8-25-2018

A Linked Coptic Dictionary Online

Frank Feder
Akademie der Wissenschaften zu Göttingen

Maxim Kupreyev
Berlin-Brandenburgische Akademie der Wissenschaften

Emma Manning
Georgetown University

Caroline T. Schroeder
University of the Pacific, cschroeder@pacific.edu

Amir Zeldes
Georgetown University, amir.zeldes@georgetown.edu

A Linked Coptic Dictionary Online

Frank Feder
Akademie der Wissenschaften zu Göttingen
frank.feder@mail.uni-goettingen.de

Maxim Kupreyev
Berlin-Brandenburgische Akademie der Wissenschaften
maxim.kupreyev@bbaw.de

Emma Manning
Department of Linguistics
Georgetown University
esm76@georgetown.edu

Caroline T. Schroeder
Department of Religious Studies
University of the Pacific
carrie@carrieschroeder.com

Amir Zeldes
Department of Linguistics
Georgetown University
amir.zeldes@georgetown.edu

Abstract

We describe a new project publishing a freely available online dictionary for Coptic. The dictionary encompasses comprehensive cross-referencing mechanisms, including linking entries to an online scanned edition of Crum's Coptic Dictionary, internal cross-references and etymological information, translated searchable definitions in English, French and German, and linked corpus data which provides frequencies and corpus look-up for headwords and multiword expressions. Headwords are available for linking in external projects using a REST API. We describe the challenges in encoding our dictionary using TEI XML and implementing linking mechanisms to construct a Web interface querying frequency information, which draw on NLP tools to recognize inflected forms in context. We evaluate our dictionary's coverage using digital corpora of Coptic available online.

1 Introduction

Coptic is the final stage of the indigenous language of Egypt, spoken in Egypt in the first millennium and used as a liturgical language of Christian Copts in Egypt and the Coptic diaspora today. Together with Ancient Egyptian, the language of the hieroglyphs, it forms part of the longest continuously attested language documentation of any language on Earth. Unlike Ancient Egyptian, Coptic is written with a script derived from the Greek alphabet, with several letters added from the earlier Demotic script for native sounds not found in Greek.

Although the Coptic corpus, attested in six main dialects, is vast, it is very much under-studied compared with materials from contemporaneous Greek and Latin sources. The amount of freely available data for Coptic online is still small (see Schroeder and Zeldes 2016), and published book editions of Coptic texts are mostly limited to the main important literary works, covering only a fraction of the data preserved in the language, including literary, documentary and epigraphic material. The situation for Coptic lexicography on paper, by contrast, is more comprehensive, with Crum (2000 [1939]) and subsequent dictionaries offering excellent coverage of native Coptic words. More recently, progress has been made in the lexicography of the abundant inventory of Greek and other loan words in Coptic (Almond et al., 2013), and work is in progress in matching existing Egyptological resources, such as the Thesaurus Linguae Aegyptiae (Seidlmayer and Hafemann, 2011) with equivalent Coptic entries (Feder, 2016).

At the same time, there has been a substantial gap in providing an openly available electronic dictionary of Coptic linked with digital corpora to provide easy look-up functionality and frequency information. The present paper describes a new project, the 'Coptic Dictionary Online' (CDO), which is freely available and fully linked with the growing digital resources and NLP tools which are becoming available for Coptic. We discuss issues in TEI XML (http://www.tei-c.org) representations for Coptic data, unique issues arising from Coptic morphology, as well as linked open data standards, such as REST API connectivity. We evaluate the coverage of our lexicon based on currently available corpus data.

2 Related work

The dataset of the Coptic Dictionary Online is based on a part of the 'Thesaurus Linguae Aegyptiae' (TLA), provided by the Berlin-Brandenburg Academy of Sciences (BBAW). TLA is a digital offspring of the long-term research project 'Wörterbuch der Ägyptischen Sprache' (Dictionary of the Egyptian Language), started in 1897 at the Prussian Acedemy of Sciences (now BBAW). Following the paper publication of the five volume Egyptian lexicon, the TLA began work on the digital edition of the dictionary, which went online in 2004, with extensions including Demotic Egyptian and Coptic lemmas, but not yet broadly covering the Coptic vocabulary targeted in this paper.

A second project, 'Database and Dictionary of Greek Loanwords in Coptic' (DDGLC), originally based in Leipzig and since 2015 at the Freie Universität Berlin, has indexed Greek words in the attested Coptic corpus. DDGLC data is organized around Greek lemmas and currently contains 4,971 Greek source lemmas, 8,406 resulting Coptic lemmas and 100,106 'attestations' of the latter. CDO plans to integrate these to overcome a major drawback of current Coptic lexicographical projects, which disregard non-Egyptian vocabulary (see Section 5).

The other most extensive lexicographical project is the lexicon by the Marcion project,[1] which contains 11,437 head words indexing 87,169 items, and closely follows the printed edition of Crum's dictionary. This resource is not linked to corpora and NLP tools and does not offer a REST API or multilingual definitions, but does include links to the scanned online version of Crum's work, much like the CDO.

Finally, the situation for lexically tagged corpora is more limited: the currently largest open collection of lemmatized and grammatically analyzed Coptic data is provided by the project Coptic Scriptorium,[2] based at Georgetown University and the University of the Pacific, and open for querying via the ANNIS web interface (Krause and Zeldes, 2016). The corpora encompass approximately 500,000 running tokens of Coptic text, just under 20% of which has been manually checked (this data will be used for the evaluation in Section 5). Scriptorium data is described in more detail further below, and is the source of both linked frequency data and example look-up in the CDO project.

3 Lexical data representation

3.1 Coptic grammar

Coptic is an agglutinative Afro-Asiatic language, descended from the earlier highly inflected system of Ancient Egyptian, which was more similar to Semitic languages. Originally, Coptic was written in manuscripts in *scriptio continua*, without spaces between words, as shown in Figure 1. However, like other languages of the Middle East, such as Arabic and Hebrew, modern conventions spell Coptic with spaces between stressed word groups, known as bound groups (see Layton 2011 in detail). Bound groups most often include only one content lexeme, usually either a noun or a verb, along with accompanying clitic articles, auxiliaries, prepositions and object or possessor pronouns, and therefore do not receive individual lexicon entries.

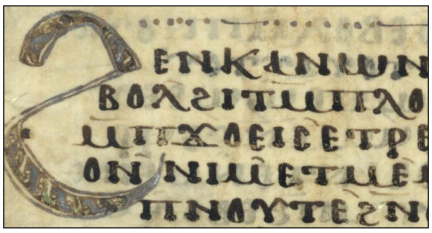


Figure 1: Excerpt from a manuscript of Shenoute's Canon 5 in Three Folios at the National Library in Vienna showing text in *scriptio continua*. Image: Österreichische Nationalbibliothek, `http://data.onb.ac.at/rec/RZ00002466`.

However, analyzing what constitutes a content item for dictionary entries is non-trivial, since within each bound group, Coptic nominal compounds are uninterrupted and spelled together (unlike Arabic or

[1] `http://marcion.sourceforge.net/dictionary/coptic.html`
[2] `http://copticscriptorium.org`

Hebrew), and verbal incorporation (see Grossman 2014) can often create complex lexical items containing multiple content morphemes (cf. English complex verbs such as 'breastfeed'), which are often listed in dictionaries as a complex item if they are frequent or have opaque senses. Both incorporation and cliticization also lead to changes to verb and noun stems, as shown in the following examples.[3] Example (1) shows a bound group with a preposition, article and the noun 'name', while (2) shows a reduced form of the same noun with a possessive 2nd person masculine singular clitic.

(1) ϩⲙ̀-ⲡ-ⲣⲁⲛ hm-p-ran 'in-the-name'

(2) ⲣⲛⲧ=ⲕ rnt=k 'name-your (SG.M)'

The noun's form in (1) is referred to as 'absolute' (*status absolutus*), while in (2) it is in a bound-pronoun state (also called *status pronominalis*; forms fused to a subsequent noun, rather than pronoun, take a third possible form called *status nominalis*). Examples (3)-(5) show a verbal bound group, with stem reduction via incorporation in (4), and nominalization of a complex verb in (5).

(3) ⲁ-ϥ-ϩⲱⲧⲃ
a-f-hōtb
PST-he-kill
'he killed'

(4) ϩⲉⲧⲃ-ⲯⲩⲭⲏ
hetb-psychē
kill-soul
'(to) soul-kill'
(incorporated)

(5) ⲙⲙⲛⲧ-ⲣⲉϥ-ϩⲉⲧⲃ-ⲯⲩⲭⲏ
mnt-ref-hetb-psychē
ness-er-kill-soul
'soul-killing'
(lit. 'soul-kill-er-ness')

The last example shows that incorporated verbs behave like normal verbs in allowing subsequent derivations. Representing these forms consistently is challenging, as we discuss in the sections below.

3.2 Properties of lexicon composition

The coverage of the dataset used in the Coptic Dictionary Online is based on W. Crum's 'Coptic Dictionary' (2000 [1939]) and currently contains 8,042 words ("entries") and 18,150 word forms. The Coptic data follows the design of the combined Egyptian-Demotic-Coptic lexicon in the TLA, with the aim of integrating the Coptic dataset to form a collection of all lexical items from 4,500 years of the language's recorded history. With regards to the coverage of the Coptic lexicon, CDO has two types of restrictions. A lexeme may be absent:

- because it is absent in Crum's dictionary (notably Greek and Arabic loan words)
- due to data model design despite being in Crum (bound forms of verbs and plural nouns; see below)

Here we outline the major principles behind the lexicon data composition and encoding, and provide details on the ongoing extension of the project.

The backbone of entries from Crum in the CDO is extended with cross-references to the later Coptic dictionaries of Westendorf (2008 [1965-1977]), Černý (1976), Vycichl (1983) and Cherix (2014). What all these have in common is the absence of loanwords – a tradition in Coptic lexicography, going back to the 18th century. Beginning from the *Lexicon Aegyptiaco-Latinum* of the Huguenot polymath Maturin Veyssière de La Croze of 1721, lexicographers were interested primarily in 'autochthonous' Egyptian vocabulary, and disregarded Greek and Arabic as 'too familiar' (Richter, 2017). Recent research has challenged traditional conceptualizations of "loan words" as a classification of vocabulary separate or distinct from "autochthonous" words, particularly in the multilingual context of Roman Egypt (Grossman 2013, Papaconstantinou 2010). The lack of loan words is one of the issues which CDO is currently working on (see Section 5 on the impact of their absence, and Section 6 on future plans). Apart from loan words, word forms available in Crum but missing in CDO include orthographic forms in Coptic dialects other than Sahidic, the classical dialect supplying the reference forms in the dictionary, and bound forms of Coptic words when used as clitics. Entries only attested in other dialects are included.[4]

[3] We follow the convention in Coptic studies of joining clitic lexical items in bound groups with hyphens, and clitic pronouns with the '=' sign.

[4] A pilot including all entries from Crum and all dialect forms for the letters *alpha* and *beta* has also been carried out, but due to time constraints has not been extended to further letters.

incorporated		multi-word expression	
ti-meeue	*give-thought, think*	ti-m-p-meeue	*give-of-the-thought, suggest*
ti-erxot	*give-wound, to wound*	ti-n-ou-erxot	*give-of-a-wound, inflict a wound*
k^{y}i-bekhe	*get-wage, get-paid*	či-m-p-beke	*receive payment*

Table 1: Incorporated complex verbs and unincorporated multi-word expressions.

Bound forms are exceptionally recorded only if the absolute (non-clitic) form is not available, as is the case with inalienable possessed nouns, e.g. ⲣⲁⲧ= *rat-* 'foot (of)'. Inflected forms are also not included, but are findable using automatic lemmatization (see Section 4.4). This means the small class of morphological plurals is not listed (pluralization is usually indicated by the article), and the same applies to stative verb forms (also called 'qualitative', distinct forms denoting a *state* resulting from the action, rather than the action itself). Statives are entered if they are the only form of a verb attested in Coptic.

Verbal clitic status also determines spelling of verbal compounds in CDO: verbs in bound state followed by incorporated objects are treated as morphemes, forming one lexical unit with the following word, and are thus written together, though purely compositional productive formations are only included if they have entries in Crum's dictionary. Verbs in absolute state are treated as separate lexical units and are written separately. Compounds in which absolute and bound state are indistinguishable are treated as bound forms if they are not followed by prepositions.

Determining whether a complex verb form is considered a single entry depends on whether a bare noun is incorporated without determiners in a generic reading (cf. 'breastfeed') vs. using a full nominal bound group, e.g. with determiners, possessives or prepositions (cf. 'feed with her breast'). Table 1 contrasts similar items: incorporated single verbs, and multi-word expressions. Although the latter entries contain information on object nouns, verb valency information is otherwise not included in the dictionary.

The guidelines for handling complex verbs correspond to the word segmentation practices in current corpora available from the project Coptic Scriptorium (see Section 2), and therefore facilitate compatibility with corpus look-up and frequency data (see Section 4.3), as well as interoperability with existing NLP tools for Coptic (Zeldes and Schroeder, 2016). However there are a number of cases in which bound forms are accompanied by explicitly or implicitly specific nouns that are represented as entries in Crum's dictionary, but are segmented apart by NLP tools and corpora based on their outputs:

- Verbs with possessed nouns: k^{y}n-rat= 'search (lit. find one's foot)', meh-hēt= 'fill one's belly'
- Fixed verb + article + noun: r-p-ke 'do another (thing)', r-t-k^{y}ot 'be like, lit. do the likeness (of)'
- Verbs bound with compound prepositions (preposition + possessed noun + suffix): r-ha-čō= 'go towards (do to head of)', r-hi-čn- 'be over (do on head of)'

We expose the lexical units in such compounds using XML encoding in the element oRef, described in the next section. Summing up, the current limitations of the lexicon are:

- no loan words, and dialect forms usually included only if not attested in Sahidic
- bound forms only if unattested in absolute form, and stative only if verb is unattested in infinitive
- no inflected forms of nouns (but see Section 4.4 on auto-redirecting from most inflected forms)

3.3 Properties of lexicon encoding

CDO is encoded in XML, managed using the Python libraries ElementTree and LXML, as well as Xpath queries. Data integrity is validated by an XSD Schema, developed as a subset of the TEI's Dictionary module.[5] The basic element is **<entry>**, defined by standard spelling, part of speech and

[5]See `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`. TEI's dictionary module is also viewed by Romary (2015) as a standard serialization of the Lexical Mark-up Framework (ISO, 2008). For the CDO's latest schema, see `https://github.com/KELLIA/dictionary/blob/master/xml/Coptic_Lemma_Schema.xsd`, which represents our underlying conceptual UML model, cf. Routledge et al. (2002).

gender information, and senses: only one **<form type="lemma">**, **<pos>** and **<gen>** element (gender) is allowed per entry (meaning feminine derivations receive separate entries) and at least one **<sense>** element is required. At the same time, multiple orthographical variants (**<form>**) and meanings (**<sense>**) are possible in one entry. Entries belong to the superordinate entity **<superentry>**, which corresponds to Crum's dictionary entry, uniting lemmas derived from the same root in Coptic. The schema is outlined in Figure 2.[6]

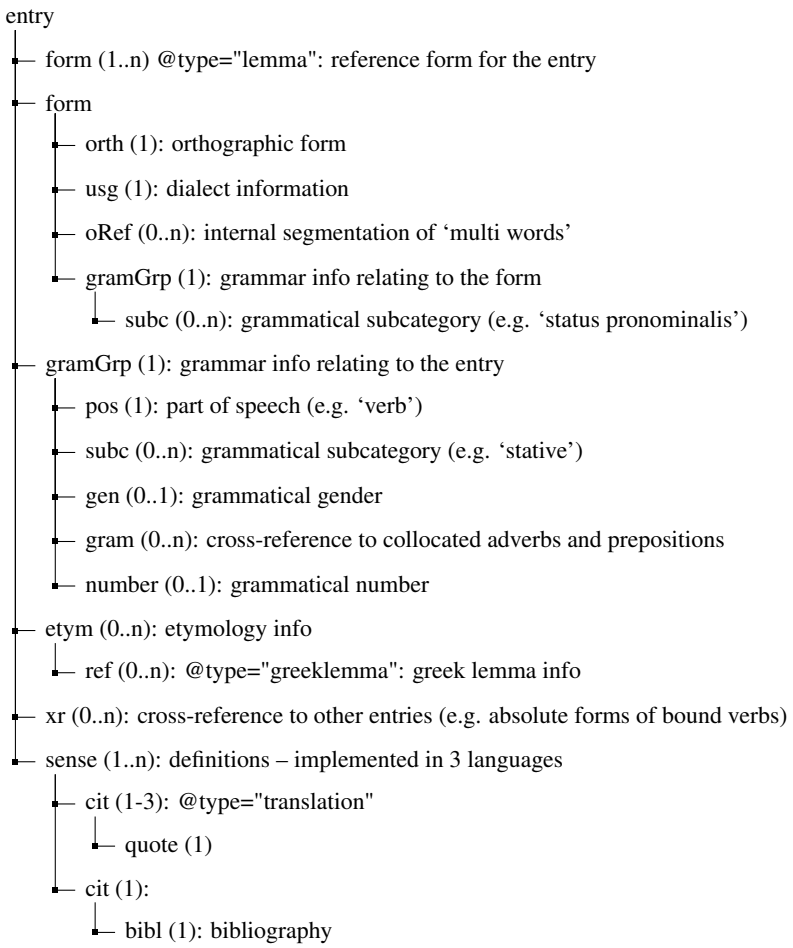


Figure 2: Hierarchy of the entry element in the XSD schema

Grammatical information can be set on both the 'entry' and on the 'form' level. The boundedness status information is encoded within form, e.g. **<subc>**Status nominalis**</subc>**, while information pertaining to inflected forms belongs to the entry level, e.g. for stative forms: **<pos>**Vb.**</pos>** with the subcategory **<subc>**Qualitativ**</subc>**. The **<gram>** tag provides information about prepositions and adverbs collocated with a given verb. The values available in elements such as **<pos>**, **<subc>** and **<usg>** (dialects) are controlled by the vocabulary set in the XSD Schema.

Each entry and each form have unique IDs. The **<xr>** tag is currently used for cross-referencing the absolute form of lexical verbs appearing as bound in complex entries. However, multi-word expressions also encode all of their constituents (including function words): the multi-word expression is listed spelled together inside **<orth>**, but separated in the element **<oRef>**, as in (6) for the expression he-p-ouō 'inquire, lit. find-the-news':

[6]We follow the mapping in Romary (2015) between LMF concepts and corresponding TEI elements as follows:

LMF component	TEI representation	LMF component	TEI representation
LexicalEntry	**<entry/>**	writtenForm	**<orth/>**
Lemma	**<form type="lemma"/>**	partOfSpeech	**<pos/>**
Word Form	**<form type="inflected"/>**	grammaticalNumber	**<number/>**

(6) <orth>ϩⲉⲡⲟⲧⲱ</orth> <oRef>ϩⲉ ⲡ ⲟⲧⲱ</oRef>

This segmentation information will be used below to search for multi-word entries in corpus data.

4 Web interface

4.1 User interface design

To facilitate efficient search, the lexicon's XML files are compiled into a SQL database providing records for each entry, with multiple orthographic forms grouped into super-entries, which correspond to Crum's entries.[7] Each record contains the dictionary information such as orthographic forms, morphological information, definitions in English, French and German, and etymological and bibliographic information. A record also contains both a unique entry number and a super-entry number used to link it to related entries based on Crum's groupings. We then designed an online interface to search this database and view entries. From the search page, users may search any combination of:

- Coptic word form/regular expression (with a virtual keyboard option to enter Coptic characters)
- Dialect (using Crum's dialect sigla, subject to the restrictions mentioned above)
- Part-of-Speech tag (using Scriptorium NLP tags, mapped onto the TLA's model; see below)
- Definition, with options to search for the exact sequence entered or for definitions containing all words entered in any order, as well as an option to search a specific one of the three definition languages, or any language. Regular expressions are also allowed.

To reconcile the different part of speech tags used by the TLA's data model and Coptic Scriptorium's model, we have collapsed tags for some of the more fine grained tag distinctions. This is necessary, among other things, for the benefit of NLP-tool based look-up (see Section 4.4). Currently, users can limit searches to the following options:

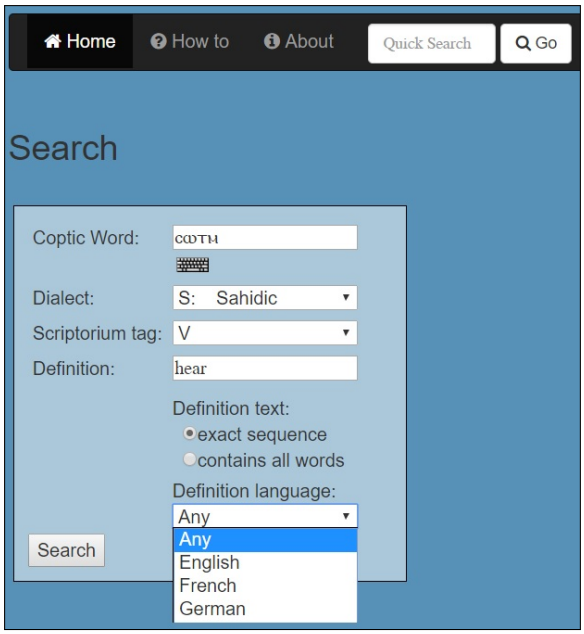
- **A** - any auxiliary (merges Scriptorium/TLA)
- **ART** - articles
- **C** - any of the so called 'converters', a morphological class of Coptic conjunctions
- **CONJ** - all other, independent conjunctions
- **N** - nouns, collapses common and proper nouns
- **NEG** - negations
- **NUM** - numerals
- **pronouns** - demonstrative (PDEM), interrogative (PINT), personal (PPER), possessive (PPOS)
- **PREP** - prepositions
- **PTC** - particles
- **V** - collapses all verbal tags, including finite, non-finite, imperative and 'verboids'

The original grammar info encoded in XML (using German terms) is displayed along with the Coptic Scriptorium part of speech tags. In addition to the main search page, shown in Figure 3a, a Quick Search is available in the navigation bar at the top of any page. Users can enter any combination of Coptic words and English, French and German words to search the word form and definitions, respectively. These, too, support regular expressions, and space-delimited search words are automatically classified as either Coptic (for search in the entries themselves) or not (for search in the definitions).[8]

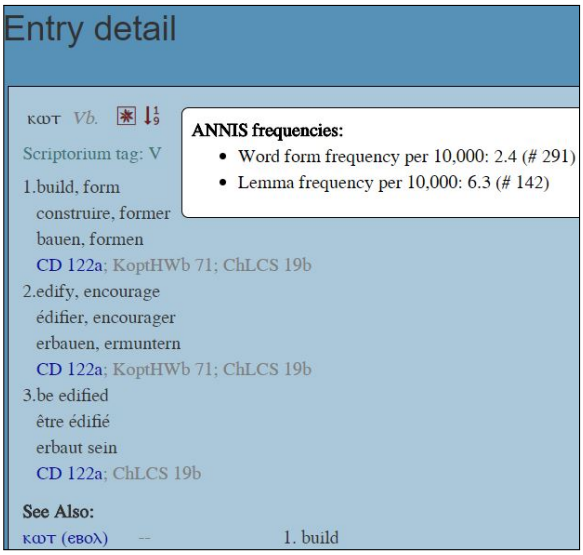
[7] An anonymous reviewer has inquired about the possibility of encoding the database version of the lexicon in RDF format – this is certainly possible and the adoption of a SQL database is in no way a principled decision, but rather one of convenience. We regard the XML representation as the primary serialization of the lexicon's data model and use the tabular SQL schema only internally as an index for the search interface.

[8] We do not currently classify non Coptic words as English, French or German, but rather match non-Coptic script words in all definition languages.

When a search uniquely identifies an entry, it takes the user directly to the entry page, which displays the word's form or forms, each with morphological information, dialect if available, and ANNIS frequency information (see Section 4.3). Each entry page also contains Scriptorium part of speech tags, the definitions in all three languages for each sense, as well as any etymological and bibliographic information, including cross-reference links. If there are other entries in the same super-entry, they are linked in a 'See Also' section at the bottom of the entry page. When a search has more than one result, users are taken to a page which lists up to 100 matches, displaying Coptic forms and English definitions for each. Users can click on any of these Coptic words to go to the associated entry page.



(a) Main search form



(b) Entry with frequency information for ⲕⲱⲧ *kōt* 'build'.

Figure 3: CDO web interface.

4.2 Corpus query link up

The dictionary is linked to freely available corpora provided by the Coptic Scriptorium project (`http://copticscriptorium.org`), which currently include over 500,000 tokens from 16 corpora containing over 700 Coptic documents. Each entry page provides a link to a lemma search in all corpora for each form specified by an **`<orth>`** element (i.e. a single entry page will have multiple search links if there are multiple variant forms). Conversely, Scriptorium corpora offer links to the CDO for all lemmas, which connect to the relevant entry pages whenever available (see Section 6 for some limitations).

One challenge in implementing the linking capacity is dealing with differences in segmentation for lemma definitions in the dictionary, and existing segmentation standards used in the corpora based on current NLP tools for Coptic. Several types of mismatch are possible – an entry may: 1. specify a variant not considered a lemma in the corpora due to normalization; 2. be a multiword expression, with more than one lemma in the corpora; or 3. contain added words not included in the head word proper, e.g. collocated prepositions.

To overcome the first difficulty, we automatically harvest the list of available lemmas in the current version of the corpora using the REST API provided by Coptic Scriptorium's ANNIS web interface. Single word dictionary items not attested in the corpora as lemmas can then fall back to searching on the word form level instead of the lemma level, which may include the desired form.[9]

The second problem is more complex, but can be addressed using the **`<oRef>`** tags in the dictionary XML files. The contents of these tags indicate a space-delimited segmentation for multiword entries,

[9] A further option of using fuzzy search to find unknown inflected forms is not currently supported, but remains a possibility for future development.

which are then searched for as a sequence of (possibly inflected) word forms. Since a multiword expression often contains inflected forms, the linked search uses the word form level of the corpora, rather than a sequence of lemmas. A reverse look-up for multiword expressions in the corpora is not yet implemented: the corpora are currently lemmatized with the Coptic NLP pipeline (Zeldes and Schroeder, 2016), which always assigns token-wise lemmas; however, we are considering how 'secondary lemmas' may be introduced whenever corpora contain sequences of words which are attested as dictionary entries.

Finally, for the third problem above, we currently suppress collocated elements in search queries. For example, for the entry ⲕⲱⲧ (ⲉⲃⲟⲗ) 'build (out)', we omit the optional adverb given in brackets, and simply search for the verb 'build' by itself. This approach produces more search results than might be intended, and a better solution for these cases in the future would be desirable (see Section 6).

4.3 Corpus frequency information

Using the same corpora and REST API above, we collect frequency information for all dictionary `<orth>` elements in their capacity as both word forms and lemmas. We then rank words by frequency and offer users, for each word form, the frequencies and ranks of the item being viewed, as shown in Figure 3b. Tied ranks are shared across items, i.e. there can be multiple entries with lemma rank 100, if multiple items are tied for this position. A limitation of using corpus frequencies is that the corpora are not truly aligned in an intelligent way to dictionary entries, and are not sense disambiguated, i.e. homonyms all contribute to one frequency pool for each orthographic string. This is unlike cross-reference links to Crum's dictionary, which point to page numbers based on actual senses.

Notwithstanding this limitation, frequencies can be useful to users and used to decide how likely it is that an unclear passage contains a possible word, or to prioritize learning high frequency items when studying the language. While frequency information cannot be used for searches yet, we plan to expose it to searches as well as building frequency visualizations to make the data more interactive and accessible.

4.4 NLP and lemma look-up

Since inflected forms are not included in the dictionary, users searching for them will not find any results. Although this is generally not an issue for users coming in from linked corpora (since links point to lemmatized forms), users who manually enter inflected forms will run into this problem.

To circumvent this issue, we use all possible outputs of the same lemmatizer used by the NLP pipeline from Zeldes and Schroeder (2016), without context information. Since morphologically inflected items in Coptic are largely closed-class, and productive stem-altering inflection is rare, the possible responses from the lemmatizer virtually always contain the correct analysis. This means that users can find singular entries for nouns with morphological plurals, or infinitive forms in searches for stative verbs, etc.

Linking from the corpora to the dictionary also depends on correct lemmatization, which, unlike the 'multiple options' look-up strategy, is deterministic (the corpora contain a single 'gold' lemma). This lemma is generally correct in manually annotated corpora, but also very likely to be correct for automatically annotated corpora, if automatic segmentation of word forms was correct. The lemmatization accuracy of the NLP pipeline on correctly segmented text is 97.23% (see Zeldes and Schroeder 2016), though automatic segmentation accuracy is currently lower, at about 94.5%.

5 Evaluation

In order to evaluate the coverage of our dictionary, we use the publicly available corpora from the Coptic Scriptorium project (`http://copticscriptorium.org/`). We restrict the evaluation to the manually annotated corpora of literary Coptic and remove items overlapping lacunae from damaged manuscripts, and items not in Coptic script, as well as punctuation. The remaining data covers over 74,000 gold segmented running lexical items in Coptic, of which about 6,500 are foreign loanwords, not currently part of the target language covered by the dictionary, and 775 of which are proper names, which are mostly not covered. Table 2 gives the rate of coverage for tokens and types, as well as the numbers when loan words and proper names are not penalized.

	tokens			types		
	covered	total	%	covered	total	%
all lemmas	65,347	74,744	87.43	1,079	2,602	41.47
names ok	66,273	74,744	88.67	1,264	2,602	48.58
foreign ok	71,763	74,744	96.01	1,976	2,602	75.94
both ok	72,689	74,744	97.25	1,991	2,602	76.52

Table 2: Lexicon coverage on 80K lexical items from Coptic Scriptorium corpora.

The table shows that, for proficient users who can identify the need to look up foreign words in a Greek dictionary, and the presence of proper names, coverage at the token level is very good, with less than 3% of tokens in the corpora that might need to be in the dictionary (native common nouns) not being found. The situation is considerably worse for beginners, who may not be able to distinguish Greek words, and we are therefore planning to integrate loan words into the lexicon soon (see Section 6).

Looking at the type coverage, the situation is more partial: about 23.5% of native types are not covered. Given the high token coverage of the native vocabulary, it is clear that the remaining types without coverage are a large number of rare items. This is due to the 'long tail' of productive formations, created via incorporation, derivation and compounding processes, as shown in Section 3.1. We are therefore considering linking the morphologically analyzed sub-parts of words, which are outputted by the Coptic NLP pipeline's morphological analyzer, which are currently not linked. An evaluation of coverage using this strategy remains outstanding.

6 Conclusion and outlook

This paper has presented the Coptic Dictionary Online, a freely available, linked lexical resource for Coptic with definitions in English, French and German, cross-references to the main paper Coptic dictionary by W. Crum as well as other dictionaries, and frequency information connected to corpus search in a collection of open access digitized texts. The coverage of the lexicon for lexicalized native items is high, corresponding to all entries in Crum's dictionary, and if foreign words and names are ignored, covering over 97% of non-punctuation tokens in running text.

At the same time, integration of loan word definitions is a high priority for future work. A total of 8,406 Coptic lemmas of Greek loanwords, recorded by the DDGLC project (Almond et al., 2013), has just been integrated into our data set in May, and is currently being prepared for release after conversion to the TLA's entry semantics (checking spelling, grammatical information, and recording the source Greek lemmas). We also intend to integrate non-Sahidic vocabulary, compiled by W. P. Funk but as yet unpublished online, in the near future.

A further goal is integrating links for multi-word entries from corpora, and linking units smaller than words for productive derivations and incorporation. This development will require separate evaluation and is expected to increase our coverage of the remaining lexical types from the native Coptic vocabulary.

Acknowledgments

This work was supported by joint funding from the National Endowment for the Humanities Office of Digital Humanities and a bilateral NEH (HG-229371)/ DFG project (273503199). We also thank the Berlin-Brandenburg Academy of Sciences (BBAW), the Göttingen Academy of Sciences and Humanities, and in particular the Digital Edition of the Coptic Old Testament Project, as well as the DDGLC project at Leipzig and Freie Universität in Berlin for their contributions to the lexicon. We thank Sonja Dahlgren, Julien Delhez, Lena Krastel, Tonio Sebastian Richter and Anne Sörgel who contributed to compiling the lexical data and Mitchell Abrams for contributions to the Web interface.

References

Mathew Almond, Joost Hagen, Katrin John, Tonio Sebastian Richter, and Vincent Walter. 2013. Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen: Lehnwort-Lexikographie im Projekt Database and Dictionary of Greek Loanwords in Coptic (DDGLC). In Ingelore Hafemann, editor, *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pages 283–315, Berlin. BBAW.

Pierre Cherix. 2014. *Lexique copte dialecte sahidique*. Pierre Cherix, Geneva.

Walter E. Crum. 2000 [1939]. *A Coptic Dictionary*. Clarendon Press, Oxford.

Frank Feder. 2016. The integration of a Coptic lexicon and text corpus into the Thesaurus Linguae Aegyptiae. In Paola Buzi, Alberto Camplani, and Federico Contardi, editors, *Coptic Society, Literature and Religion from Late Antiquity to Modern Times. Proceedings of the Tenth International Congress of Coptic Studies, Rome, September 17th-22nd, 2012, and Plenary Reports of the Ninth International Congress of Coptic Studies, Cairo, September 15th-19th, 2008*, Orientalia Lovaniensia Analecta 247, pages 1375–1382.

Eitan Grossman. 2013. Greek loanwords in Coptic. In Georgios K. Giannakis, editor, *Encyclopedia of Ancient Greek Language and Linguistics*.

Eitan Grossman. 2014. Transitivity and valency in contact: The case of Coptic. In *47th Annual Meeting of the Societas Linguistica Europaea*, Poznań, Poland.

ISO. 2008. ISO24613:2008: Language resource management - lexical markup framework (LMF).

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Bentley Layton. 2011. *A Coptic Grammar*. Porta linguarum orientalium 20. Harrassowitz, Wiesbaden, third edition, revised and expanded edition.

Arietta Papaconstantinou, editor. 2010. *The Multilingual Experience in Egypt, from the Ptolemies to the Abbasids*. Ashgate Publishing, Farnham, Surrey and Burlington, Vermont.

Tonio Sebastian Richter. 2017. Whatever in the Coptic language is not Greek, can wholly be considered Ancient Egyptian: Recent approaches toward an integrated view of the Egyptian-Coptic lexicon. *Journal of the Canadian Society for Coptic Studies*, 9.

Laurent Romary. 2015. TEI and LMF crosswalks. *Journal for Language Technology and Computational Linguistics*, 30:47–70.

Nicholas Routledge, Linda Bird, and Andrew Goodchild. 2002. UML and XML schema. In *Proceedings of the 13th Australasian database conference*, pages 157–166, Melbourne.

Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

Stephan J. Seidlmayer and Ingelore Hafemann. 2011. *Handbuch zur Benutzung des Thesaurus Linguae Aegyptiae (TLA). Auf der Grundlage der Hilfetexte des Thesaurus Linguae Aegyptiae (TLA)*. BBAW, Berlin.

Jaroslav Černý. 1976. *Coptic Etymological Dictionary*. Cambridge University Press, Cambridge.

Werner Vycichl. 1983. *Dictionnaire étymologique de la langue copte*. Peeters, Leuven.

Wolfhart Westendorf. 2008 [1965–1977]. *Koptisches Handwörterbuch*. Carl Winter, Heidelberg.

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.