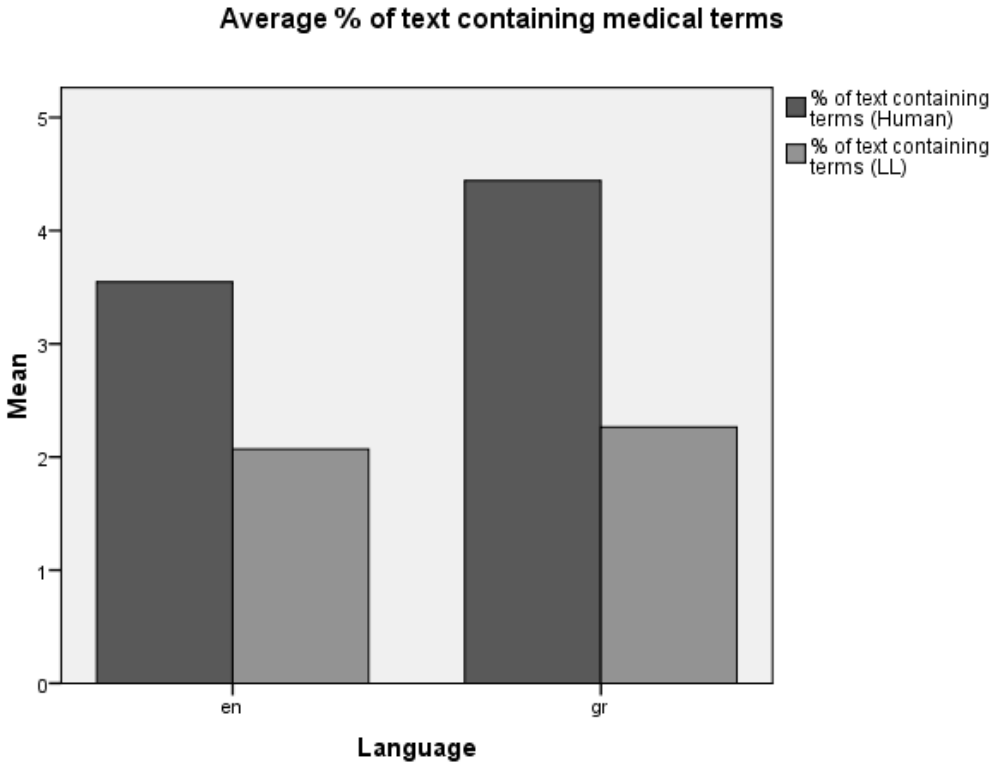
- terminology extraction precision for each language separately: the number of terms extracted using the statistical procedure is divided by the number of terms manually detected.
- bilingual terminology extraction precision: the number of pairs of translated terms extracted via the statistical procedure is divided by the number of existing pairs of translated terms.

[3] HNC is a 33 Mwords general language corpus developed by the *Institute for Language and Speech Processing* (ILSP; see Hatzigeorgiu et al. 2000).

4.3 Results

We compared the average percentage of text containing terms identified by human annotators with the average percentage of text containing terms detected by the automatic procedure, so as to determine whether there is significant difference in term detection between English and Greek texts. The results are presented in Chart 3:

Chart 3: Comparison between the terminology load between the manual and the automated terminology extraction procedure in English and Greek texts



In Chart 3, the average terminology load of the corpus is measured as the average percentage of the text size including medical terms. The dark grey bars represent the terminology load as measured by the human annotators and the light grey bars represent the terminology load as measured by the statistical procedure. Although there is a visual difference between languages in the terminology load measure coming from the annotators, a t-test showed that both human and statistical procedures across languages are statistically non significant. This means that text pairs in both languages share the same terminology load, which can be interpreted as further evidence for good content comparability.

The precision of the automated procedure was evaluated using two different conditions. The first condition relates to the precision obtained in each language separately. The results of this evaluation are presented in Chart 4:

Chart 4: Single language terminology extraction precision

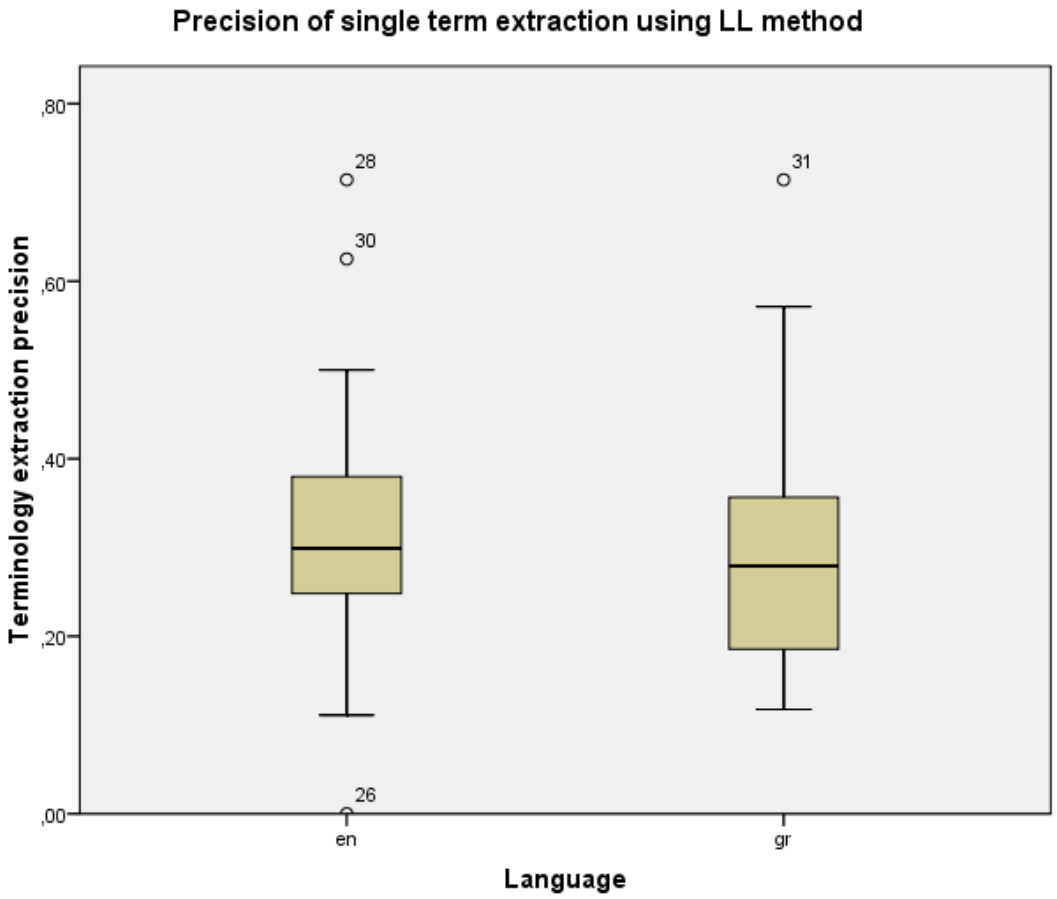


Chart 4 shows the precision of single term extraction using the LL method. This method appears to perform equally well in both languages resulting in an average precision of 0.31 in the English data and 0.30 in the Greek data. This similarity in precision can be interpreted as an indirect indication of the method's robustness.

The second evaluation of the automated method relates to the bilingual terminology extraction precision. The results of this evaluation appear in Chart 5:

Chart 5: The precision of bilingual terminology extraction

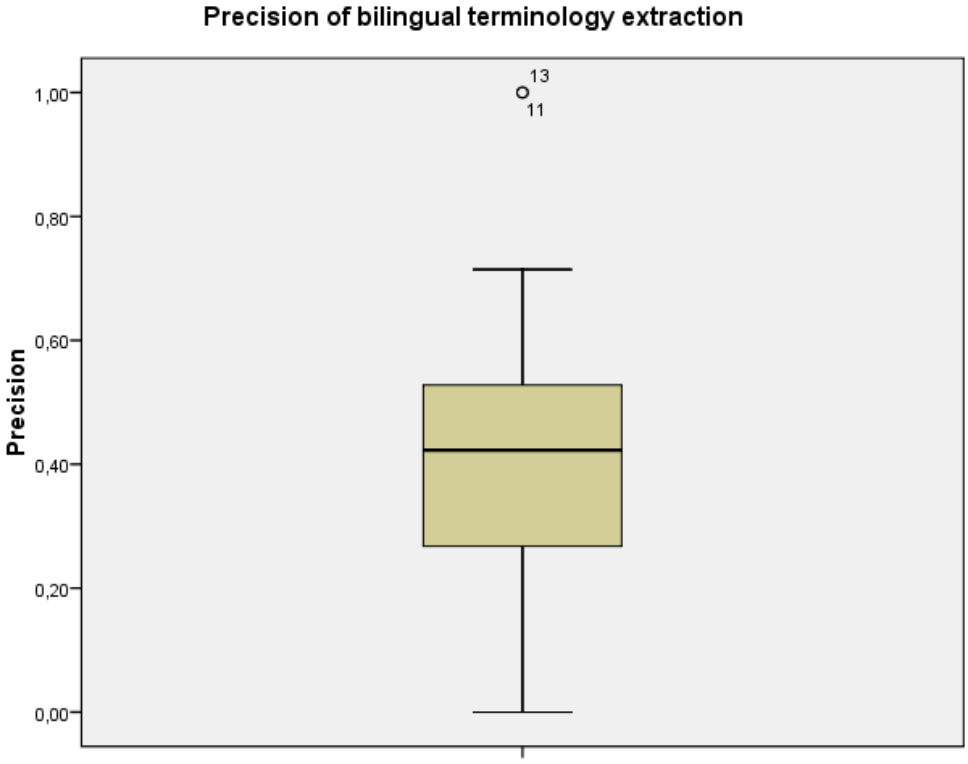


Chart 5 presents the extraction precision of bilingual term pairs. The average is 0.47, which means that nearly half of the bilingual term pairs of the corpus are successfully detected. This result can also be considered quite satisfactory. However, it should be noted that the specific task exhibits considerable variation (see the 5th and 95th percentile of the distribution). In some cases we have zero precision, while in other cases we have 0.7 or even 1, as the two outliers (no 11 and 13) show.

5. Summary

The aim of this paper was to offer a brief description of the EGCC and to explore some quantitative methods for using it to retrieve bilingual terminology.

In order to develop a comparable corpus of English and Greek web texts including a variety of topics and genres, we have defined a sampling frame including 14 different topics and 2 genres. The corpus is compiled from raw texts and so far amounts to 517,799 words and 618 texts (309 for each language). Metadata information has also been stored in a separate database file. The present version of the corpus is designed for off-line use with software tools such as *WordSmith Tools* and *Monoconc*.

A quantitative assessment of the two language subcorpora reveals that both English and Greek texts are of similar size. However, a preliminary investigation of inter-rater agreement in the evaluation of comparable bilingual text pairs shows that the specific task is prone to subjectivity. This result should be taken into serious consideration, especially in case more than one compilers work on the same text topic.

Bilingual single term extraction from the corpus was obtained with satisfactory precision, given the facts that we used only raw textual data without any kind of grammatical annotation and that we employed a statistical keyword extraction method without any language-specific rules.

EGCC is work in progress. Although the current version contains mostly informative and academic texts, an extension of the corpus is planned, so as to include other genres. Furthermore, an on-line version of the corpus is planned, so that it becomes accessible to all researchers and students who would be interested in using it.

Acknowledgments

The authors would like to thank the coordinator of the project Prof. Eleni Antonopoulou for her constant support and her insightful comments on the present paper.

References

Atkins, B. T. S. and Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bekavac, B., Osenova, P., Simov, K. and Tadić, M. 2004. Making monolingual corpora comparable: A case study of Bulgarian and Croatian. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa and R. Silva (eds), *4th International Conference on Language Resources and Evaluation LREC 2004*. Pariz-Lisabon: ELRA, 1187-1190. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.2105 (8 October 2008).

British National Corpus. http://www.sketchengine.co.uk/auth (7 October 2008).

Granger, S. 2003. The corpus approach: A common way forward for Contrastive Linguistics and Translation Studies? In S. Granger, J. Lerot and S. Petch-Tyson (eds), *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam/New York: Rodopi, 17-30.

Hanks, P. 2008. The lexicographical legacy of John Sinclair. *International Journal of Lexicography* 21: 219-229.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari, E., Papageorgiou, H. and Demiros, I. 2000. Design and implementation of the online ILSP Greek Corpus. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, ELRA, 1737-1742. http://www.sdjt.si/bib/lrec00/ps/336.ps (8 October 2008).

Hellenic National Corpus. http://hnc.ilsp.gr (8 October 2008). [in Greek]

Kilgarriff, A. 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of the 5th ACL Workshop on Very Large Corpora*. Beijing and Hong Kong, 231-245. http://acl.ldc.upenn.edu/W/W97/W97-0122.pdf (8 October 2008).

Kilgarriff, A. and Rose, T. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*. Granada, Spain, 46-52. ftp://ftp.itri.bton.ac.uk/reports/ ITRI-98-07.ps (8 October 2008).

Krishnamurthy, R. 2008. Corpus-Driven Lexicography. *International Journal of Lexicography* 21: 231-242.

Maia, B. 2003. What are comparable corpora? In *Proceedings of pre-conference workshop* Multilingual Corpora: Linguistic Requirements and Technical Perspectives, at *Corpus Linguistics 2003*. Lancaster, 27-34. http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc (8 October 2008).

Monoconc. http://www.athel.com/mono.html (8 October 2008).

Olohan, M. 2002a. Leave it out! Using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução* IX : 153-169. http://www.cadernos.ufsc.br/online/cadernos9/maeve.pdf (8 October 2008).

_____2002b. Comparable corpora in translation research: Overview of recent analyses using the *Translational English Corpus*. In *LREC Language Resources in Translation Work and Research Workshop Proceedings*. Las Palmas, Canary Islands, Spain, 27 May-2 June 2002, 5-9. http://www.ifi.uzh.ch/cl/yuste/postworkshop/repository/proceedings.pdf (8 October 2008).

Scott, M. and Tribble, C. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam/Philadelphia: John Benjamins.

Sidiropoulou, M. (ed.) 2008. *Greek-English Dictionary of Political Discourse Collocations*. Athens: Diavlos. [in Greek]

Sinclair, J. 2003a. Corpora for lexicography. In P. van Sterkenburg (ed.), *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, 167-178.

_____2003b. Corpus processing. In P. van Sterkenburg (ed.), *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, 179-193.

WordSmith Tools. http://www.lexically.net/wordsmith (8 October 2008).

Zanettin, F. 1998: Bilingual comparable corpora and the training of translators. *Meta* XLIII: 616-630. http://www.erudit.org/revue/meta/1998/v43/n4/004638ar.pdf (8 October 2008).