Typologie des langues automatique à partir de treebanks

Philippe Blache Grégoire de Montcheuil Stéphane Rauzy

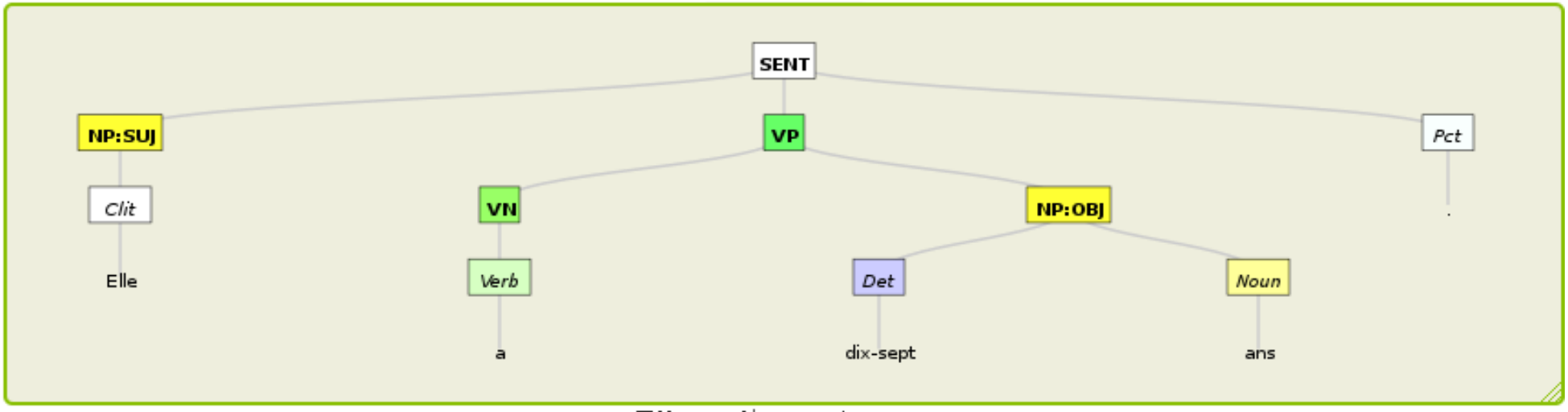
TALN - Caen - 23 Juin 2015

Plan

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

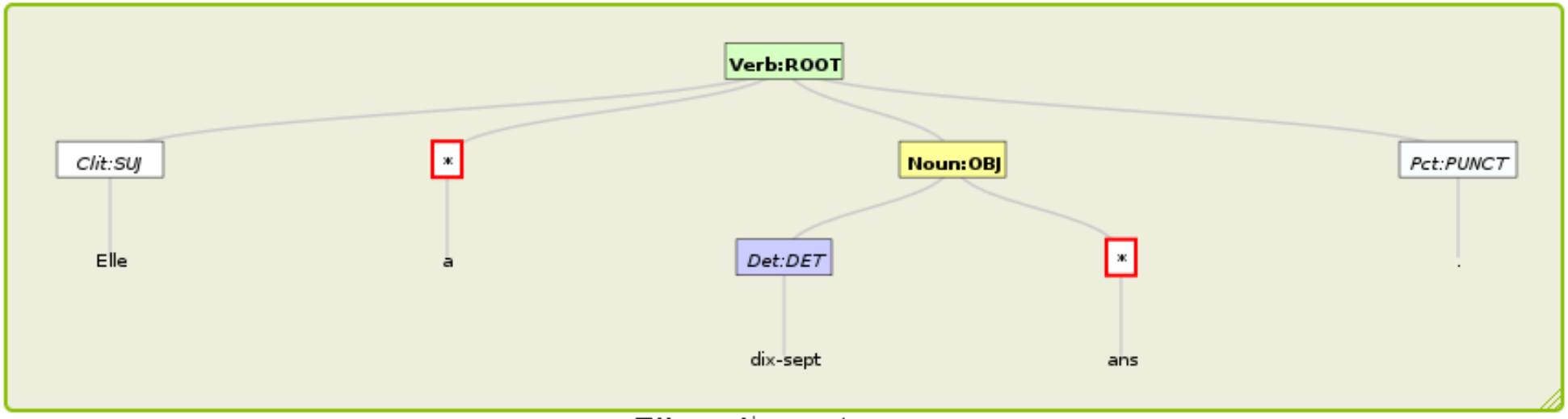
Treebanks

- 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

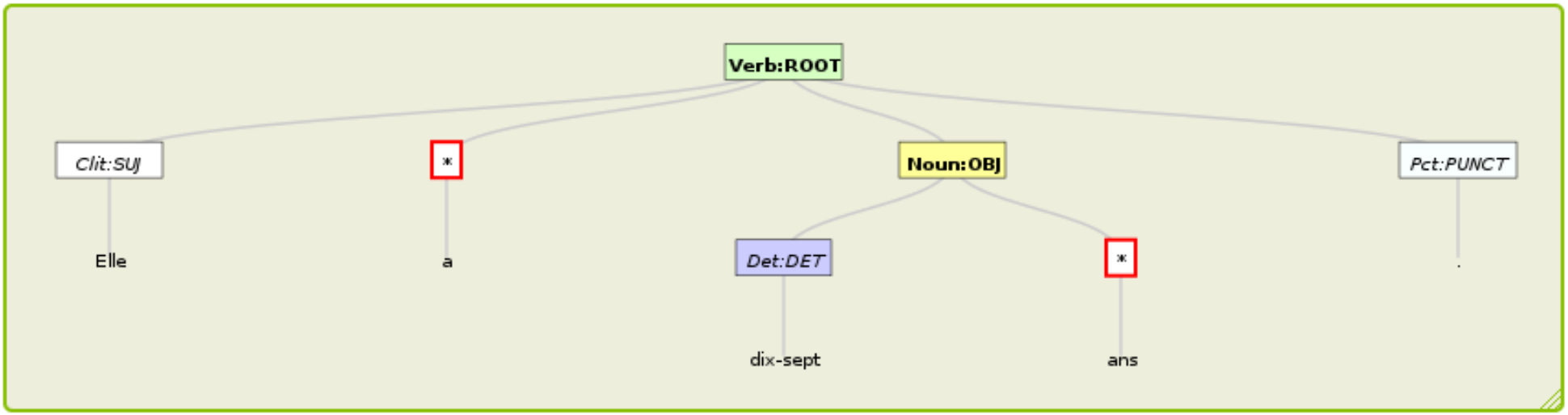
Linéarité : l'ordre des composants est toujours le même
precede(A,B) *(i.e. A précède toujours B)*

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'apparait jamais plusieurs fois dans la partie droite
unicity(A)

Extraction de la CFG implicite



Elle a dix-sept ans .

`Verb:ROOT` $\rightarrow$ `Clit:SUJ` `Verb:ROOT` `Noun:OBJ Pct:PUNCT`

`Noun:OBJ` $\rightarrow$ `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

(...)

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

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

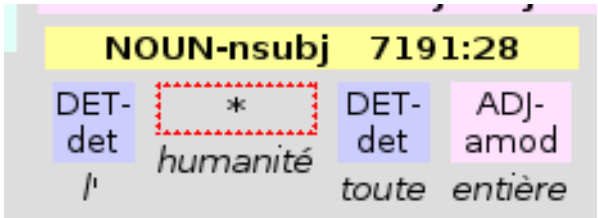
(...) Validating

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

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

Violating

(1 occ.)



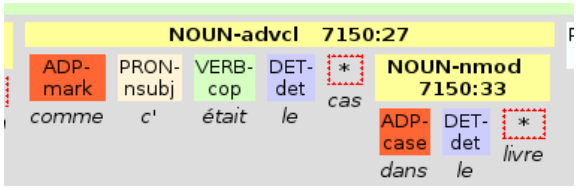
l' *humanité* *toute* *entière*

Propriété “plus faible”

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

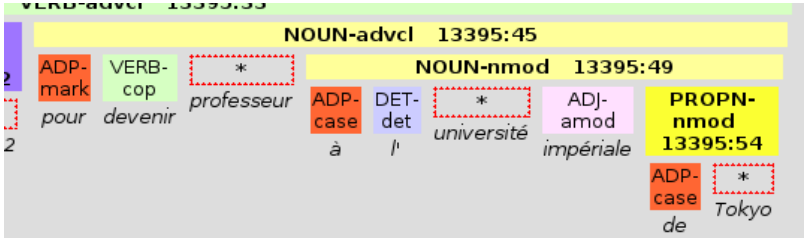
copula *nominal modifier*

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



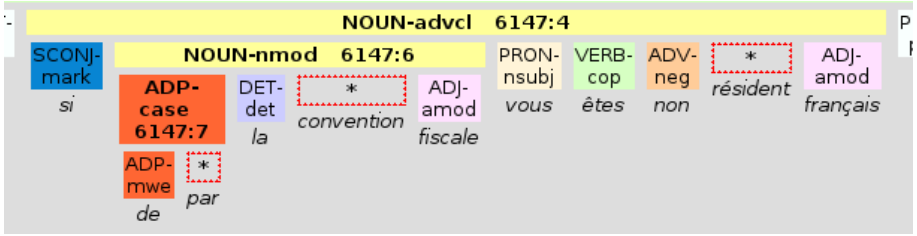
comme c'était le cas dans le livre

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



pour devenir professeur
à l'université impérial de Tokyo

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



si de part la convention fiscale
vous êtes non résident français

Poids des propriétés

- Ratio des occurences verifiant la propriété

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

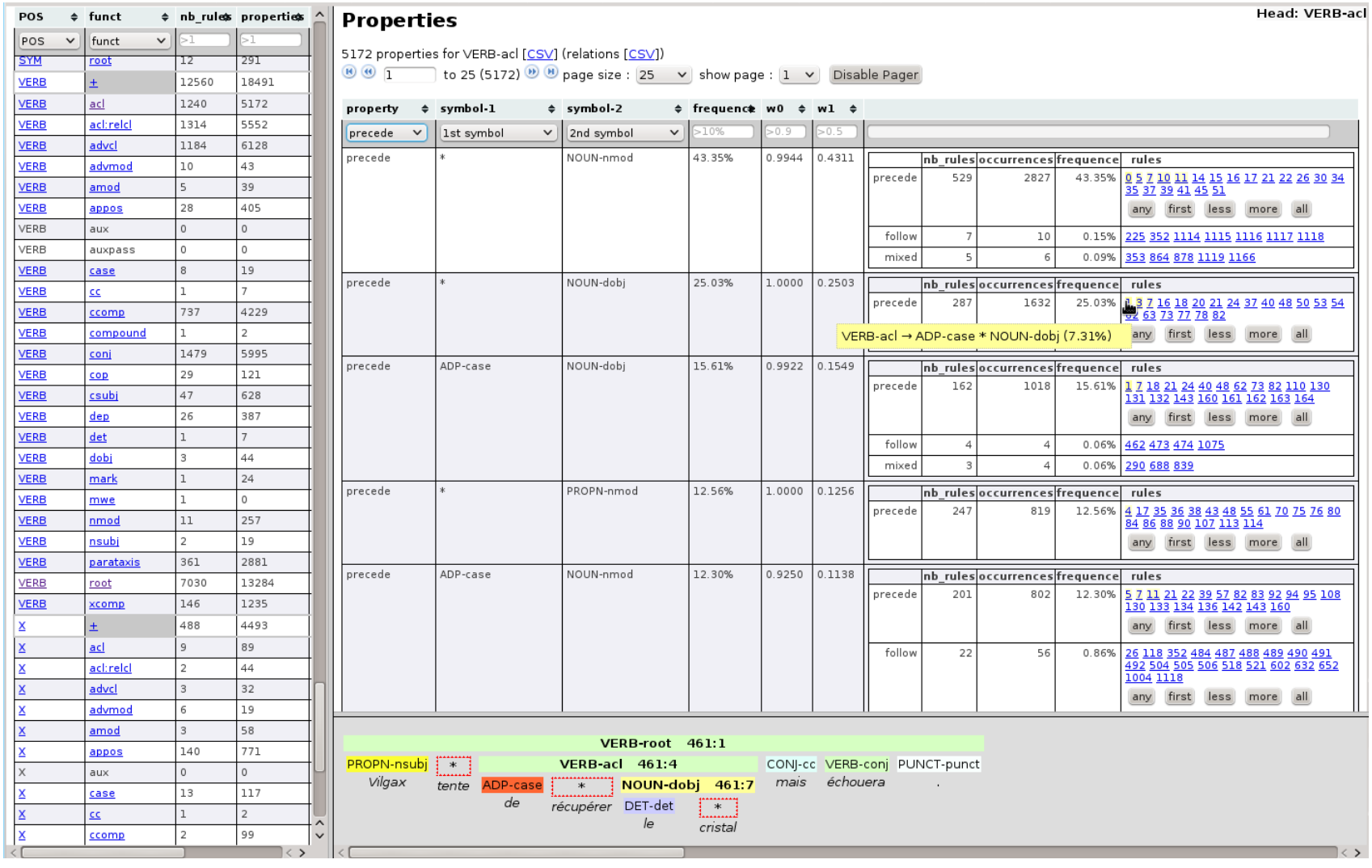
  - $w_0 = 1$: propriété toujours vérifiée
  - $w_0 > 0.5$: plus souvent vérifiée que contredite
  - $w_0 \geq \alpha/(\alpha + 1)$: α-*plus souvent* vérifiée que contredit

- Pondération par la fréquence

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

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

MarsaGram



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

Typologie des langues

Taille de la grammaire de propriétés

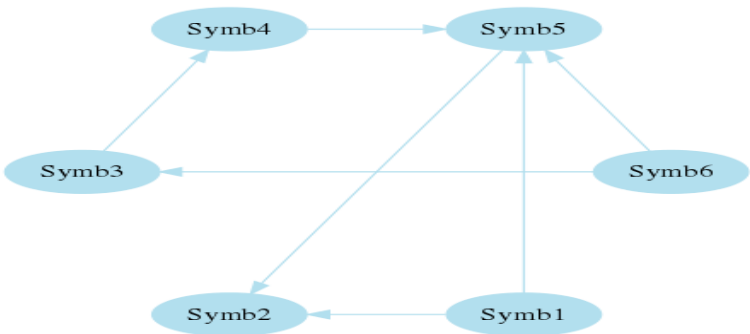
Nombre de propriétés ($w_0 = 1$, sans les relations de dépendance)

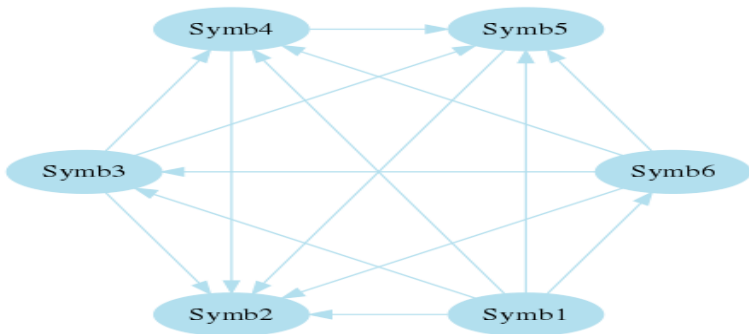
cs	de	en	es	fi	fr	ga	hu	it	sv
598	683	755	708	523	716	547	448	750	547

Ordre des mots libre (le tchèque, le finnois et le hongrois)
$\Rightarrow$ 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 (linératité, exigence, exclusion)





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

$$Densité = 1$$

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

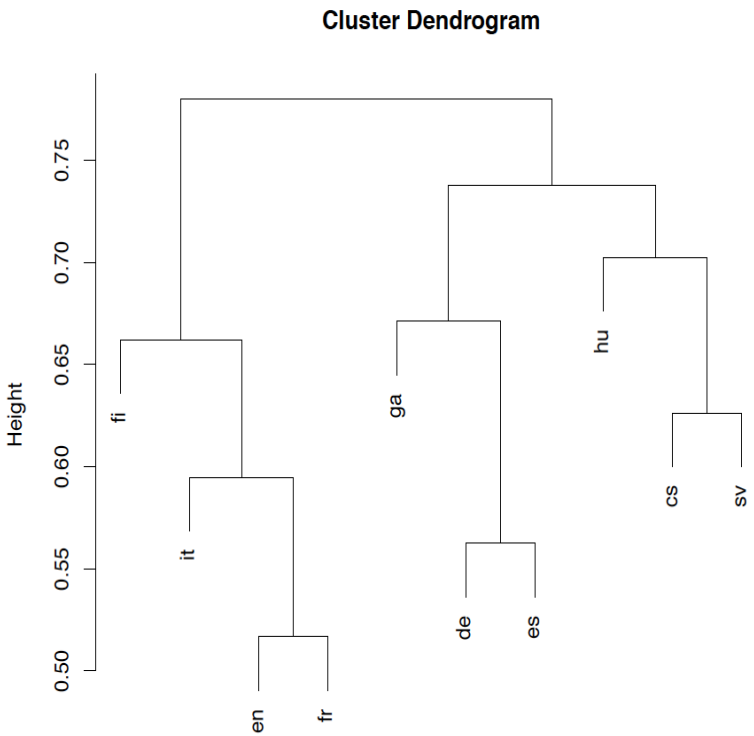
Propriété : quadruplet $< C, tp, A, B >$
- C : contexte (tête)
- tp : type de propriété (*precede*, *require*, *exclude*, *unicity*)
- A et B : composants (dépendants)

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

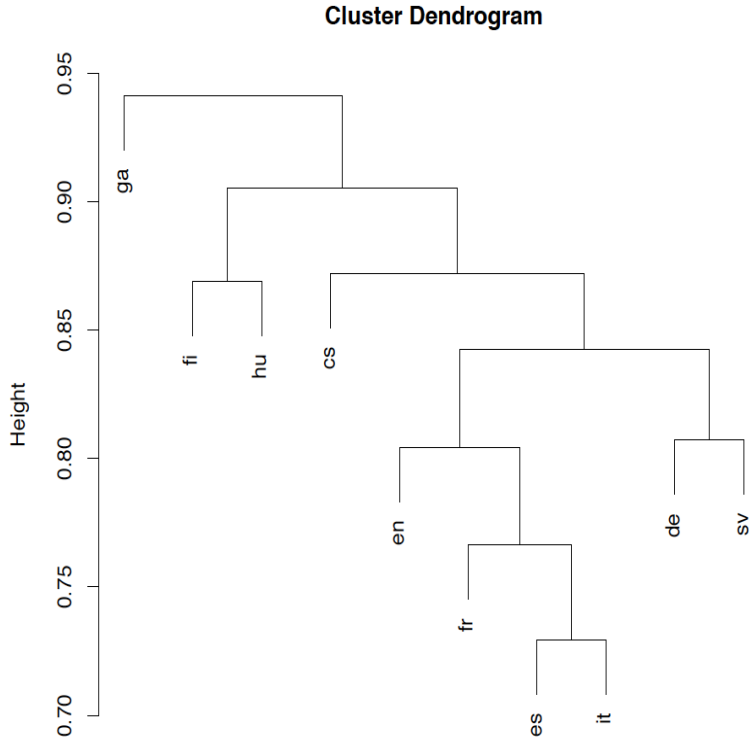
Nombre de propriétés communes entre l'italien et les autres langues :

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

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

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