Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH)

10 December 2015
Warsaw, Poland

Editors:
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Sponsors

CLARIN-PL
Common Language Resources and Technology Infrastructure

PAN

Zespół Inżynierii Lingwistycznej

ISBN: 978-83-63159-19-1

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warszawa, Poland

http://ipipan.eu/

Preface

The workshop on *Corpus-Based Research in the Humanities* (CRH) is a direct descendant of the workshop on *Annotation of Corpora for Research in the Humanities* (ACRH), which was held three times: in Heidelberg (5.1.2012), Lisbon (29.11.2012), and Sofia (12.12.2013).
All three editions were co-located with the international workshop on *Treebanks and Linguistic Theories* (TLT), a tradition which we continue with CRH.

The new name was motivated by the wish to change the focus slightly, towards corpus-based research in the humanities in general. While the earlier editions focused on questions related to annotation and a number of papers in the current proceedings do so as well, we wanted to visibly broaden the scope of the workshop, as even the earlier editions of the workshop had attracted submissions that did not centre on the question of annotation. In fact, there are many scholars in the humanities who use textual corpora in their everyday work but are not interested in or just do not need to deal with annotation issues. This is partly due to the fact that many corpora still lack linguistic annotation at all, thus requiring scholars to use just the raw text for their research purposes. As our original motivation for initiating the ACRH workshop series was to bring together the often separate communities of (digital) humanities and computational linguistics and to foster communication and collaboration between them, we felt that the focus on annotation in the name of the workshop was undermining our intention by discouraging humanities researchers working with corpora to submit papers.

In addition to changing the name of the workshop, we made several smaller adjustments. First, we included several scholars from digital humanities in the programme committee. While such a mixed committee is not entirely without problems due to different reviewing cultures in digital humanities and computational linguistics, we still believe this a step in the right direction for bringing both communities closer together and assessing submissions from both areas fairly. Second, this year's call asked for long abstracts (up to six pages) rather than full papers. This reflects common practices in the digital humanities better and did help to attract more proposals. Finally, we decided to organise the workshop on a biannual basis instead of an annual one in order to reduce the workload of the organisers and reviewers and avoid competing with too many similar workshops too frequently.

In total we received 17 long abstracts by authors from 12 different countries in Europe and South and North America. Each submission was reviewed independently by three members of the programme committee in a double-blind fashion. After the reviewing process, we accepted 11 submissions. One further submission was moved from TLT to CRH because it was a better fit to the topics of CRH than those of TLT. The overall acceptance rate was 70.6%. This reflects the fact that the average quality of the abstracts was high and most of them received favourable reviews. Another positive observation is that a number of the workshop speakers are promising young scholars.

We hope you will enjoy the workshop and the proceedings and wish to thank all authors who submitted papers, the 19 members of the programme committee, Reinhard Förtsch, who kindly agreed to give the invited talk, and last but not least the local and non-local organisers of TLT-14 and in particular the chair of the local organisation committee, Adam Przepiórkowski.

The CRH Co-Chairs and Organisers
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Program Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Members:
Monica Berti (Germany)
Federico Boschetti (Italy)
David Bouvier (Switzerland)
Neil Coffee (USA)
Lonneke van der Plas (Malta)
Dag Haug (Norway)
Neven Jovanovic (Croatia)
Mike Kestemont (Belgium)
John Lee (Hong Kong)
Alexander Mehler (Germany)
Roland Meyer (Germany)
Willard McCarty (UK)
John Nerbonne (The Netherlands)
Bruce Robertson (Canada)
Neel Smith (USA)
Uwe Springmann (Germany)
Melissa Terras (UK)
Sara Tonelli (Italy)
Martin Wynne (UK)

Organising Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Göttingen, Germany)

Local Committee:
Adam Przepiórkowski (chair)
Michał Ciesiołka
Konrad Gołuchowski
Mateusz Kopeć
Katarzyna Krasnowska
Agnieszka Patejuk
Marcin Woliński
Alina Wróblewska

Contents

The Digital Fragmenta Historicorum Graecorum and the Ancient Greek-Latin Dynamic Lexicon

Tariq Yousef and Monica Berti

Alexander von Humboldt Lehrstuhl für Digital Humanities
Institut für Informatik - Universität Leipzig
E-mail: name.surname@uni-leipzig.de

Abstract

This paper describes a model that provides training data for a word alignment system that will be used to identify the translation relationships among the words in the parallel texts (Greek/Latin) of the bilingual corpus of the Digital Fragmenta Historicorum Graecorum (DFHG).

1 Introduction

Statistical machine translation uses alignment models to extract and identify translation correspondences between words and phrases in two parallel texts in two different languages. The aligned pairs of words or phrases are used as training data for machine translation systems.

The goal of this paper is to provide training data for a word alignment system and these data will be used to identify the translation relationships among the words in the parallel texts (Greek/Latin) of the bilingual corpus of the Digital Fragmenta Historicorum Graecorum (DFHG). The biggest challenge is that large digitized Ancient Greek/Latin lexica are publicly unavailable. The only available dictionary is Glosbe, which contains about 1600 Ancient Greek/Latin phrases entered by users[1]. Glosbe is an unreliable source, because anyone can update the dictionary without supervision. Moreover, the size of the database is not large enough to build an alignment system only relying on it. This paper investigates a simple and effective method for automatic bilingual lexicon extraction (Ancient Greek/Latin) from the available aligned bilingual texts (Ancient Greek/English and Latin/English) produced by the Dynamic Lexicon project of the Perseus Digital Library.

[1] Glosbe contains thousands of dictionaries for every existing pair of languages: www.glosbe.com/grc/la

2 The DFHG Corpus and the Dynamic Lexicon

The DFHG is a project of the Alexander von Humboldt Chair of Digital Humanities at the University of Leipzig that is producing a digital edition of the five volumes of the Fragmenta Historicorum Graecorum edited by Karl Müller in the 19th century[2]. This corpus includes extracts from ancient sources that preserve quotations and text reuses of Greek authors and works that are now lost. More than 600 fragmentary authors are collected in the volume and the sources range from the 6th century BC through the 7th century CE. The content is arranged by author and the volumes provide scholars with the Greek texts of the fragments (or Latin texts when the witnesses are Latin) and their modern Latin translations produced by Müller. Introductions and commentaries are in Latin [2]. The Dynamic Lexicon is a project of the Perseus Digital Library to automatically create bilingual dictionaries (Greek/English and Latin/English) using parallel texts (source texts in Greek or Latin aligned with their English translations) along with the syntactic data encoded in treebanks[3]. The final goal is to enrich the Perseus Dynamic Lexicon with Greek/Latin pairs and to extend the work also to other sources beyond the fragmentary ones.

3 Previous Work

There are many approaches to construct bilingual lexica by using a third language (usually English). Tanaka and Umemura [8] uses an Inverse Consultation (IC) method to produce a Japanese/French lexicon using English as a bridge language. Ács [1] extends the IC method up to 53 pivot languages to improve the accuracy of the lexicon, which relies on the fact that pairs found via several intermediate languages are more correct. Bond [3] uses semantic classes along with an intermediate language to produce Japanese/Malay dictionary. Paik [7] improves a method (multi-pivot criterion) to produce a Korean/Japanese lexicon using English as an intermediate language and shared Chinese characters among Japanese and Korean words. Noisy translations are a big problem and therefore Kaji [4] introduces distributional similarity (DS) as a measure to avoid noisy translations produced by triangulation. In the next sections we introduce our proposed approach to produce Ancient Greek/Latin lexicon via English as a bridge language, and JACCARD Index as a similarity metric to measure the quality of translation pairs in order to eliminate noisy translations.

4 Proposed Approach

The starting point of our approach is to provide as much parallel texts as possible to extract all possible translation candidates. The Perseus Digital Library con-

[2] http://www.dh.uni-leipzig.de/wo/dfhg/

[3] http://nlp.perseus.tufts.edu/lexicon/

tains approximately 10.5 million words of Latin source texts, 13.5 million words of Greek, and 44.5 million words of English. The texts are all public-domain materials that have been scanned, OCR'd, and formatted into TEI-compliant XML [4]. The Perseus Digital Library contains at least one English translation for most of its Latin and Greek prose and poetry texts. Our Corpus consists of 163 parallel documents aligned at the word level (104 Ancient Greek/English documents and 59 Latin/English documents). The Greek/English dataset consists approximately of 210 thousand sentence pairs with 4.32 million Greek words, whereas the Latin/English dataset consists approximately of 123 thousand sentence pairs with 2.33 million Latin words. The parallel texts are aligned on a sentence level using Moore's Bilingual Sentence Aligner [6], which aligns the sentences with a very high precision (one-to-one alignment). The Giza++ toolkit is used to align the sentence pairs at the level of individual words.

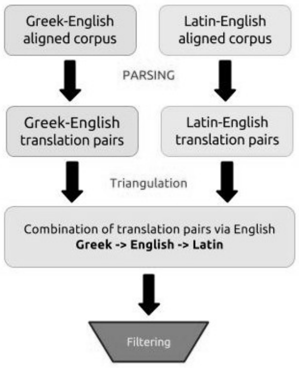
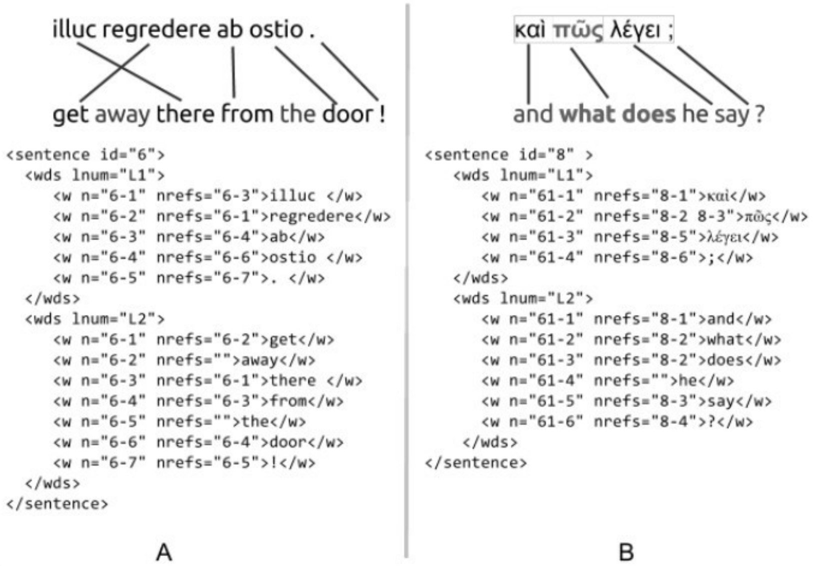

Greek-English aligned corpus
Latin-English aligned corpus
PARSING
Greek-English translation pairs
Latin-English translation pairs
Triangulation
Combination of translation pairs via English
Greek -> English -> Latin
Filtering

Figure 1: Explanation of the method

4.1 Preprocessing

In this stage we are going to parse the data sets we have in XML format (Fig. 2). Each document has a Perseus-id and consists of sentences in the original language



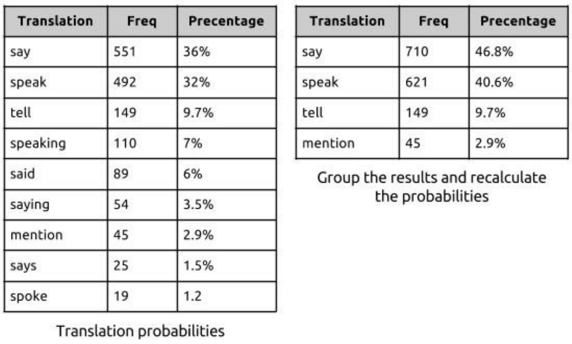
```
<sentence id="6">
  <wds lnum="L1">
     <w n="6-1" nrefs="6-3">illuc </w>
     <w n="6-2" nrefs="6-1">regredere</w>
     <w n="6-3" nrefs="6-4">ab</w>
     <w n="6-4" nrefs="6-6">ostio </w>
     <w n="6-5" nrefs="6-7">. </w>
  </wds>
  <wds lnum="L2">
     <w n="6-1" nrefs="6-2">get</w>
     <w n="6-2" nrefs="">away</w>
     <w n="6-3" nrefs="6-1">there </w>
     <w n="6-4" nrefs="6-3">from</w>
     <w n="6-5" nrefs="">the</w>
     <w n="6-6" nrefs="6-4">door</w>
     <w n="6-7" nrefs="6-5">!</w>
  </wds>
</sentence>
```

```
<sentence id="8" >
   <wds lnum="L1">
      <w n="61-1" nrefs="8-1">καὶ</w>
      <w n="61-2" nrefs="8-2 8-3">πῶς</w>
      <w n="61-3" nrefs="8-5">λέγει</w>
      <w n="61-4" nrefs="8-6">;</w>
   </wds>
   <wds lnum="L2">
      <w n="61-1" nrefs="8-1">and</w>
      <w n="61-2" nrefs="8-2">what</w>
      <w n="61-3" nrefs="8-2">does</w>
      <w n="61-4" nrefs="">he</w>
      <w n="61-5" nrefs="8-3">say</w>
      <w n="61-6" nrefs="8-4">?</w>
    </wds>
</sentence>
```

Figure 2: The aligned sentences in XML format

[4] https://github.com/PerseusDL

(Ancient Greek or Latin) and its translation in English (Fig. 2A). Each Latin or Greek word is aligned to one word in the English text (one-to-one Alignment), but in some cases a word in the original language can be aligned to many words (one-to-many / many-to-one) or not aligned at all (Fig. 2B). Lemmatization of English translations will produce better results, because that will reduce the number of translation candidates as we can see in this example: The Greek word λέγειν is translated with "say", "speak", "tell", "speaking", "said", "saying", "mention", "says", "spoke". Many of the translation candidates share the same lemma (*say* for "said", "saying", "says"), (*speak*, "speaking", "spoken"). Before the lemmatization there were nine translation candidates and after the lemmatization there are only four candidates, showing therefore the change of frequencies.



Translation	Freq	Precentage
say	551	36%
speak	492	32%
tell	149	9.7%
speaking	110	7%
said	89	6%
saying	54	3.5%
mention	45	2.9%
says	25	1.5%
spoke	19	1.2

Translation probabilities

Translation	Freq	Precentage
say	710	46.8%
speak	621	40.6%
tell	149	9.7%
mention	45	2.9%

Group the results and recalculate the probabilities

Figure 3: Lemmatization of English translations

4.2 Triangulation

Triangulation is based on the assumption that two expressions are likely to be translations if they are translations of the same word in a third language. We will use triangulation to extract the Greek/Latin pairs via English. In order to do that, we query our datasets to get the Greek and Latin words that share the same English translation along with their frequencies. See Figure 4.

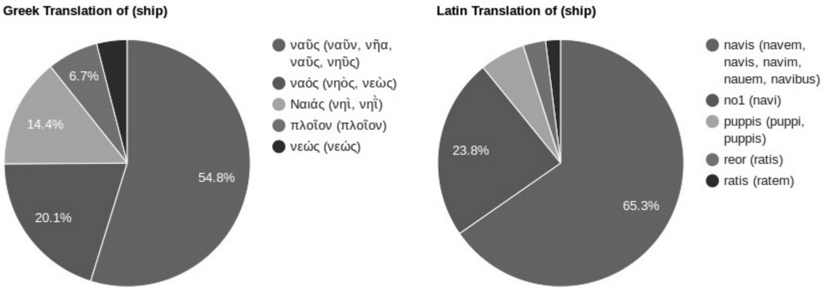


Figure 4: An example of triangulation

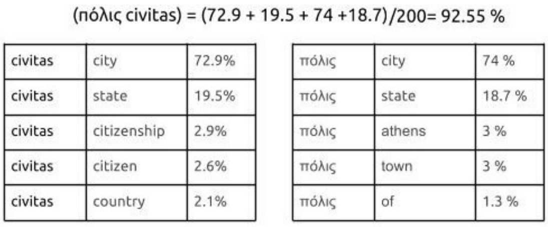
The English word *ship* is translated with the Greek word ναῦς (54.8%), with ναός (21.5%) and so on; the same English word ship is translated with the Latin word navis (65.3%), with no (23.8%), and so on. The extracted pairs via triangulation are the following (ναῦς, navis), (ναῦς, no), (ναός, navis), (ναός, no). These pairs don't have the same level of relatedness, therefore we have to filter the results to keep only strong related pairs.

4.3 Translation-Pairs filtering

The translation pairs are not completely correct, because there are still some translation errors. In order to eliminate incorrect pairs, we will use a similarity metric to measure the similarity or the relatedness between every Greek/Latin pairs. The Jaccard coefficient [5] measures the similarity between finite sample sets (in our case two sets), and it is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

A and B in equation 2 are two vectors of translation probabilities (Greek/English, Latin/English). For example, the relatedness between the Greek word πόλις and the Latin word *civitas* is reported in figure 5



(πόλις civitas) = (72.9 + 19.5 + 74 +18.7)/200= 92.55 %

civitas	city	72.9%
civitas	state	19.5%
civitas	citizenship	2.9%
civitas	citizen	2.6%
civitas	country	2.1%

πόλις	city	74 %
πόλις	state	18.7 %
πόλις	athens	3 %
πόλις	town	3 %
πόλις	of	1.3 %

Figure 5: Use of Jaccard algorithm

5 Evaluation

The quality evaluation of translations candidates extracted by the proposed method is done manually with the help of humanists. We have randomly selected 200 translation pairs obtained via the proposed method with different frequencies (high and low) and different JACCARD Co values. Each pair should be assigned into one of four categories: Correct, small difference, big difference and incorrect. We employed the mean reciprocal rank (MRR) [9] to assess the performance. We assigned each category a score (Reciprocal Rank) (Table 1).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|} RR_i \tag{2}$$

Category	**Reciprocal Rank (RR)**
Correct	1
Small difference	0.75
Big difference	0.25
Incorrect	0

Table 1: Group the results and recalculate the probabilities

We have to determine a threshold to classify the translation pairs as accepted or not accepted. High threshold yields high accuracy lexicon but with less number of entries, whereas low threshold produces more translation pairs with lower accuracy, as we can see in the table above (Table 2).

Jaccard Co	**0.60 <**	**0.70 <**	**0.80 <**	**0.90 <**
# Pairs	200	150	100	50
MRR	%61.25	%74	%87.50	%94.50

Table 2: MRR Scores

6 Conclusion

The proposed method is language-independent and it can be used to build a bilingual lexicon between any language pairs with aligned corpora that share a pivot language. The accuracy of the method depends on two factors: **1) The size of the aligned-parallel corpora** plays an important role to improve the accuracy of the lexicon: bigger corpora produce better translation probability distribution and more translation candidates which yield a more accurate lexicon, and they also cover more words; **2) The quality of the aligner** used to align the parallel corpora: manually aligned corpora yield more accurate results, whereas automatic alignment tools produce some noisy translations; in our case Giza++ has been used to align the parallel corpora.

References

[1] Judit Ács (2014). Pivot-based Multilingual Dictionary Building using Wiktionary. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland

[2] Berti, M. et al. "The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors". In *Journal of the Text Encoding Initiative 8 (2014-2015)* (selected papers from the 2013 TEI Conference) doi 10.4000/jtei.1218

[3] Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. *Design and construction of a machine-tractable Japanese Malay dictionary*. In *MT Summit VIII*, pp. 53– 58, Santiago de Compostela, Spain

[4] Kaji, H., Tamamura, S., and Erdenebat, D. (2008). Automatic construction of a Japanese-Chinese dictionary via English. In *LREC, volume 2008*, pp. 699–706.

[5] Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(140), pp. 241–272.

[6] Moore, Robert. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of 5th Conference of the Association for Machine Translation* in the Americas, pp. 135–244.

[7] Kyonghee Paik, Francis Bond, and Satoshi Shirai. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *NLPRS-2001* Tokyo, Japan, pp. 63–70.

[8] Tanaka, K., Umemura, K. 1994. Construction of a bilingual dictionary intermediated by a third language, *Proceedings of COLING- 94*, pp. 297-303.

[9] Voorhees, E.M. 1999. The trec-8 question answering track report. In *Proceedings of the 8 th Text Retrieval Conference*.

[10] Yousef, Tariq. 2015. Word Alignment and Named-Entity Recognition applied to Greek Text Reuse. *MSc's Thesis. Alexander von Humboldt Lehrstuhl für Digital Humanities, Universität Leipzig*.