

22nd International Workshop on Treebanks and Linguistic Theories (TLT 2024)

Developing the Egyptian-UJaen Treebank

Hamburg KorpusLab — 5 December 2024

Roberto Antonio Díaz Hernández
University of Jaén



Marco Carlo Passarotti
Università Cattolica del Sacro Cuore



Sections

- 1) Introduction.
- 2) Issues.
- 3) Training a model.
- 4) Conclusion.

I. Introduction

Egyptian in Universal Dependencies



Egyptian

1

14K



Afro-Asiatic, Egyptian

Egyptian treebanks



Ujaen

14K



Egyptian-Ujaen is the first dependency treebank created for the morphosyntactic annotation of pre-Coptic Egyptian. Its current state (UD v2.15) consists of 1,573 sentences and 14,650 words manually annotated from texts written in Old Egyptian, mainly from the Pyramid Texts.

- Contributors: Roberto Antonio Díaz Hernández
- Repository [master](#) [dev](#)
- [README](#)
- [Treebank hub page](#)
- [Download](#)



Coptic

1

57K



Afro-Asiatic, Egyptian

Coptic treebanks



Scriptorium

57K



UD Coptic contains manually annotated Sahidic Coptic texts, including Biblical texts, sermons, letters, and hagiography.

- Contributors: Mitchell Abrams, Elizabeth Davidson, Amir Zeldes
- Repository [master](#) [dev](#)
- [README](#)
- [Treebank hub page](#)
- [Download](#)

Egyptian

Earlier Egyptian

Old Egyptian
(ca. 2700–2000 BC)

Middle Egyptian*
(ca. 2000–1550BC)

Later Egyptian

Late Egyptian
(ca. 1550–700 BC)

Demotic
(ca. 7th century BC to
5th century AD)

Coptic
(4th century to 14th
century AD)

* Middle Egyptian became a standardised and classical language from 1550 BC onwards

Egyptian text corpora for the EUJA treebank

Old Egyptian

Pyramid Texts

Old Kingdom and
First Intermediate
Period biographical
texts

Middle Egyptian

Coffin Texts

Middle Kingdom
biographical texts

Literary texts

Classical Egyptian

The Book of the
Dead

18th Dynasty
biographical texts

Literary texts

Late Egyptian

New Kingdom
biographical texts

Literary texts



Administrative texts






Demotic

Literary texts



Administrative texts



The Egyptian-Ujaen Treebank


 UniversalDependencies / UD_Egyptian-UJaen







[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

 [master](#) [UD_Egyptian-UJaen / README.md](#) 

 **UD-Egyptian** Update README.md fc84df4 · 2 weeks ago [History](#)

[Preview](#) [Code](#) [Blame](#) 122 lines (97 loc) · 6.64 KB [Raw](#)    

Summary

Egyptian-UJaen is the first dependency treebank created for the morphosyntactic annotation of pre-Coptic Egyptian. Its current state (UD v2.15) consists of 1,573 sentences and 14,650 words manually annotated from texts written in Old Egyptian, mainly from the Pyramid Texts.

Composition

1	# sent_id = EUJA-1	text	origin / place	date	genre	language
2	# ref =	EdeI 2008 (eds. Seyfried/Vieler): pl. LV, line 5	Qubbet el-Hawa (QH35e)	6th Dynasty	histbio	OE
3	# text =	iw rč.n (i) t' n ḥkr				
4	1	iw iw AUX Particle Aspect=Perf 2 aux _ Hiero=𓂏				
5	2	rč.n rči VERB SPC=Past-2 Type=Pred Tense=Past VerbForm=Fin 0 root _ Hiero=𓂏:𓂏				
6	3	(i) i PRON Pron=SFP Gender=Com Number=Sing Person=1 PronType=Prs 2 nsubj _ Hiero=No				
7	4	t' t' NOUN Hierocl=Yes Gender=Masc Number=Sing 2 obj _ Hiero=𓂏:𓂏				
8	5	n n ADP Status=Cons Case=Dat 6 case _ Hiero=𓂏				
9	6	ḥkr ḥkr NOUN _ Gender=Masc Number=Sing 2 iobj _ Hiero=𓂏(𓂏:𓂏)				
10						
11	# sent_id = EUJA-2					
12	# ref =	Pyramid Texts § 1775a N	Saqqara	6th Dynasty	rel	OE
13	# text =	t3 m 3w.t-ib				
14	1	t3 t3 NOUN _ Gender=Masc Number=Sing 3 nsubj _ Hiero=𓂏:𓂏				
15	2	m m ADP Status=Cons Case=Ess 3 case _ Hiero=𓂏				
16	3-4	3w.t-ib _ _ _ _ _ Hiero=(𓂏:𓂏)𓂏				
17	3	3w.t 3wi NOUN _ Definite=Cons Gender=Masc 0 root _ Hiero=(𓂏:𓂏)𓂏				
18	4	ib ib NOUN MWE=Yes Case=Gen Gender=Masc Number=Sing 3 compound _ Hiero=𓂏				
19						
20	# sent_id = EUJA-3					
21	# ref =	TPPI § 15, 13, Dra Abu el-Naga, First Intermediate Period, histbio, OE				
22	# text =	ḥkn ṣč m <nh m 3w.t-ib				
23	1	ḥkn ḥkn VERB SPC=Sub Mood=Sub 0 root _ Hiero=𓂏(𓂏:𓂏:𓂏)				
24	2	ṣč č PRON Pron=SFP Gender=Fem Number=Sing Person=2 PronType=Prs 1 nsubj _ Hiero=𓂏				
25	3	m m ADP Status=Cons Case=Ess 4 case _ Hiero=𓂏				
26	4	<nh <nh NOUN _ Gender=Masc Number=Sing 1 obl _ Hiero=𓂏(𓂏:𓂏)				
27	5	m m ADP Status=Cons Case=Ess 6 case _ Hiero=𓂏				
28	6-7	3w.t-ib _ _ _ _ _ Hiero=(𓂏:𓂏)𓂏				
29	6	3w.t 3wi NOUN _ Gender=Masc 1 obl _ Hiero=(𓂏:𓂏)				
30	7	ib ib NOUN MWE=Yes Case=Gen Gender=Masc Number=Sing 6 compound _ Hiero=𓂏				

Annotation of the Pyramid Texts

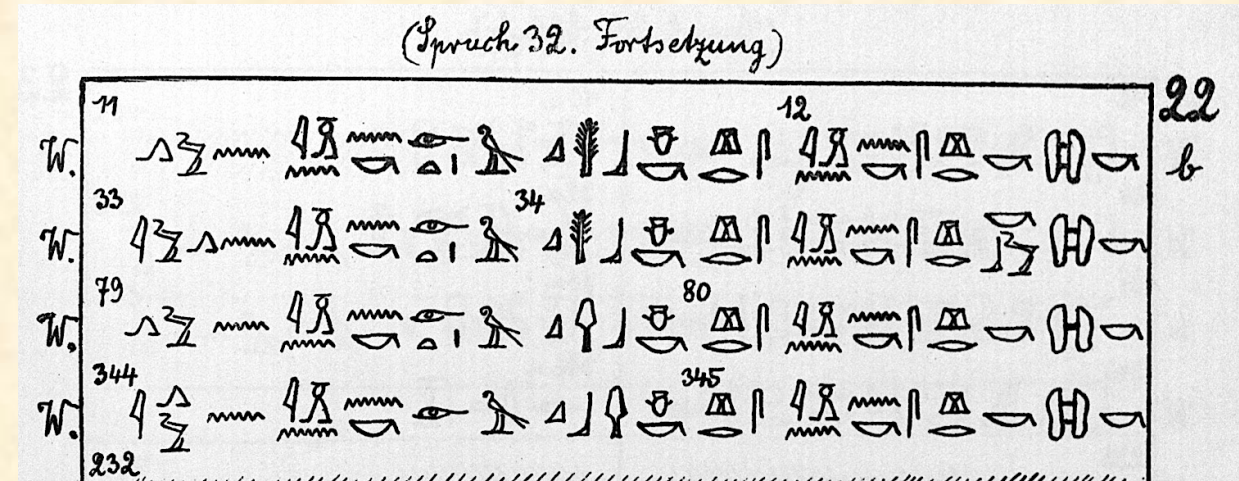
What are the Pyramid Texts?

The Pyramid Texts are a collection of spells recorded on the walls of the pyramids of Old Kingdom kings. They contain more than eight hundred magic formulae recited during mortuary rituals for the king.

The **Pyramid of Unas** at Saqqara contains the oldest version of the Pyramid Texts, dated to 2353-2323 BC.



Unas's Pyramid Texts



Sethe's edition of the Pyramid Texts (1908-1922)

II. Issues

1. One treebank for pre-Coptic Egyptian or a treebank for each stage of this language?

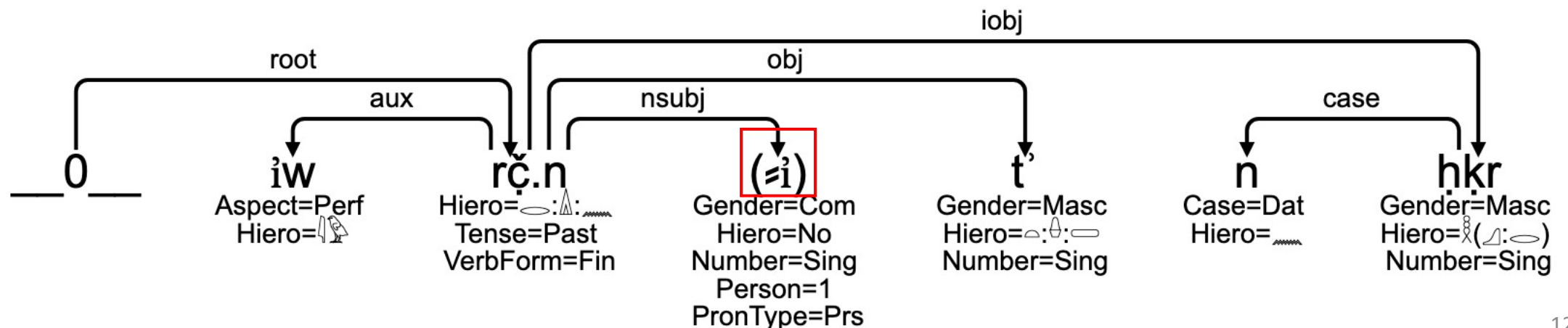
2. Hieroglyphs

```

1 # sent_id = EUJA-1
2 # ref = Edel 2008 (eds. Seyfried/Vieler): pl. LV, line 5, Qubbet el-Hawa (QH35e), 6th Dynasty, histbio, OE
3 # text = iw rç.n (≠i) t' n ḥkr
4 1 iw iw AUX Particle Aspect=Perf 2 aux _ Hiero=𓂏
5 2 rç.n rçἱ VERB SPC=Past-2|Type=Pred Tense=Past|VerbForm=Fin 0 root _ Hiero=𓂏:𓂏:𓂏
6 3 (≠i) i PRON Pron=SFP Gender=Com|Number=Sing|Person=1|PronType=Prs 2 nsubj _ Hiero=No
7 4 t' t' NOUN Hierocl=Yes Gender=Masc|Number=Sing 2 obj _ Hiero=𓂏:𓂏:𓂏
8 5 n n ADP Status=Cons Case=Dat 6 case _ Hiero=𓂏
9 6 ḥkr ḥkr NOUN _ Gender=Masc|Number=Sing 2 iobj _ Hiero=𓂏(𓂏:𓂏)

```

trans = (I) have given bread to the hungry



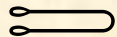
3. Unicode Font for Hieroglyphs

1085 8 s3č s3č VERB _ Gender=Masc|VerbForm=Inf 1 parataxis _ Hiero=(:≡)UC_14386

13360

Egyptian Hieroglyphs

1342F

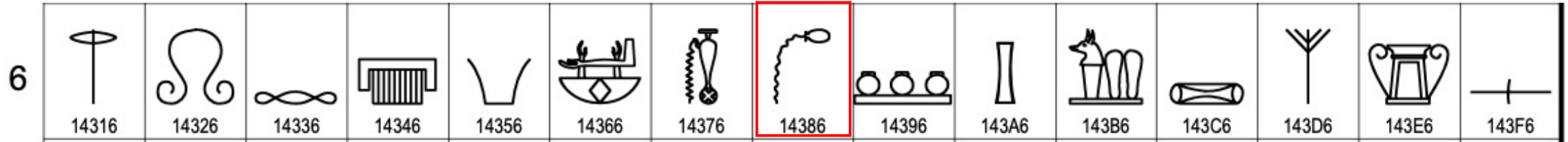


Hiero=(:≡)UC_14386

14310

Egyptian Hieroglyphs Extended-A

143FF



Hiero=(:≡)UC_14386



























Unicode Control Characters


1. Colon (:) to indicate subordination of signs, for example $\square : \text{~~~~~}$ corresponds to $\begin{smallmatrix} \square \\ \text{~~~~~} \end{smallmatrix} pn$ “this”.


2. Brackets () to segment groups of hieroglyphs.

3. Asterisk (*) to indicate the juxtaposition of hieroglyphs, for example $(\square * \triangle :)\text{=====}$ corresponds to $\begin{smallmatrix} \square & \triangle \\ \text{=====} \end{smallmatrix} p.t$ “sky”.

4. Transcription System

	LUT	Tübingen	Unicode
	ʒ	ʒ	A723
	l	l	A7BD
	y	y	
	ï	ï	00EF
	‘	‘	A725
	w	w	
	b	b	
	p	p	
	f	f	
	m	m	
	n	n	
	r	r	
	h	h	
	ḥ	ḥ	1E25
	ḥ	ḥ	1E2B
	ḥ	ḥ	1E96
	z	s	
	s	ś	015B
	š	š	0161
	q	ḵ	1E33
	k	k	
	g	g	
	t	t	
	<u>t</u>	č	010D
	d	ṭ	1E6D
	ḏ	č	010D+0323

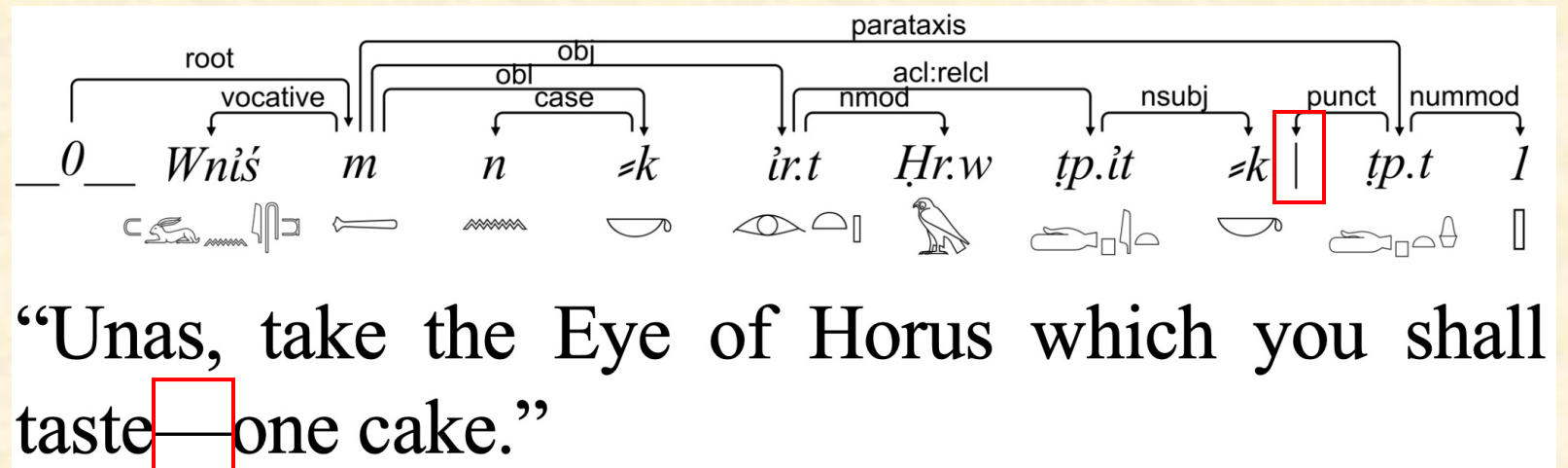
 = [t̪] = č voiceless postalveolar affricate like in Czech
 ≠ t used for /θ/ (ث) in Arabic

 = [t̪ʰ] = ṭ alveolar and dental ejective stops like in Ge'ez
 ≠ d

Recitation text

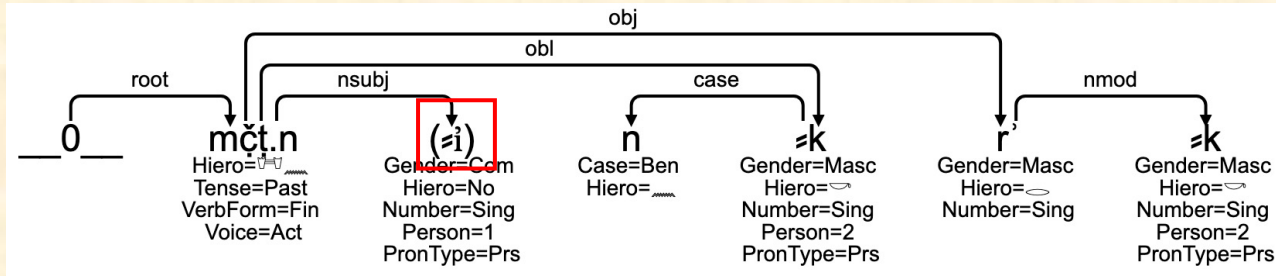
Line

Ritual remark



Leiden System for editing ancient texts

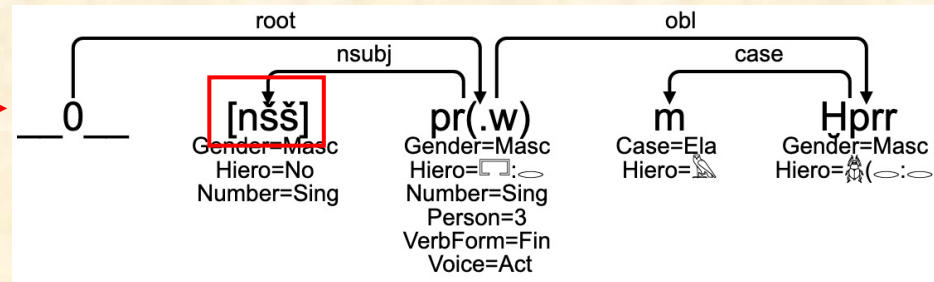
1) Brackets ()



trans (I) hit your mouth for you.

(PT 11b N = EUJA-70)

2) Square brackets []

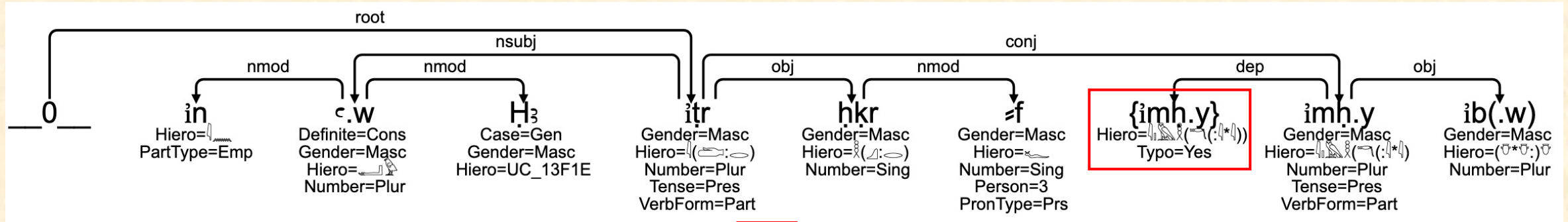


trans: [The spittle] came from Kheprer.

(PT 199a W = EUJA 642)

3) Curly brackets {}

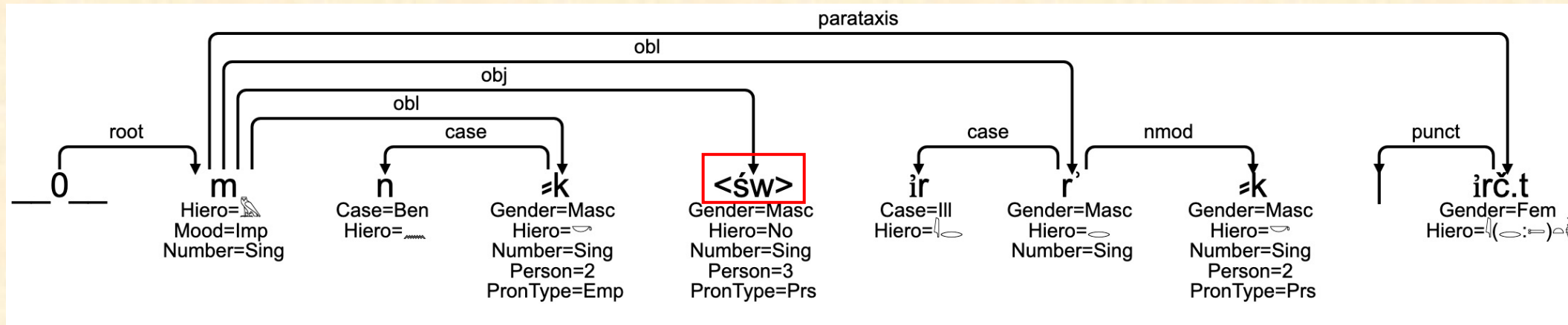
(PT 119b W = EUJA-381)



trans: It is the arms of Ha that drive away his hunger and {fill} fill the hearts.

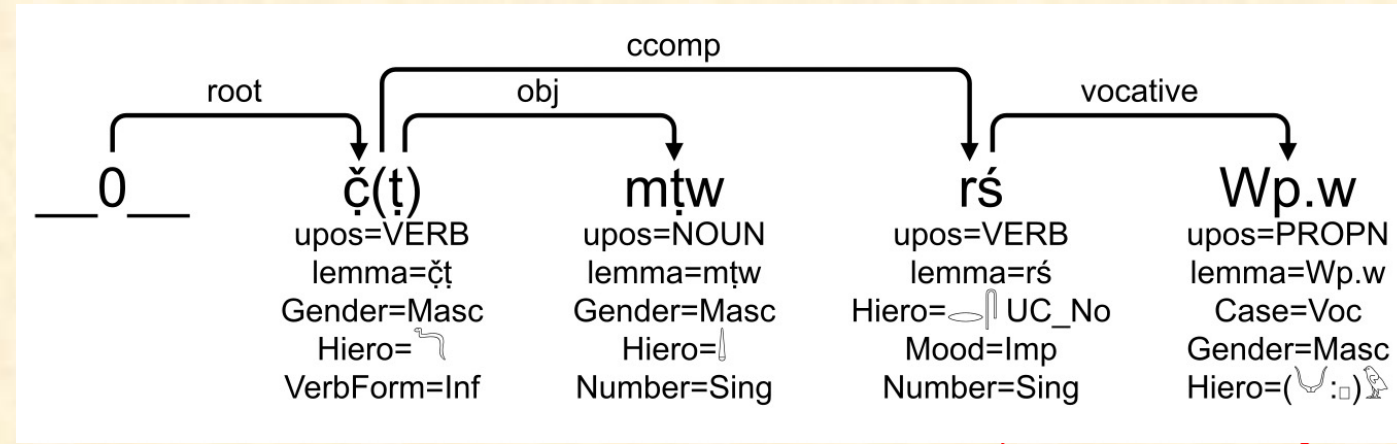
4) Angle brackets <>

(PT 32a W = EUJA-154)



trans: Take <it> to your mouth - milk.

5. Sentence segmentation

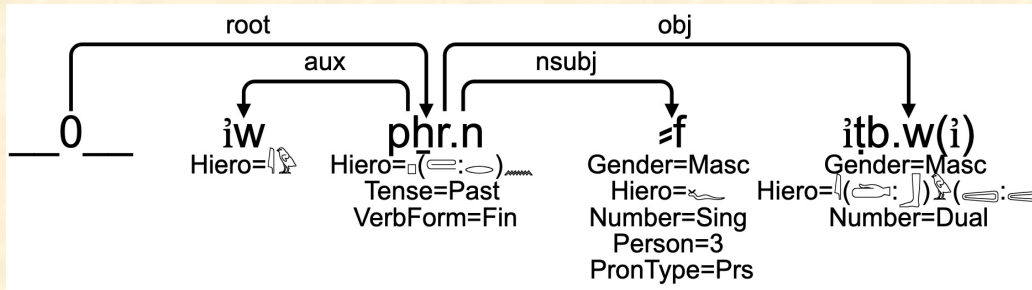


trans = Saying a speech (i.e. recitation): "Awake, (O) Wepu!"

(PT 126a W = EUJA-410)

Three types of sentences

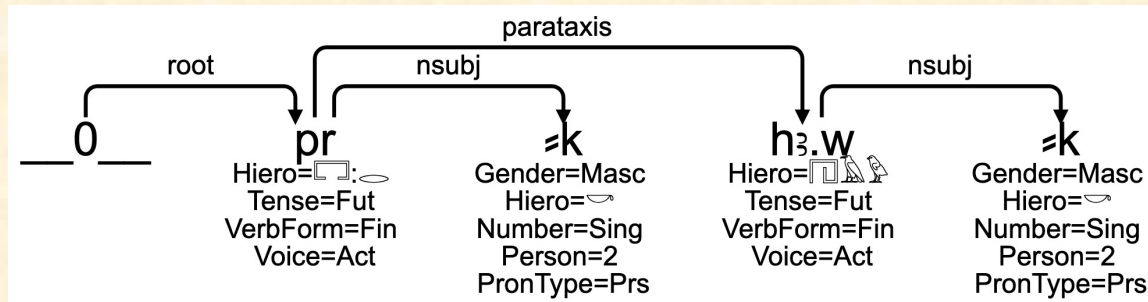
1) Independent sentences



trans = He has circled the Two Banks.

(PT 406c W = EUJA-1268)

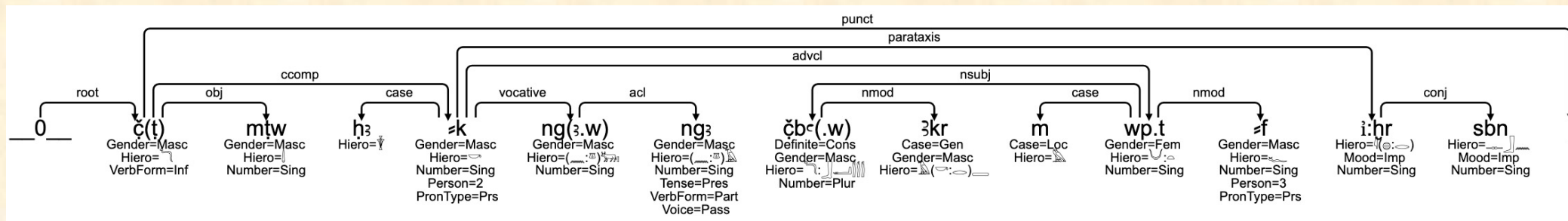
2) Sentences consisting of two or more clauses linked by parataxis or coordination



trans = You will go up — you will go down.

(PT 209b/210b W = EUJA-682)

3) Paragraphs or complex sentences



(PT 504a W = EUJA-1521)

trans = Saying a speech: “Get back, O Long-horn, that will be slaughtered, while the fingers of the Earth-god are on his vertex; fall down and crawl away.”

6. Lemmatisation of words

1) **Nouns** = *Thesaurus Linguae Aegyptiae* (TLA) and *Wörterbuch der ägyptischen Sprache* (Wb.):

```
# sent_id = EUJA-407
# ref = Pyramid Texts § 125a W, Saqqara, 5th Dynasty, rel, OE
# text = č(ṯ) mṭw w₃č šw n ič.n Wniš iš.t ≠f
```

1	č(ṯ)	čṯ	VERB	_	Gender=Masc VerbForm=Inf	0	root	_	Hiero=
2	mṭw	mṭw	NOUN	_	Gender=Masc Number=Sing	1	obj	_	Hiero=
3	w₃č	w₃č	VERB	SPC=Pres Type=Abstrel VerbTop	Tense=Pres VerbForm=Fin Voice=Act	6	dislocated	_	Hiero=:
4	šw	šw	NOUN	_	Gender=Masc Number=Sing	3	nsubj	_	Hiero=
5	n	n	PART	Neg PartType=Neg		6	advmod	_	Hiero=
6	ič.n	ičḷ	VERB	SPC=Past-2 Mood=Pot Tense=Pres VerbForm=Fin Voice=Act		1	ccomp	_	Hiero=:
7	Wniš	Wniš	PROPN	_	Gender=Masc	6	nsubj	_	Hiero=c(:)
8	iš.t	iḥ.t	NOUN	_	Gender=Fem Number=Sing	6	obj	_	Hiero=(:)
9	≠f	f	PRON	Pron=SFP Poss=Yes	Gender=Masc Number=Sing Person=3 PronType=Prs	8	nmod	_	Hiero=

2) **Derivates** ≠ *Thesaurus Linguae Aegyptiae* (TLA) and *Wörterbuch der ägyptischen Sprache* (Wb.):

a) Nisbas:

2 **im(.i)** **m** **ADJ** Nisba=Prep Case=Equ | Gender=Masc | Number=Sing **1** **amod** **_** **Hiero=𓂏𓂐** EUJA-103

jm.j (Lemma ID 25130)

Hieroglyphic spelling: 𓂏𓂐𓂏  Copy Unicode  Copy MdC

jm.j 𓂏𓂐𓂏 (ID 25130)

adjective (deprepositional nisbe)

 **befindlich in (lokal); befindlich in (temporal):** TLA

b) Participles:

EUJA-45

1 **mr.y** **mrj** **NOUN** **_** Gender=Masc | Number=Sing | Tense=Past | VerbForm=Part | Voice=Pass **0** **root** **_** **Hiero=𓂏𓂐**

mr.y (Lemma ID 400005)

Hieroglyphic spelling: 𓂏𓂐  Copy Unicode  Copy MdC

mr.y 𓂏𓂐 (ID 400005)

common noun (masc.)



 **Geliebter**  **the beloved (of)** TLA

c) Causative verbs:

EUJA-109

7 **ś:wꜥb** **wꜥb** **VERB** SPC=Sub | Voice=Cau | Clause=Final Mood=Sub **1** **advcl** **_** **Hiero=𓂏𓂐𓂏𓂏**

swꜥb (Lemma ID 130010)

Hieroglyphic spelling: 𓂏𓂐𓂏𓂏  Copy Unicode  Copy MdC

swꜥb 𓂏𓂐𓂏𓂏 (ID 130010)

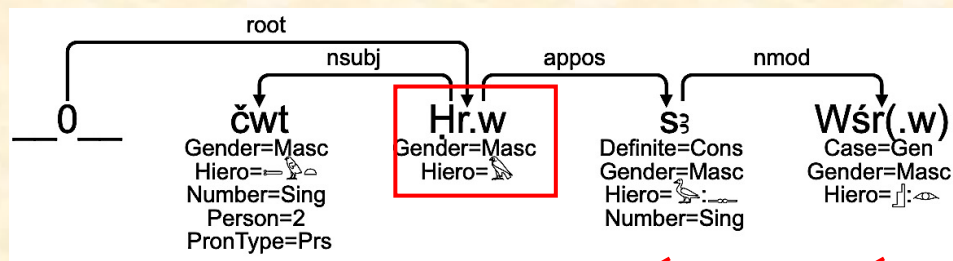
verb (caus. 3-rad.)

 **reinigen**  **to cleanse; to purify** TLA

7. Egyptian morphosyntactic idiosyncrasy

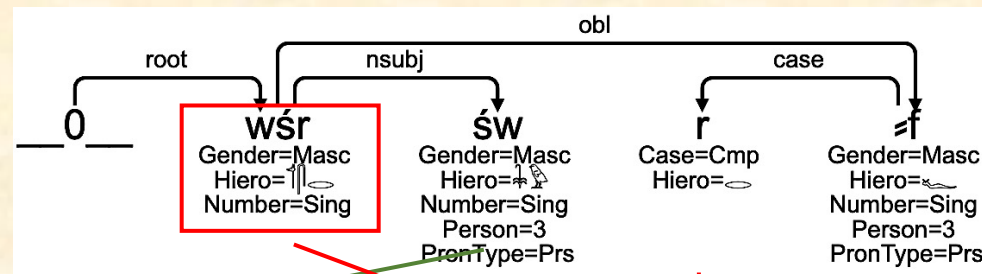
Examples: 1) The root of non-verbal sentences

a) Nominal sentences: (PT 466a W = EUJA-1429)



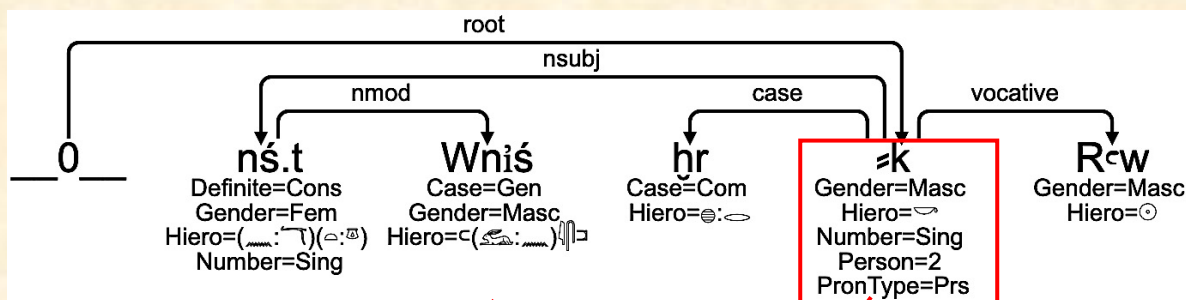
trans = You are Horus, son of Osiris.

b) Adjectival sentences: (PT 395b W = EUJA-1237)



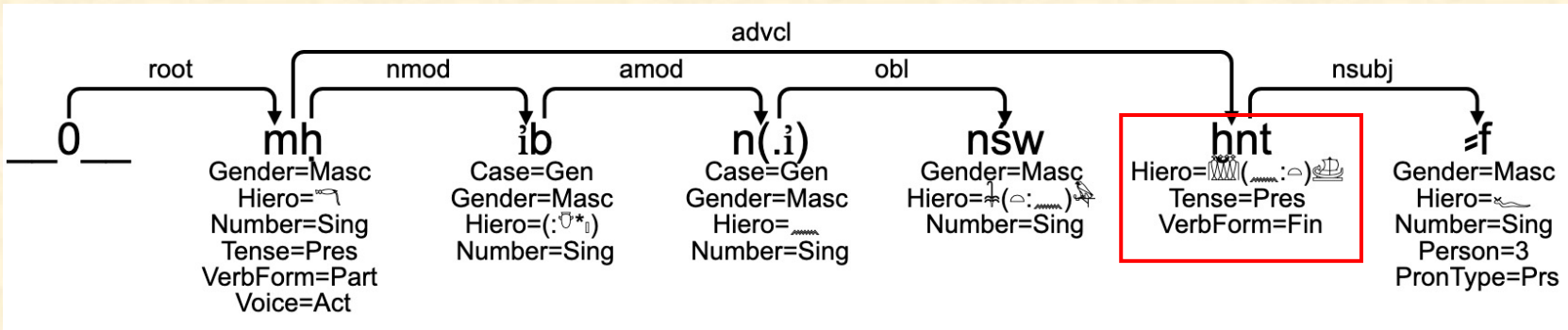
trans = He is more powerful than him

c) Adverbial sentences: (PT 460c W = EUJA-1416)



trans = The throne of Unas is with you, O Re

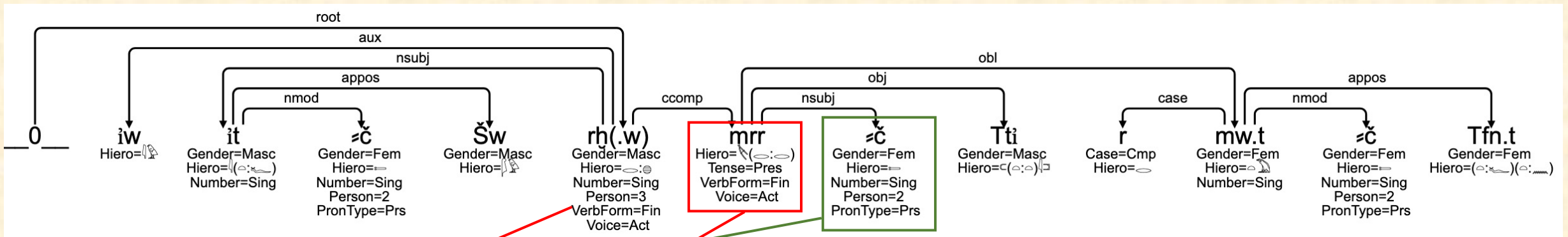
2) Adordination i.e. the syntactic dependency relation caused by a temporal reference of the verb form in the “adordinate” clause, e.g.:



(Asyut V, 18-19, Khety = EUJA-32)

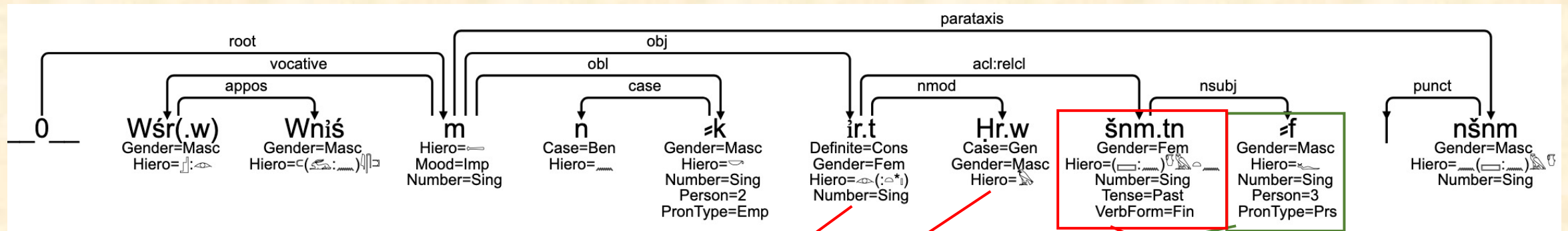
trans = One who fills the heart of the king (when) he sails upstream

3) Nominal finite verb forms (PT 5d T = EUJA-62)



trans = Your father, Shu, knows (that) you love Teti more than your mother, Tefnut.

4) Adjectival finite verb forms (PT 51b W = EUJA-222)



trans = Osiris Unas, take the eye of Horus, (which) he rejoined—a sacred oil

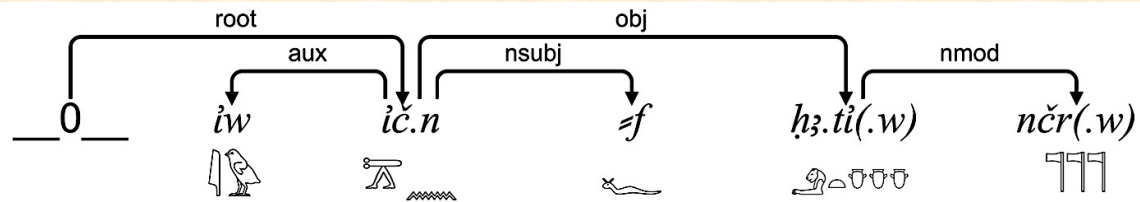
III. Training a model

Evaluation of an NLP model

Metric	F1 Score
UPOS	90.30
XPOS	76.01
UFeats	75.87
AllTags	65.39
Lemmata	89.38
UAS	82.52
LAS	71.97
CLAS	69.13
MLAS	56.14
BLEX	63.27

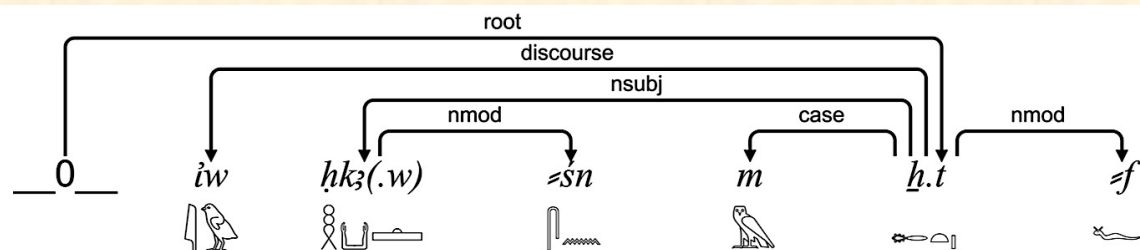
Semi-automatic annotation using the trained model

(PT 409c W = EUJA-1280)



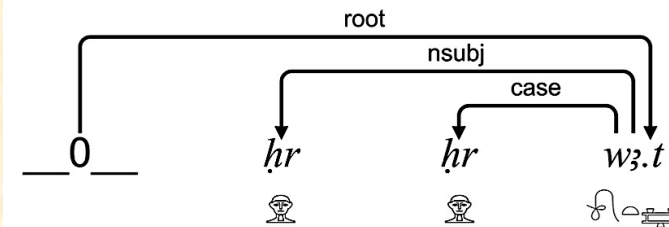
“He (≠f) has (iw) taken (ic.n) the hearts (h3.ti(.w)) of the gods (ncr(.w)).”

(PT 411b W = EUJA-1287)



“Their (≠sn) magic (hk3(.w)) is now (iw) in (m) his (≠f) belly (h.t).”

(PT 429b W = EUJA-1324)



“The face (hr) is on (hr) the way (w3.t) (i.e. the head looks down).”

IV. Conclusion

- 1) The EUJA treebank will be an auxiliary digital source for the study of Egyptian grammar, facilitating the synchronic and diachronic parsing of structures and words.
- 2) It can be used for the development of other tools concerning machine learning, such as the automatic translation of Egyptian texts.
- 3) The next two phases in the development of the EUJA treebank are:
 - a) the annotation of the remaining part of the Pyramid Texts.
 - b) the annotation of the Old Kingdom and First Intermediate Period biographical texts.
- 4) As a result, the treebank will hold over 100,000 Old Egyptian words and the annotation of the Middle Egyptian corpus will begin.

Acknowledgements

- 1) UniDive (COST Action 21167)



- 2) Flavio Cecchini, Amir Zeldes and Daniel Zeman



**Cordial be the hearing of the lady and the lord,
may they live, prosper and be healthy.**

Thank you for your attention!

Roberto Antonio Díaz Hernández
University of Jaén



Marco Carlo Passarotti
Università Cattolica del Sacro Cuore

