Ancient Greek WordNet meets the Dynamic Lexicon: the example of the fragments of the Greek Historians

Monica Berti
Gregory R. Crane
Tariq Yousef
Institute of Computer Science,
University of Leipzig
Leipzig, Germany
{name.surname}@uni-leipzig.de

Yuri Bizzoni
Federico Boschetti
Riccardo Del Gratta
CNR-ILC "A. Zampolli",
Via Moruzzi 1
Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

The Ancient Greek WordNet (AGWN) and the Dynamic Lexicon (DL) are multilingual resources to study the lexicon of Ancient Greek texts and their translations. Both AGWN and DL are works in progress that need accuracy improvement and manual validation. After a detailed description of the current state of each work, this paper illustrates a methodology to cross AGWN and DL data, in order to mutually score the items of each resource according to the evidence provided by the other resource. The training data is based on the corpus of the Digital Fragmenta Historicorum Graecorum (DFHG), which includes ancient Greek texts with Latin translations.

1 Introduction

The Ancient Greek WordNet (AGWN) and the Dynamic Lexicon (DL), which will be illustrated in detail in the next sections (see sections 2 and 4), are complementary resources to study the Ancient Greek lexicon. AGWN is based on the paradigmatic axis provided by bilingual dictionaries, while DL is based on the syntagmatic axis provided by historical and literary texts aligned to their scholarly translations. Both of them have been created automatically and they need to be corrected and extended. In this specific case the data is taken from the Digital Fragmenta Historicorum Graecorum (DFHG), which is a corpus of quotations and text reuses of ancient Greek lost historians and their Latin translations provided by the editor Karl Müller (Berti et al., 2014 2015; Yousef, 2015)[1]. This corpus is part of LOFTS (Leipzig Open Fragmentary Texts Series) at the Humboldt Chair of Digital Humanities at the University of Leipzig. We have been using this collection because it is big enough to include many different sources preserving information about Greek historians. Instead of working with extant authors, the DFHG allows us to focus on specific topics related to ancient Greek lost historiography and on the language of text reuse within this domain. The working hypothesis is that the evidence provided by Dynamic Lexicon Greek - Latin pairs is relevant to score the Greek word - conceptual node (synset) associations in the Ancient Greek WordNet and, on the other hand, that the evidence provided by AGWN Greek word - Latin translations is relevant to score the DL Greek - Latin pairs.

2 Ancient Greek WordNet

The creation of the Ancient Greek WordNet has been outlined in (Bizzoni et al., 2014). It is based on digitized Greek-English bilingual dictionaries (in particular the Liddell-Scott-Jones and the Middle Liddell provided by the Perseus Project[2]): first, Greek-English pairs (Greek words and English translations) are extracted from the dictionaries; then, the English word is projected onto the Princeton WordNet (PWN) (Fellbaum, 1998). If the English word is in PWN, then its synsets are assigned to the Greek word; the same goes for its lexical relations with other lemmas. Thus AGWN is created "bootstrapping" data from different datasets. As a bootstrapped process, its result is quite inaccurate. For example, induced polysemy (from English) maps the Greek verb ἔχω *-échō-* over 170 English words (including "cut", "make", "brake" ...). On the contrary, when the English word is not in PWN, the Greek word of the pair is excluded from AGWN, thus strongly reducing the coverage of AGWN for the entire Greek lexicon to c.a 30%.

[1] http://opengreekandlatin.github.io/dfhg-dev/

[2] http://www.perseus.tufts.edu

Currently, AGWN is linked not only to PWN, but also to other WordNets, in particular to the Latin WordNet (LWN) (Minozzi, 2009) and to the Italian WordNet (IWN) (Roventini et al., 2003). The way these WordNets are interconnected follows the guidelines illustrated in (Vossen, 1998; Rodríguez et al., 2008), by using English as the bridge language. As a consequence, Greek and Latin and/or Greek and Italian are linked through the common sense(s) in English.

3 The conceptual structure of Ancient Greek WordNet

Sharing a unique conceptual network among different languages is a good solution when the civilizations expressed by those languages are very similar, due to the effects of the globalization. In this case, only few conceptual nodes must be inserted when a concept is lexicalized in the source language but not in the target language, and few nodes must be deactivated when a concept is only lexicalized in the target language, but not in the source language.

On the contrary, when the civilizations expressed by the source and the target languages are highly dissimilar, the conceptual network needs to be heavily restructured.

As illustrated in the introduction, the conceptual network of AGWN is originally based on PWN, but the glosses of the synsets and the semantic relations can be modified through a web interface.[3]

4 Dynamic Lexicon

The Dynamic Lexicon is an increasing multilingual resource constituted by bilingual dictionaries (Greek/English, Latin/English, Greek/Latin), which have been created through the direct automated alignment of original texts with their translations or through a triangulation with a bridge language.

The first version of the DL[4] is a National Endowment for the Humanities (NEH)[5] co-funded project developed at Tufts University (Medford, MA) by the Perseus Project, whereas the second version is under development at the University of Leipzig by the Open Philology. Project[6]

[3] http://www.languagelibrary.eu/new_ewnui

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

[5] http://www.neh.gov/about

[6] http://www.dh.uni-leipzig.de

5 Bilingual Dictionary Extraction

This section investigates a simple and effective method for automatic extraction of a bilingual lexicon (Ancient Greek/Latin) from the available aligned bilingual texts (Greek/English and Latin/English) in the Perseus Digital Library using English as a bridge language.

The data comes from the corpus of the DFHG and consists of 163 parallel documents aligned at a word level (104 Ancient Greek/English files and 59 Latin/English). The Greek-English dataset consists approximately of *210K* sentence pairs with *4,32M* Greek words, whereas the Latin-English dataset consists approximately of *123K* sentence pairs with *2,33M* Latin words. The parallel texts are aligned on a sentence level using Moore's Bilingual Sentence Aligner (Moore, 2002), which aligns the sentences with a very high precision (one-to-one alignment).[7] Then the GIZA++ toolkit[8] is used to align the sentence pairs at the level of individual words. Table 1 introduces statistics about the DFHG parallel corpus, while Figure 1 displays the used workflow. Note that the number of words in Table 1 is the total number of words in the documents, whereas the aligned pairs are the number of aligned words in the documents. Some words are not aligned at all, therefore the number of aligned words is smaller than the total number of words.

	Ancient Greek	Latin
Files	104	59
Sentences	$210K$	$132K$
Words	$4,32M$	$2,33M$
Aligned words	$3,34M$	$1,71M$
Distinct words	$872K$	$575K$

Table 1: Size of the corpora.

5.1 Preprocessing

The data sets provided by the workflow in Figure 1 are available in XML format. Each document is identified (through an *id*) in the Perseus Digital Library and consists of sentences in the orig-

[7] Sentences have been segmented using punctuation marks excluding commas.

[8] GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team at the Center for Language and Speech Processing at Johns-Hopkins University.

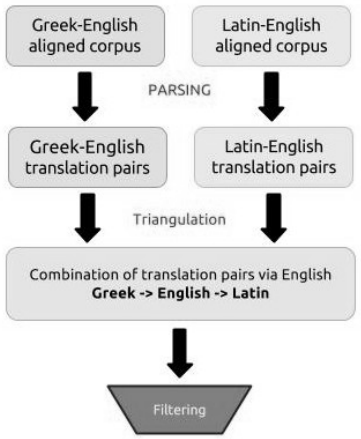

Figure 1: Explanation of the method

inal language (Ancient Greek or Latin) and their translation in English, as reported in Figure 2 (A). 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 could be aligned to many words (one-to-many / many-to-one) or not aligned at all, cf. Figure 2 (B).

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 *-légein-* 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 9 translation candidates and after the lemmatization there are only four candidates, showing therefore the change of frequencies.

Table 2 shows how the lemmatization process recalculates the frequencies and percentages of each single translation.

5.2 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 3.

The English word *ship* is associated to the Greek word ναῦς *-naûs-* (54.8%), to ναός *-naós-* (21.5%) and so on; the same English word *ship* is associated to the Latin word *navis* (65.3%), to *no* (23.8%), and so on.

Lemma	Freq.	%	Word	Freq.	%
say	719	46.8	say	551	36
			said	89	6
			saying	54	3.5
			says	25	1.5
speak	621	40.6	speak	492	32
			speaking	110	7
			spoke	19	1.2
tell	149	9.7	tell	149	9.7
mention	45	2.9	mention	45	2.9

Table 2: Lemmas and words:frequencies and percentages

The extracted pairs via triangulation are the correct association {ναῦς, *navis*} and the wrong associations {ναῦς, *no*} (*ship-to swim*), {ναός, *navis*} (*temple-ship*), {ναός, *no*} (*temple-to swim*). These pairs don't have the same level of relatedness, therefore we have to filter the results to keep only strong related pairs, as exposed in Section 5.3.

5.3 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 (Jaccard, 1901) measures the similarity between finite sample sets (in our case two sets), and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \quad (1)$$

A and B in equation 1 are two vectors of translation probabilities (Greek-English, Latin-English). For example, the relatedness[9] between the Greek word πόλις and the Latin word *civitas* is reported in Figure 4.

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 produce more translation pairs with lower accuracy. The accuracy of the method depends on two factors:

[9]In the calculation we use the fact that *city* and *state* are shared English translation between πόλις *-pólis-* and *civitas*

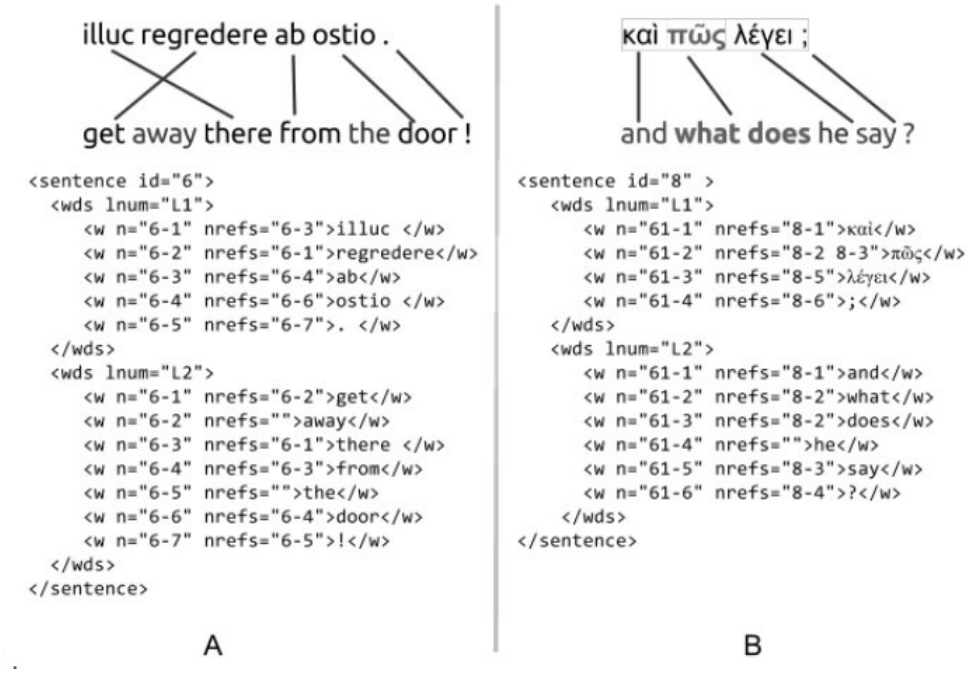

Figure 2: The aligned sentences in XML format

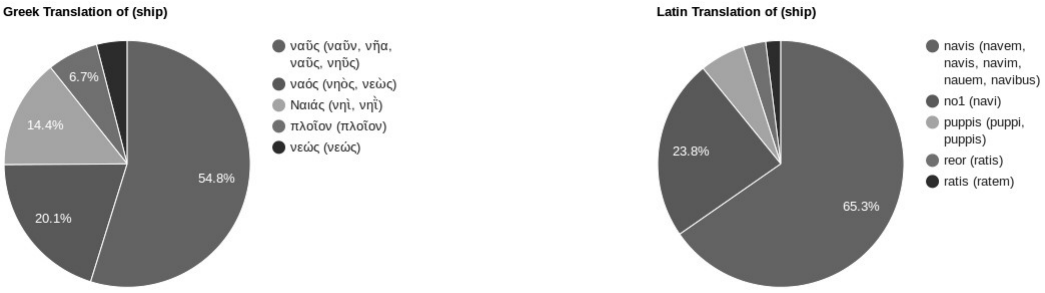

Figure 3: An example of triangulation

(πόλις civitas) = (72.9 + 19.5 + 74 +18.7) = 185.1
(πόλις civitas) = (100 + 100)=200
J(πόλις , civitas) = 185.1/200= 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 4: Use of Jaccard algorithm for aligning πόλις to *civitas*

The size of aligned-parallel corpora plays important role to improve the accuracy of the produced lexicon: bigger corpora produce better translation probability distribution and more translation candidates which yield a more accurate lexicon. In addition to that bigger corpora cover more words

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.

6 Evaluating and extending the AGWN through evidence provided by the Dynamic Lexicon and vice versa

Students and scholars that evaluate and extend the AGWN synset items need to compare online dictionaries and other lexical resources. The DL can provide evidence for this purpose, especially to discover relevant missing correspondences. An example should clarify.

In AGWN we can find the association *minister* (eng) / *minister* (lat) / διάκτορος *-diáktoros-* (grc), but not *minister* (eng) / *minister* (lat) / διάκονος *-diákonos-* (grc), which is instead provided by the DL. If we consult the bilingual dictionary Liddell-Scott-Jones, we find out that διάκτορς "taken as minister, =διάκονος". The automatic parser used to bootstrap AGWN from bilingual dictionaries has not processed this information, so the DL provides a hint for the integration of this missed item in the correct synset of AGWN.

Complementary, the DL is missing the triplet *minister* (eng) / *minister* (lat) / διάκτορος (grc), which would be a relevant translation, even if not attested by the aligned bilingual texts of the training corpus. Moreover, AGWN can be used to add scoring criteria to the DL system, by tuning the results with a further piece of evidence, which reinforces the Jaccard score.

For example, the score of the correct association {ναῦς, *navis*}, discussed in Section 5.2 is reinforced, due to its presence in AGWN, whereas the scores of the wrong associations {ναῦς, *no*}, {ναός, *navis*} and {ναός, *no*} are weakened, due to their absence in AGWN.

7 Future work

The next step is the creation of a gold standard both for AGWN and for DL, in order to quantify the gain in terms of precision and recall that we can obtain by crossing AGWN and DL data.

8 Conclusion

In conclusion, we think that the paradigmatic approach, by extraction of bilingual pairs from dictionaries, and the syntagmatic approach, by extraction of bilingual pairs from aligned texts, are complementary for the study of Ancient Greek semantics and that they can be integrated, in order to mutually improve the performances of both of them.

References

Monica Berti, Bridget Almas, David Dubin, Greta Franzini, Simona Stoyanova, and Gregory Crane. 2014-2015. The Linked Fragment: TEI and the Encoding of Text Reuses of Lost Authors. *Journal of the Text Encoding Initiative*, 8.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The Making of Ancient Greek WordNet. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, MA, USA.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Stefano Minozzi. 2009. The Latin WordNet Project. In Peter Anreiter and Manfred Kienpointner, editors, *Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik*, volume 137 of *Innsbrucker Beiträge zur Sprachwissenschaft*, pages 707–716.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, M. Antonia Martí, William Black, Sabri Elkateb, James Kirk, Piek Vossen, and Christiane Fellbaum. 2008. Arabic Wordnet: Current State and Future Extensions. In *Proceedings of the Fourth International Global WordNet - Conference – GWC 2008*, pages 387–406, January.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. ItalWordNet: building a large semantic database for the automatic treatment of italian. *Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI*, 2:745–791.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Tariq Yousef. 2015. Word Alignment and Named-Entity Recognition applied to Greek Text Reuse, school = Alexander von Humboldt Lehrstuhl für Digital Humanities, Universität Leipzig. Master's thesis.