

Typologie des langues automatique à partir de treebanks

Philippe Blache Grégoire de Montcheuil Stéphane Rauzy

TALN - Caen - 23 Juin 2015



UMR 7309 - CNRS - AMU
PAROLE ET
LANGAGE



BLRI
Brain & Language Research Institute

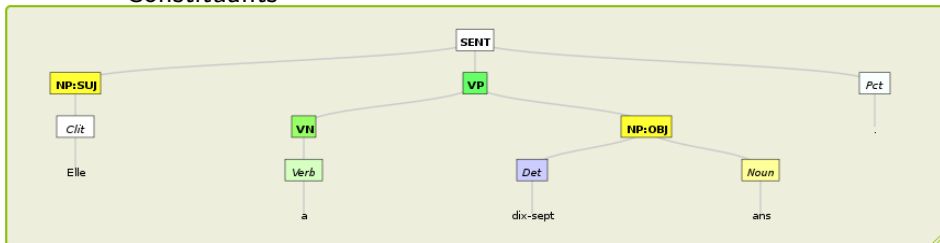


Plan

- ▶ Treebanks, l'Universal Dependencies Treebank
- ▶ Extraction de propriétés
- ▶ Typologie des langues et classification

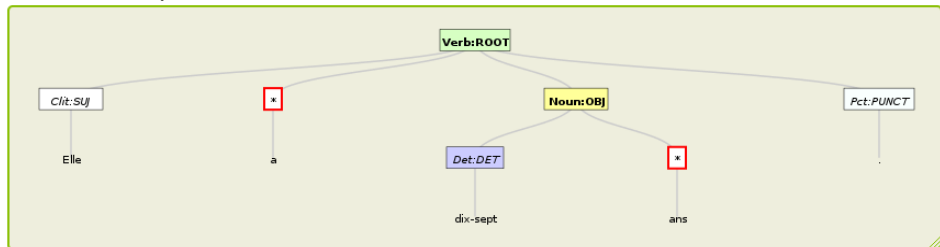
Treebanks

► Constituants



Elle a dix-sept ans .

► Dépendances



Elle a dix-sept ans .

Universal dependencies treebank

- ▶ 17 POS tags *universels*
- ▶ 40 relations de dépendance *universelles*
- ▶ 10 langues (version 1.0)

L'Universal Dependencies Treebank est disponible sur LINDAT,
identifiant pérenne : hdl:11234/1-1464 (version 1.0)

10 langues

	famille - genre	langue	#arbres	#tokens	caractéristiques typologiques
Indo-Européenne	Slave	cs tchèque	87,913	1,482,147	SVO, accentuelle, ordre des mots libre
	Germanique	de allemand	15,918	297,985	V2 et SOV, flexionnelle, accusative, accentuelle, à accent d'intensité
		en anglais	16,622	254,930	SVO, flexionnelle, accusative, accentuelle, à accent d'intensité
		sv suédois	6,026	96,699	SVO, flexionnelle, accusative, accentuelle, à accent de hauteur
	Romane	es espagnol	16,006	430,764	SVO, syllabique
		fr français	16,418	398,964	SVO, flexionnelle, accusative, syllabique
		it italien	10,077	214,748	SVO, syllabique
	Celte	ga irlandais	1,020	23,686	VSO, flexionnelle, accusative, accentuelle, à accent d'intensité
Ouralienne	Fenique	fi finnois	13,581	181,022	SVO, ordre libre
	Ougrienne	hu hongrois	1,299	25,064	SVO, ordre libre, agglutinante, accusative

Propriétés

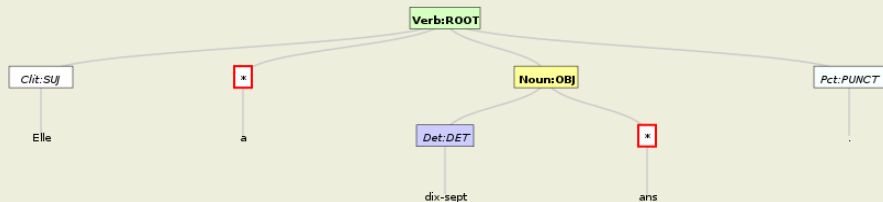
Linéarité : l'ordre des composants est toujours le même
precede(A,B) (i.e. *A précède toujours B*)

Exigence : la présence de l'un exige la présence de l'autre
require(A,B) (i.e. *si A est présent, B est aussi présent*)

Exclusion : la présence de l'un exclut la présence de l'autre
exclude(A,B) (i.e. *si A est présent, B ne l'est pas*)

Unicité : n'apparaît jamais plusieurs fois dans la partie droite
unicity(A)

Extraction de la CFG implicite



Elle a dix-sept ans .

Verb:ROOT → Clit:Suj Verb:ROOT Noun:OBJ Pct:PUNCT

Noun:OBJ → Det:DET Noun:OBJ

nota : méthode identique pour les formalismes de dépendances et de constituants

Extraction des propriétés

NOUN:nsubj → DET:det NOUN:nsubj
NOUN:nsubj → DET:det NOUN:nsubj NOUN:nmod
NOUN:nsubj → DET:det NOUN:nsubj ADJ:amod
(...)

Validating

⇒ [NOUN:nsubj] precede(DET:det, NOUN:nsubj)

NOUN:nsubj → DET:det NOUN:nsubj DET:det ADJ:amod

Violating

(1 occ.)

NOUN-nsubj 7191:28			
DET-det	*	DET-det	ADJ-amod
l'	humanité	toute	entière

l' humanité toute entière

Propriété "plus faible"

[NOUN:advcl] precede(VERB:cop, NOUN:nmod)

copula nominal modifier

NOUN:advcl → ADP:mark PRON:nsubj VERB:cop
DET:det NOUN:advcl NOUN:nmod

NOUN-advcl 7150:27			
ADP:mark	PRON:nsubj	VERB:cop	DET:det
comme	c'	était	le

NOUN-nmod 7150:33		
ADP:case	DET:det	*
dans	le	livre

comme c'était le cas dans le livre

NOUN:advcl → ADP:mark VERB:cop NOUN:advcl NOUN:nmod

NOUN-advcl 13395:45		
ADP:mark	VERB:cop	*
pour	devenir	professeur

NOUN-nmod 13395:49		
ADP:case	DET:det	*
a	l'	université

ADJ-amod 13395:54	
PROPN-nmod	*
impériale	

ADP:case	
*	
de	Tokyo

pour devenir professeur
à l'université impériale de Tokyo

NOUN:advcl → CONJ:mark NOUN:nmod PRON:nsubj
VERB-cop ADV-neg NOUN:advcl ADJ-amod

NOUN-advcl 6147:4						
CONJ:mark	NOUN-nmod 6147:6		PRON:nsubj	VERB:cop	ADV:neg	*
si	ADP:case	DET:det	*	vous	êtes	non
	6147:7	la	convention			résident
			fiscale			français

ADP:mwe	
*	
de	par

si de part la convention fiscale
vous êtes non résident français

Poids des propriétés

- ▶ Ratio des occurrences vérifiant la propriété

$$w_0 = \frac{Occ(\textit{Validating}(p))}{Occ(\textit{Validating}(p)) + Occ(\textit{Violating}(p))}$$

- ▶ $w_0 = 1$: propriété toujours vérifiée
 - ▶ $w_0 > 0.5$: plus souvent vérifiée que contredite
 - ▶ $w_0 \geq \alpha/(\alpha + 1)$: α -plus souvent vérifiée que contredit
- ▶ Pondération par la fréquence

$$w_1 = w_0 \frac{Occ(\textit{Validating}(p))}{\sigma_{Occ}}$$

Où :

- $Occ(R)$ somme des occurrences du sous-ensemble de règles R
- σ_{Occ} somme des occurrences des règles de même tête

MarsaGram

POS	funcnt	nb_rules	properties
RNM	root	12	291
VERB	±	12560	18491
VERB	act	1240	5172
VERB	act-relcl	1374	5552
VERB	adocl	1194	6128
VERB	adomad	10	43
VERB	amod	5	39
VERB	areaa	28	405
VERB	aux	0	0
VERB	auxpass	0	0
VERB	case	8	19
VERB	cc	1	7
VERB	ccame	737	4229
VERB	compound	1	2
VERB	conj	1479	5995
VERB	cop	29	121
VERB	csubi	47	628
VERB	dep	26	387
VERB	det	1	7
VERB	dobj	3	44
VERB	mark	1	24
VERB	ntmod	1	0
VERB	nmmod	11	257
VERB	nsubi	2	19
VERB	parataxis	361	2881
VERB	root	7030	13284
VERB	scame	146	1235
X	±	488	4493
X	act	9	89
X	act-relcl	2	44
X	adocl	9	92
X	adomad	6	19
X	amod	3	58
X	areaa	140	771
X	aux	0	0
X	case	13	117
X	cc	1	2
X	ccame	2	99

Properties

5172 properties for VERB-act (CSV) (relations (CSV))

1 to 25 (5172) page size: 25 show page: 1 Disable Pager

property	symbol-1	symbol-2	frequency	w0	w1
precede	*	NOUN-nmod	43.35%	0.9944	0.4311
precede	*	NOUN-dobj	25.03%	1.0000	0.2503
precede	ADP-case	NOUN-dobj	15.61%	0.9922	0.1549
precede	*	PROPN-nmod	12.56%	1.0000	0.1256
precede	ADP-case	NOUN-nmod	12.30%	0.9250	0.1138

nb_rules	occurrences	frequency	rules
precede	529	2827	43.35%
follow	7	10	0.15%
mixed	5	6	0.09%

nb_rules	occurrences	frequency	rules
precede	287	1692	25.03%

VERB-act → ADP-case + NOUN-dobj (7.31%)

nb_rules	occurrences	frequency	rules
precede	162	1018	15.61%
follow	4	4	0.06%
mixed	3	4	0.06%

nb_rules	occurrences	frequency	rules
precede	247	819	12.56%

nb_rules	occurrences	frequency	rules
precede	201	802	12.30%
follow	22	56	0.86%

VERB-root 461:1

PROPN-nsubi * VERB-act 461:4 CONJ-cc VERB-conj PUNCT-punct

Vilgax tante de récupérer NOUN-dobj 461:7 mais échouera .

le cristal

Disponible sur le SLDR, identifiant pérenne : hdl:11041/ortolang-000917

Typologie des langues

Taille de la grammaire de propriétés

Nombre de propriétés ($w_0 = 1$, sans les relations de dépendance)

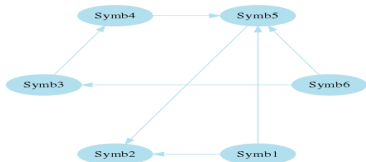
cs	de	en	es	fi	fr	ga	hu	it	sv
598	683	755	708	523	716	547	448	750	547

Ordre des mots libre (le tchèque, le finnois et le hongrois)
⇒ grammaires de propriétés de taille plus réduite

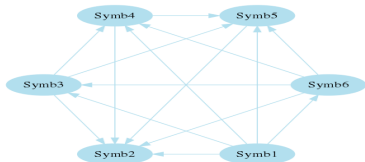
Densité des contraintes

Graphe d'une propriété

- sommets : les dépendants d'une tête donnée
- arêtes : une des propriétés binaires (linéarité, exigence, exclusion)



$$\text{Densité} = \frac{\text{nb.prop}}{t_c * (t_c - 1)/2} = \frac{7}{15}$$



$$\text{Densité} = 1$$

Densité de la relation de linéarité en français et finnois :

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X	Moy
fr	0,40	0,33	0,33	0,32	0,53	0,24	0,30	0,86	0,43	0,27	0,24	0,67	0,35	0,75	0,60	0,46	0,40	0,42
fi	0,04	0,30	0,08	0,36	0,42	0,00	0,46	0,06	0,07	0,00	0,06	0,14	0,29	1,00	0,06	0,05	0,15	0,22

Propriétés communes

Propriété : quadruplet $\langle C, tp, A, B \rangle$

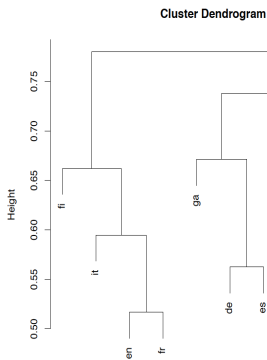
- C : contexte (tête)
- tp : type de propriété (*precede, require, exclude, unicity*)
- A et B : composants (dépendants)

A , B et C appartiennent au même *tagset* pour toutes les langues

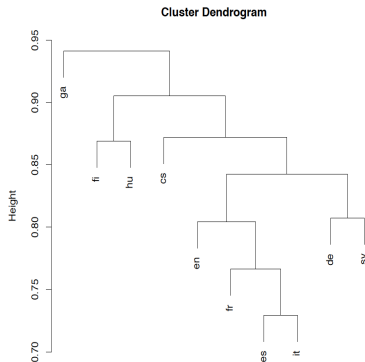
Nombre de propriétés communes entre l'italien et les autres langues :

propriétés	cs	de	en	es	fi	fr	ga	hu	sv
toutes	706	740	1022	972	755	1023	566	505	630
linéarité	114	142	192	225	111	186	97	86	142

Classification hiérarchique



toutes propriétés



linéarité

$$\text{simil}(lg_1, lg_2) = \frac{\text{card}(\mathcal{P}(lg_1) \cap \mathcal{P}(lg_2))}{\text{card}(\mathcal{P}(lg_1) \cup \mathcal{P}(lg_2))}$$

Conclusion

- ▶ Outil d'extraction de propriétés et d'exploration de treebanks (indépendant du formalisme)
- ▶ Mesures permettant de comparer les langues
- ▶ Méthode de classification des langues