The Coptic Universal Dependency Treebank

Amir Zeldes & Mitchell Abrams

{amir.zeldes, MJA284}@georgetown.edu

UDW18@ EMNLP Brussels. Belgium

Image: Österreichische Nationalbibliothek

GEORGETOWN UNIVERSITY

C@GU Corpling@GU

Language Overview

- *Coptic is an **Afro-asiatic language** and part of the longest continuously documented human language on Earth (over 4000 years!)*
- *Written in the early first millennium CE using a variant of the Greek alphabet, with added letters for Egyptian sounds absent from Greek.*
- *Text separated into multi-word units known as **bound groups:***

  ϩⲙ-ⲡ-ⲣⲁⲛ hm-p-ran 'in-the-name'

- *Fusional morphology, classified as **agglutinative/polysynthetic** (Loprieno 1995)*
- ***Sahidic dialect** - the classical form of the language*

Corpus contents

source	genre	documents	tokens	sents
translated				
Apophthegmata Patrum	hagiography	1–6, 18–19, 23–26	1,318	62
Gospel of Mark	Bible (narrative)	Chapters 1–6	7,087	248
1 Corinthians	Bible (epistle)	Chapters 1–6	3,571	124
original				
Shenoute, Discourses 4	sermons	Not Because a Fox Barks	2,553	97
Shenoute, Canons 3	sermons	Abraham our Father (XL93-94)	579	26
		Acephalous 22 (YA421-28)	1,703	43
Letters of Besa	letters	Letters 13, 15, 25	1,981	93
Martyrdom of Victor	martyrdom	Chapters 1–6	1,985	88
total			20,777	781

Tagged source data from the **Coptic Scriptorium** project (http://copticscriptorium.org/, Schroeder & Zeldes 2016)

Segmentation

12-14	ϩⲙⲡⲉϥϩⲃⲏⲩⲉ							
12	ϩⲛ	in	ADP	PREP	...	14	case	Orig=ϩⲛ|SpaceAfter=No
13	ⲡⲉϥ	his	DET	PPOS	...	14	det	SpaceAfter=No
14	ϩⲃⲏⲩⲉ	deeds	NOUN	N	.	9	obl	
15-16	ⲙⲁⲁⲧⲡⲉⲩⲯⲩⲭⲏ							
15	ⲙ	of	ADP	PREP	...	16	case	Orig=ⲙ|SpaceAfter=No
16	ⲙⲁⲧⲣⲉⲩⲅⲉⲧⲃⲯⲩⲭⲏ	soul-killing	NOUN	N	.	14	nmod	Morphs=ⲙⲁⲧ-ⲣⲉⲩ-ϩⲉⲧⲃ-ⲯⲩⲭⲏ

Three levels of granularity

1. Bound groups at the highest level
2. POS tag bearing units- corresponding to CoNLL-U tokens (normed + orig in MISC)
3. Units below POS tag level for incorporated components (in MISC field)

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

POS tags

- **Mapping of *Coptic Scriptorium* (CS) Fine tags to UPOS**

CS	UPOS	CS	UPOS
A*	AUX	FUT	AUX
ACAUS	VERB	IMOD	ADV
ACOND	SCONJ	N	NOUN
ADV	ADV		
ALIM	SCONJ	NEG	ADV
APREC	SCONJ	NPROP	PROPN
ART	DET	NUM	NUM
CCIRC	SCONJ	PDEM	DET
CCOND	SCONJ	P*	PRON
CFOC	PART	PPOS	DET
CONJ	CCONJ	PREP	ADP
COP	PRON	PTC	PART
CPRET	AUX	PUNCT	PUNCT
CREL	SCONJ	UNKNOWN	X
EXIST	VERB	V*	VERB
FM	X		

Dependencies

Absent relations

- UD Coptic uses all UD relations except **expl** and **clf** – Coptic does not have expletive pronouns or classifiers
- ***:pass subtypes** (i.e. nsubj:pass and aux:pass) not used because of ambiguous nature of passives.*

Problematic Constructions

Damaged manuscripts

- *Excerpt of papyrus (located on header) demonstrates challenge of handling **lacunae***
- *Attached normally if high confidence POS tag is known*
- *Incomprehensible phrases, tagged CS UNKNOWN (UPOS: X), attach all tokens to root as **dep** relation*

Appositions

- *enclitic particles (e.g.de 'but, and') appear in 2nd position in sentence and break up two appositional nominals*
- *analyze as **appos***
  - *particle always in 2nd position*
  - *prefer same treatment in all positions*
  - *same words as appos if not sentence initial ("he saw the king, Diocletian … ")*

But Diocletian, the king, went

ⲡ	ⲣⲣⲟ	ⲇⲉ	ⲇⲓⲟⲕⲗⲏⲧⲓⲁⲛⲟⲥ	ⲁ	ϥ	ϩⲣⲟⲕⲉⲩⲉ
the	king	but	Diocletian	PST	he	went
DET	NOUN	PART	PROPN	AUX	PRON	VERB

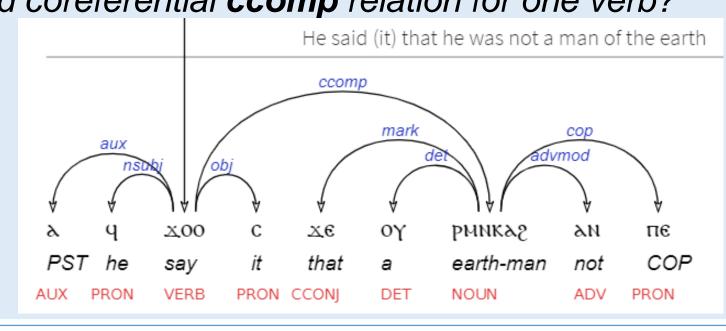#### Direct Objects with Preposition ⲛ-

- *Variation in object marking depending on tense:*
  - *object mediated by ⲛ-/ⲙⲙⲟ= in durative tenses*
  - *no case marker in tenses with auxiliary before subject*
- *Both analyzed as **obj**, distinguishes usage from actual preposition ⲛ-/ⲙⲙⲟ= meaning 'of'*
- Similar to differential object marking in Hebrew & others

Double object construction?

Cf. Zeman 2017, Przepiórkowski & Patejuk 2018

- *Object clauses can co-occur with correlate pronouns*
- *Use **obj** and coreferential **ccomp** relation for one verb?*

He said (it) that he was not a man of the earth

dislocated?
expl?
advcl?

ⲁ	ϥ	ϫⲟⲟ	ⲥ	ϫⲉ	ⲟⲩ	ⲣⲙⲛⲕⲁϩ	ⲁⲛ	ⲡⲉ
PST	he	say	it	that	a	earth-man	not	COP
AUX	PRON	VERB	PRON	CCONJ	DET	NOUN	ADV	PRON

Evaluation

annotators	tokens	UAS (% agreement)		LS (kappa)	
		punctuation	no punctuation	punctuation	no punctuation
Group A: Pre-Adjud.	276	81.1%	79.0%	0.78	0.75
Group A: Post-Adjud.	319	87.7%	86.5%	0.88	0.86
Group B: Pre-Adjud.	287	84.3%	82.9%	0.79	0.76
Group B: Post-Adjud.	297	86.5%	84.6%	0.81	0.79
Expert	703	96.0%	95.8%	0.93	0.92

- *Three pairs of annotators:*
  - *two undergraduate student groups*
  - *one expert group (grad student/faculty)*
- *Annotated using Arborator (Gerdes 2013)*
- *Results based only on reading guidelines (Pre-Adjud) and after comparing errors in first round (Post-Adjud)*
- *Very good inter-annotator agreement scores, especially with the challenge of being non-native Coptic speakers*

Common Errors

- *Coordination scope- often ambiguous in translation*
- *Confusion over labeling objects (**obj**) and obliques (**obl**)*
- *Assigning an (**obl**) relation to a verb or an (**nmod**) relation to its dependent noun*

Future work

- *Expand the treebank!*
  - *More texts*
  - *New genres*
  - *Reach shared-task size*
- *Train and evaluate automatic parsers*
- *Get subcategorization info from electronic dictionary (Feder et al. 2018)*
- *Use parser output for entity recognition (WIP)*
- *Add similar analyses for other Coptic dialects*

QR scan: e-poster available

References

- Feder, F. , Kupreyev, M., Manning, E., Schroeder, C., Zeldes, A. 2018. A linked Coptic dictionary online. LaTeCH 2018. Santa Fe, NM, 12-21.
- Gerdes, K. 2013. Collaborative dependency annotation. *DepLing 2013*. Prague, 88–97.
- Loprieno, A. 1995. *Ancient Egyptian. A Linguistic Introduction.* Cambridge University Press, Cambridge.
- Przepiórkowski , A. & Patejuk, A. 2018. Arguments and adjuncts in Universal Dependencies. *COLING2018*, Santa Fe, NM, 3837–3852.
- Schroeder, C. T. & Zeldes, A. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).
- Zeman, D. 2017. Core arguments in Universal Dependencies. *Depling 2017*, Pisa, 287–296.