The Making of Coptic Wordnet

Laura Slaughter,[♠] Luis Morgado da Costa,[♦] So Miyagawa,[♣]
Marco Büchler,[♣] Amir Zeldes,[♥] Hugo Lundhaug,[♠] Heike Behlmer [♣]

[♠] University of Oslo, Norway
[♦] Nanyang Technological University, Singapore
[♣] Georg-August-Universität Göttingen, Germany
[♥] Georgetown University, USA

Abstract

With the increasing availability of wordnets for ancient languages, such as Ancient Greek and Latin, gaps remain in the coverage of less studied languages of antiquity. This paper reports on the construction and evaluation of a new wordnet for Coptic, the language of Late Roman, Byzantine and Early Islamic Egypt in the first millenium CE. We present our approach to constructing the wordnet which uses multilingual Coptic dictionaries and wordnets for five different languages. We further discuss the results of this effort and outline our on-going/future work.

1 Introduction

This paper reports on the process of constructing a wordnet(WN) for the Coptic language. Coptic belongs to the Egyptian branch of the Afroasiatic language family, spoken in Egypt mainly in the first millennium CE and written in an extended form of the Greek alphabet (see Section 1.2). Together with its precursor Ancient Egyptian written in Hierolgyphic, Hieratic and Demotic scripts, Coptic forms part of the longest continuously documented language on Earth, spanning over four millennia. Despite its importance for historical and comparative linguistics, as well as ancient history, Coptic remains comparatively low in digital resources when compared to contemporary languages of the Ancient and Early Medieval Mediterranean such as Latin and Ancient Greek. With the recent launch of an open source Coptic Dictionary Online (Feder et al., 2018) with an interface for human reading, this project aims to follow with the next logical step in machine readable resources for Coptic: providing a wordnet for the language, which will also be the first wordnet for the Egyptian branch of the Afroasiatic languages.

Wordnet projects aim to provide a machine-tractable lexical resource for automated processing of texts. The purpose of a wordnet for the Coptic language is in the first instance to support digital scholarship on the language. The Coptic language has fewer lexical resources than Greek and Latin and the manuscripts written in Coptic (mainly between the 4th and 12th centuries CE) have received less attention, meaning there is much room for studying their transmission history, an effort that can benefit from a wordnet, for example in recognizing non-verbatim textual reuse.

In this paper, we will present our work on constructing the Coptic Wordnet and outline the goals for this on-going project, as well as an evaluation of its current coverage.

1.1 Background

A number of wordnets already exist for ancient languages: Ancient Greek (Bizzoni et al., 2014, AGWN), Latin (Minozzi, 2009), Sanskrit (Kulkarni et al., 2010), Middle Ancient Chinese (Zhang et al., 2014, MidacWN), and Pre-Qin Ancient Chinese (Zhang et al., 2017, PQACWN). Constructing a wordnet can be extremely time-consuming when done manually, so most wordnets are bootstrapped using another existing wordnet which is referred to as the "pivot language"; usually this is done using the English language Princeton WordNet (Fellbaum, 1998, PWN). The bootstrapping approach to construction is called the "expansion" approach and manual construction is referred to as the "merge" approach (Vossen, 1998). The ancient language wordnets listed above were all boot-

strapped using PWN with the exception of Sanskrit which used the Hindi Wordnet (Bhattacharyya, 2017). Latin Wordnet used two wordnets as pivot languages, Italian WN (Bhattacharyya, 2017) and PWN.

There are both advantages and disadvantages to using pivot languages to bootstrap new wordnets (Bond et al., 2016). One primary advantage is that the 'expand' approach provides immediate multilingual links. The disadvantage of the approach is that the concepts which are not in the pivot language(s) cannot be expressed and are omitted until they are added manually. This problem could be exacerbated for ancient languages since concepts that were expressed in ancient times can lack modern-day equivalents. Conversely, linking to modern terminology can result in a connection to a modern idea that is misleading or has no relevance. Some synsets in modern wordnets do not fit ancient living environments such as those having to do with modern science and technology. This particular challenge is covered in the paper describing Ancient Greek Wordnet issues concerning modern concepts that evolved from ancient concepts (Bizzoni et al., 2014).

Due to the limited number of contexts attested in ancient languages, we expect not to cover a hierarchy of terms as rich as the one that can be seen in modern language resources. To illustrate, PWN has over 10 levels of hypernyms, including terms available to discuss the taxonomy of "sheep" using modern rank-based scientific classification. Many of these categories are informed by the modern understanding of biology, as we have the benefit of scientific contributions impacting how we talk about the world, starting with Linnaeus' work on taxonomies in the 1750s. In the ancient world, we do not have evidence that words were available to cover all of these levels, differentiating, e.g. between placental mammals, monotremes, and marsupials. This issue surrounding the hierarchies is addressed in the paper describing the construction of the Sanskrit wordnet (Kulkarni et al., 2010), which points to the challenge of traditional Sanskrit texts on philosophy and medicine containing many discussions on ontological categories and hierarchies that differ from those in the modern Hindi wordnet.

Even though we see that the issues presented above could provide motivation for choosing the "merge" approach, the immediate multilingual links do provide the needed resources to applications and research within the Digital Humanities, particularly with an aim to study the relationship between Coptic texts and parallel or contemporary texts in other ancient languages. In addition, using a pivot language (such as English, through PWN) is an intermediate step to link directly to the Collaborative Interlingual Index (Bond et al., 2016, CILI), which allows concepts to link across languages without necessarily subscribing to any one wordnet's hierarchy.

1.2 The Coptic Language

The Coptic language is the last stage of the Egyptian language which has been recorded in writing for more than 5,000 years. Pre-Coptic Egyptian language was the vehicle of the culture, politics and religions of the Ancient Egyptian civilization and written in three scripts: Hieroglyphic, Hieratic and Demotic (the latter from 700 BCE).

After the conquest of Egypt in 332 BCE, the Egyptian language borrowed a considerable number of words from Ancient Greek. As early as the 1st and 2nd centuries CE, there had been attempts to write the Egyptian language with the Greek alphabet.

From the 2nd-3rd century, writing the Egyptian language with the Greek alphabet and several Demotic phonograms became common and standardized. This writing system is now known as the Coptic alphabet, and a variant of the Egyptian language which is written in this alphabet is called the Coptic language. The major Coptic dialects include: Sahidic, Boharic, Fayyumic, Mesokemic, Akhmimic, and Lycopolitan. The current version of the Coptic WN contains only the Sahidic dialect, which was the main vehicle of Coptic literature in the first millennium CE and is often considered the 'classical' form of the language. However, there are plans are to extend it to include other dialects in the future. This dialect was chosen primarily based on immediate research needs for processing text reuse cases.

Typologically, Coptic departs from earlier synthetic (highly inflectional) Middle Egyp-

tian, and more analytic (or periphrastic) Late Egyptian, developing instead an agglutinative morphology, in which pronouns and auxiliaries are fused to associated verbs, substantially complicating morphological analysis and the ability to recognize variant forms of Coptic words in running text. The language also allows object incorporation into verbs (similar to English forms such as 'to name-call', but much more frequent), as well as fusion of Greek-origin and native Egyptian lexical items (Grossman, 2014).

There is generally no word division in Coptic writing (*scripto continua*) in Late Antiquity, though modern conventions spell Coptic with spaces between word groups known as bound groups. A bound group contains a content lexeme that is usually a noun or a verb, along with clitic articles, auxiliaries, prepositions and object or possessor pronouns. Coptic is a head-initial, Subject-Verb-Object (SVO), in which nouns carry grammatical gender (M/F), and adjectival senses are generally supplied by nouns ('person of wisdom' means 'wise person') or verbs (e.g. for color terms, a verb meaning 'become white' or 'be white'), with a very small closed class of lexical adjectives remaining from older Egyptian.

As of April 2019, there are 22,777 known Coptic sources (e.g. fragments, codices, epigraphical items, etc.) indexed by the Trismegistos database.[1] The effort to digitize these sources is still on-going and the volume of available digitized text is steadily growing. While most Coptic manuscripts are still waiting to be digitized, a number of projects/sites are contributing to this effort, including: Coptic Scriptorium (Schroeder and Zeldes, 2016), the Corpus dei Manoscritti Coptic Letterari [2], the St. Shenouda the Archimandrite Coptic Society, the Editio Critica Maior of the Greek New Testament,[3] the Digital Edition of the Coptic Old Testament[4], the Marcion project[5], and the Marc Multilingue project[6].

[1] http://www.trismegistos.org/
[2] http://www.cmcl.it/
[3] https://www.uni-muenster.de/intf/ecm.html
[4] http://coptot.manuscriptroom.com/
[5] http://marcion.sourceforge.net/
[6] http://www.safran.be/marcmultilingue/

1.3 Motivation

Like most other wordnets, the motivation behind this project is to perform automatic analysis of texts, including: classic uses in NLP, word similarity tasks, classification of texts, and enhancing the performance of information retrieval. One of the major motivations behind the construction of the Coptic wordnet in particular was to use the hierarchies for text reuse in TRACER (Büchler et al., 2014), but applications for searching and hyperlemmatization using senses (discussed further in Kučera (2007)) are conceivable as well. The currently available NLP pipeline for Coptic (Zeldes and Schroeder, 2016) already offers lemmatization to base dictionary entries, but automatically linking word forms to wordnet entries could make comparisons of automatically analyzed texts to existing texts in Coptic, as well as other languages with aligned wordnets, much easier.

2 Methods

Our automated method for building a new wordnet requires two main types of resources: (1) bilingual dictionaries or any other source providing candidate lemmas aligned with translations, and (2) matching wordnets, sharing a common structure – PWN, in our case. Ideally there should be at least one high coverage wordnet for each of the languages that candidate lemmas are aligned to. Unfortunately, we know that this is rarely the case, and different languages have wordnets of different sizes, which can be a bottleneck for our automated method.

2.1 Dictionaries

The lemma alignments for Coptic were extracted from three sources: the Coptic Dictionary Online (Feder et al., 2018, CDO)[7], Marcion's dictionary[8], and a subset of data from the Database and Dictionary of Greek Loan Words in Coptic (DDGLC)[9] to which we were granted access, and which contains Greek loan words used in Coptic and their respective translations/definitions in English. Both the CDO and Marcion are based on Crum's Coptic

[7] https://coptic-dictionary.org/
[8] http://marcion.sourceforge.net
[9] https://www.geschkult.fu-berlin.de/en/e/ddglc

dictionary (Crum, 1939). The CDO provides trilingual translations in English, French, and German. Less is known about the construction of Marcion, however, which provides translations in English, Czech and Greek.

A summary of the number of Coptic lemmas and the number of translations available in each language is provided in Table 1. These numbers include several preprocessing steps of cleaning and splitting data (e.g. translations often contained multiple lemmas separated by commas or semicolons that were split; parenthetical notes were removed; etc.).

2.2 Wordnets

Concerning the second type of resources, wordnets, we were fortunate to be able to find resources for all languages available in our translations. The automated process (see Section 2.3, below) was done in two stages. For the first stage we collated wordnet data for English, Greek, Czech, German and French from multiple sources, namely: the Princeton Wordnet (Fellbaum, 2017), GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), the Open German Wordnet[10], WOLF: Wordnet Libre du Français (Sagot and Fišer, 2008), and the Greek Wordnet (Stamou et al., 2004). In addition, data for these five languages was also collected from the Extended Open Multilingual Wornet (Bond and Foster, 2013, OMW), which offers automatically collected linked-data from Wiktionary and the Unicode Common Locale Data Repository (CLDR), and from the English subset of the NTUMC Wordnet (Tan and Bond, 2014; Seah and Bond, 2014; Morgado da Costa and Bond, 2016), which includes a few thousand new senses for English, including pronouns, exclamatives and number of other basic senses missing from the Princeton Wordnet.

All this data was linked through a locally built copy of the OMW, linking all wordnets through the structure of the Princeton Wordnet. Table 2 shows the number of senses available for each language in the small multilingual wordnet built for this project, at Stage I and Stage II of the building process.

The second stage of the construction of the Coptic WN consisted of applying the same method over an improved collection of data. This included both better preprocessing of the dictionary data and the addition of two new wordnets to the local multilingual wordnet used for the automated construction: the Ancient Greek Wordnet (Bizzoni et al., 2014) and an unreleased open and improved version of the Czech Wordnet (Pala and Smrž, 2004). Although technically different languages (with different language codes), the Ancient Greek Wordnet and the Greek Wordnet were merged into a single 'Greek' lexicon to facilitate the linking process. Table 2 shows that the addition of these two wordnets significantly boosted the number of available senses for both Greek and Czech which, in turn, helped to produce an improved version of the Coptic WN (see Section 3).

2.3 Automated Construction Method

Our method follows the basic assumptions of the expansion approach, leveraging on the structure of the Princeton Wordnet as reference, but gathering new senses through a naive algorithm inspired by the idea of multilingual sense intersection (Bonansinga and Bond, 2016; Bond and Bonansinga, 2015) to determine potential senses of a new wordnet.

The idea of multilingual sense intersection has a simple logical foundation. Through this approach, the semantic space of a polysemous word in any language can be constrained by aligned translations of the same word in other languages. This technique has been used for Word Sense Disambiguation (WSD) of parallel text, and words alignments across an increasing number languages have been shown to incrementally constrain the semantic space of a word. Figure 1 shows a conceptualization of this logic, for three languages.

In our case, instead of parallel text (which often requires statistical methods to produce word alignments), we use the word-aligned dictonary data produced between Coptic and the five other languages mentioned above: English, Greek, French, German, and Czech (see section 2.1).

The data produced by this technique can be sorted in multiple ways. One of the most meaningful ways to sort this data is by the number of languages that suggest any given

[10] https://github.com/hdaSprachtechnologie/odenet

Resource	Coptic	English	Greek	Czech	German	French
Marcion	7,069	15,748	9,674	13,726	-	-
CDO	4,362	10,021	-	-	10,021	10,435
DDGCL	4,850	9,227	4,854	-	-	-

Table 1: Lemma Alignments by Resource

Language	Senses (Stage I)	Senses (Stage II)
Czech	16,079	**63,198**
English	209,787	209,787
French	130,420	130,420
German	145,420	145,420
Greek	37,765	**114,383**

Table 2: Wordnet Senses

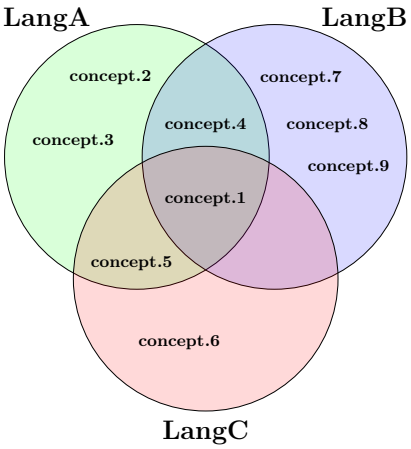

Figure 1: Sense Intersection

concept (i.e. in Figure 1 *concept.1* would be suggested by three languages, while *concept.4* and *concept.5* would be suggested by alignments in two languages). Concepts suggested by more languages have, empirically, a higher likelihood of being correct.

Within concepts suggested by the same number of languages, the algorithm we used employs other metrics to rank candidates: number of individual lemmas matched in each language; part-of-speech congruency, ambiguity of each lemma, and lemma-concept saturation level (i.e. for each concept being suggested, what percentance of lemmas was seen to inform the same concept, per language). This algorithm also performs some language specific string normalization (removal of the infinitival 'to'; removal of determiners preceeding nouns such as 'a' or 'the', case normalization – i.e. for English but not for German).

The development of this system is still ongoing and a full description of its workings is outside the scope of this paper.

2.4 Output and Data Sampling

The output of our system is exemplified in Table 3. In addition to the columns shown in Table 3, the system also outputs a sum-score of multiple other checks mentioned above. Each result row shows, in order, a reserved space for the human validation, the number of languages used to inform this result, the lemma that will be added to the candidate concept, all the translations that were matched to the candidate concept, the PWN offset of the candidate concept and English lemmas, definitions and examples, provided by the PWN.

Two Coptic scholars examined 300 rows (i.e. senses) from our results, with the goal of clarifying the true relationship between the scoring assigned and the mapping of senses to the wordnet. The evaluation task consisted of a three-way decision to be recorded in the first column of each row. This three-way decision comprised: attesting the existence of the candidate sense (i.e. the lemma was known to include the meaning proposed by the Candidate Synset) – marked with *1*; revealing uncertainty about whether the Candidate Lemma could have the proposed sense – marked with *?*; and rejecting the possibility that the Candidate Lemma could be used in the candidate sense – marked with *0*.

The initial sample of 300 senses was done under the assumption that the sum-score mentioned in Section 2.3 would outperform the simple metric of 'number of languages that suggested the concept'. Under this assumption, we selected two groups of 150 sequential sense candidates – one group with high ranked sum-scores and another with medium ranked sum-scores. Upon a closer inspection of the results (which will be discussed in detail in Section 3), we realized that the simple metric of calculating the number of overlapping languages suggesting any given concept was actu-

0/1/?	No. Langs	Candidate Lemma	Matched Translations	Candidate Synset	English Lemmas, Definitions and Examples
1	2	ϭωπϵ	‘fra|saisir|n’, ‘fra|saisir|v’, ‘eng|seize|v’, ‘eng|seize|n’	02273293-v	confiscate; attach; impound; seize; sequester [take temporary possession of as a security, by legal authority] The police confiscated the stolen artwork

Table 3: Manual Checking (example)

ally a better predictor of correct senses.

3 Results

3.1 Data Sampling

The human evaluation task (detailed in Section 2.4) focused on a blind review of 300 senses. The agreement of both reviewers over this task was 68% (i.e. 204/300 senses). This number refers to agreement in either accepting or rejecting a candidate sense.

To better understand these numbers, one important note to take into consideration, for this evaluation, is the fact that there are no native speakers of Coptic. Because of this, the Coptic knowledge of even the most expert scholar must be considered fragmentary. The amount of exposure to the language most certainly leads to some assumptions about how the language works, including the possible senses a word can have. In addition, the Zipfian nature of language distribution further corroborates our empirical understanding that being exposed to different Coptic texts most certainly has an impact on sense knowledge. In other words, some obscure senses for a given Coptic word might appear so rarely that only scholars who have read certain documents can know about it. This is also why many wordnet projects resort to sense-tagging corpora in order to further evaluate and improve their wordnets. Unfortunately, in such an early stage of our project, we have not yet been able to include this method in our evaluation.

Following the discussion in the paragraph above, we calculated two different measures to evaluate our automated construction method: the percentage of senses accepted by either of the reviewers (i.e. union), and a stricter measure reporting only the percentage of senses accepted by both reviewers (i.e. intersection). These results are presented in Table 4.

No. Langs	Correct(%) Union	Correct(%) Intersect.
1	(n=119) 25%	7%
2	(n=134) 89%	49%
3	(n=40) 98%	63%
4	(n=7) 100%	100%
Total	62% (n=300)	34% (n=300)

Table 4: Human evaluation of the results (union and intersection), by language overlap

Union was calculated by identifying when either of the reviewers assigned a *1* (correct), regardless if the second reviewer assigned *0* (incorrect) or *?* (uncertain). This measure always rewards the user who claims to know the existence of a sense, since the other reviewer might assume or not know of its existence. Intersection was calculated by only counting answers when both reviewers provided answers compatible with the inclusion of that sense. In both measures, when one reviewer assigned a *?* (uncertain), the second reviewer's response was considered the default – in other words, the answer *?* (uncertain) is compatible with both accepting or rejecting an answer, taking the other reviewer's response as final. For example, if one reviewer attested the existence of a sense, but the second reviewer was uncertain, we counted this as “correct” (for both union and intersection measures). In this sample, there was no instance where both reviewers were uncertain.

In addition to the total scores, Table 4 also presents scores grouped by the number of intersected languages that informed each candidate sense. We consider these numbers to be very positive, as they show that the overlap of two or more languages gives a union baseline score of 89%. The intersection of 3

or more languages gives a baseline score of 98% for union (and 63% for intersection). Finally, senses informed by four languages, predict candidate senses 100% of the time.

Despite an unbalanced sample, the numbers still show that our method is principled. The higher the number of intersected languages, the better the prediction accuracy of our method. Furthermore, the overlap of just two languages appears to already be quite informative – reaching a high boundary union score of 89% and a low boundary intersection score of 49%. Even assuming that the union score might include some false positives, a value within this range would suggest a prediction well above chance.

3.2 Wordnet Statistics and Coverage

A summary of the final results produced by our method can be found in Table 5. In total, the second stage of our wordnet includes 218,677 automatically inferred Coptic senses, which is a decent increase from what was generated during the first stage (with less data). In addition, and following the discussion on confidence scores in the section above, Table 5 also shows the number of available senses sorted by the number of languages that intersected that sense.

No. Langs	No. Senses (Stage I)	No. Senses (Stage II)
1	182,883	184,657
2	19,967	30,207
3	3,329	3,575
4	183	238
Total	206,362	218,677

Table 5: Senses per Language Overlap

While the majority of senses was informed by only one language, 34,020 senses (Stage II) are the result of the intersection of two or more languages. If the numbers from Table 4 are confirmed in our ongoing evaluation experiment, then these senses would be expected to have a confidence score of 89% and above.

Table 6 presents how these 218,677 senses are distributed among synsets and parts of speech. In total, the senses are distributed among 25,871 synsets, and fairly well distributed across different parts of speech. On average, there are 7 senses per nominal synset,

POS	No. synsets	No. senses
nouns	13,904	97,527
verbs	7,491	92,019
adjective	3,488	20,723
satellite adj	229	587
adverb	737	7,373
non-referential	22	448
Total	25,871	218,677

Table 6: WN Coverage: Coptic (Sahidic)

and about 12.2 senses per verbal synset. Although many of these senses might not be correct, the high number of senses might also be explained by the many forms a single Coptic lemma can take – which were listed in the dictionaries we used. Many of these forms are, in fact, motivated by morphology, while others are motivated mostly by spelling variation. In the future we would like to dedicate some time to better classify and tag these forms.

The 25,871 synsets cover about 77.4% of the list of 5000 "core" word senses in Princeton WordNet (Boyd-Graber et al., 2006) – a usual measure for coverage of wordnet resources. Further evaluations of coverage at such an early stage of our project might be somewhat difficult. Nevertheless, we decided to test how our wordnet fared in a task of sense matching over open text. A small corpus of $52,789$ word tokens was used, and $20,235$ ($38,3\%$) out of all tokens were able to find a compatible entry in the Coptic WN. While this coverage may seem low, it fits with other similar experiments done for Ancient Greek (34%) and Latin (33%) (Moritz et al., 2016).

3.3 Release

This Coptic Wordnet is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[11]. We have produced OMW tsv files, which can also be used in the Python Natural Language Toolkit (Bird et al., 2009). In addition, and keeping up with the recent requirements to belong to the OMW, we will also release this data using the WN-LMF format[12]. The use of WN-LMF will be essential to access the new Collaborative Interlingual Index (CILI) (Bond et al., 2016) – a language agnostic, flat-structured in-

[11] https://creativecommons.org/licenses/by/4.0/
[12] https://github.com/globalwordnet/schemas

dex to link wordnets across languages. The Coptic WN data can be found on GitHub at https://github.com/coptic-wordnet.

4 Discussion and Future Work

The results from the method of constructing the Coptic WN are promising. We have introduced the method of sense intersection to construct a wordnet which jumpstarts the process of producing a wordnet that is useful for digital humanities tasks. One of the current limitations relates to the expansion approach that uses only dictionary sources. We plan to create and annotate a sense-tagged corpus, alongside the wordnet, so that we can also gain word frequency information, test for coverage and review concepts in context.

We have also argued for the use of union between reviewers as a valid metric since reviewers will not have the same experience with the language. Several different reviewers can positively identify attested concepts and this is in no way a reflection that they do not agree. It can indicate, however, that there is debate within the scholarly community. Because of this, we would like to invite more Coptic scholars into this project, so that the full lexical semantic knowledge can be captured within this resource.

We are currently discussing ways to link to the Coptic Dictionary Online (CDO). This would require following practices of Linked Open Data, where the Coptic WN can be connected to CDO's entries (e.g. via URIs) and, conversely, CDO could be extended to link to related entries from Coptic WN.

Additional on-going work relates to the Collaborative Interlingual Index (CILI). Within the domain of Religious Studies, the PWN has shown numerous shortcomings, including badly formed definitions and an inconsistent hierarchical structure (Slaughter et al., 2018). Following this, we believe that the development of the Coptic WN can be used to contribute to on-going Digital Humanities work within the domain of Religious Studies. This is especially true since much of the content of Coptic sources is primarily religious or theological in nature.

We also believe that the Coptic WN can be a useful resource to further inform multiple Coptic (pre-)processing tools, and help in tasks such as part-of-speech tagging and lemmatization. One such example would be the tools available through the Coptic Scriptorium (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016) which includes multiple Coptic processing tools.

The Coptic WN is relevant to the study of purely linguistic research topics, including but not limited to research in lexical semantics. We would like to extend the work of the Etymological Wordnet (de Melo, 2014) to provide a tool for the study of Coptic-related language evolution – including the problems of concept drift (Fokkens et al., 2016) and diachronous meaning shift, concerning how concepts travel through space and time (crossing dialects and even languages), taking slightly different meanings as they move.

Finally, as it was mentioned above, one of the major motivations behind the construction of the Coptic WN was to use its hierarchy for text reuse. In essence, this task is designed to capture short snippets of text similarity (e.g. quoting, summarizing, paraphrasing, translation). TRACER is a system capable of using multiple algorithms to find text reuse across large corpora – which is accomplished by word replacement. Our wordnet can be used to generate possible word replacements including synonyms, hypernyms, hyponyms, or cohyponyms. We are currently exploring hierarchy traversal and replacement strategies that best produce accurate examples of text reuse.

Acknowledgments

We would like to thank the Database and Dictionary of Greek Loanwords in Coptic (DDGLC) Project Coordinator, Tonio Sebastian Richter and Scientific-Technical Staff, Katrin John.

Many thanks to Adam Rambousek for providing the work *in progress* of the future version of the Czech Wordnet.

We would like to thank Milan Konvicka for providing the data from MARCION.

The TRACER part of this work has been made available by the early career research group eTRAP (No. 01UG1409, 01UG1509) funded by the German Ministry of Education and Research.

We would also like to acknowledge the project "A Linked Digital Environment for Coptic Studies" (NEH Grant No. HAA-261271-18).

References

Pushpak Bhattacharyya. 2017. IndoWordNet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit NLTK*. O'Reilly Media, Inc.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1140–1147, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proc. of the 8th Global WordNet Conference*, pages 44–49.

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 56–61, Trento.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In Christiane Fellbaum Verginica Barbu Mi-titelu, Corina Forăscu and Piek Vossen, editors, *Proceedings of the Global WordNet Conference*, pages 50–57, Bucharest, Romania.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of the Third International WordNet Conference*, pages 29–36.

Marco Büchler, Philip R Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In *Text Mining*, pages 221–238. Springer.

Walter E. Crum. 1939. *A Coptic Dictionary*. Oxford University Press, Oxford.

Gerard de Melo. 2014. Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1148–1154, Paris, France, May. European Language Resources Association (ELRA).

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 12–21, Santa Fe, NM.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Christiane Fellbaum. 2017. Wordnet: An electronic lexical resource. *The Oxford Handbook of Cognitive Science*, pages 301–314.

A.S. Fokkens, S. ter Braake, E. Maks, and D. Ceolin. 2016. On the semantics of concept drift: Towards formal definitions of semantic change. In S. Darányi, L. Hollink, A. Meroño Peñuela, and E. Kontopoulos, editors, *Proceedings of Drift-a-LOD*, pages 247–265.

Eitan Grossman. 2014. Transitivity and valency in contact: The case of Coptic. In *47th Annual Meeting of the Societas Linguistica Europaea*, Poznań, Poland, 9. Talk given at a workshop on Transitivity and Valency in Contact: A Cross-Linguistic Perspective (convened by Susanne Michaelis and Eitan Grossman).

Birgit Hamp and Helmut Feldweg. 1997. GermaNet-a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT-the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

Karel Kučera. 2007. Hyperlemma: A concept emerging from lemmatizing diachronic corpora. *In: Levicka, J., Garabik, R. (eds): Computer Treatment of Slavic and East European Languages*, pages 121–125.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing Sanskrit Wordnet. In *Proceedings on the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai*, pages 287–294.

Stefano Minozzi. 2009. The Latin WordNet Project. In *In Anreiter, P. and Kienpointner, M., editors, Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur*

Lateinischen Linguistik, Innsbrucker Beitrage zur Sprachwissenschaft, volume 137, pages 707––716.

Luís Morgado da Costa and Francis Bond. 2016. Wow! What a useful extension! Introducing non-referential concepts to WordNet. In *Proceedings of the 10th edition of the International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4323–4328, Portorož, Slovenia.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas, November. Association for Computational Linguistics.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Caroline Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *DHQ: Digital Humanities Quarterly*, 10(2). `http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html(visited:2019-06-28)`.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 82–88.

Laura Slaughter, Wenjie Wang, Luis Morgado da Costa, and Francis Bond. 2018. Enhancing the collaborative interlingual index for digital humanities: Cross-linguistic analysis in the domain of theology. In P. Vossen (Eds.) F. Bond, C. Fellbaum, editor, *The 9th Global WordNet Conference (GWC 2018)*, pages 8–12.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 781–784, Lisbon.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89.

Piek Vossen. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, Dordrecht: Kluwer Academic Publishers.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th ACL SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH2016)*, pages 146–155, Berlin.

Yingjie Zhang, Bin Li, Xiaoyu Wang, Xueyang Liu, and Jiajun Chen. 2014. Mapping word senses of Middle Ancient Chinese to WordNet. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 446–450. IEEE.

Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. PQAC-WN: Constructing a wordnet for Pre-Qin Ancient Chinese. *Language Resources and Evaluation*, 51(2):525–545.