



ALEXANDRU IOAN CUZA
UNIVERSITY of IAȘI



Proceedings of the Eighth Global WordNet Conference

Editors:

Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen

Bucharest, Romania, January 27-30, 2016

ISBN 978-973-0-20728-6

Preface

This eighth meeting of the international Wordnet community coincides with the 15th anniversary of the Global WordNet Association and the 30th anniversary of the Princeton WordNet. We are delighted to welcome old and new colleagues from many countries and four continents who construct wordnets, ontologies and related tools, as well as colleagues who apply such resources in a wide range of Natural Language Applications or pursue research in lexical semantics.

The number of wordnets has risen to over 150 and includes – besides all the major world languages – many less-studied languages such as Albanian and Nepali. Wordnets have become a principal tool in computational linguistics and NLP, and *wordnet*, *SemCor* and *synset* have entered the language as common nouns. Coming together and sharing some of the results of our work is an important part of the larger collaborative effort to better understand both universal and particular properties of human languages.

Many people have donated their time and effort to make this meeting possible: the review committee, the local organizers and their helpers (Eric CUREA, Maria Mitrofan, Elena Irimia), our sponsors (PIM, QATAR Airways, Oxford University Press), EasyChair and our host, the Romanian Academy. Above all, thanks go to you, the contributors, for traveling to Bucharest to present your work, listen and discuss.

January, 2016

Bucharest

Christiane Fellbaum, Corina Forăscu,

Verginica Mititelu, Piek Vossen

Table of contents

Preface	i
Program and Organising Committees	vi
The Awful German Language: How to cope with the Semantics of Nominal Compounds in GermaNet and in Natural Language Processing <i>Erhard Hinrichs</i>	1
Adverbs in Sanskrit Wordnet <i>Tanuja Ajotikar and Malhar Kulkarni</i>	2
Word Sense Disambiguation in Monolingual Dictionaries for Building Russian WordNet <i>Daniil Alexeyevsky and Anastasiya V. Temchenko</i>	10
Playing Alias - efficiency for wordnet(s) <i>Sven Aller, Heili Orav, Kadri Vare and Sirli Zupping</i>	16
Detecting Most Frequent Sense using Word Embeddings and BabelNet <i>Harpreet Singh Arora, Sudha Bhingardive and Pushpak Bhattacharyya</i>	22
Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary <i>Martin Benjamin</i>	27
Ancient Greek WordNet Meets the Dynamic Lexicon: the Example of the Fragments of the Greek Historians <i>Monica Berti, Yuri Bizzoni, Federico Boschetti, Gregory R. Crane, Riccardo Del Gratta and Tariq Yousef</i>	34
IndoWordNet::Similarity- Computing Semantic Similarity and Relatedness using IndoWordNet <i>Sudha Bhingardive, Hanumant Redkar, Prateek Sappadla, Dharendra Singh and Pushpak Bhattacharyya</i>	39
Multilingual Sense Intersection in a Parallel Corpus with Diverse Language Families <i>Giulia Bonansinga and Francis Bond</i>	44
CILI: the Collaborative Interlingual Index <i>Francis Bond, Piek Vossen, John McCrae and Christiane Fellbaum</i>	50
YARN: Spinning-in-Progress <i>Pavel Braslavski, Dmitry Ustalov, Mikhail Mukhin and Yuri Kiselev</i>	58
Word Substitution in Short Answer Extraction: A WordNet-based Approach <i>Qingqing Cai, James Gung, Maochen Guan, Gerald Kurlandski and Adam Pease</i>	66
An overview of Portuguese WordNets <i>Valeria de Paiva, Livia Real, Hugo Gonalo Oliveira, Alexandre Rademaker, Cl�udia Freitas and Alberto Sim�es</i>	74
Towards a WordNet based Classification of Actors in Folktales <i>Thierry Declerck, Tyler Klement and Antonia Kostova</i>	82

Extraction and description of multi-word lexical units in plWordNet 3.0 <i>Agnieszka Dziob and Michał Wendelberger</i>	87
Establishing Morpho-semantic Relations in FarsNet (a focus on derived nouns) <i>Nasim Fakoornia and Negar Davari Ardakani</i>	92
Using WordNet to Build Lexical Sets for Italian Verbs <i>Anna Feltracco, Lorenzo Gatti, Elisabetta Jezek, Bernardo Magnini and Simone Magnolini</i>	100
A Taxonomic Classification of WordNet Polysemy Types <i>Abed Alhakim Freihat, Fausto Giunchiglia and Biswanath Dutta</i>	105
Some strategies for the improvement of a Spanish WordNet <i>Matias Herrera, Javier Gonzalez, Luis Chiruzzo and Dina Wonsever</i>	114
An Analysis of WordNet's Coverage of Gender Identity Using Twitter and The National Transgender Discrimination Survey <i>Amanda Hicks, Michael Rutherford, Christiane Fellbaum and Jiang Bian</i>	122
Where Bears Have the Eyes of Currant: Towards a Mansi WordNet <i>Csilla Horváth, Ágoston Nagy, Norbert Szilágyi and Veronika Vincze</i>	130
WNSpell: a WordNet-Based Spell Corrector <i>Bill Huang</i>	135
Sophisticated Lexical Databases - Simplified Usage: Mobile Applications and Browser Plugins For Wordnets <i>Diptesh Kanojia, Raj Dabre and Pushpak Bhattacharyya</i>	143
A picture is worth a thousand words: Using OpenClipArt library for enriching IndoWordNet <i>Diptesh Kanojia, Shehzaad Dhuliawala and Pushpak Bhattacharyya</i>	149
Using Wordnet to Improve Reordering in Hierarchical Phrase-Based Statistical Machine Translation <i>Arefeh Kazemi, Antonio Toral and Andy Way</i>	154
Eliminating Fuzzy Duplicates in Crowdsourced Lexical Resources <i>Yuri Kiselev, Dmitry Ustalov and Sergey Porshnev</i>	161
Automatic Prediction of Morphosemantic Relations <i>Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Tsvetana Dimitrova and Maria Todorova</i>	168
Tuning Hierarchies in Princeton WordNet <i>Ahti Lohk, Christiane Fellbaum and Leo Vohandu</i>	177
Experiences of Lexicographers and Computer Scientists in Validating Estonian Wordnet with Test Patterns <i>Ahti Lohk, Heili Orav, Kadri Vare and Leo Vohandu</i>	184
African WordNet: A Viable Tool for Sense Discrimination in the Indigenous African Languages of South Africa <i>Stanley Madonsela, Mampaka Lydia Mojapelo, Rose Masubelele and James Mafela</i>	192

An empirically grounded expansion of the supersense inventory <i>Hector Martinez Alonso, Anders Johannsen, Sanni Nimb, Sussi Olsen and Bolette Pedersen</i>	199
Adverbs in plWordNet: Theory and Implementation <i>Marek Maziarz, Stan Szpakowicz and Michal Kalinski</i>	209
A Language-independent Model for Introducing a New Semantic Relation Between Adjectives and Nouns in a WordNet <i>Miljana Mladenović, Jelena Mitrović and Cvetana Krstev</i>	218
Identifying and Exploiting Definitions in Wordnet Bahasa <i>David Moeljadi and Francis Bond</i>	226
Semantics of body parts in African WordNet: a case of Northern Sotho <i>Mampaka Lydia Mojapelo</i>	233
WME: Sense, Polarity and Affinity based Concept Resource for Medical Events <i>Anupam Mondal, Dipankar Das, Erik Cambria and Sivaji Bandyopadhyay</i>	242
Mapping and Generating Classifiers using an Open Chinese Ontology <i>Luis Morgado Da Costa, Francis Bond and Helena Gao</i>	247
IndoWordNet Conversion to Web Ontology Language (OWL) <i>Apurva Nagvenkar, Jyoti Pawar and Pushpak Bhattacharyya</i>	255
A Two-Phase Approach for Building Vietnamese WordNet <i>Thai Phuong Nguyen, Van-Lam Pham, Hoang-An Nguyen, Huy-Hien Vu, Ngoc-Anh Tran and Thi-Thu-Ha Truong</i>	259
Extending the WN-Toolkit: dealing with polysemous words in the dictionary-based strategy <i>Antoni Oliver</i>	265
A language-independent LESK based approach to Word Sense Disambiguation <i>Tommaso Petrolito</i>	273
plWordNet in Word Sense Disambiguation task <i>Maciej Piasecki, Paweł Kędzia and Marlena Orlńska</i>	280
plWordNet 3.0 -- Almost There <i>Maciej Piasecki, Stan Szpakowicz, Marek Maziarz and Ewa Rudnicka</i>	290
Open Dutch WordNet <i>Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen and Piek Vossen</i>	300
Verifying Integrity Constraints of a RDF-based WordNet <i>Alexandre Rademaker and Fabricio Chalub</i>	309
DEBVisDic: Instant Wordnet Building <i>Adam Rambousek and Ales Horak</i>	317
Samāsa-Kartā: An Online Tool for Producing Compound Words using IndoWordNet <i>Hanumant Redkar, Nilesh Joshi, Sandhya Singh, Irawati Kulkarni, Malhar Kulkarni and Pushpak Bhattacharyya</i>	322

Arabic WordNet: New Content and New Applications <i>Yasser Regragui, Lahsen Abouenour, Fettoum Krieche, Karim Bouzoubaa and Paolo Rosso</i>	330
Hydra for Web: A Browser for Easy Access to Wordnets <i>Borislav Rizov and Tsvetana Dimitrova</i>	339
Towards a methodology for filtering out gaps and mismatches across wordnets: the case of plWordNet and Princeton WordNet <i>Ewa Rudnicka, Wojciech Witkowski and Łukasz Grabowski</i>	344
Folktale similarity based on ontological abstraction <i>Marijn Schraagen</i>	352
The Predicate Matrix and the Event and Implied Situation Ontology: Making More of Events <i>Roxane Segers, Egoitz Laparra, Marco Rospocher, Piek Vossen, German Rigau and Filip Ilievski</i>	360
Semi-Automatic Mapping of WordNet to Basic Formal Ontology <i>Selja Seppälä, Amanda Hicks and Alan Ruttenberg</i>	369
Augmenting FarsNet with New Relations and Structures for verbs <i>Mehrnoush Shamsfard and Yasaman Ghazanfari</i>	377
High, Medium or Low? Detecting Intensity Variation Among polar synonyms in WordNet <i>Raksha Sharma and Pushpak Bhattacharyya</i>	384
The Role of the WordNet Relations in the Knowledge-based Word Sense Disambiguation Task <i>Kiril Simov, Alexander Popov and Petya Osenova</i>	391
Detection of Compound Nouns and Light Verb Constructions using IndoWordNet <i>Dhirendra Singh, Sudha Bhingardive and Pushpak Bhattacharyya</i>	399
Mapping it differently: A solution to the linking challenges <i>Meghna Singh, Rajita Shukla, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia and Pushpak Bhattacharyya</i>	406
WordNet-based similarity metrics for adjectives <i>Emiel van Miltenburg</i>	414
Toward a truly multilingual GlobalWordnet Grid <i>Piek Vossen, Francis Bond and John McCrae</i>	419
This Table is Different: A WordNet-Based Approach to Identifying References to Document Entities <i>Shomir Wilson, Alan Black and Jon Oberlander</i>	427
WordNet and beyond: the case of lexical access <i>Michael Zock and Didier Schwab</i>	436
Author index	445

Program Committee

Eneko Agirre, University of the Basque Country, Spain

Eduard Barbu, Translated.net, Italy

Francis Bond, Nanyang Technological University, Singapore

Sonja Bosch, University of South Africa, South Africa

Alexandru Ceașu, Euroscript Luxembourg S.à r.l., Luxembourg

Dan Cristea, Alexandru Ioan Cuza University of Iași, Romania

Agata Cybulska, VU University Amsterdam / Oracle Corporation, the Netherlands

Tsvetana Dimitrova, Institute for Bulgarian Language, Bulgaria

Marieke van Erp, VU University Amsterdam, the Netherlands

Christiane Fellbaum, Princeton University, USA

Darja Fiser, University of Ljubljana, Slovenia

Antske Fokkens, VU University Amsterdam, the Netherlands

Corina Forăscu, Alexandru Ioan Cuza University of Iași & RACAI, Romania

Ales Horak, Masaryk University, Czech Republic

Florentina Hristea, University of Bucharest, Romania

Shu-Kai Hsieh, Graduate Institute of Linguistics at National Taiwan University, Taiwan

Radu Ion, Microsoft, Ireland

Hitoshi Isahara, Toyohashi University of Technology, Japan

Ruben Izquierdo Bevia, VU University Amsterdam, the Netherlands

Kaarel Kaljurand, Nuance Communications, Austria

Kyoko Kanzaki, Toyohashi University of Technology, Japan

Svetla Koeva, Institute for Bulgarian Language, Bulgaria

Cvetana Krstev, University of Belgrade, Serbia

Margit Langemets, Institute of the Estonian Language, Estonia

Bernardo Magnini, Fondazione Bruno Kessler, Italy

Verginica Mititelu, RACAI, Romania

Sanni Nimb, Society for Danish Language and Literature, Denmark

Kemal Oflazer, Carnegie Mellon University, Qatar

Heili Orav, University of Tartu, Estonia

Karel Pala, Masaryk University, Czech Republic

Adam Pease, IPsoft, USA

Bolette Pedersen, University of Copenhagen, Denmark

Ted Pedersen, University of Minnesota, USA

Maciej Piasecki, Wroclaw University of Technology, Poland

Alexandre Rademaker, IBM Research, FGV/EMAp Brazil

German Rigau, University of the Basque Country, Spain

Horacio Rodriguez, Universitat Politècnica de Catalunya, Spain

Shikhar Kr. Sarma, Gauhati University, India

Roxane Segers, VU University Amsterdam, the Netherlands

Virach Sornlertlamvanich, SIIT, Thammasart University, Thailand

Dan Ștefănescu, Vantage Labs, USA

Dan Tufiș, RACAI, Romania

Gloria Vasquez, Lleida University, Spain

Zygmunt Vetulani, Adam Mickiewicz University, Poland

Piek Vossen, VU University Amsterdam, the Netherlands

Additional Reviewers

Anna Feltracco

Filip Ilievski

Vojtech Kovar

Egoitz Laparra

Simone Magnolini

Zuzana Neverilova

Adam Rambousek

Conference Chairs

Christiane Fellbaum, Princeton University, USA

Piek Vossen, VU University Amsterdam, the Netherlands

Organising Committee

Verginica Mititelu, RACAI, Romania

Corina Forăscu, Alexandru Ioan Cuza University of Iași & RACAI, Romania

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 {name.surname}@ilc.cnr.it

Yuri Bizzoni
Federico Boschetti
Riccardo Del Gratta
CNR-ILC “A. Zampolli”,
Via Moruzzi 1
Pisa - Italy

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 Project²): 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 ἐχῶ -*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 DL⁴ 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. Project⁶

³http://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++ toolkit⁸ 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-

⁷Sentences have been segmented using punctuation marks excluding commas.

⁸GIZA++ 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.

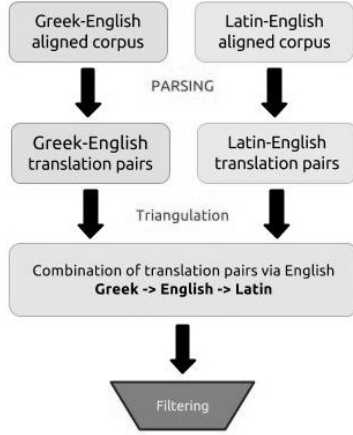


Figure 1: Explanation of the method

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 *-lēgein-* 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*

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

(23.8%), and so on.

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:

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

A and B in equation 1 are two vectors of translation probabilities (Greek-English, Latin-English). For example, the relatedness⁹ 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:

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

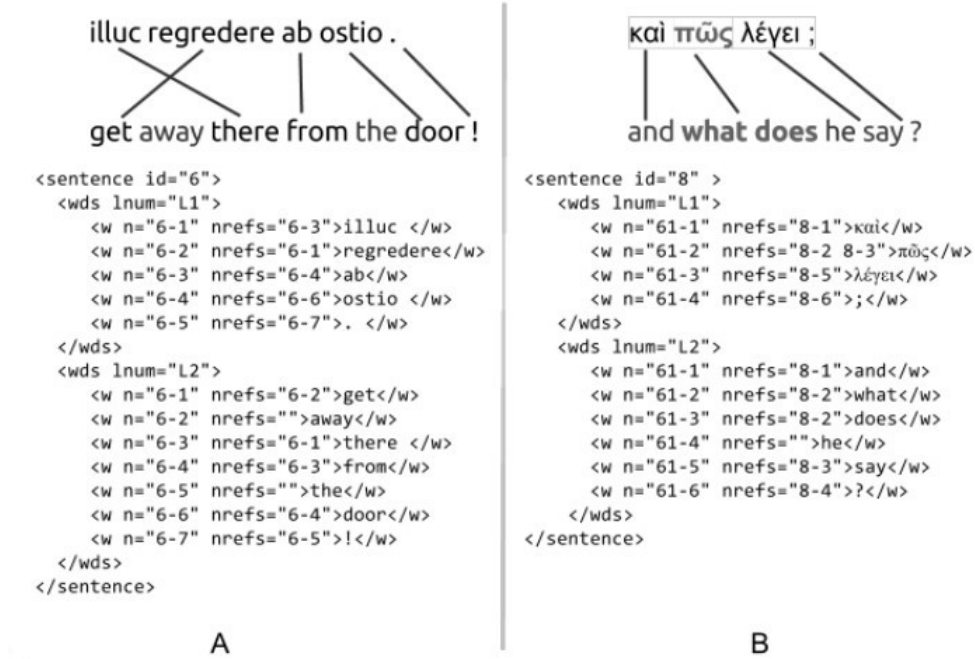


Figure 2: The aligned sentences in XML format

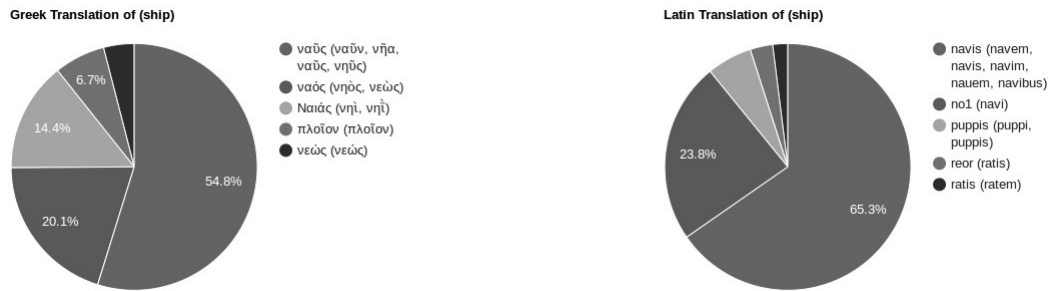


Figure 3: An example of triangulation

$(\text{πόλις civitas}) = (72.9 + 19.5 + 74 + 18.7) = 185.1$
 $(\text{πόλις civitas}) = (100 + 100) = 200$
 $J(\text{πόλις, civitas}) = 185.1/200 = 92.55\%$

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

allel 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.

The quality of the aligner used to align the par-

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 *διάκτορος* “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. Internationalen 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.