

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.

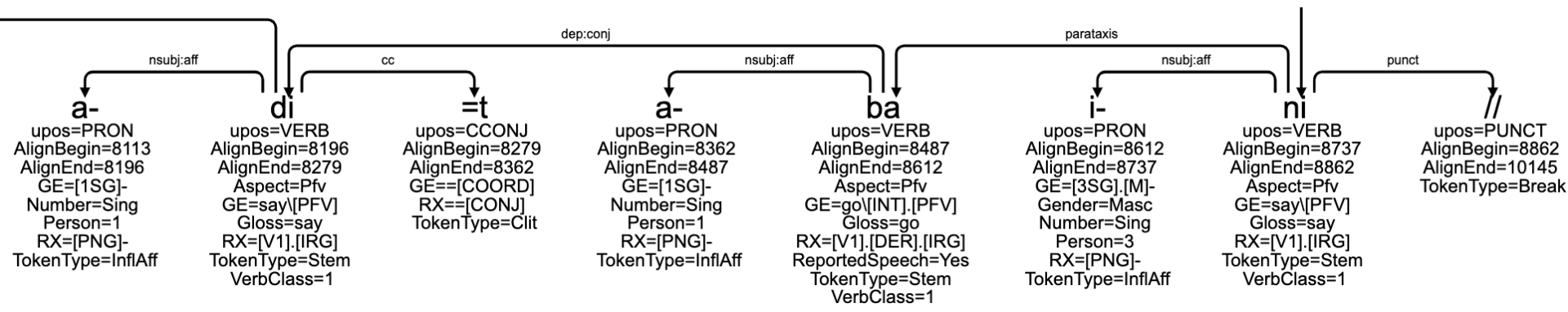
- stage de M2 Rayan Ziane (6 mois en 2021)

- 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

- T1
- T2
- T3 (word)
- T4 (morph)
- T5 (GE)
- T6 (RX)
- T7
- T8 (MFT)

00:00:07.000	00:00:07.500	00:00:08.000	00:00:08.500
BEJ_MV_NARR_03_camel_010		BEJ_MV_NARR_03	BEJ_MV_NARR_03_camel_012
w?i:d arrafi: /		248	adi:d ab ini //
w?i:d	arrafi /	248	adit aba ini //
w= ?i:d arraf -i /		a- di =t a- ba i- ni //	
DEF.SG.M= Aid_feast congratulate -AOR.1SG .		1SG- say\IP, =COO 1SG- go\INT.P, 3SG.M- say\PFV .	
DET= N.M V2 -TAM.PNG .		PNG- V1.IR, =CON PNG- V1.DER.I, PNG- V1.IRG .	
I will wish the Aid		I said, I went", he said.	
I went to wish him a blessed Aid", he said.			



Show corpora list

SUD_Beja-NSC@latest



1

Clustering 1: No Key Whether

lemma upos xpos features textform/wordform context

sentences order:

Search

Save

Export

Basic n-grams Clustering Misc

- Search for a form
- Search for a lemma (does not exist in all languages)
- Search for a POS (upos)
- Search for a dependency relation
- Search for both relations and tags
- Filter with NAP (Negative Application Patterns)

Complex edges (see **Grew-doc**)

- Search on one edge feature
- Use disjunction on edge features
- Use negation on edge features
- Request for absence of an edge feature
- Request for presence of an edge feature

56 occurrences [0.00s]

More results

1 / 10

- BEJ_MV_NARR_03_camel_001-008
- BEJ_MV_NARR_03_camel_009-013
- BEJ_MV_NARR_03_camel_014-019
- BEJ_MV_NARR_03_camel_020-021
- BEJ_MV_NARR_03_camel_022-024
- BEJ_MV_NARR_03_camel_025-034
- BEJ_MV_NARR_03_camel_035-042
- BEJ_MV_NARR_03_camel_043-053
- BEJ_MV_NARR_03_camel_054-063
- BEJ_MV_NARR_03_camel_064-067

0:00 / 3:49

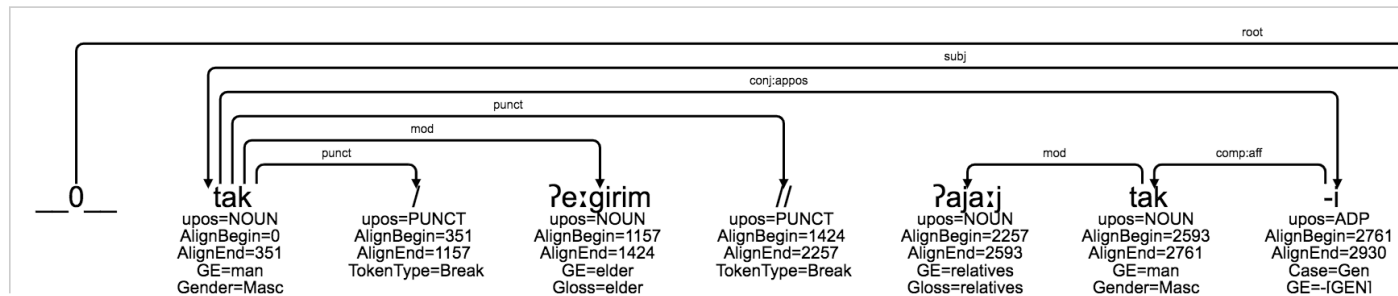
Speech rate: 0.5x

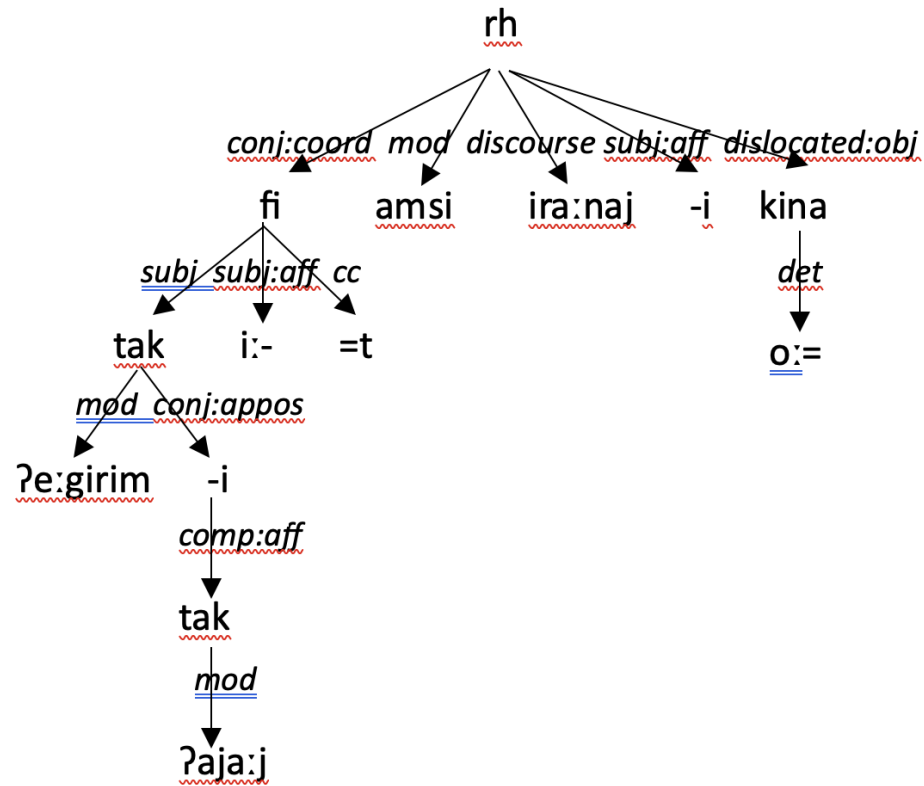
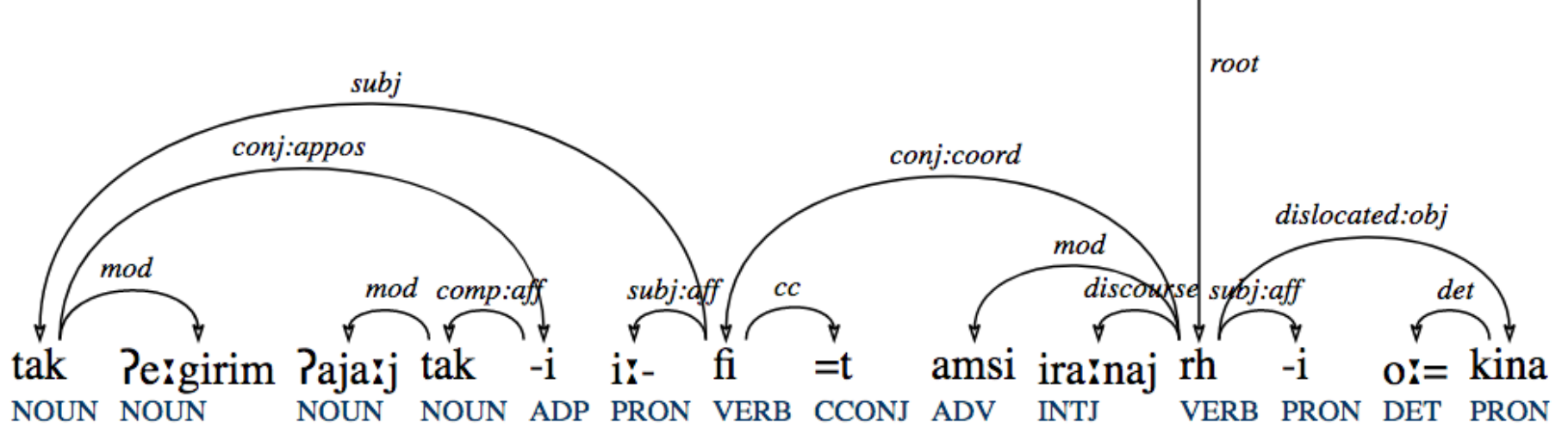
Metadata

CoNLL

SVG

0 tak / ?e:girim // ?ajaj tak -i i: -fi =t amsi ir:naj rh -i / o:= kina /





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

grewmatch Tutorial UD 2.9 SUD 2.9 UD Latest SUD Latest

Show corpora list **SUD_Beja-NSC@latest**

1

Clustering 1: No Key Whether

lemma upos xpos features textform

Search Save Export

56 occurrences [0.00s]

More results

- BEJ_MV_NARR_03_camel_001-008
- BEJ_MV_NARR_03_camel_009-013
- BEJ_MV_NARR_03_camel_014-019
- BEJ_MV_NARR_03_camel_020-021
- BEJ_MV_NARR_03_camel_022-024
- BEJ_MV_NARR_03_camel_025-034
- BEJ_MV_NARR_03_camel_035-042
- BEJ_MV_NARR_03_camel_043-053
- BEJ_MV_NARR_03_camel_054-063
- BEJ_MV_NARR_03_camel_064-067

Copy in clipboard

```
# sent_id = BEJ_MV_NARR_03_camel_001-008
# text = tak / ?e:girim // ?aja:j tak -i i:- fi =t amsi ir:nej rh -i / o:= kina /
# text_wb = tak / ?e:girim // ?aja:j taki i:fi =t amsi ir:nej rhi / o:= kina /
# text_en = "There was an old man of my family and, gosh! I even saw him today."
# phonetic_text = tak / e:girim // ?aja:j taki i:fi:t amsi ir:nej rhi / o:kna /
# sound_url = https://api.nakala.fr/data/10.34847/nkl.dfa3fw9v/9b76a83067e87e8edb370d60a6cd1e7c50b69506

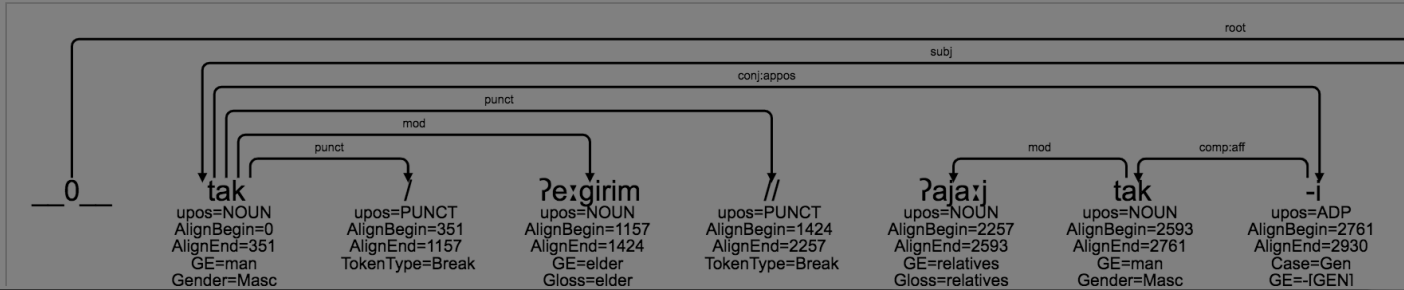
1 tak _ NOUN N Gender=Masc 9 subj _ AlignBegin=0|AlignEnd=351|GE=man|Gloss=man|RX=[SBJ].[N].
[M]|TokenType=Stem
2 / _ PUNCT PUNCT _ 1 punct _ AlignBegin=351|AlignEnd=1157|TokenType=Break
3 ?e:girim _ NOUN CN _ 1 mod _ AlignBegin=1157|AlignEnd=1424|GE=elder|Gloss=elder|RX=
[CN]|TokenType=Stem
4 // _ PUNCT PUNCT _ 1 punct _ AlignBegin=1424|AlignEnd=2257|TokenType=Break
5 ?aja:j _ NOUN N _ 6 mod _ AlignBegin=2257|AlignEnd=2593|GE=relatives|Gloss=relatives|RX=
[N]|TokenType=Stem
6 tak _ NOUN N Gender=Masc 7 comp:aff _ AlignBegin=2593|AlignEnd=2761|GE=man|Gloss=man|RX=[N].
[M]|TokenType=Stem
7 -i _ ADP CASE Case=Gen 1 conj:appos _ AlignBegin=2761|AlignEnd=2930|GE=-[GEN]|RX=-
[CASE]|TokenType=InflAff
8 i:- _ PRON PNG Gender=Masc|Number=Sing|Person=3 9 subj:aff _
AlignBegin=2930|AlignEnd=3042|GE=[3SG].[M]-|RX=[PNG]-|TokenType=InflAff
9 fi _ VERB VI, IRG Aspect=Aor|VerbClass=1 13 conj:coord _ AlignBegin=3042|AlignEnd=3154|GE=be_there\
[AOR]|Gloss=be_there|RX=[V1].[IRG]|TokenType=Stem
10 =t _ CONJN CCONJ _ 9 cc _ AlignBegin=3154|AlignEnd=3267|GE=[[COORD]]|RX=[[CONJ]]|TokenType=Clit
11 amsi _ ADV ADV _ 13 mod _ AlignBegin=3267|AlignEnd=3603|GE=today|Gloss=today|RX=
[ADV]|TokenType=Stem
```

0:00 / 3:49

Speech rate: 0.5x

Metadata CoNLL SVG

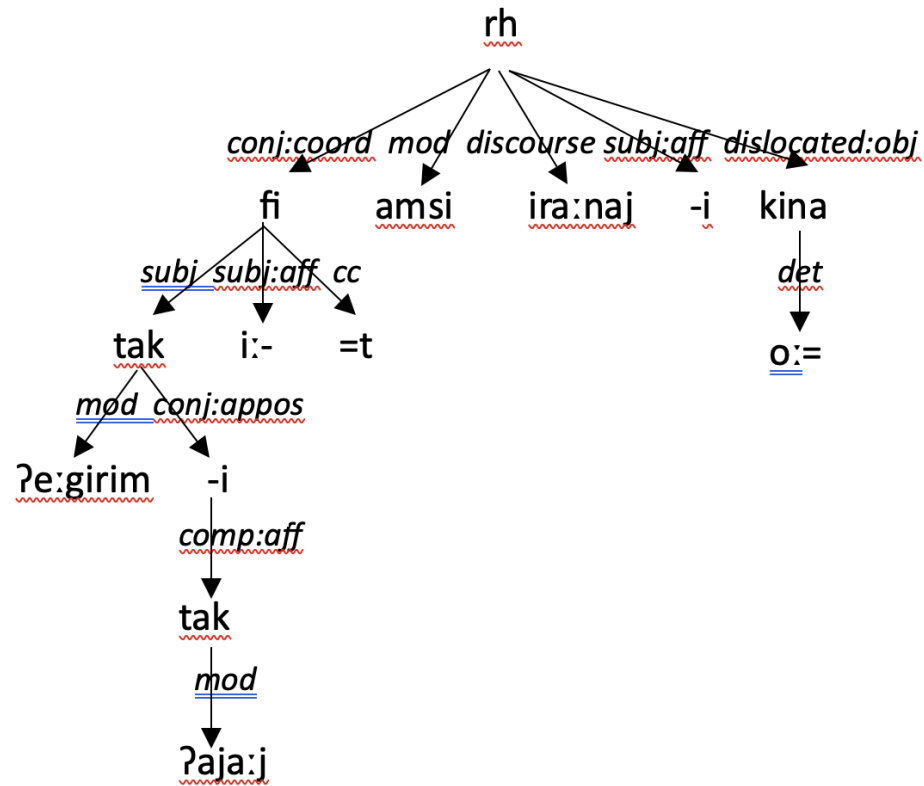
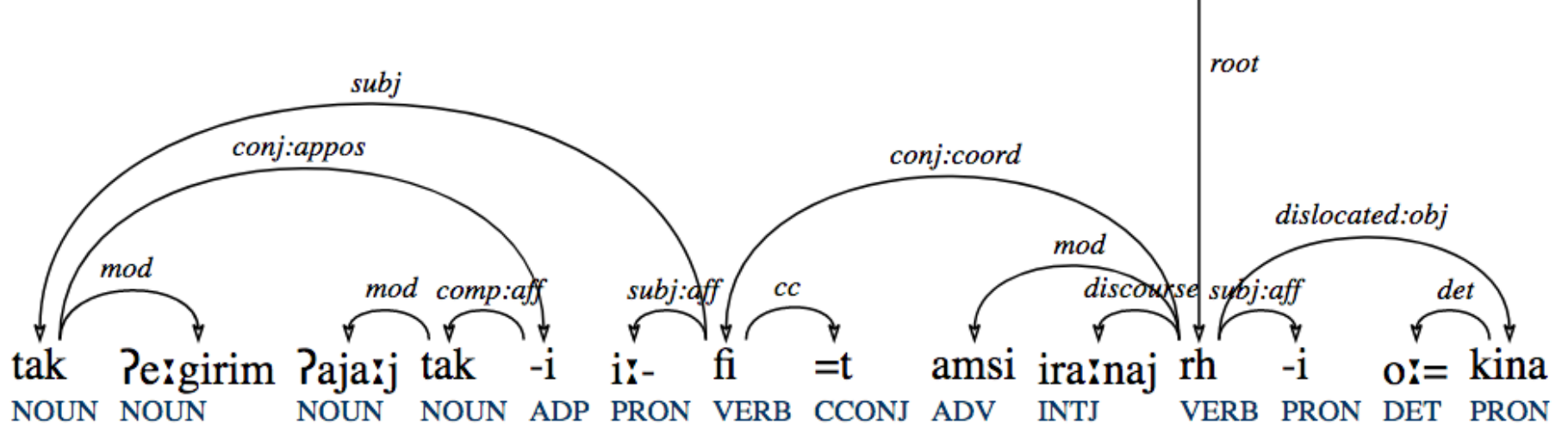
0 tak / ?e:girim // ?aja:j tak -i i:- fi =t amsi ir:nej rh -i / o:= kina /



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 / ?e:girim // ?aja:j tak -i i:- fi =t amsi ira:naj rh -i / o:= k				
3	# text_wb =	tak / ?e:girim // ?aja: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 // ?aja:j taki i:fi:t amsi ira:nej rhi / o:k				
6	# sound_url =	https://api.nakala.fr/data/10.34847/nkl.dfa3fw9v/9b76a85007e87e8cc8570d00a6cc1e7c				
7	1	tak		NOUN	N	Gender=Masc 9 subj
8	2	/		PUNCT	PUNCT	1 punct
9	3	?e:girim		NOUN	CN	1 mod
10	4	//		PUNCT	PUNCT	1 punct
11	5	?aja: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							



Affixes dérivationnels

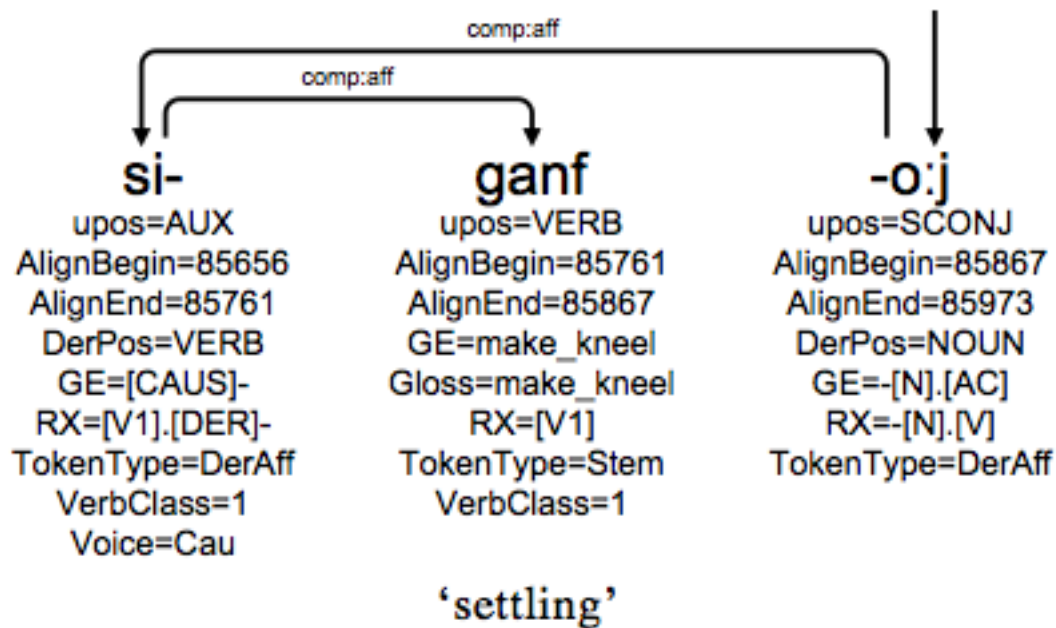


Figure 4. Derivational affixes (SUD-style)

Affixes dérivationnels

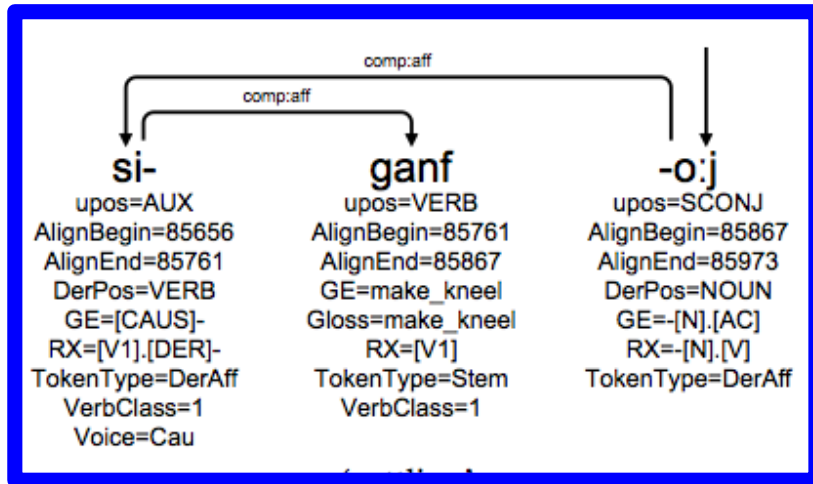


Figure 4. Derivational affixes (SUD-style)

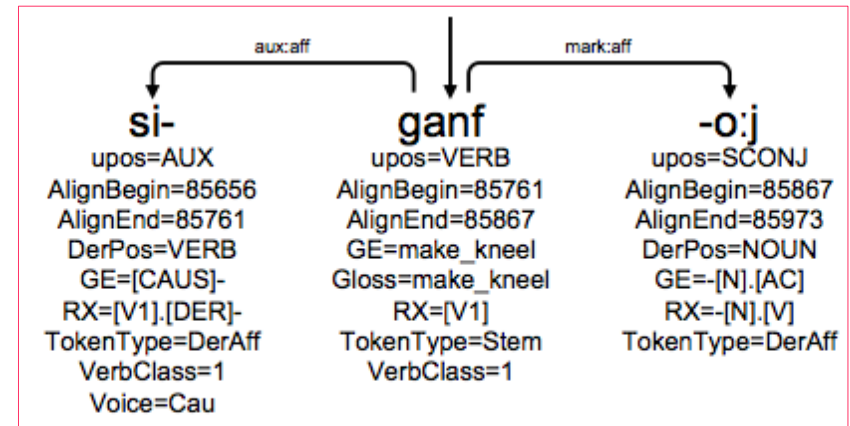


Figure 5. Derivational affixes (UD-style)

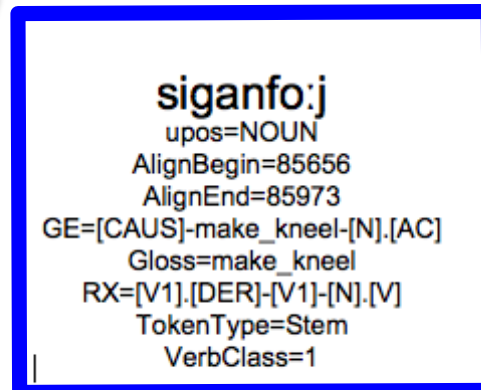


Figure 6. The word-based annotation of the derived word *siganfo:j*

information
manquante

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

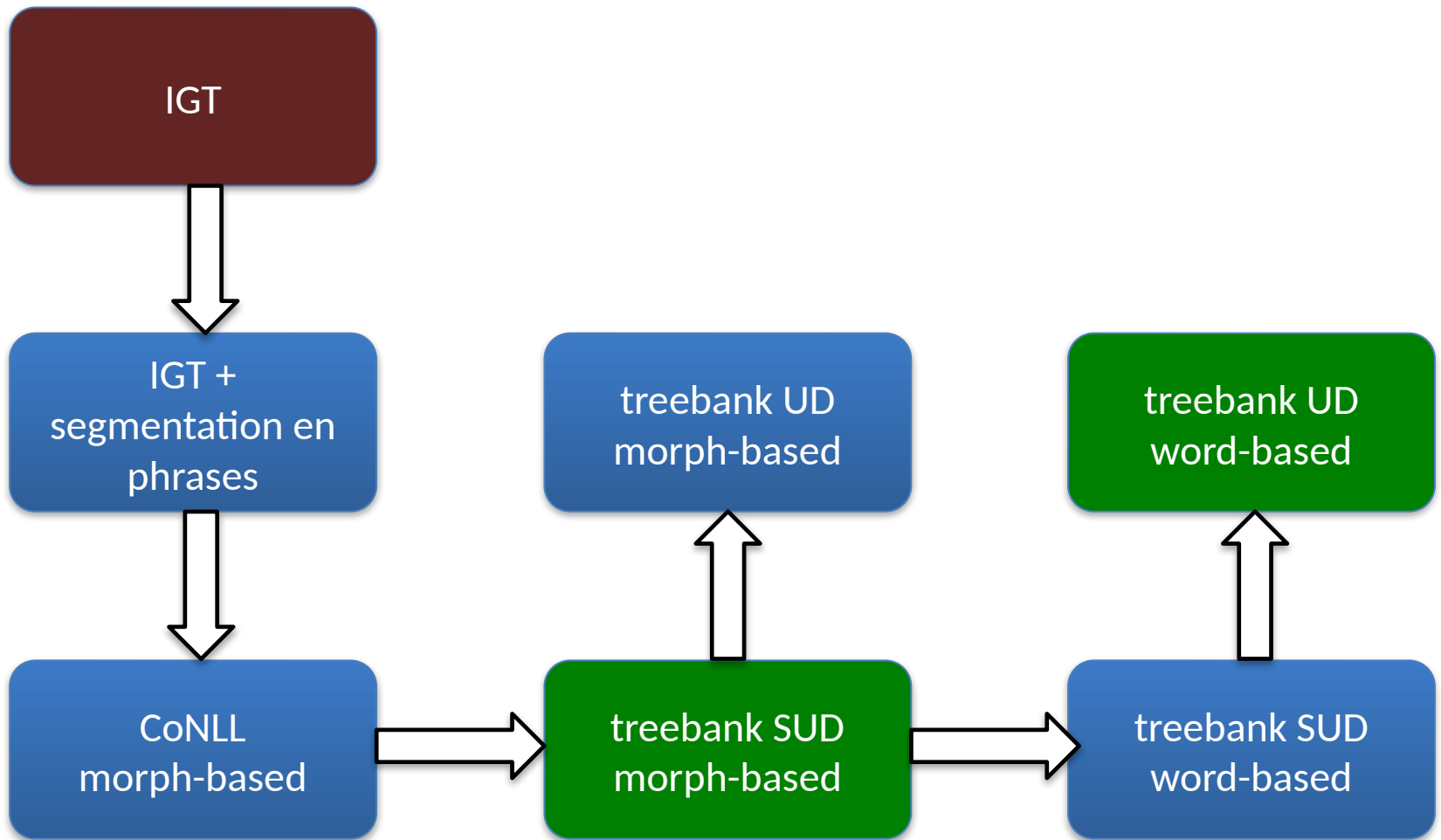
IGT vs UD

- 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 fonctionnelle : ADP est la tête d'un groupe adpositionnel ...
- Convertible en UD word-based

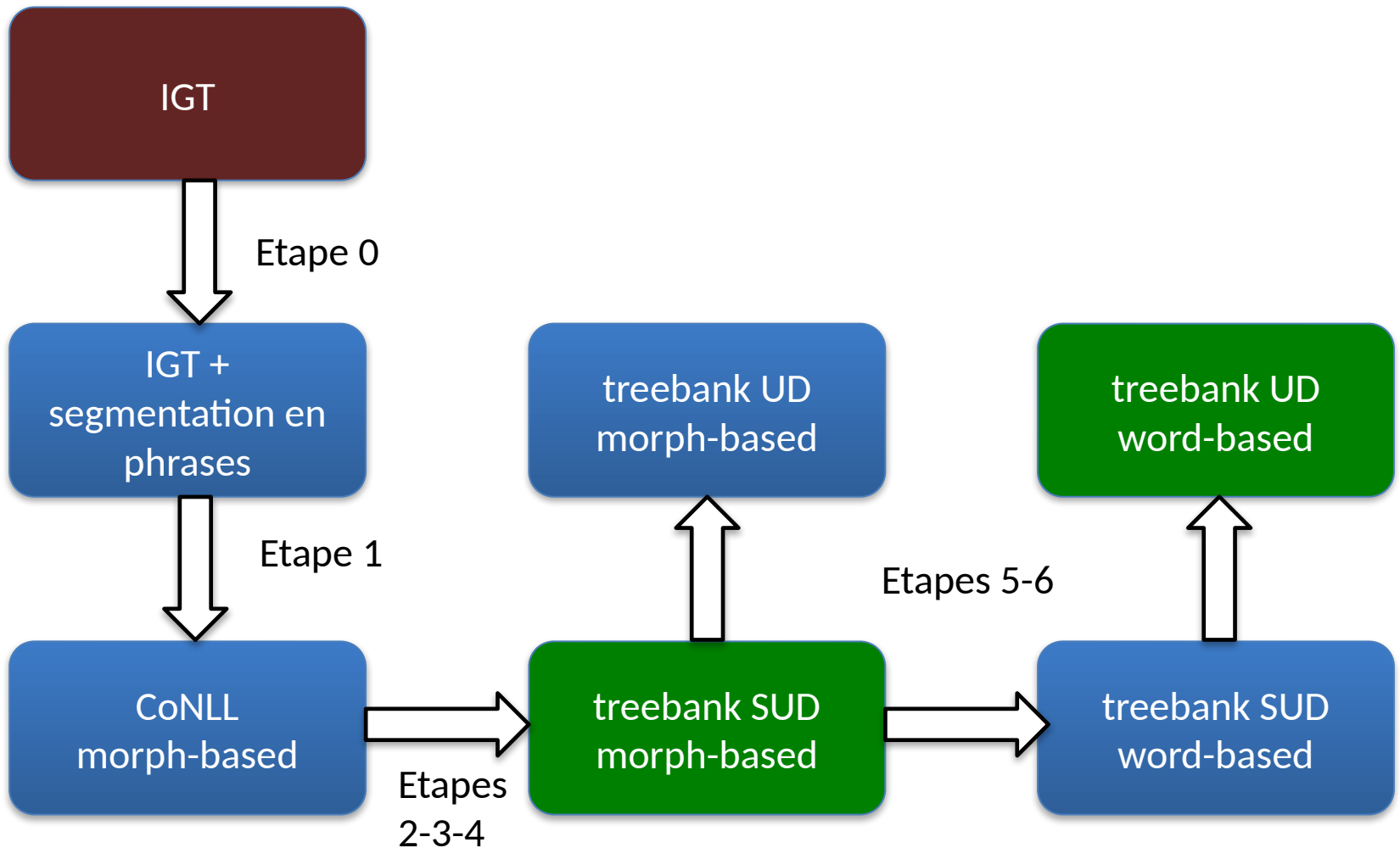
La chaîne de traitement



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 : préanalyse 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