Langage, Langues et Cultures d'Afrique Noire
Llacan

DES GLOSES MORPHÉMIQUES À L'ANNOTATION SYNTAXIQUE ET PRAGMATIQUE

Martine Vanhove – martine.vanhove@cnrs.fr

Journée d'étude "Corpus Glosés: de la construction à l'exploitation automatique", Paris, 28 juin 2023
GDR LIFT et TAL

PLAN

- Gloses morphémiques
- Aperçu historique
- Arabes vernaculaires
- Bedja
- GRAID

- **But des gloses morphémiques**

Lehmann, Christian. 1982. Directions for Interlinear Morphemic Translations. *Folia Linguistica* 16: 199-224.

Lehmann, Ch. 2004, "Interlinear morphemic glossing". In Geert Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas (eds.), *Morphologie. Ein internationals Handbuch zur Flexion und Wortbildung*. 2. Halbband. Berlin: de Gruyter, 1834-1857.

« Interlinear morphemic gloss »

"Its primary aim is to make the reader understand the grammatical structure of the L1 text by identifying aspects of the free translation with meaningful elements of the L1 text. The ultimate purpose may be to aid the reader in grasping the spirit of the language, to control the linguistic argument the author is making by means of the L1 example or to scan a corpus for a certain gloss in order to find relevant examples."

- **But des gloses morphémiques**

✓ Lehmann (2004)

"The aim of the following treatment is a standardization of an aspect of linguistic methodology on the basis of widespread usage as developed in the 20th century"

✓ LGR preamble (February 2008) (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf)

"Linguists by and large conform to certain notational conventions in glossing, and the main purpose of this document is to make the most widely used conventions explicit."

"The main purpose that is assumed here is the presentation of an example in a research paper or book."

Lehmann (2004: 1835-1836)

- 19e et 1ère moitié 20ème s.: pas de gloses morphémiques; grammaires scientifiques et études comparatives pour initiés (pas besoin de gloses)
- Précurseurs:
  - parfois une traduction littérale pour faire comprendre
    - « l'esprit d'une langue »,
    - « une vision du monde propre à une langue »
  - G. Gabelentz (1901:460), sur l'origine possessive des indices personnels du verbe en sémitique (ex. arabe):

    ya-kfī-ka-hùm er genügt dir gegen sie (eig. er-genügt-dein-ihr) (il-suffisant-ton-eux)

- Arrivée tardive des gloses morphémiques
  - dans les années 1960
  - Pratique courante à partir des années 1980

Rappel

- Gloser est (était) coûteux en temps

- Améliorations avec
  - des logiciels d'interlinéarisation type Shoebox, Toolbox, FLEX, ELAN

  - les progrès de la typologie (cf. EUROTYP)

- Quelques principes (Lehmann 2004: 1841)

word class	instead of	use
copulas, auxiliaries	*be* *have* (except to mean 'possess, own')	COP, PASS, PROG ... PF, OBLG ...
prepositions	*by* *with* *for* *as* *from* *to* *of*	AG, ERG ... INST, COM, ASSOC ... BEN, DEST ... EQT, ESS ... ABL, DEL ... DAT, ALL, DEST, TERM, INF ... GEN, ASSOC ...
subordinators	*that* *if*	COMP, SR (, D3) INT, COND.SR
relativizers	*that* *who* *which*	REL REL.HUM.NOM ... REL.NHUM.NOM ...

Tab. 169.2: Free grammatical morphemes

- Quelques principes (suite)

label	intended meaning	comment
A	transitive subject	in morphemic glosses, the abbreviation is ERG
ADV	adverb	specify meaning
AGR	agreement	specify agreement categories
AGT	agent	this is not a value of a morphological category
ART	article	only if it has no determinative properties
ASP	aspect	specify particular aspect
AUX	auxiliary	only if there is only one auxiliary morpheme in the language
CARD	cardinal	only if it is a morpheme or grammatical feature
CLF	classifier	this is a word class
CLT	clitic	this is neither a morphological category nor a value of one
EP	epenthetic	has no morphological status, should not be separated in the first place
EVID	evidential	specify particular evidential
PAT	patient	this is not a value of a morphological category
PREP	preposition	this is a word class
PTL	particle	this is (at best) a word class
TNS	tense	specify particular tense

Tab. 169.3: Labels to be avoided

Vicente *et al.* 2015. Glossing in Semitic languages. A comparison of Moroccan Arabic and Modern Hebrew. In Mettouchi et al. (eds) *Corpus-based studies of lesser-described languages. The CorpAfroAs corpus of spoken AfroAsiatic languages*, pp. 173-206, Amsterdam-Philadelphia: Benjamins.

- Introduction des gloses encore plus lente et tardive, pas généralisée
- En augmentation depuis le début du 21e siècle
- Méconnaissance des règles LGR et des principes de Lehmann
- E.g. « Rule 2: Morpheme-by-morpheme correspondence »

fa	ga‘	yigūl	luhum	ənna	“’āna	miḥtār”
so	PROG	he-says-IPFV	to = them	that	“I	confused-A.PTCP”

“So he’s telling them that “I’m confused”, (Tsukanova 2008: 448)

ah, maši	*bḥal*	*l-mġaṛba*	*lli*	*ka-ne-ʕṛef-hŭm*	*ana*	*f*	*l-.. eh*
oh neg	like	def-Moroccan.pl	rel	asp-1-know-3pl	1sg	in	def-er

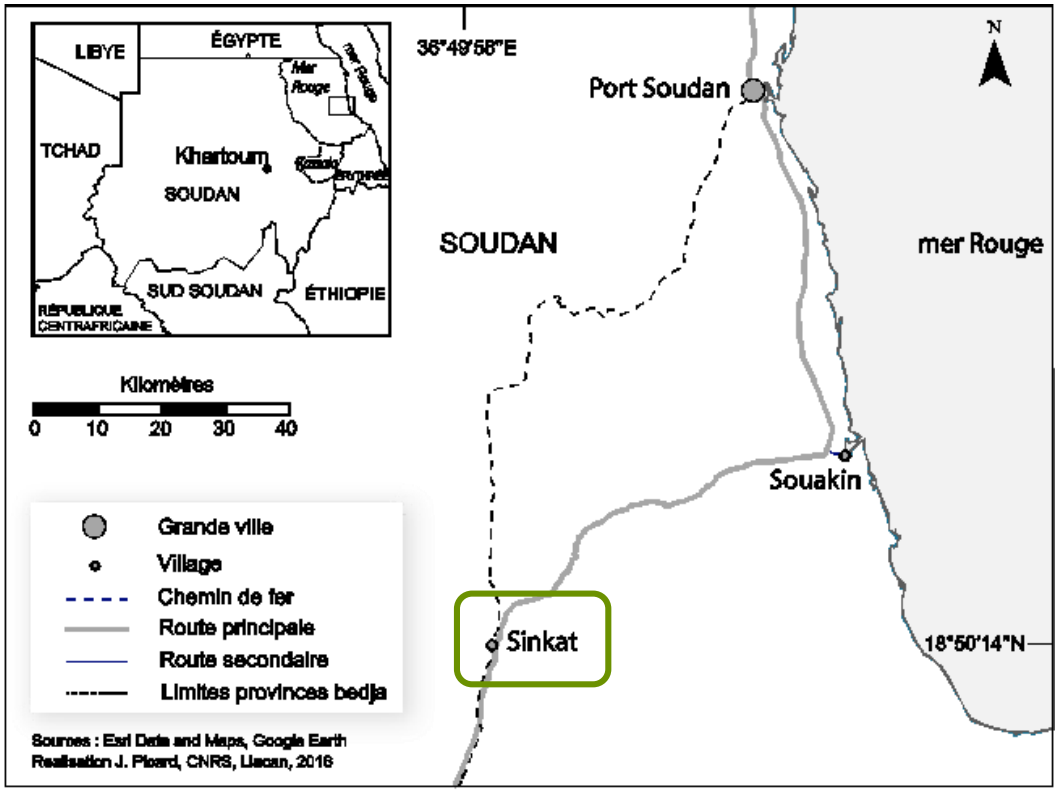
omgeving	*dyal-i*
environment	of-1sg

“Oh, it is not at all like the Moroccans I know in my er environment”, (Boumans 1998: 190).

- Vicente *et al.* 2015. = Harmonisation pour arabes vernaculaires sur la base des LGR
- Projet ANR *CorpAfroAs* (2007-2012, PI A. Mettouchi)
- ELANCorpA (C. Chanard)
- 6 lignes d'annotation
  - \tx = transcription phonétique large en mots prosodiques
  - \mot = découpage en lexèmes
  - \mb = découpage phonologique en morphèmes
  - \ge = glose sémantique ou grammaticale
  - \rx = parties du discours, etc.
  - \ft = traduction libre

\mot	ma	kaixərʒuʃ
\mb	ma	ka=j-xrəʒ-u=ʃ
\ge	NEG	REAL=3-go_out\IPFV-PL=NEG
\rx	PTCL	TAM=PNG-V-PNG=CL
\ft	they do not go out (ARY_AB_NARR_01_279)	

- Couchitique-nord (Afroasiatique)
- Soudan oriental
- Approx. 2.000.000 locuteurs



- **Historique**
- Pas de gloses dans les grammaires anciennes (Almkvist 1881-1885; Reinisch 1893-1894; Roper 1928; Hudson 1964)
- 1ères gloses morphémiques (non LRG) chez Morin (1995)

i-sakána *úun* *naa-ná-at-i*
la-nouvelle celle-ci quoi-chose-r-préd.
"qu'est-ce que c'est que cette histoire-ci? (2.18.)"

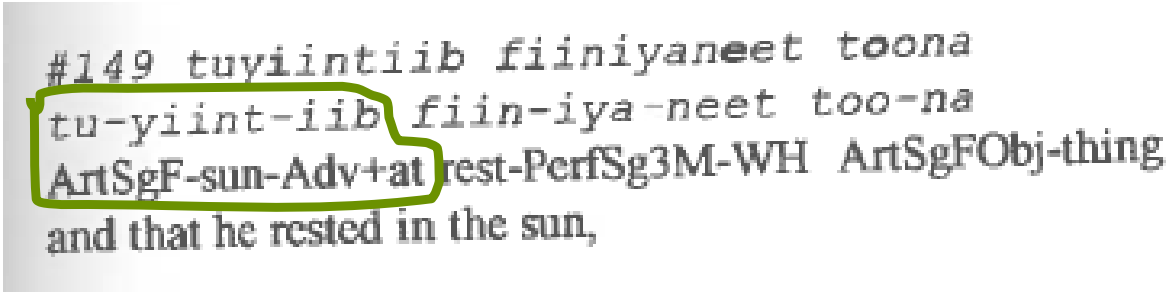
Morin, Didier. 1995. *« Des paroles douces comme la soie ». Introduction aux contes dans l'aire couchitique (bedja, afar, saho, somali)*. Paris: Peeters. (p, 26)

NB: r = relateur de fonction « non sujet » indéfini

- **Historique**

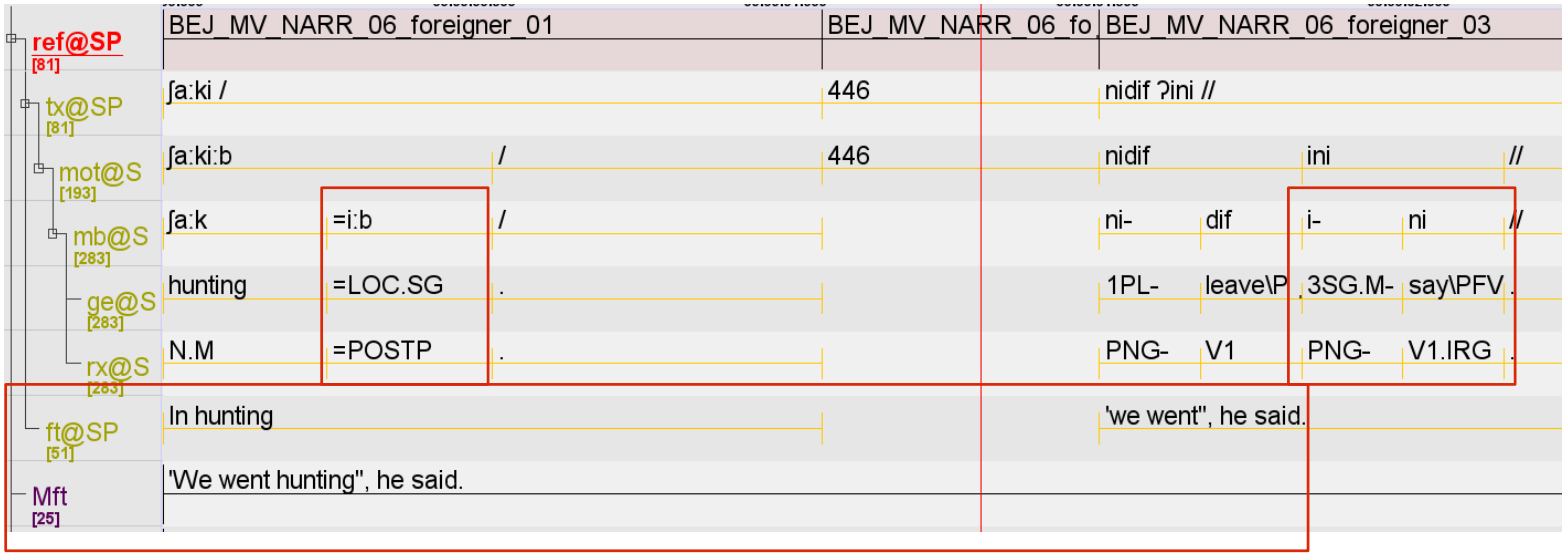
Gloses morphémiques (non LRG) chez Wedekind

Wedekind, Klaus & Charlotte and Abuzeinab Musa. 2007. *A Learner's grammar of Beja (East Sudan)*. Köln: Rüdiger Köppe.


#149 tuyiintiib fiiniyaneet toona
tu-yiint-iib fiin-iya-neet too-na
ArtSgF-sun-Adv+at rest-PerfSg3M-WH ArtSgFObj-thing
and that he rested in the sun,

tu=yiin=t=iib
DEF.SG.F=sun=INDF.F=LOC.SG

- ANR CorpAfroAs et CorTypo: 142 mn corpus glosé (57 monologues narratifs, 1 conversation; 15.859 mots; 33.199 morphèmes)
- Découpage en unités intonatives (2382 UIs, dont 1363 pauses)
- Langue SOV → 1 ligne de traduction Mft pour regrouper de manière intelligible le sens des UIs



- Avec B. Comrie, mise au point d'une liste commune de gloses sur la base des LRG: http://corpafroas.huma-num.fr/glosses.html
  - 341 pour la ligne ge@SP
    - (340 in Lehmann, pas tous identiques, 84 in LRG + 10 règles permettant d'augmenter la liste de base)
  - 131 pour la ligne rx@SP

- Idiosyncratismes pour l'Afroasiatique et ses familles (comme chez Lehmann et autres)

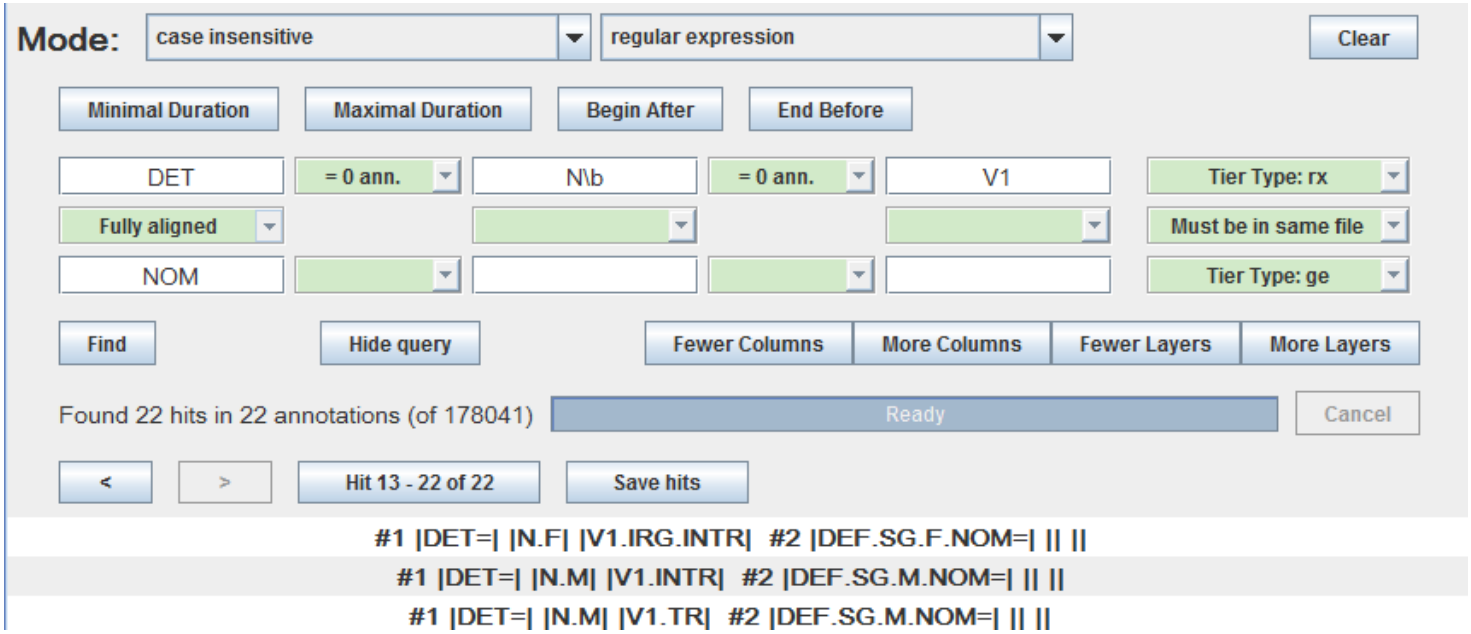
- OK pour les recherches sur la morphologie
- Plus compliqué pour la syntaxe et la structure de l'information

BEDJA: STRUCTURE DE L'INFORMATION

- Langue SOV non stricte (pragmatique)
- Présence obligatoire d'une marque casuelle (nominatif, accusatif) sur l'article défini (sauf exceptions – nombreuses – morpho-phonologiques)
- Pas de marque segmental de topicalisation de S
  - Exception: si un démonstratif avant GN au lieu d'après GN
- Anti-topiques et post-rhèmes postposés au verbe
- Question: comment démêler les rôles de
  - sujet et de topique?
  - anti-topique et post-rhème?

BEDJA: STRUCTURE DE L'INFORMATION

- **Critères:**
  - Ordre des mots (rx + ge)
  - Prosodie (/ et //)
  - Cas nominatif marqué ou (CorTypo) SBJ dans rx



- **Résultats**
  - 94% de 605 énoncés S(O)V
  - Suffisant pour dégager des tendances sur interface avec prosodie, mais:
    - si grande distance entre GN_{NOM} et V, vérifications manuelles

	NP_{NOM} +V in same IU	NP_{NOM} +V in ≠ IUs	V+NP_{NOM}	Total
# NP_{NOM}	360	210	35	605
% of NP_{NOM}	59.5%	34.7%	5.8%	100%

Table 8: Intonation units and word order of NP_{NOM} + V

- Analyse des contours prosodiques dans PRAAT
- Analyse pragmatique et sémantique

- **Résultats**
- Analyse des contours prosodiques dans PRAAT
  - GNs_{NOM} et V dans 1 UI: descendants: 295/360 (82%)
  - GNs_{NOM} et V dans 2 UI: montants ou hauts sur GNs_{NOM}: 149/210 (71%)
  - Exceptions dues en majorité aux hésitations

- + Analyse pragmatique et sémantique
  - GNs_{NOM} dans même UI que V = nouveau référent = S
  - GNs_{NOM} dans 2 UIs = topique contrastif ou topique sélectif

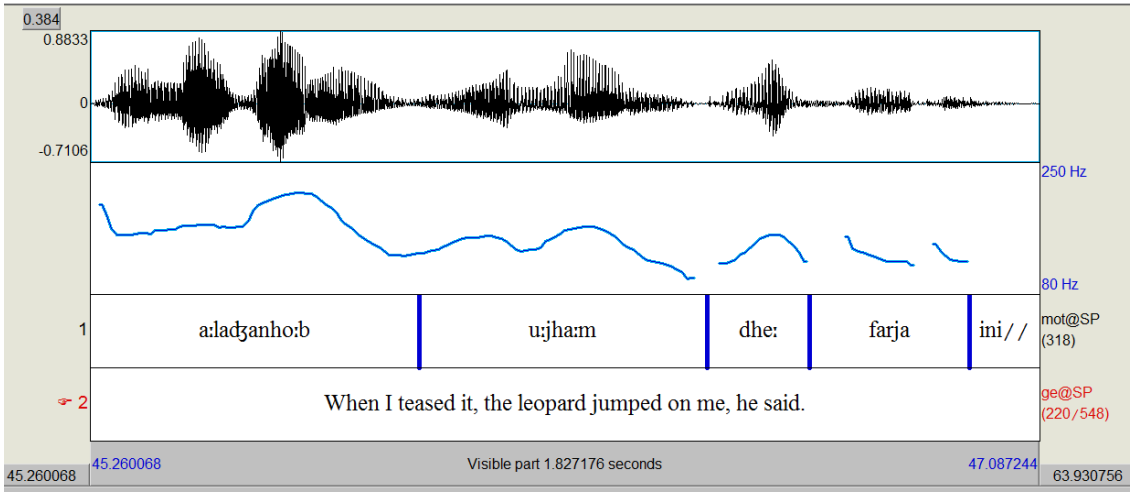
BEDJA: DISCOURS RAPPORTÉ

- **Discours rapporté**
  - Toujours direct
  - Pas de complémenteur
  - Pas de pronoms logophoriques
  - Donc pas de glose morphologique
  - Recherche uniquement à partir du verbe ‘dire’
  - Exclu les DR sans verbe quotatif

- Recherche typologique et prosodique sur interface entre DR et cadre quotatif
  - Analyse dans PRAAT + longueur pause notée

BEDJA: DISCOURS RAPPORTÉ

- **Résultats**
  - Très forte intégration prosodique du verbe quotatif avec la fin du DR (90% des 317 exemples de CorpAfroAs)
  - Exceptions: si le DR est un énoncé exclamatif, s'il contient un impératif ou une onomatopée.
  - Rupture prosodique marque début du DR



- Haig, Geoffrey & Stefan Schnell. 2014. Annotations using GRAID (Grammatical Relations and Animacy in Discourse). Manual Version 7.0. https://d-nb.info/1069290009/34

« GRAID is a system of symbols and conventions for glossing the grammatical relations and overt forms (noun phrases, pronouns etc.) of major clause constituents in texts. The purpose of GRAID annotations is to facilitate crosscorpus research in language typology. »

« In addition to the syntactic function and morphological form, GRAID annotations also register animacy features of referential expressions. Hence, GRAID-annotated text corpora facilitate additional research questions in the area of animacy and referential hierarchies in discourse. »

« morphemic glosses provide no direct or consistent means of identifying syntactic constituents »

Table 1: Glosses for the form of referential expressions

np	noun phrase
pro	free pronoun in full form
=pro	'weak' clitic pronoun
-pro	pronominal affix, cf. 3
0	covert argument / phonologically null argument
refl	reflexive or reciprocal pronoun, cf. Section 4.2
adp	adposition
w	'weak' (optional symbol), indicates a phonologically lighter form, it precedes the form symbol, e.g. <wpro>
x	'non-referential', see below for explanation
other	used for expressions 1. that are not of a type listed above 2. the form of which is not considered relevant
ln	NP-internal subconstituent occurring to the left of NP head
rn	NP-internal subconstituent occurring to the right of NP head
lv	subconstituent of verb complex occurring to the left of verbal head
rv	subconstituent of verb complex occurring to the right of verbal head

Table 2: Glosses for the properties of referents

1	1st person referent(s)
2	2nd person referent(s)
h	human referent(s)
d	anthropomorphized referent(s); the use of this symbol is optional

Table 3: Glosses for major syntactic functions

s (or: S)	intransitive subject
a (or: A)	transitive subject
p (or: P)	transitive object
ncs	non-canonical subject
g	goal argument of a goal-oriented verb of motion, but also: recipient of verb of transfer, and addressee of verb of speech
l	locative argument of verbs of location
obl	oblique argument, excluding goals and locatives
p2	secondary object
dt	dislocated topic (right or left-dislocated)
voc	vocative
poss	possessor
appos	appositional
other	other function

Table 4: Form and function glosses for predicates

v	verb or verb complex (cf. Section 2.5.1)
vother	non-canonical verb-form (cf. Section 2.5.5)
cop	(overt) copular verb (cf. Section 2.5.2)
aux	auxiliary (cf. Section 2.5.2)
-aux	suffixal auxiliary
=aux	clitic auxiliary
pred	predicative function
predex	predicative function in existential / presentational constructions

Table 7: Glosses for irrelevant and non-classifiable elements

other	forms / words / elements which are not relevant for the analysis
nc	'not considered' / 'non-classifiable'

Table 5: Glosses for clause boundaries, embedded clauses, and clausal operators

##	boundary of independent clause, inserted at left edge
#	boundary of dependent clause, inserted at left edge, further specified
rc	relative clause
cc	complement clause
ac	adverbial clause
ds	direct speech
neg	negative polarity
%	end of a dependent clause (if not coinciding with the end if its main clause)

##	leftward-boundary of main clause
#	leftward-boundary of a dependent clause, type and function not specified
#ds	dependent clause, direct speech, not negated,
##ds	independent clause, direct speech, not negated function not specified
#ds_rc	relative clause rendering direct speech
#ac.neg	adverbial clause, negated
#ds_cc.neg:p	complement clause of a transitive verb, negated, rendering direct speech

- Complexité et limitation des gloses morphologiques (et PoS) pour les recherches syntaxiques
- Ex. Kabyle

```
QUERY : [rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ANN < mot =.] {rx <
3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge <
3 & mot=1} [ rx=ADV & ge=. < mot =.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot
=.]OR[rx = \bV & ge=. < mot=.] {rx < 3 & ge < 3 & mot=1} [ rx=N & ge=POSTNEG < mot =.]
{rx < 3 & ge < 3 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.] {rx < 3 &
ge < 3 & mot=1} [ ge = HESIT | # & rx=. < mot =.] {rx =1 & ge =1 & mot=1} [ rx=/ & ge=. <
mot =.]{rx =1 & ge = 1 & mot=1} [ rx=N & ge=ABSL < mot =.]OR[rx = \bV & ge=. < mot=.]
{rx < 3 & ge < 3 & mot=1} [ ge = FS & rx=. < mot =.]{rx =1 & ge = 1 & mot=1} [ rx=## &
ge=## < mot =##]{rx =1 & ge = 1 & mot=1} [ rx=N & ge=ABSL < mot =.]
```

Figure 16: Complete query for Direct Object

CONCLUSION

- Intérêt d'une annotation de la syntaxe et de la structure de l'information, en particulier pour des langues qui ignorent certains types de marquage segmental
  - GRAID permet cela (et +), mais l'annotation est entièrement manuelle et donc chronophage
  - SUD permet des annotations syntaxiques et de la structure de l'information, y compris à partir des morphes, rapidement semi-automatisables, par ex. pour les « disloqués » (pas suffisamment fines) et le « discours rapporté »

  - → Recherches automatiques plus faciles et possibles, quelle que soit la morphologie de la langue

  - → Si pas de morphologie ou de syntaxe marquée segmentalement, toujours le problème de la recherche prosodique

MERCI!