Integration of Greek-Coptic dataset (DDGLC) into the existing Coptic lexicon (TLA).

1. Data preparation

Input file: **0. DDGLC_Export_TLA_Vorab.xlsx**
Number of rows: **21452**

1.1 Source data clean-up

Exchange “coptic_usage::cl_type” values with TLA “pos” values

DDGLC ->	CLL	Commentary
(empty) (497 total) ->	?	προτρεπε, to urge forwards, urge on
Noun ->	Subst.	

[gen.] ->	Konj.	ϩⲟⲧⲁⲛ, when, whenever
Adjective ->	Adj.	
Adverb ->	Adv.	
Conjunction ->	Konj.	
Exclamation ->	Interjektion	ⲉⲩⲣⲉ, good!, well said!
impersonal Verb ->	unpersönlicher Ausdruck	ⲟⲩⲕ, it is not permitted, right for s.o. to do s.th. compare: ⲙⲛⲍⲟⲙ (ⲙⲙⲟ=), es ist nicht möglich
language island ->	l.i.	ⲕⲩⲣⲓⲱⲥ "valid and without hindrance" ⲧⲉⲧⲣⲁⲅⲱⲛⲟⲛ "from the four sides (defining the borders of a house or plot)"
Particle – Interrogative ->	Fragepartikel	
Particle – Discourse ->	Partikel	
Particle ->	Partikel	
Preposition ->	Präp.	
TBA (44 total) ->	?	
TBD (79 total) ->	?	
Verb ->	Vb.	
word combination ->	Subst.	

Delete rows where "coptic_usage::cl_lemma" (Coptic lemma) value is unclear

Values:

- values in brackets (e.g. (ⲅⲣⲁⲙⲙⲁ)) -> 12 total
- "Gloss" -> 1 total
- "incertum" -> 337 total
- "Unc. wh. & how abbr." -> 3 total
- (empty) -> 10 total

Deleted rows are saved in the second spreadsheet ("coptic_usage_cl_lemma_unclear": 364 rows)

Delete rows where "coptic_usage::cl_lemma" (Coptic lemma) value is unclear

Values

- (language island)
- (unknown)

Deleted rows are saved in the third spreadsheet ("greek_lemma_grl_lemma_unclear": 364 rows)

Insert values where Coptic Form (“attest_orthograhy”) is not set

- inserted from "attest_orthograhy_BU" (all in brackets): 9 rows

Exchange Greek characters with Coptic in Coptic forms (“attest_orthograhy”)

- use the sorting function
- not everything was exchanged, e.g.(ϊ for lemma н and ιδ for lemma ντεκτιανος is in).

Output File: 1. DDGLC_Export_TLA_Cleaned_Up
Sort: according to "attest_orthograhy"
No of Rows: 21079

1.2 Sorting the rows for conversion

Background information: DDGLC and TLA data models.

On top of DDGLC data hierarchy is a Greek lemma, to which then Coptic “usages” are assigned, formally represented by “attest_orthograhy_BU” (the diplomatic spelling of the usage) and “attest_orthograhy” (the normalized spelling of the usage). Apart from the spelling each “usage” has the following properties:

- “coptic_usage::cl_type” is part of speech information assigned to the usage;
- “coptic_usage::cl_lemma” is Coptic lemma assigned to the usage;
- “coptic_usage::cl_meaning” is Coptic meaning assigned to the usage;
- “coptic_usage::cu_hierarchy represents the internal hierarchy of Coptic meanings attributed to the given “usage”

Part of speech information in DDGLC is not a property of Coptic lemma, but of Coptic usage and Greek lemma.

TLA data, on the other hand, has Coptic lemma (standard spelling + part of speech information) on top of the hierarchy. To Coptic lemma multiple forms and senses (meanings) can be assigned.

Sorting the rows of the DDGLC: possible options

Option1: columns are sorted according to the **spelling of Greek lemmata and Coptic usages**:

1)	greek_lemma::grl_lemma	Greek lemma
2)	attest_orthography	normalized orthography of the Coptic usage
3)	coptic_usage::cu_hierarchy	semantic hierarchy of senses within the Coptic usage

Setting “attest_orthography” as sort criterium removes the internal hierarchy of senses (coptic_usage::cu_hierarchy), which is important for DDGLC (i.e. in the XML output the forms will be listed orthographically, but the relevant senses will be messed up).

Option 2: columns are sorted according to the **spelling of Greek lemmata and semantic hierarchy of Coptic senses**:

1) greek_lemma::grl_lemma — Greek lemma
2) coptic_usage::cu_hierarchy — semantic hierarchy of senses within the Coptic usage
3) attest_orthography — normalized orthography of the Coptic usage

Setting “coptic_usage::cu_hierarchy” as a major criterium, messes up the order of “attest_orthography”.

Decision: produce two versions:

- “2. DDGLC_Export_TLA_Cleaned_Up_Sort_Orthography.xlsx” sorted according to the orthography of Coptic usages
- “2. DDGLC_Export_TLA_Cleaned_Up_Sort_cuHierarchy.xlsx” sorted according to the semantic hierarchy of Coptic senses.
- For the conversion **“2. DDGLC_Export_TLA_Cleaned_Up_Sort_Orthography.xlsx”** should be used, the senses will be corrected in the XML output file manually.

2. Data conversion

2.1 XSLX to CSV

Xpath would get confused with (') and (") in step (2.3), therefore:

- Replace all (') with (s_qq): 592 instances
- Replace all (") with (d_qq): 1254 instances (necessary, because otherwise in Step 4.

Output File: 3. DDGLC_Export_TLA_Cleaned_Up_Sort_Orthography_export.csv
Rows: 21079

2.2 CSV to TLA XML

Each CSV row becomes an XML entry. Model:

```
<entry>
 <form type="lemma">
  <orth>ⲁⲃⲁⲥⲓⲗⲉⲩⲧⲟⲥ</orth>
 </form>
 <form>
  <orth>ⲁⲃⲁⲥⲓⲗⲉⲩⲧⲟⲥ</orth>
  <usg type="geo">S</usg>
 </form>
 <gramGrp>
```

```
 <pos>Adj.</pos>
</gramGrp>
<etym>
 <ref type="greek_lemma::grl_ID">3742</ref>
 <ref type="greek_lemma::grl_lemma">ἀβασίλευτος</ref>
 <ref type="greek_lemma::grl_meaning">kingless, unruled</ref>
 <ref type="greek_lemma::grl_pos">adjective</ref>
 <ref type="greek_lemma::grl_ref">LSJ 2a</ref>
</etym>
<sense>
 <cit type="translation" xmllang="en">
  <quote>unruled </quote>
 </cit>
 <cit>
  <bibl/>
 </cit>
</sense>
</entry>
```

NB:

- <Sense> combines „coptic_usage::cl_meaning“ + „coptic_usage::cu_criterion““.
- Greek info goes into <entry> <etym>
- "Attest_ID" is not recorded (DDGLC decision)
- For the detailed list of concordances see document “DDGLC_CLL_Concordance”.

Correct Unicode errors (if any): “Format and Indent” (Oxygen Editor: select “Document / Source / Format and Indent”) should be successful as soon as the errors are corrected.

Scripts:

- CSV to XML: "1_read_cvs.py"
- Debug (number of XML entries = number of CSV rows): "0_read_cvs.py":.

Output File: **“1.simple_xml_sqq_dqq.xml”**
Number of entries: **21079**

2.3 Merge identical XML elements

Task: compare and merge XML elements according to the following criteria:

- Entries (<entry>) are merged if values in <form type = „lemma“> <orth> AND <gramGrp><pos> are identical.
- Forms (<form>) are merged if values in <form><orth> AND <form><usg type="geo"> are identical
- Senses (<sense>) are merged if values in <sense><cit type="translation" xml:lang="en"><quote> are identical.

NB:

- Greek info is taken from the first of the merged entries.

Script: **"1.merge.py".**
Output File: **“2.xml_out_unsorted.xml”**

2.4 Sort elements in the output XML

There is a problem with the order of elements in the output files (e.g. “forms” follow “senses”), so we need to:

- Sort the elements
- Find and replace (s_qq) with ('): 185 matches
- Find and replace (d_qq) with (“): 321 matches

Script: **"2.allign.py".**
Output file: **3.xml_out_sorted.xml**
No of entries: **3972**
No of forms: **17101**

3. Data debugging and proofreading

3.1 Create CSV / XLS for QA purposes

Script: “3.debug.py".
Output files: “4. CLL.csv”, “4. DDGLC-CLL Debug.xlsx”

NB:

- Open “CLL.csv” in Notepad and save as UTF-8
- Import this file in Excel, select all and set the font to Antinou.
- Save this file as “4. DDGLC-CLL Debug.xlsx”

3.2 Proofread “4. DDGLC-CLL Debug.xlsx”

- DDGLC team went through the export file and checked the concordances of values of Coptic lemmata, forms, pos and senses.
- Types of corrections that were made:
  - Missing POS values inserted
  - Removed verbal compounds ('ⲣⲁⲓⲧⲓ' removed; ⲁⲓⲧⲉⲓ remained)
  - Removed other compounds (ϣⲁ ⲉⲛⲉⲣ ⲁⲙⲏⲛ from ⲣⲁⲙⲏⲛ)

- Removed forms with definite and indefinite articles (ⲟⲩⲁⲛⲁⲅⲕⲁ̈ⲟⲛ removed, ⲁⲛⲁⲅⲕⲁⲓⲟⲛ remained; ⲧⲁⲛⲁⲅⲕⲏ removed, ⲁⲛⲁⲅⲕⲏ remained)
- Some wrongly assigned forms removed and re-assigned to proper lemmata (e.g. ⲡⲣⲟⲥ was in ⲁⲓⲧⲏⲥⲓⲥ)
- All expressions called “language island” removed

Output file: **“4. DDGLC-CLL 03.xlsx”**

3.3 Export “4. DDGLC-CLL 03.xlsx” as TLA XML

Corrected *.xlsx file was used to generate *.xml according to TLA XML Schema. The following addition

- New lemmata and forms, which appeared in the DDGLC DB after December 2017 (the export date of the source file for conversion), inserted. Outdated lemmata and forms removed.
- “Coptic_usage::cu_ID” (information pertaining to Coptic usages in DDLC database) inserted in each “sense” as <ref type="coptic_usage::cu_ID">4562</ref> . This might be useful for later TLA –DDGLC cross-references.
- The order of senses in each lemma corrected
- Bibliographic information pertaining to Coptic lemma encoded as <cit> <bibl>DDGLC</bibl>
- Lemma and form IDs assigned

NB:

- Nouns deriving from Greek roots differentiated in in gender, e.g. ⲁⲅⲅⲉⲗⲓⲕⲏ (fem. in Geek), ⲁⲅⲅⲉⲗⲓⲕⲟⲥ (masc. in Greek), ⲁⲅⲅⲉⲗⲓⲕⲟⲛ (fem. In Greek), are treated as separate lemmata, but possess no gender information in Coptic.
- Verbs deriving from Greek roots using different borrowing mechanisms, e.g. (ⲉⲣ-) + ⲁⲅⲁⲡⲁⲛ, B (aux + Greek infinitive) vs. ⲁⲅⲁⲡⲁ, S (Greek imperative); ⲁⲛⲁⲭⲱⲣⲓⲛ (B, infinitive) vs. ⲁⲛⲁⲭⲱⲣⲉⲓ (S, imperative) are treated as separate lemmata.

4. Data release

4.1 Extend the TEI Header

- **Coptic Lexicon TEI Header scheme:**
  1. <fileDesc> file description

     1.1 titleStmt

     1.1.1 title

     1.1.2 author

1.1.3 editor

1.1.4 sponsor

1.1.5 funder

1.1.6 principal

1.1.7 respStmt

1.1.7.1 resp

1.1.7.2. name

1.2 editionStmt

1.2.1 edition

1.2.2 respStmt

1.2.2.1 resp

1.2.2.2 name

1.3 publicationStmt

1.3.1 publisher

1.3.1.1 pubPlace

1.3.1.2 address

1.3.1.3 idno

1.3.1.4 availability

1.3.1.5 date

1.3.1.6 license

1.3.2 distributor

1.3.3 authority

1.4 seriesStmt

1.5 notesStmt

1.6 sourceDesc

16.1 head

1.6.2 bibl

1.6.3 biblFull

1.6.4 biblStruct

1.6.5 listBibl

1.6.6 msDesc

2. <encodingDesc> encoding description

2.1 projectDesc

2.2 schemaSpec

2.3 schemaRef

3. <profileDesc> text profile

3.1 abstract

3.2 creation

3.3 languageUsage

3.3.1 language

3.4 textClass

3.4.1 keywords

4. <revisionDesc> revision history

4.1 listChange

4.4 change

4.2 Adopt XML-Schema

To accommodate the new types of information the XML Schema was adopted:

- Element <etym> should allow Greek info:
  <ref type="greek_lemma::grl_ID">
  <ref type="greek_lemma::grl_lemma">
  <ref type="greek_lemma::grl_meaning">
  <ref type="greek_lemma::grl_pos">
  <ref type="greek_lemma::grl_ref">
- “MinOccurs” of element "cit" in “sense” was changed from 4 to 1: <xs:element name="cit" minOccurs="1" maxOccurs="unbounded">. Explanation: “sense” in TLA Coptic lexicon contains 4 “cits” – three languages and one bibliography. Loan word entries contain two "cit"s only - one with English translation and another with bibliography. . It was checking that each entry has 3 translation + bibliography. I think it is Ok to change it to 1, as, the missing translations and bibliography can be checked with python script.
- Allow <usg type="geo"> to include V ("South Fayyumic") and W ("Crypto-Mesokemic")

DDGLC	CLL	Commenraty
A	A	
B	B	
F	F	

L	L	
M	M	
S	S	
V	Extend XML Schema with “V”	“South Fayyumic”
W	Extend XML Schema with “W”	“Crypto-Mesokemic”[1]
	P	
	Ak	

- Attribute <xs:attribute name="change"> with value <xs:enumeration value="added"/> allowed to mark the forms, added to DDGLC database after December 2017 (after the initial export of DDGLC database).

4.3 Prepare Release notes

	Publication (article)	Raw-Data	Processed data
Coptic dataset	Coling DOI	Github URL, Refubium DOI	
Greek-Coptic dataset	-	Github, Refubium	

[1] “Fayyumic, together with Mesokemic (M) and two minor dialects (W and V), is considered to form the so called “Middle Coptic major group” (Kasser 1991a and 1991c). Their inter-relation is not entirely clear, but they are closely related. The dialect W is attested in a single fragmentary papyrus (P. Mich 3521. published by Elinor Husselmann (1962)), which holds chapters from the Gospel of John. It has Fayyumic characteristics in orthography, though without lambdacism, but its morphosyntax is closer to Middle- Egyptian. Sometimes it is also described as “Crypto-Mesokemic”. (25: Note that formerly it was labeled as V by Rodolphe Kasser (1981: 115)) The dialect V4 looks more like Fayyumic, without lambdacisms again. It is also called South Fayyumic, and sometimes considered a subdialect of F4. Biblical texts (Ecclesiastes, 1 John, 2 Peter) in this dialect can be found in P. Mich 3520, which was published only recently by Hans-Martin Schenke and Rodolphe Kasser (2003). Because of its neutral phonology, Rodolphe Kasser (1991a: 99) hints to the possibility that it functioned as a regional vernacular. The whole group is said to be subject to a strong influence of Sahidic (26) For other minor varieties that can be related to this group, see Kasser (1990: 147). For Fayyumic in general, the main reference is the article of Rodolphe Kasser in The Coptic Encyclopedia (Kasser 1991c). (B. Egedi 2012: 24-25)