Guy Perrier

D’un corpus glosé à un treebank en dépendance : exemple du beja

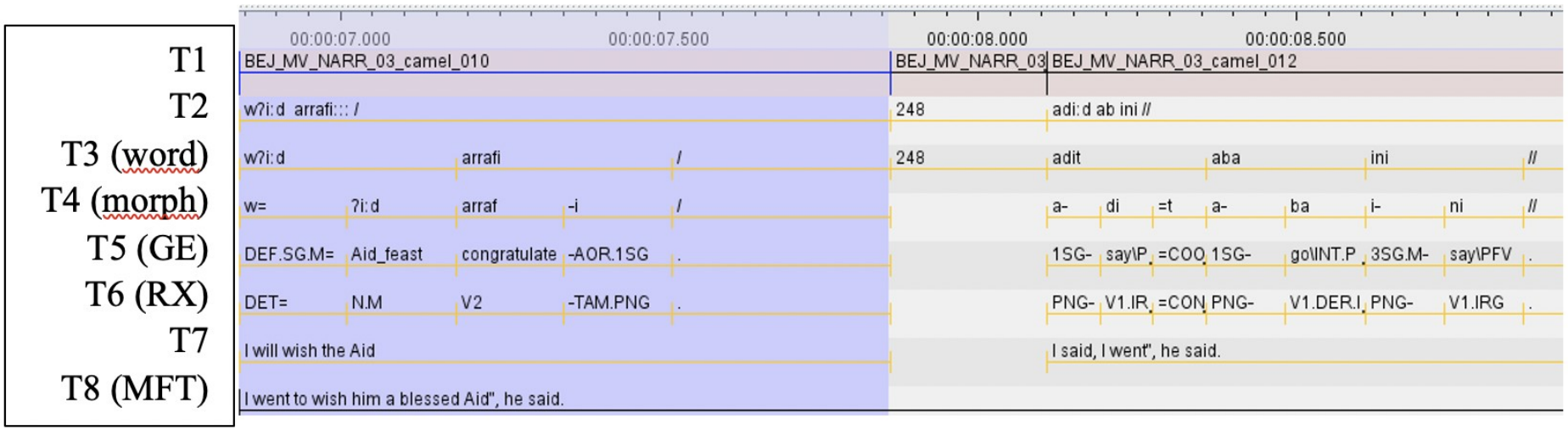
Sylvain Kahane
avec **Martine Vanhove, Rayan Ziane, Bruno Guillaume**

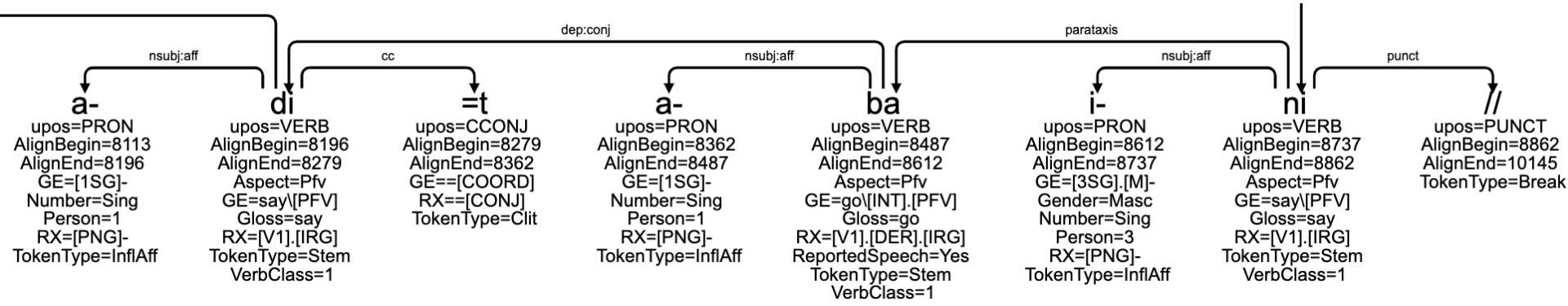
Journée Corpus Glosés, 28 juin 2023

Treebank du beja

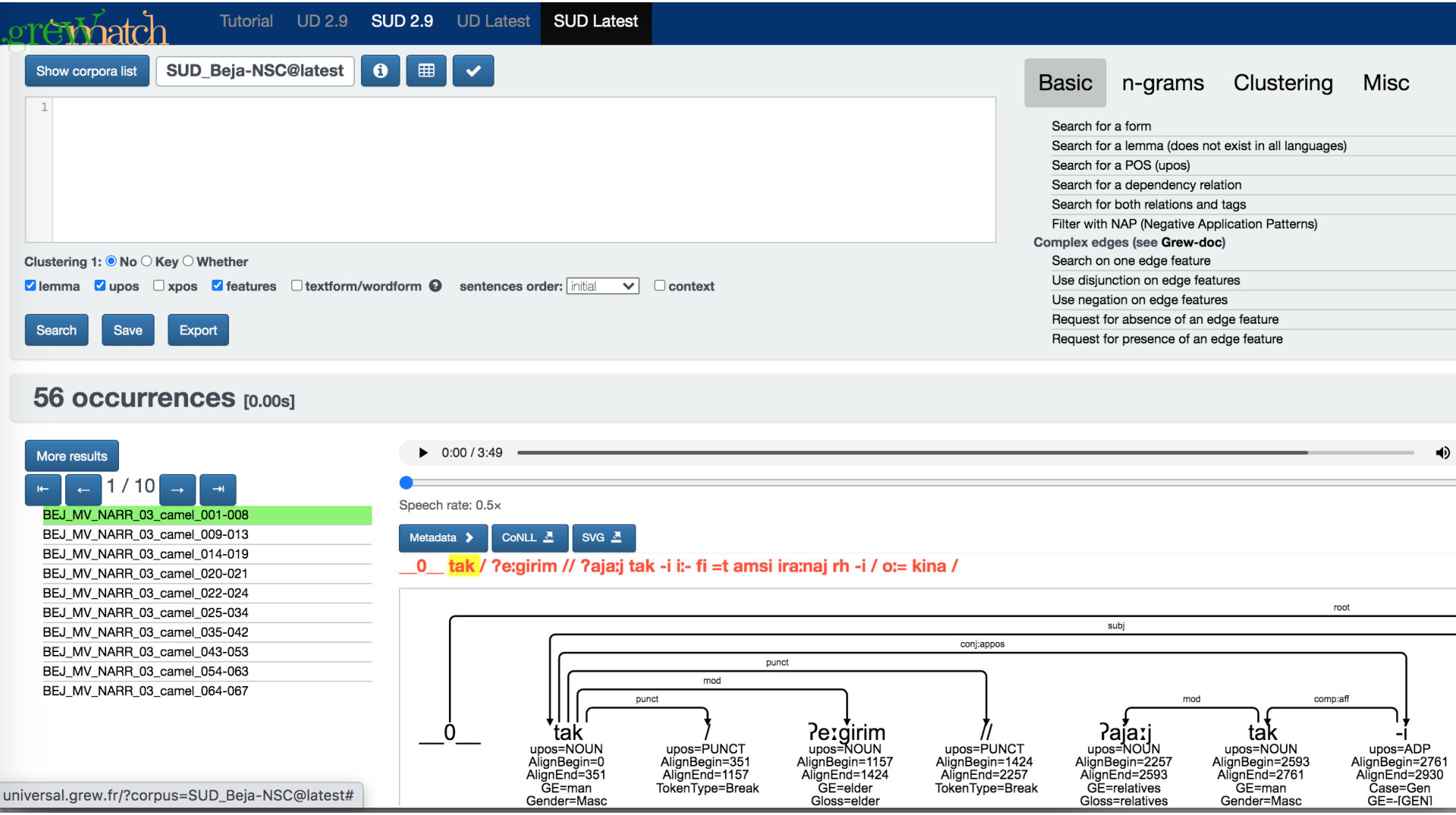
- corpus de Martine Vanhove
  - Vanhove, M. 2014. The Beja Corpus. In Mettouchi, A. and C. Chanard (eds.). *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*.
  - Comrie, B. 2015. From the Leipzig Glossing Rules to the GE and RX lines. In A. Mettouchi, M. Vanhove & D. Caubet (eds.), *Corpus-based Studies of Lesser-described Languages*. John Benjamins, 207-219.
- stage de M2 Rayan Ziane (6 mois en 2021)
- A morph-based and a word-based treebank for Beja (Kahane, Vanhove, Ziane, Guillaume, *Proceedings of Treebanks and linguistic theories (TLT)*, 2021)

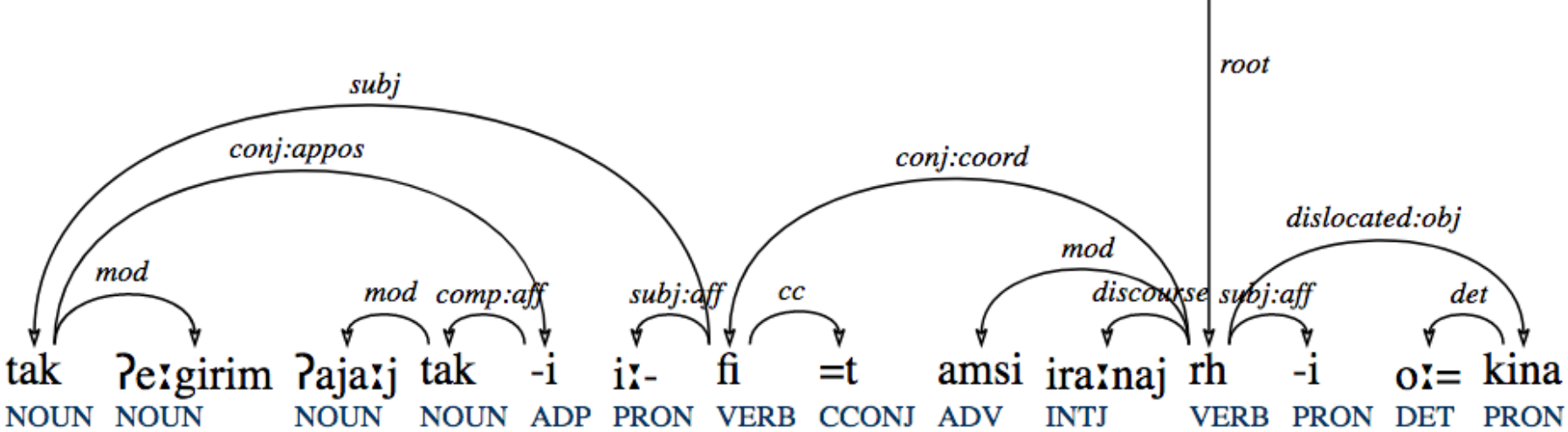
IGT vs treebank

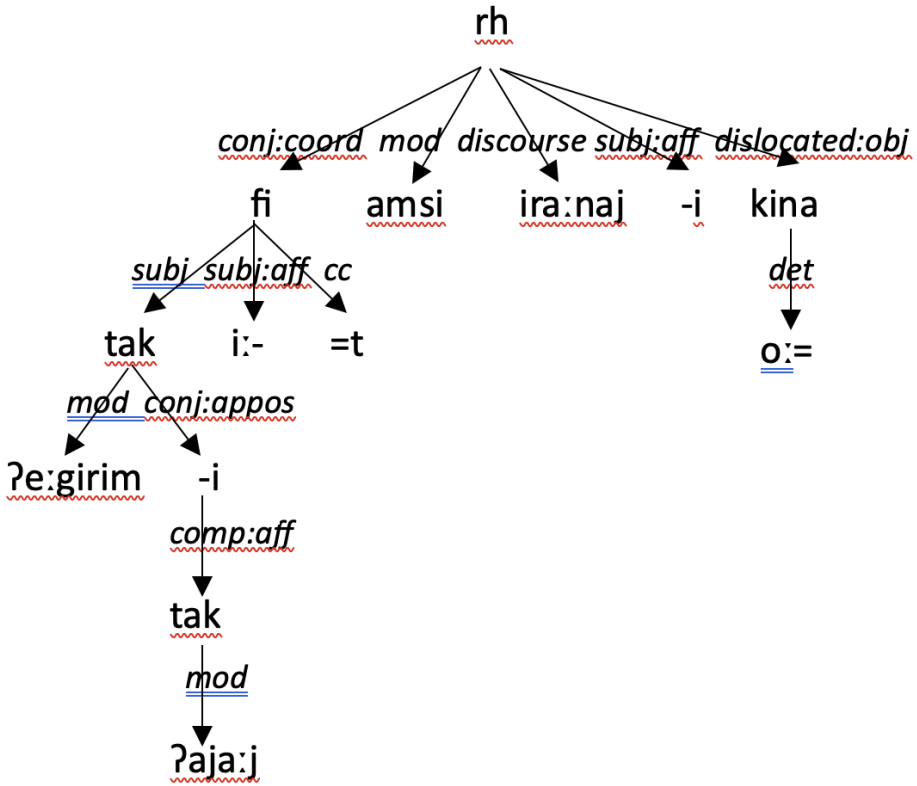




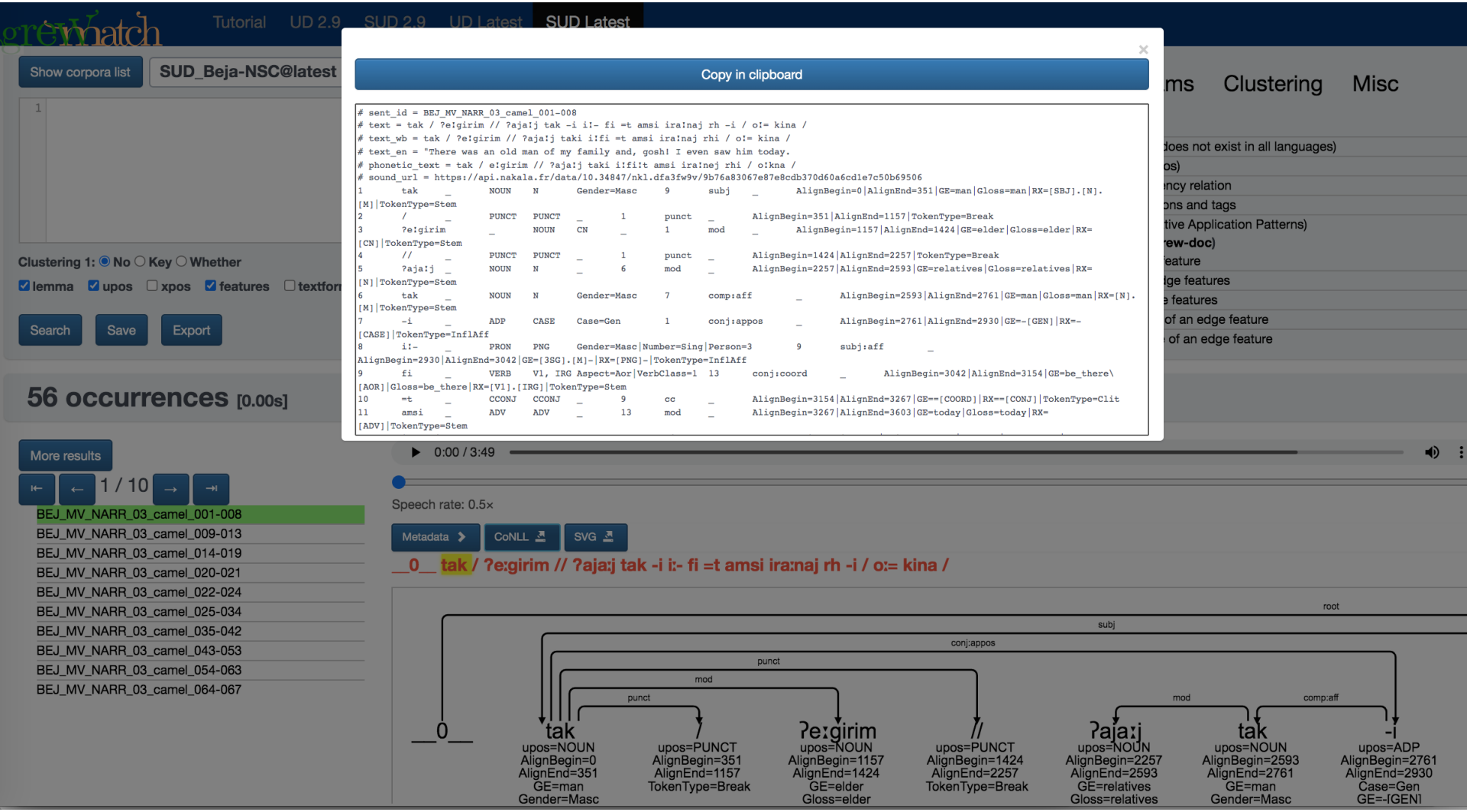
http://**universal.grew.fr**/?corpus=SUD_Beja-NSC@latest#




root
subj
conj:appos
conj:coord
dislocated:obj
mod
mod
discourse
mod
comp:aff
subj:aff
cc
subj:aff
det
tak ʔeːgirim ʔajaːj tak -i iː- fi =t amsi iraːnaj rh -i oː= kina
NOUN NOUN NOUN NOUN ADP PRON VERB CCONJ ADV INTJ VERB PRON DET PRON


rh
conj:coord mod discourse subj:aff dislocated:obj
fi amsi iraːnaj -i kina
subj subj:aff cc
det
tak iː- =t
oː=
mod conj:appos
ʔeːgirim -i
comp:aff
tak
mod
ʔajaːj

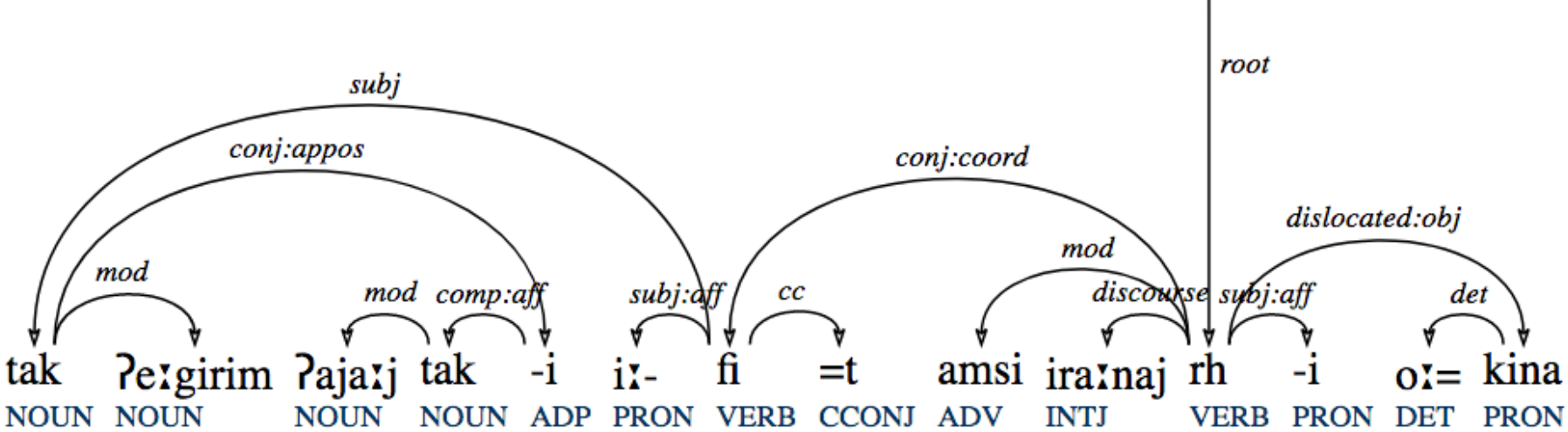
http://**universal.grew.fr**/?corpus=SUD_Beja-NSC@latest#

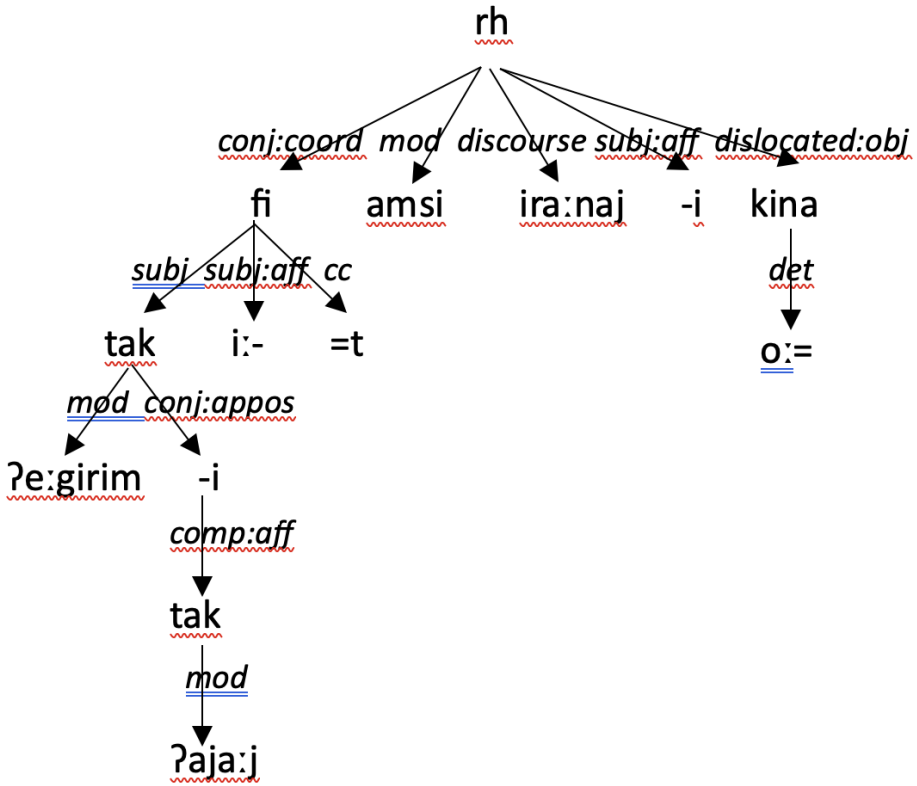


format tabulaire
dit format CoNLL
(Conference in Natural Language Learning)

	A	B	C	D	E	F		
1	# sent_id = BEJ_MV_NARR_03_camel_001-008							
2	# text = tak / ʔe:girim // ʔaja:j tak -i i:- fi =t amsi ira:naj rh -i / o:= kir							
3	# text_wb = tak / ʔe:girim // ʔaja:j taki i:fi =t amsi ira:naj rhi / o:= kir							
4	# text_en = "There was an old man of my family and, gosh! I even							
5	# phonetic_text = tak / e:girim // ʔaja:j taki i:fi:t amsi ira:nej rhi / o:k							
6	# sound_url = https://api.nakala.fr/data/10.34847/nkl.dfa3fw9v/9b76							
7	1	tak	_	NOUN	N	Gender=Masc	9	subj
8	2	/	_	PUNCT	PUNCT	_	1	punct
9	3	ʔe:girim	_	NOUN	CN	_	1	mod
10	4	//	_	PUNCT	PUNCT	_	1	punct
11	5	ʔaja:j	_	NOUN	N	_	6	mod
12	6	tak	_	NOUN	N	Gender=Masc	7	comp:aff
13	7	-i	_	ADP	CASE	Case=Gen	1	conj:appos
14	8	i:-	_	PRON	PNG	Gender=Masc|Number=Sing|Person=3	9	subj:aff
15	9	fi	_	VERB	V1, IRG	Aspect=Aor|VerbClass=1	13	conj:coord
16	10	=t	_	CCONJ	CCONJ	_	9	cc
17	11	amsi	_	ADV	ADV	_	13	mod
18	12	ira:naj	_	INTJ	_	_	13	discourse
19	13	rh	_	VERB	V2	VerbClass=2	0	root
20	14	-i	_	PRON	TAM, PNG	Aspect=Aor|Gender=Masc|Number=Sin	13	subj:aff
21	15	/	_	PUNCT	PUNCT	_	13	punct
22	16	o:=	_	DET	DET	Case=Acc|Definite=Def|Gender=Masc|N	17	det
23	17	kina	_	PRON	PRO	Reflex=Yes	13	dislocated:obj
24	18	/	_	PUNCT	PUNCT	_	13	punct

I	J	K	L	M	N	O	P
c50b69506							
_	AlignBegin=0|AlignEnd=351|GE=man|Gloss=man|RX=[SBJ].[N].[M]|TokenType=Stem						
_	AlignBegin=351|AlignEnd=1157|TokenType=Break						
_	AlignBegin=1157|AlignEnd=1424|GE=elder|Gloss=elder|RX=[CN]|TokenType=Stem						
_	AlignBegin=1424|AlignEnd=2257|TokenType=Break						
_	AlignBegin=2257|AlignEnd=2593|GE=relatives|Gloss=relatives|RX=[N]|TokenType=Stem						
_	AlignBegin=2593|AlignEnd=2761|GE=man|Gloss=man|RX=[N].[M]|TokenType=Stem						
_	AlignBegin=2761|AlignEnd=2930|GE=-[GEN]|RX=-[CASE]|TokenType=InflAff						
_	AlignBegin=2930|AlignEnd=3042|GE=[3SG].[M]-|RX=[PNG]-|TokenType=InflAff						
_	AlignBegin=3042|AlignEnd=3154|GE=be_there\[AOR]|Gloss=be_there|RX=[V1].[IRG]|TokenType=Stem						
_	AlignBegin=3154|AlignEnd=3267|GE==[COORD]|RX==[CONJ]|TokenType=Clit						
_	AlignBegin=3267|AlignEnd=3603|GE=today|Gloss=today|RX=[ADV]|TokenType=Stem						
_	AlignBegin=3603|AlignEnd=3940|GE=gosh|Gloss=gosh|RX=[EXCL]|TokenType=Stem						
_	AlignBegin=3940|AlignEnd=4108|GE=see|Gloss=see|RX=[V2]|TokenType=Stem						
_	AlignBegin=4108|AlignEnd=4277|GE=-[AOR].[3SG].[M]|RX=-[TAM].[PNG]|TokenType=InflAff						
_	AlignBegin=4277|AlignEnd=4764|TokenType=Break						
_	AlignBegin=4764|AlignEnd=4921|GE=[DEF].[SG].[M].[ACC]=|RX=[DET]=|TokenType=Clit						
_	AlignBegin=4921|AlignEnd=5079|GE=owner|Gloss=owner|RX=[PRO].[REFL]|TokenType=Stem						
_	AlignBegin=5079|AlignEnd=5895|TokenType=Break						


root
subj
conj:appos
conj:coord
dislocated:obj
mod
mod
discourse
mod
comp:aff
subj:aff
cc
subj:aff
det
tak ʔeːgirim ʔajaːj tak -i iː- fi =t amsi iraːnaj rh -i oː= kina
NOUN NOUN NOUN NOUN ADP PRON VERB CCONJ ADV INTJ VERB PRON DET PRON


rh
conj:coord mod discourse subj:aff dislocated:obj
fi amsi iraːnaj -i kina
subj subj:aff cc
det
tak iː- =t
oː=
mod conj:appos
ʔeːgirim -i
comp:aff
tak
mod
ʔajaːj

Affixes dérivationnels

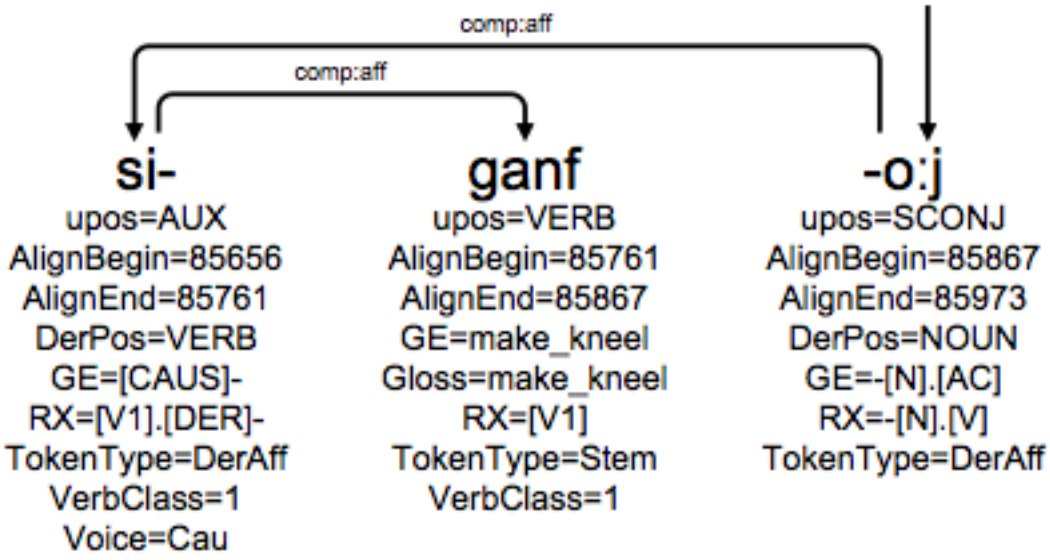


Figure 4. Derivational affixes (SUD-style)

Affixes dérivationnels

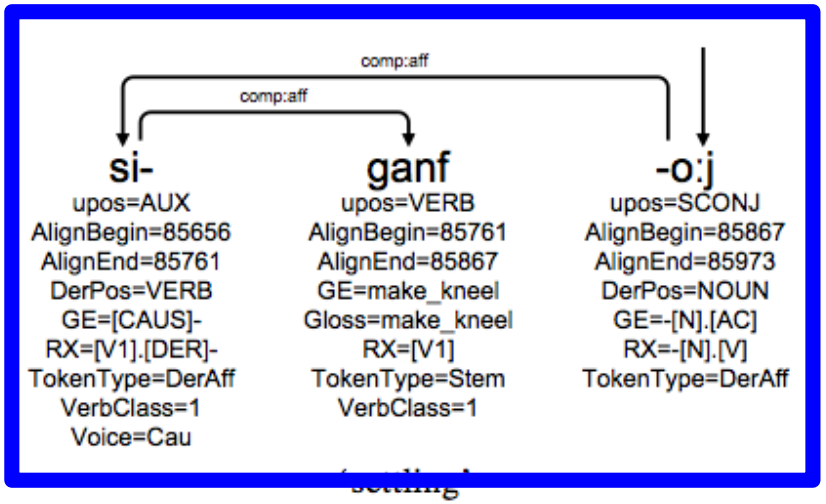


Figure 4. Deri onal affixes (SUD-style)

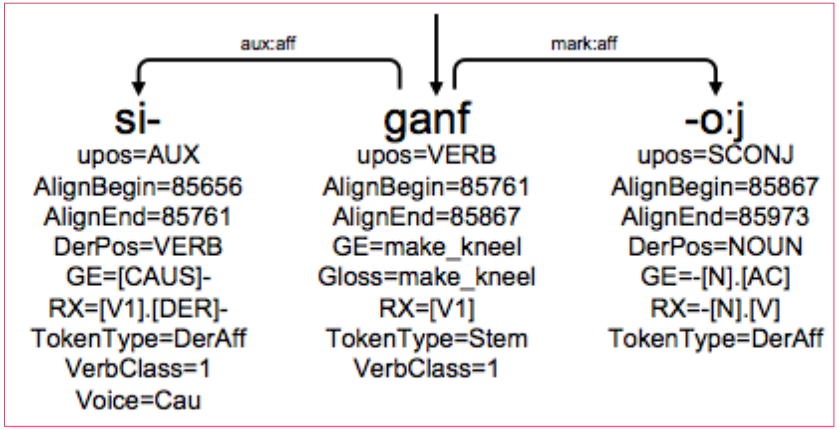


Figure 5. Derivational affixes (UD-style)

information manquante

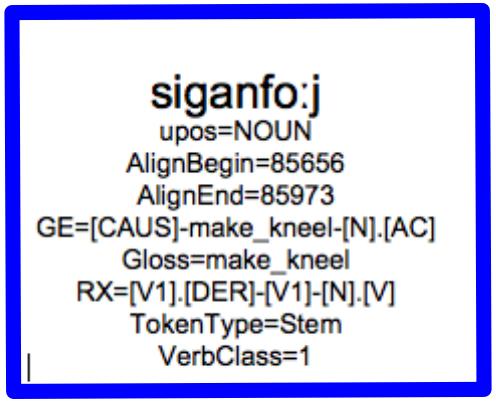


Figure 6. The word-based annotation of the derived word *siganfo:j*

IGT vs treebank

informations supplémentaires dans le treebank

- unités maximales
  - phrases, unités illocutoires
- unités minimales (tokens)
  - mots, morphes, syntaxèmes
- dépendances syntaxiques
  - structure arborescente => structure de constituants
- relations/fonctions syntaxiques
  - typage des constructions (par l'étiquetage des relations)
- traits morpho-syntaxiques
  - gloses sous forme de structure de traits
    (avantage pour les requêtes)

Méthodologie

- Conservez au maximum l'information contenue dans l'IGT
  - Métadonnées
  - Alignement temporel au son
  - Traduction
  - Analyse morphosyntaxique
  - Gloses
- Schéma d'annotation universel
  - Discussion en groupe sur les issue trackers UD et SUD
  - Possibilité de personnaliser son schéma SUD

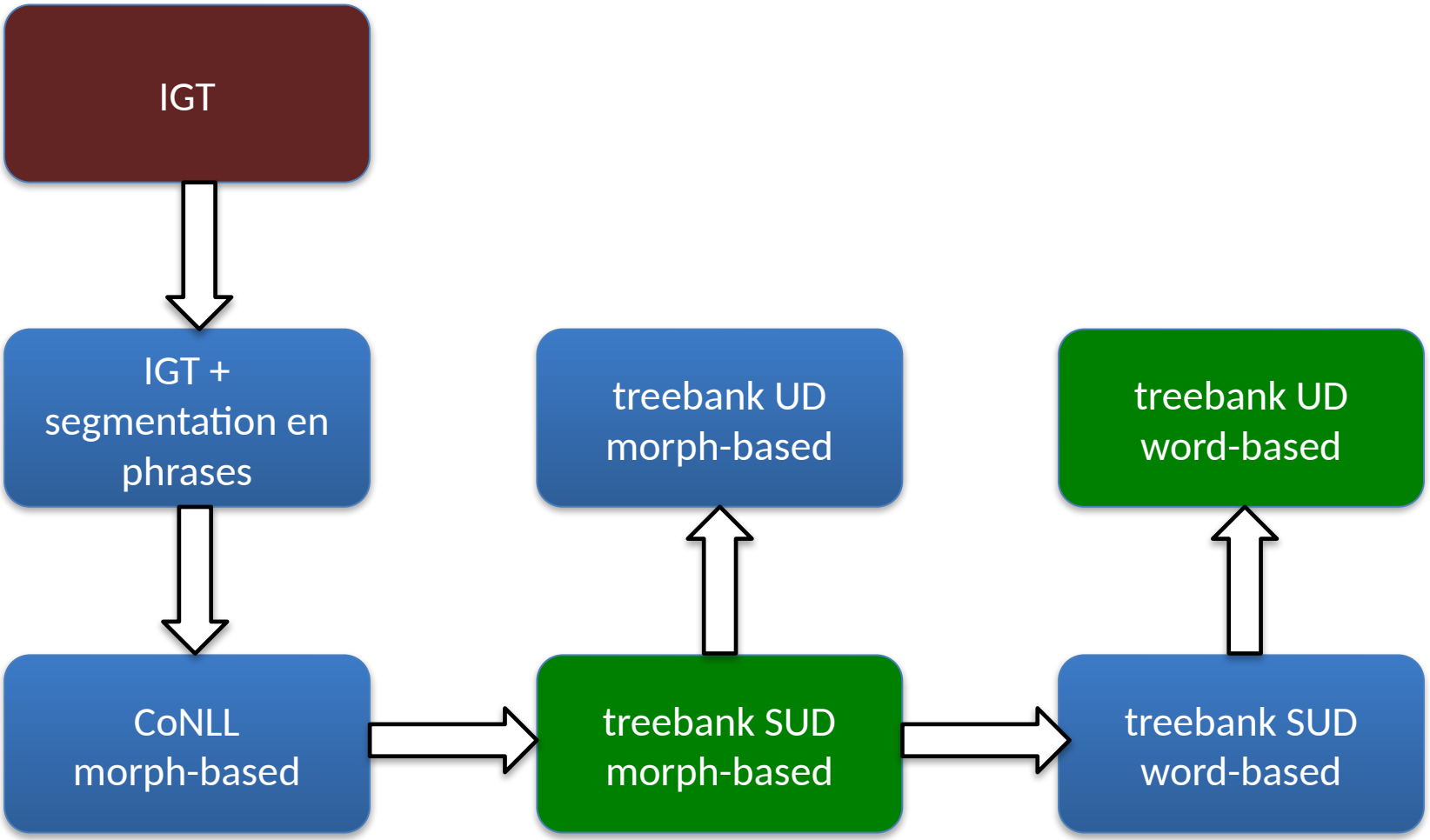
IGT vs UD

- Universal Dependencies
  - Un même schéma d'annotation pour toutes les langues
  - Word-based : le mot est l'unité minimale
  - Un jeu de 15 POS (non négociable)
  - Un jeu de traits morphosyntaxiques (ajouts possibles)
  - Un jeu de relations syntaxiques
    - Non négociable
    - Mais on peut ajouter des extensions (nmod:poss, obj:agent …)

Notre choix

- Un treebank morph-based
- Notre jeu de relations syntaxiques (SUD)
  - à la fois plus simple
    - subj au lieu de nsubj vs csubj
    - mod au lieu de nmod, amod, advmod, acl, advcl …
  - plus riche
    - comp:obl vs mod au lieu de obl
  - et basé sur des principes théoriques (distributionnels)
    - une relation correspond à un paradigme positionnel
      je veux une banane/manger/que tu manges
      => comp:obj au lieu de obj vs xcomp vs ccomp
    - tête fonctionelle : ADP est la tête d'un groupe adpositionnel …
- Convertible en UD word-based

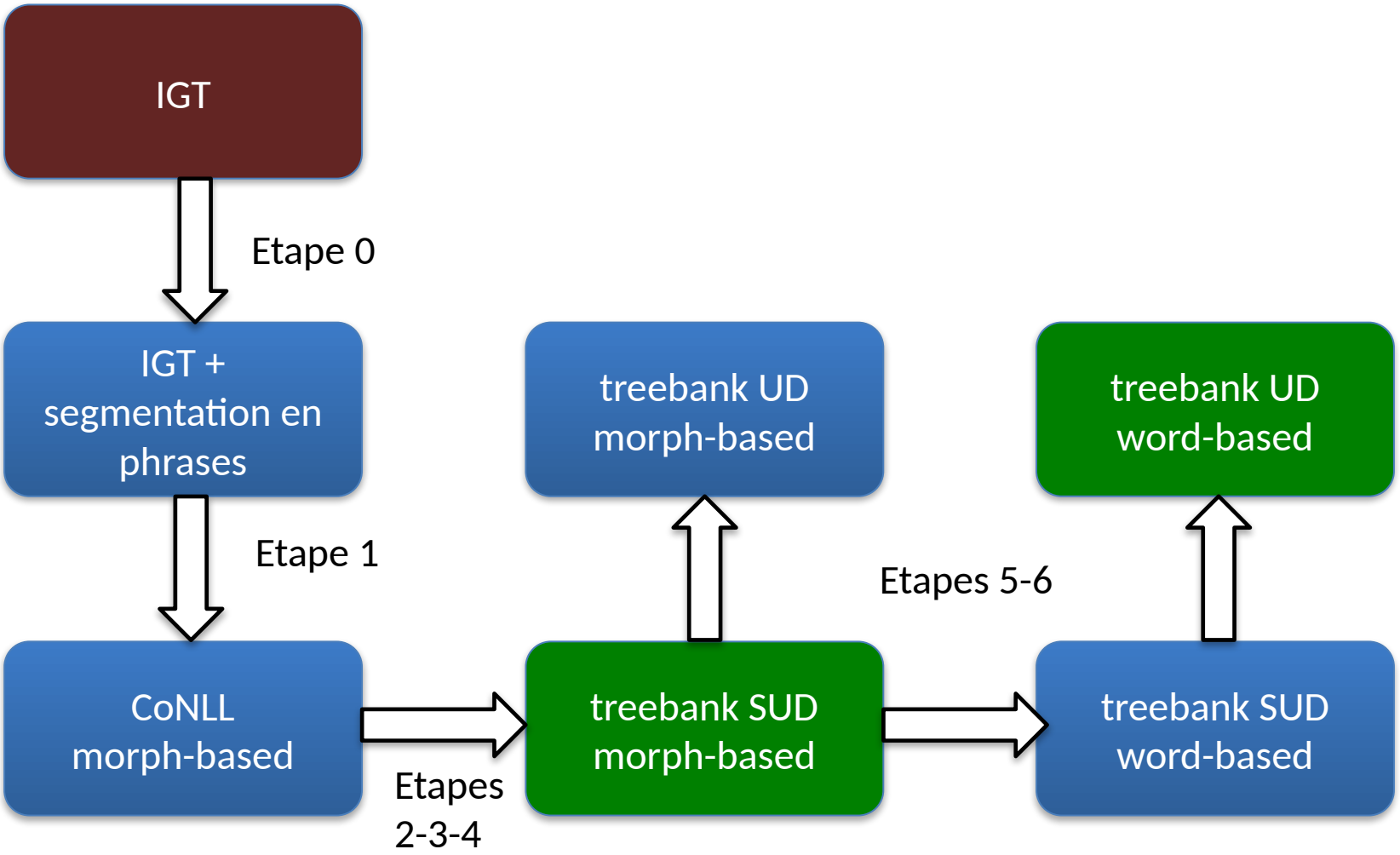
La chaîne de traitement



De l’IGT à un treebank

- Etape 0 : ajout des délimiteurs de phrase dans l'IGT, choix des tiers à exporter en CoNLL
- Etape 1 : glose to conll
- Etape 2 : enrichissement du conll avec les tables de conversion des gloses
- Etape 3 : préanalayse automatique
- Etape 4 : création d’un projet sous ArboratorGrew et annotation manuelle, guide d'annotation
- Etape 5 : conversion SUD-MB to UD-WB
- Etape 6 : mise en ligne sur grew-match et UD
- + validation de la cohérence des annotations

La chaîne de traitement



Conclusion

- enrichir une IGT pour obtenir un treebank S.UD
  - pas de perte d'information
  - structure de traits : avantage pour les requêtes
  - normalisation de certains traits
    - concepts comparatifs
- informations supplémentaires dans le treebank
  - unités maximales (phrases), dépendances syntaxiques, relations/fonctions syntaxiques, traits morpho-syntaxiques
- base de données UD
  - visibilité accru
  - comparaison de treebanks, typologie