Automatic Detection of Coptic Text Reuse: Applying Coptic Wordnet to Intertextuality Studies in Selected Coptic Monastic Writings

So Miyagawa
University of Tsukuba
miyagawa.so.kb@u.tsukuba.ac.jp

Luis Morgado da Costa
Vrije Universiteit Amsterdam
lmorgado.dacosta@gmail.com

Laura Slaughter
University of Oslo
l.a.slaughter@dscience.uio.no

Heike Behlmer
Georg-August-Universität Göttingen
hbehlme@uni-goettingen.de

Abstract

This study explores the application of Coptic Wordnet to intertextuality studies in Coptic literature, focusing on works of the 4th/5th century Egyptian abbots Shenoute and Besa, and on the Bible in Coptic. Using the semantic relations captured in Coptic Wordnet, we enhance the automatic detection of text reuse, including quotations, allusions, and paraphrases. Our findings demonstrate that incorporating Coptic Wordnet data enables the identification of previously undetected intertextual relationships, providing a more comprehensive understanding of the interconnected nature of Coptic texts. This research highlights the potential of wordnets in tracing the traveling of words, phrases, and concepts from authoritative texts into the wider language, offering new insights into the linguistic and conceptual landscape of Coptic-speaking Egyptian Christianity in Late Antiquity.

1 Introduction

Intertextuality studies in the Coptic literature have traditionally relied on manual analysis and expert knowledge to identify text reuse. However, the advent of digital tools and resources has opened up new possibilities for automating and enhancing this process. This study explores the application of Coptic Wordnet to intertextuality studies, focusing on works of the 4th/5th century Egyptian abbots Shenoute and Besa, and on the Bible in Coptic. Coptic Wordnet, a lexical database that captures semantic relations beyond direct synonymity between Coptic words, serves as a powerful tool to improve automatic detection of text reuse. By leveraging this resource, we aim to uncover not only verbatim quotations but also more subtle forms of intertextuality, such as allusions and paraphrases.

The significance of this research lies in its potential to enhance our understanding of the intertextual relationships within Coptic literature, provide insights into the transmission of religious and cultural concepts in Late Antique Egypt, and demonstrate the value of computational approaches in the study of ancient languages and literatures. Through this work, we seek to bridge the gap between traditional philological methods and modern computational techniques, offering new perspectives on the rich tapestry of Coptic literary production.

2 Background

2.1 Intertextuality in Coptic Literature

Coptic literature, particularly the works of prominent monastic leaders like Shenoute (4--5th century) and Besa (5th century), is characterized by its rich intertextual relationships with the Bible and other authoritative texts by patristic authors such as Athansius, or earlier monastic writings. These connections offer valuable insights into the religious, cultural, and linguistic landscape of Coptic-speaking Egyptian Christianity in Late Antiquity. Intertextuality in Coptic literature manifests itself in various forms, including direct quotations, allusions, paraphrases, thematic parallels, and structural echoes.

Previous studies, such as Karl Heinz Kuhn's work on Besa's *Letters and Sermons* (Kuhn, 1956), have provided attestations of these intertextual relationships. However, the manual nature of these studies limits their scalability and the ability to uncover more subtle forms of text reuse. Our approach aims to complement and extend these traditional methods by applying computational techniques to a wider corpus of texts, potentially revealing patterns and connections that might otherwise remain hidden.

2.2 The Coptic Language

Coptic, the final stage of the Egyptian language, has a recorded history spanning over 5,000 years.

Written in a script derived from the Greek alphabet supplemented with characters from Demotic Egyptian, Coptic was primarily used from the 2nd to the 14th centuries CE. It served as the language of Christian Egypt during the Roman, Byzantine, and early Islamic periods, playing a crucial role in the preservation and transmission of early Christian literature.

The Coptic language is characterized by several key features that make it both fascinating and challenging to study. It comprises multiple dialects, with Sahidic being the main literary dialect in the first millennium CE and the focus of the Coptic Wordnet. Coptic's morphology is primarily agglutinative, with some fusional features, allowing for complex word formations. Its lexicon is a rich blend of native Egyptian words and Greek loanwords, reflecting the cultural and linguistic interactions between Greeks and Egyptians starting with the conquest of the country by Alexander the Great in 332 BCE. The verbal system is particularly complex, with numerous tenses, aspects, and moods expressed through a combination of auxiliaries and affixes. Syntactically, Coptic employs a system of bound groups[1] and converbs, which can pose challenges for automatic analysis and translation.

Understanding these linguistic features is crucial for developing effective tools and resources for Coptic language processing, including the construction and application of Coptic Wordnet. Our work takes into account these unique characteristics of Coptic to ensure that our computational approaches are sensitive to the language's specific structures and nuances.

2.3 Coptic Wordnet

Coptic Wordnet (Slaughter et al., 2019) is a lexical database that organizes Coptic words into synsets (synonym sets) and captures various semantic relations between them. It is part of the larger family of Wordnets, which originated with the development of Princeton Wordnet for English in the mid-1980s. The core structure of Coptic Wordnet comprises synsets, senses, as well as semantic relations established by the Princeton Wordnet. This rich semantic network enables the identification of conceptual connections beyond simple synonymity relations, making it a valuable resource for intertextuality studies.

The development of Coptic Wordnet represents a significant step forward in the digital humanities approach to Coptic studies. By providing a structured representation of the Coptic lexicon and its semantic relationships, it opens up new possibilities for computational analysis of Coptic texts. This resource not only aids in the automatic detection of intertextuality but also serves as a valuable tool for linguists, historians, and scholars of early Christianity studying the Coptic language and its literature.

2.3.1 Construction of Coptic Wordnet

A team of Coptologists and computer scientists built Coptic Wordnet automatically and evaluated it manually. Its construction was a complex process that drew upon multiple data sources and linguistic resources, leveraging them through an algorithm based on multilingual sense intersection.

It used lemma alignments from three primary sources: the Coptic Dictionary Online (CDO) (Feder et al., 2018), the MySQL version of Crum's *Coptic Dictionary* (Crum, 1939) in Milan Konvicka's Marcion Software,[2] and the Database and Dictionary of Greek Loan Words in Coptic (DDGLC).[3] These sources provided a rich foundation of Coptic words with translations in multiple languages, including English, French, German, Czech, and Greek.

The construction of the Coptic Wordnet followed the 'expand' approach, using other wordnets as a foundation for its structure. These wordnets included the Princeton Wordnet (English; Fellbaum, 2017), GermaNet and Odenet (German; Hamp and Feldweg, 1997; Siegel and Bond, 2021), WOLF (French; Sagot and Fišer, 2008), Greek Wordnet (Bizzoni et al., 2014), Ancient Greek Wordnet (Bizzoni et al., 2014) and Czech Wordnet (Pala and Smrž, 2004).

This automated process was followed by a manual evaluation phase, where Coptic scholars reviewed a sample of the automatically generated senses to assess accuracy and refine the results. Based on this evaluation, the Coptic Wordnet has data distributed over four levels of confidence, based on the number of interesected language dur-

[1] A Coptic bound group is a sequence of morphs, likely connected by a shared stress, that consists of structural elements with grammatical meaning and lexical elements, and the order of constituents within the group is fixed based on dependency classes Layton, 2011, 22--27, Haspelmath, 2014, 123--126.

[2] `http://marcion.sourceforge.net/` (accessed October 16, 2024)

[3] `https://www.geschkult.fu-berlin.de/en/e/ddglc/index.html` (accessed October 16, 2024)

ing its construction phase (for more details, see Slaughter et al., 2019). For this expriment, we used the wordnet as a whole (i.e. without filtering by levels of confidence). The size of the Coptic Wordnet used for this experiment can be see in Table 1.

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

Table 1: Coptic Wordnet Coverage (taken from: Slaughter et al., 2019)

3 Methodology

Our approach combines the text reuse detection capabilities of the text reuse detection tool TRACER (see 3.1) with the semantic richness of Coptic Wordnet to enhance the identification of intertextual relationships in Coptic literature. This integration allows us to move beyond simple string matching and consider the semantic context and relationships between words, thereby improving our ability to detect non-verbatim text reuse.

3.1 Text Reuse Detection with TRACER

TRACER, developed by the eTRAP research group at the University of Göttingen (Büchler et al., 2018; Büchler et al., 2014), is a versatile tool for detecting text reuse. Its workflow consists of text preprocessing, feature selection, link generation, scoring, and post-processing steps. While effective at identifying verbatim quotations and idiomatic expressions, TRACER initially struggled with more subtle forms of intertextuality (Büchler, 2013). This limitation prompted us to explore ways to enhance its capabilities through the integration of semantic information from Coptic Wordnet.

3.2 Integration of Coptic Wordnet

To address the limitations of purely lexical matching, we integrated Coptic Wordnet into TRACER's workflow. TRACER relies on word-to-word mappings to find text reuse. We used the Coptic Wordnet to create these mappings in several steps. We started by collecting standard synonymity mappings based on synsets. We then expanded these word sets based on the semantic relations captured in the Wordnet including hypernymy/hyponymy, and co-hyponymy. Details on the generation of word pairings are provided below, for each semantic relation:

- **Synonymy:** All senses belonging to a synset were paired with all other senses belonging to the same synset.

- **Hypernymy/hyponymy:** All senses belonging to a synset were paired with all senses belonging to its hypernyms/hyponyms (multiple inheritance allowed); Hypernym/hyponym chains were limited to 12 levels of recursion. This creates mappings of the type "dog ↔ animal" (and vice versa).

- **Co-hyponymy:** All senses belonging to a synset were paired with all senses belonging to hyponym of the first synset's hypernym. Hypernym chains were limited to 3 level of recursion. This creates mappings of the type "dog ↔ cat" (and vice versa), because both share a hypernym.

This greatly expanded semantically related words incorporated into TRACER's feature selection process. A summary of the number of relations extracted can be see in Table 2.

Relation	No. pairs
synonymy	4,103,650
hypernymy	9,797,575
hyponymy	9,867,688
cohyponymy	164,789,665
Total	188,558,578

Table 2: Lexical pairs provided to TRACER

TRACER's similarity calculation algorithm was adjusted to consider these semantic relationships when comparing text segments. This allowed us to detect potential text reuse even when the exact wording differed, as long as the concepts expressed were semantically related. Finally, we adjusted the thresholds for identifying potential text reuse to accommodate this expanded feature set, striking a balance between sensitivity and precision in our detection of intertextual relationships.

4 Results and Discussion

4.1 Improved Intertextual Detection

Previously, TRACER had been employed to detect text reuse from the Psalms in works of the two monastic authors mentioned above: Shenoute and Besa (Miyagawa, 2022). The choice had focused on Shenoute's *Canon* 6, a collection of writings on monastic discipline, and Besa's letters to monks and nuns, because these works exist in digital editions. The size of the corpora are follows: Shenoute, *Canon* 6 (49,412 words), Besa, *Letters and Sermons* (60,628 words), and the Sahidic Coptic translation of the Psalms (104,815 words).[4]

The integration of Coptic Wordnet with TRACER significantly enhanced our ability to detect non-verbatim text reuse. Table 3 shows the number of text reuse candidates generated by TRACER with and without the use of Coptic Wordnet (CWN) for various works.

	With CWN?	
Work	**No**	**Yes**
Besa		
Letters and Sermons	629	42,542
Shenoute, *Canon* 6 works		
He Who Sits Upon His Throne	84	5,535
Remember, O Brethren	31	2,582
I Am Not Obliged	207	11,293
Is It Not Written	98	8,235
People Have Not Understood	3	115

Table 3: TRACER's Text Reuse Candidates between Shenoute/Besa's Works and Psalms

As evident in Table 3, the use of Coptic Wordnet dramatically increased the number of candidate text reuse identified by TRACER. For instance, in Besa's *Letters and Sermons*, the number of candidates increased from 629 to 42,542 when using CWN.

Our analysis of the works of Shenoute's *Canon* 6 and Besa's *Letters and Sermons*, and the Sahidic Coptic translation of the Psalms revealed a rich tapestry of intertextual relationships that had previously gone unnoticed.

For this paper, our evaluation focused on the candidates generated for Psalm 1:1 in Besa's *Letters and Sermons*. We identified 18 possible text reuses out of 82 candidates. These allusions were not direct quotations, but rather semantic echoes that our enhanced system was able to capture through the recognition of related concepts and themes. Checking candidates of text-reuse is time-consuming and requires a high level of expertise. For this same reason, there is no gold standard against which we can evaluate our results. Regardless, in our expert point of view, we deem our work of great value to support the detection of text reuse.

In Shenoute's works, we discovered semantic clusters that shed light on the author's conceptual framework. For example, we found that the concept of "righteousness / justice" (ⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ *dikaiosunê* was frequently associated with related ideas such as "judgement" (ϩⲁⲡ *hap* or ⲕⲣⲓⲥⲓⲥ *krisis*) and "truth" (ⲙⲛⲧⲙⲉ *mntme*). This clustering of concepts provides insights into Shenoute's theological and ethical thinking, revealing patterns that might not be immediately apparent through traditional close reading methods.

4.2 Discussion and Limitations

The construction and application of Coptic Wordnet revealed several challenges specific to working with ancient languages. The limited textual evidence available for Coptic means that certain word senses or usage patterns are difficult to verify. By using an automatically created resource, such as the Coptic Wordnet, we are also constrained by the limitations of this resource. We know, for example, that many of the pairings used in this study are of low confidence. This means that other parameters, such as TRACER's sensibility, had to be tuned to filter out many other potentially interesting instances. We understand the value that human curation would bring to the Coptic Wordnet. But, at the same time, the lack of Coptic native speakers eliminates the possibility of intuition-based verification, a method often employed in developing Wordnets for modern languages.

One problem that most certainly arises from the automatic methods used to create the Coptic Wordnet is the challenge of capturing diachronic changes in word meanings over the many centuries of Coptic's use. Another problem posed by the current version of the Coptic Wordnet is the cultural and conceptual gaps between the ancient Coptic-speaking world and our modern context assumed by most wordnets. Many concepts in Coptic texts are deeply rooted in ancient Egyptian, Hellenis-

[4] The corpora were processed using the Coptic NLP Service (available at `https://tools.copticscriptorium.org/coptic-nlp/`, accessed October 18, 2024). For our analysis, we utilized the "word" unit as defined by this tool.

tic, and early Christian cultures, making them difficult to map onto modern conceptual frameworks. We had to contend with dialectal variations within Coptic, although our current focus on Sahidic Coptic mitigated this issue to some extent.

4.3 Future Work

Looking to the future, we have identified several key areas for improvement and expansion of our work. Refinement of Coptic Wordnet through manual curation and expansion will be a priority, as will enhancing our text reuse detection methods using more advanced NLP techniques. We plan to develop a larger, manually annotated corpus of Coptic texts, which will provide valuable training data for machine learning approaches and allow for more robust evaluation of our methods.

Generating and comparing different types of word-to-word mappings would be important to explore in future experiments. This could be done, for example, by filtering mappings by confidence level (which would likely generate much fewer but higher quality candidates of text reuse). It would also be worth exploring how different methods or thresholds for hypernym/hyponym recursion would affect TRACER's performance. For this experiment, we opted for a broad coverage (up to 12 levels of recursion) -- which should help detect less literal instances of text reuse -- such as allusions and paraphrases. However, this comes with the cost of generating many spurious text reuse candidates.

Another exciting avenue for future research would be exploring cross-linguistic intertextual relationships, particularly between Coptic and contemporaneous languages like Greek and Syriac. This comparative approach could shed light on the transmission and adaptation of ideas across linguistic and cultural boundaries in the Late Antique world.

Finally, we aim to integrate our work with other digital humanities resources, creating a more comprehensive ecosystem of tools and data for Coptic studies. This integration will not only enhance the utility of Coptic Wordnet but also contribute to the broader goal of making Coptic literature more accessible to researchers and the public alike.

5 Conclusion

This research demonstrates the potential of applying Coptic Wordnet to intertextuality studies in Coptic literature. By integrating semantic relations from Coptic Wordnet into text reuse detection tools, we have significantly enhanced our ability to identify and analyze intertextual relationships in Coptic texts. This approach not only advances our understanding of Coptic literature but also provides a model for similar studies in other ancient languages, showcasing the value of digital humanities approaches in classical and religious studies.

Our work bridges the gap between traditional philological methods and modern computational techniques, offering new perspectives on the rich tapestry of Coptic literary production. As we continue to refine our methods and expand the scope of our research, we anticipate that this approach will yield further insights into the interconnected nature of ancient texts and the transmission of ideas in the early Christian world. The development and application of Coptic Wordnet represents a significant step forward in the digital study of Coptic, opening up new possibilities for research and analysis. As we move forward, we hope that this work will not only contribute to Coptic studies but also inspire similar efforts in other areas of historical linguistics and digital humanities.

References

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).

Marco Büchler. 2013. Informationstechnische Aspekte des Historical Text Re-use.

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.

Marco Büchler, Greta Franzini, Emily Franzini, Maria Moritz, and Kirill Bulert. 2018. TRACER-a multilevel framework for historical text reuse detection.

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

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.

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

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.

Martin Haspelmath. 2014. A grammatical overview of Egyptian and Coptic. In Eitan Grossman, Tonio Sebastian Richter, and Martin Haspelmath, editors, *Egyptian-Coptic Linguistic in Typological Perspective*, Empirical Approaches to Language Typology 55, pages 103--144. De Gruyter Mouton, Berlin.

K. H. Kuhn. 1956. *Letters and Sermons of Besa*. Corpus Scriptorum Christianorum Orientalium, vol. 157. Scriptores Coptici, tomus 21. Imprimerie Orientaliste L.Durbecq, Louvain.

Bentley Layton. 2011. *A Coptic grammar: With chrestomathy and glossary: Sahidic dialect*, 3 edition. Harrassowitz Verlag, Wiesbaden.

So Miyagawa. 2022. *Shenoute, Besa and the Bible Digital Text Reuse Analysis of Selected Monastic Writings from Egypt*. SUB Göttingen.

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 *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a German wordnet from other resources. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*, pages 192--198.

Laura Slaughter, Luis Morgado Da Costa, So Miyagawa, Marco Büchler, Amir Zeldes, and Heike Behlmer. 2019. The making of coptic wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 166--175.