Computational Typology: How to Compare Languages from Treebanks

Philippe Blache, Grégoire de Montcheuil, Stéphane Rauzy

Laboratoire Parole et Langage
CNRS & Aix-Marseille Université

Introduction

- Typology: language comparison
  - Structural and functional features for classifying languages
  - Quantitative typology: distribution of structural patterns in the world's languages
  - Resources:

THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE

Ethnologue
Languages of the World

- Computational Typology
  - Only few works
  - Some machine learning
  - Difficulty to rationalize feature identification and quantification

Language comparison: equi-complexity?

19^{th} century

No. Higher complexity = higher cultures

20^{th} century

Yes. Structuralists (anthropologists) and generative linguists (innateness, universal grammar).
The *equi-complexity hypothesis* [Hockett58]: *what is complex in one domain makes easier the others*.

21^{st} century

May be not. Language evolution

[McWhorter01] Creoles have the simplest grammars

(a) little or no inflectional morphology
(b) no lexical or morphosyntactic tone
(c) no non-transparent derivational morphology

During the pidgin phase all 'ornamental' marking has been lost, creoles are too young to have been able to develop complex features of older languages

$\Rightarrow$ Necessity to identify and quantify features

[Hockett58] Hocket C.F. (1958) *A Course in Modern Linguistics*, Macmillan
[McWhorter01] Mc Whorter J.H. (2001) "The world's simplest grammars are creole grammars", in *Linguistic Typology*

Structural Complexity: Language Comparison

[Parkvall08] Parkvall, M. (2008) *The simplicity of creoles in a cross-linguistic perspective*, in Miestamo and al. (eds), Benjamins

01	Size of consonant inventories
02	Size of vowel quality inventories
04	Complexity of syllable structure
12	Number of genders
13	Gender Distinctions in Independent Personal Pronouns
15	Inclusive/Exclusive Distinction in Independent Personal Pronouns
18	Politeness Distinctions in Second Person Pronouns
22	Occurrence of Nominal Plural Markers
28	Definite Articles
29	Indefinite Articles
33	Distance Contrasts in Demonstratives
36	Numeral Classifiers
58	Alignment of Case Marking of Full Noun Phrases
...	

Numerical values:

- Boolean features : 1/0 (0,5 for the case maybe)
- Numerical values : normalization to 0-1
- Features expressed in prose, subjective: simple, moderately complex, complex

Structural Complexity: Ranking Languages

[Parkvall08] Parkvall, M. (2008) *The simplicity of creoles in a cross-linguistic perspective*, in Miestamo and al. (eds), Benjamins

Language	R1	R2	Language	R1	R2
Hindi	9		Finnish	55	
Spanish	11		Guarani	65	
Basque	14		Lakhota	69	
French	16		Russian	73	
German	27		English	78	
Slave	30		Mandarin	89	
Georgian	32		Arabic Egyptian	94	
Nenets	39		Hungarian	97	
Japanese	46		Thai	132	
Armenian	48		Vietnamese	147	

R1: *Parkvall ranking*

Structural Complexity: Ranking Languages

[Nichols07] Nichols, Johanna (2007). "The distribution of complexity in the world's languages", in procs of the Annual Meeting of the LSA

Language	R1	R2	Language	R1	R2
Hindi	9	7	Finnish	55	17
Spanish	11	14	Guarani	65	10
Basque	14	18	Lakhota	69	1
French	16		Russian	73	6
German	27	12	English	78	
Slave	30	3	Mandarin	89	16
Georgian	32	4	Arabic Egyptian	94	2
Nenets	39	15	Hungarian	97	5
Japanese	46	11	Thai	132	8
Armenian	48	9	Vietnamese	147	13

R1: *Parkvall ranking*
R2: *Nichols ranking*

Summary

Goal

- Rationalizing language comparison
- Providing mechanisms for a numerical comparison

Method

- Identifying more general features
- Comparison directly from data, not from linguistic studies
- Evaluating feature weights

Outline

- Treebank browsing
- Syntactic features extraction
- Language comparison

Summary

Goal

- Rationalizing language comparison
- Providing mechanisms for a numerical comparison

Method

- Identifying more general features
- Comparison directly from data, not from linguistic studies
- Evaluating feature weights

Outline

- Treebank browsing
- Syntactic features extraction
- Language comparison

Summary

Goal

- Rationalizing language comparison
- Providing mechanisms for a numerical comparison

Method

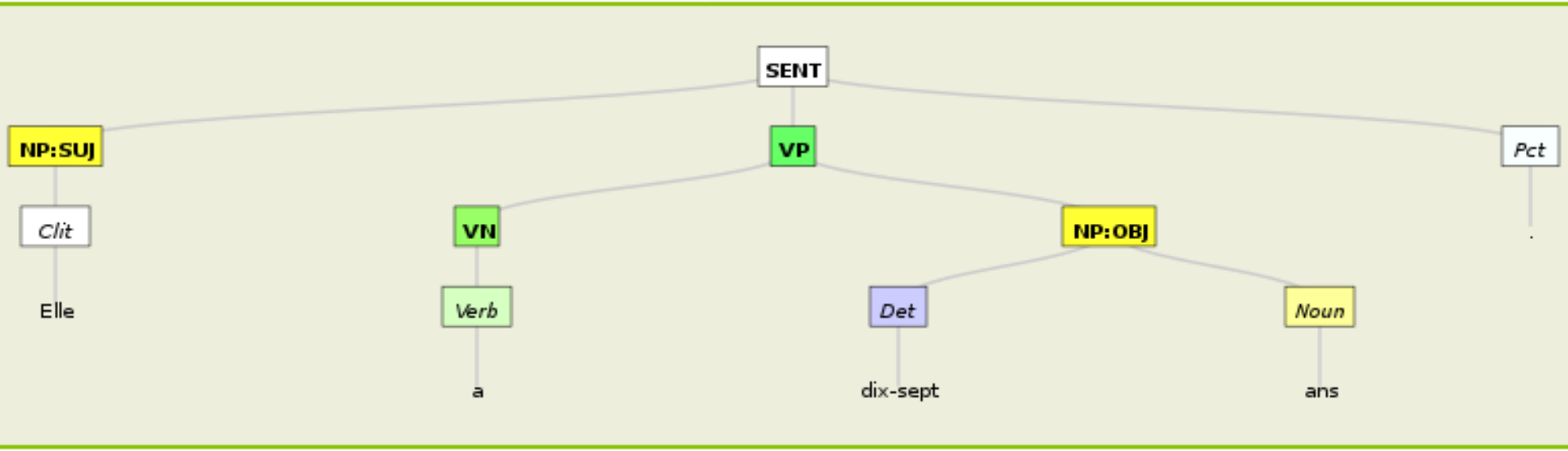
- Identifying more general features
- Comparison directly from data, not from linguistic studies
- Evaluating feature weights

Outline

- Treebank browsing
- Syntactic features extraction
- Language comparison

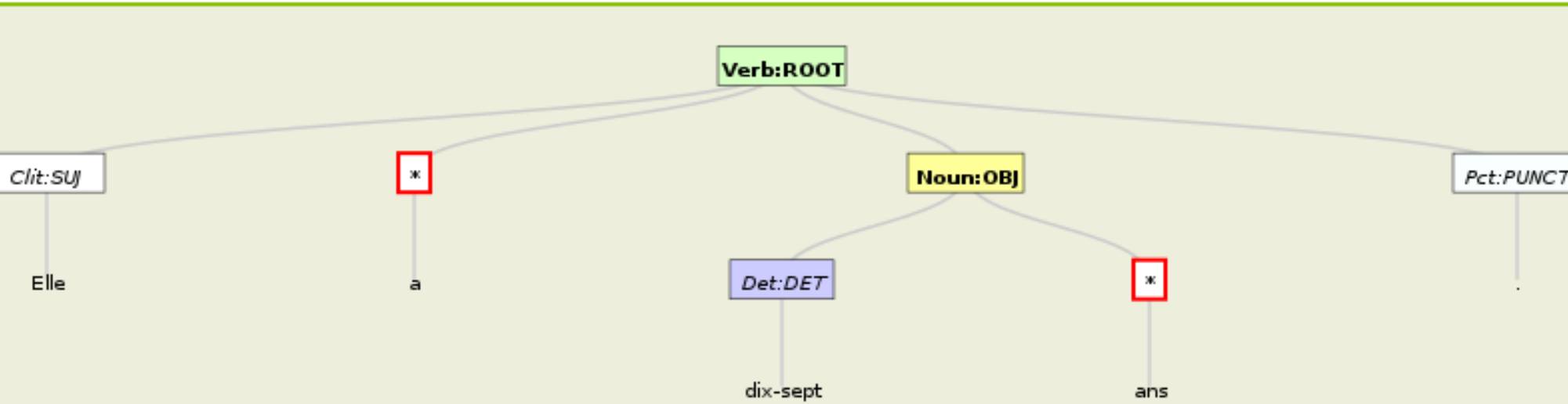
Treebanks

- Constituency



Elle a dix-sept ans .

- Dependency



Universal Dependencies Treebank

- 35 languages (10 languages considered in this work, version 1.0)

Amharic	–		–	–	–		
Ancient Greek	244K	LF					
Ancient Greek-PROIEL	206K	LF	–				
Arabic	282K	F	–				
Basque	121K	LF					
Bulgarian	156K	LF					
Catalan	–			–	–		
Croatian	87K	LF	–				
Czech	1,503K	LF					
Danish	100K	LF					
Dutch	200K	LF	–				
English	254K	LF					

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

- 17 *universal* POS tags
- 40 *universal* dependency relations

Core dependents of clausal predicates

Nominal dep	*Predicate dep*	
nsubj	csubj	
nsubjpass	csubjpass	
dobj	ccomp	xcomp
iobj		

Non-core dependents of clausal predicates

Nominal dep	*Predicate dep*	*Modifier word*
nmod	advcl	advmod
		neg

Noun dependents

Nominal dep	*Predicate dep*	*Modifier word*
nummod	acl	amod
appos		det
nmod		neg

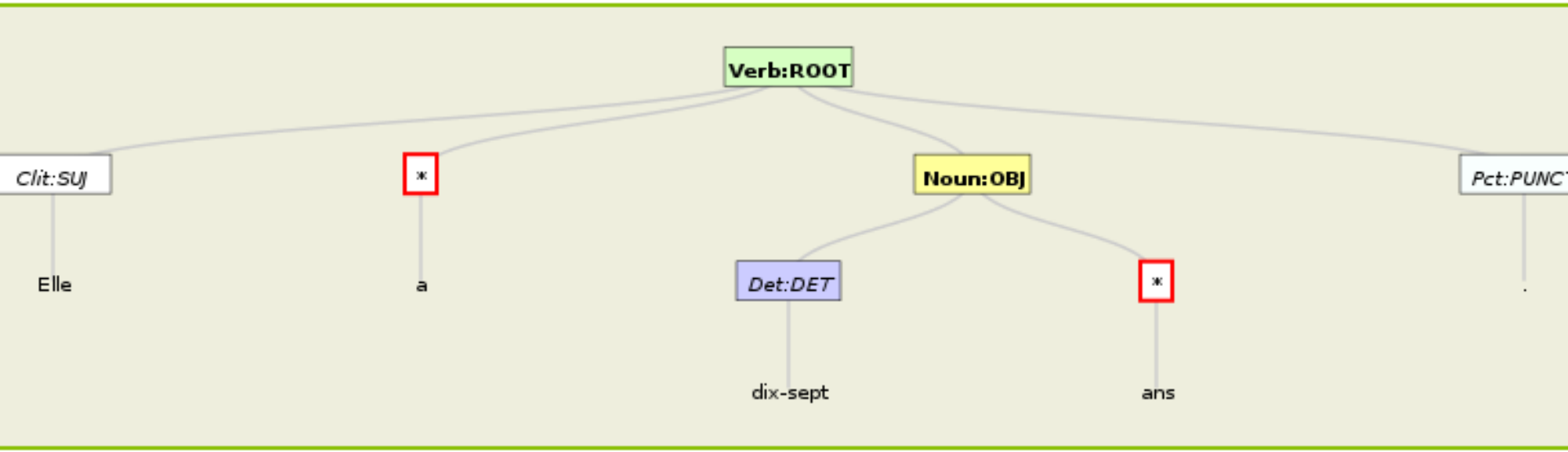
Compounding and unanalyzed

compound	mwe	goeswith
name	foreign	

Universal Dependencies Treebank

Family		Language		#Trees	#Tokens	Typological features
Indo-European	Slavic	cs	Czech	87,913	1,482,147	SVO, stress timing, free word-order
	Germanic	de	German	15,918	297,985	V2, SOV, inflected, accusative, stress timing, stress accent
		en	English	16,622	254,930	SVO, inflected, accusative, stress timing, stress accent
		sv	Swedish	6,026	96,699	SVO, inflected, accusative, stress timing, pitch accent
	Romance	es	Spanish	16,006	430,764	SVO, syllabic
		fr	French	16,418	398,964	SVO, inflected, accusative, syllable timing
		it	Italian	10,077	214,748	SVO, syllable timing
	Celtic	ga	Irish	1,020	23,686	VSO, inflected, accusative, stress timing, stress accent
Uralic	Finnic	fi	Finnish	13,581	181,022	SVO, stress timing
	Hungarian	hu	Hungarian	1,299	25,064	SVO, FWO, agglutinative, accusative

Extraction of the implicit CFG



Elle a dix-sept ans .

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

`Noun:OBJ` $\rightarrow$ `Det:DET` `Noun:OBJ`

General Syntactic Features

Linearity: the word order is always the same
precede(A,B) *(i.e. A always precedes B)*

Co-occurrence: the realization of a given category implies that of another one
require(A,B) *(i.e. if A is realized, then B has too)*

Exclusion: the realization of a given category excludes that of another one
exclude(A,B) *(i.e. if A is realized, then B cannot)*

Uniqueness: a given category is never realized more than once
uniqueness(A)

Feature Extraction

- PS rules extracted from the TB

  NOUN:nsubj → DET:det NOUN:nsubj

  NOUN:nsubj → DET:det NOUN:nsubj NOUN:nmod

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

  (...)

- Inferred feature

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

Feature Extraction

- Rules inferring a feature

NOUN:nsubj → DET:det NOUN:nsubj

NOUN:nsubj → DET:det NOUN:nsubj NOUN:nmod

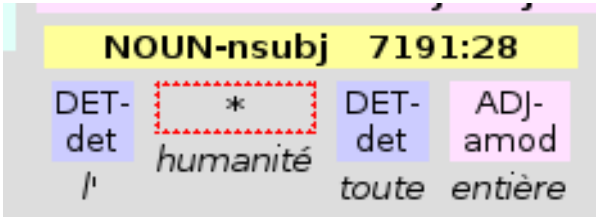
NOUN:nsubj → DET:det NOUN:nsubj ADJ:amod

↓

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

- Rule violating the inferred feature

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

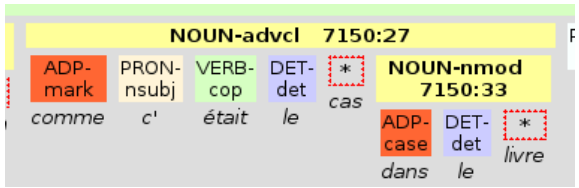


Hard vs. Soft Features

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

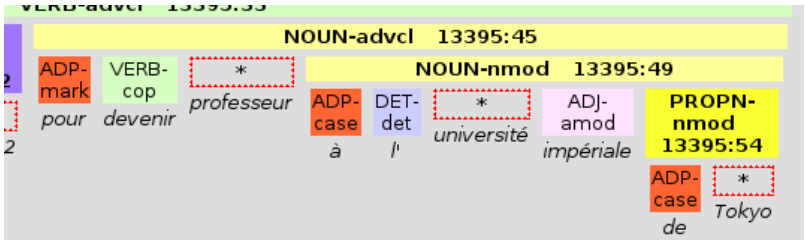
copula *nominal modifier*

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



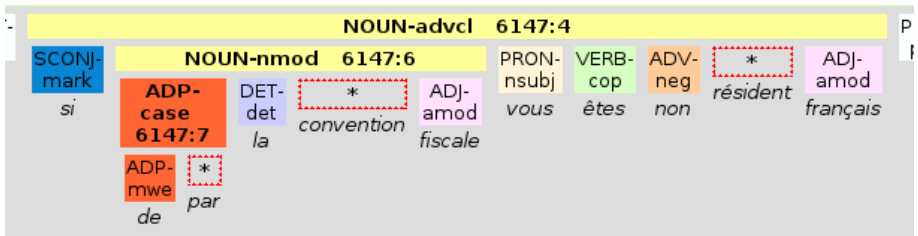
comme c'était le cas dans le livre

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



pour devenir professeur à l'université impériale de Tokyo

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



si de par la convention fiscale vous êtes non résident français

Feature Weights

- Ratio of the number of rules satisfying the feature to the total number of rules with this feature

$$w_0 = \frac{|\ (rules\_sat(f)\ |}{|\ rules\_sat(f)\ | + |\ rules\_viol(f)\ |}$$

  - $w_0 = 1$: feature always satisfied
  - $w_0 > 0.5$: feature more often satisfied than violated
  - $w_0 \geq \alpha/(\alpha + 1)$: α-*more often* satisfied than violated

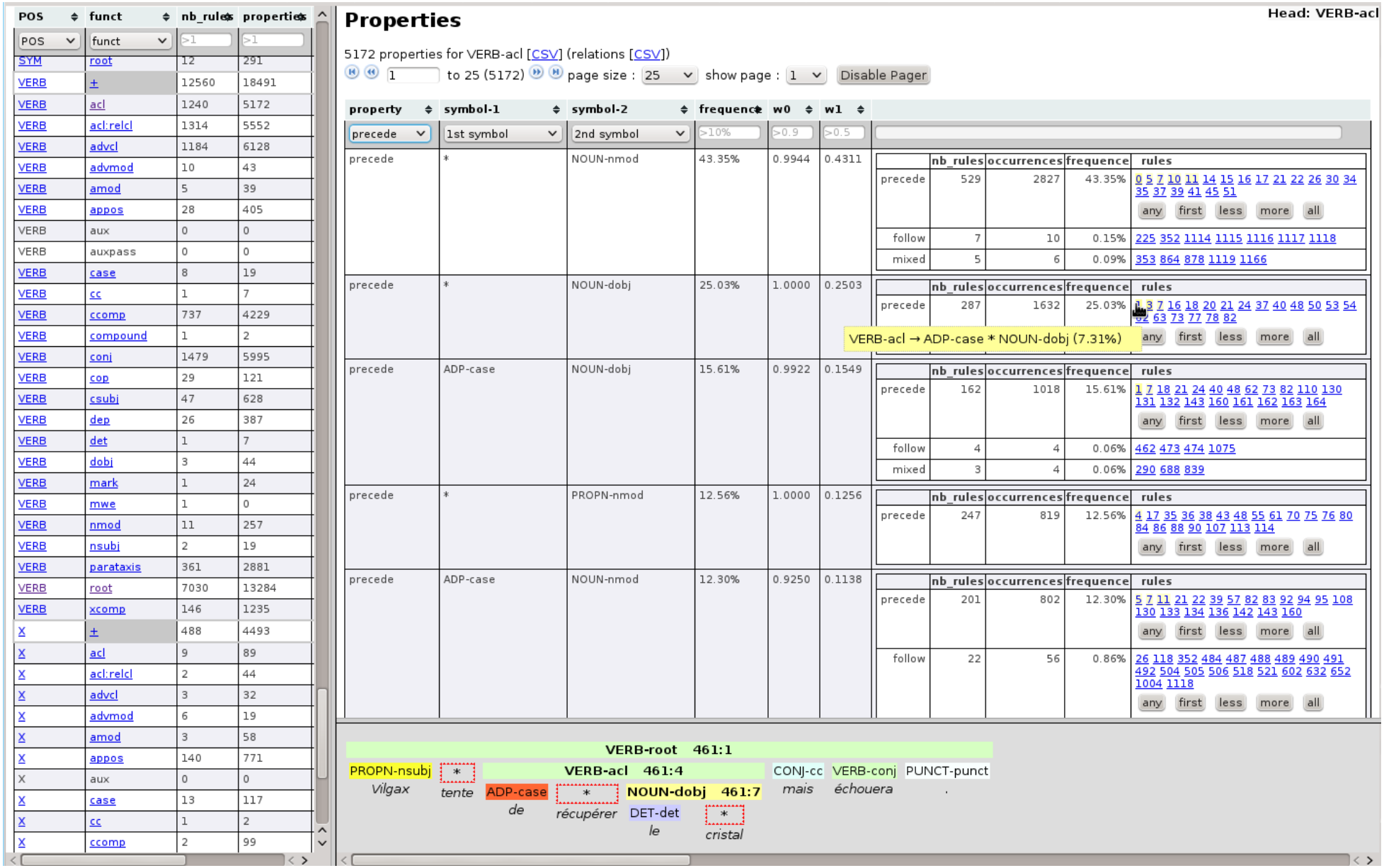
- Balanced with frequency of the rules statisfying the feature

$$w_1 = w_0 \frac{|\ rules\_sat(f)\ |}{\sigma_f}$$

  Where:

  - σ_f total number of rules with the same head as f

MarsaGram



POS	funct	nb_rules	properties
SYM	root	12	291
VERB	±	12560	18491
VERB	acl	1240	5172
VERB	acl:relcl	1314	5552
VERB	advcl	1184	6128
VERB	advmod	10	43
VERB	amod	5	39
VERB	appos	28	405
VERB	aux	0	0
VERB	auxpass	0	0
VERB	case	8	19
VERB	cc	1	7
VERB	ccomp	737	4229
VERB	compound	1	2
VERB	conj	1479	5995
VERB	cop	29	121
VERB	csubj	47	628
VERB	dep	26	387
VERB	det	1	7
VERB	dobj	3	44
VERB	mark	1	24
VERB	mwe	1	0
VERB	nmod	11	257
VERB	nsubj	2	19
VERB	parataxis	361	2881
VERB	root	7030	13284
VERB	xcomp	146	1235
X	±	488	4493
X	acl	9	89
X	acl:relcl	2	44
X	advcl	3	32
X	advmod	6	19
X	amod	3	58
X	appos	140	771
X	aux	0	0
X	case	13	117
X	cc	1	2
X	ccomp	2	99

Available at SLDR (hdl:11041/ortolang-000917)

Elements of comparison between languages

Size of the grammar

Number of features ($w_0 = 1$, excluding dependency relations)

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

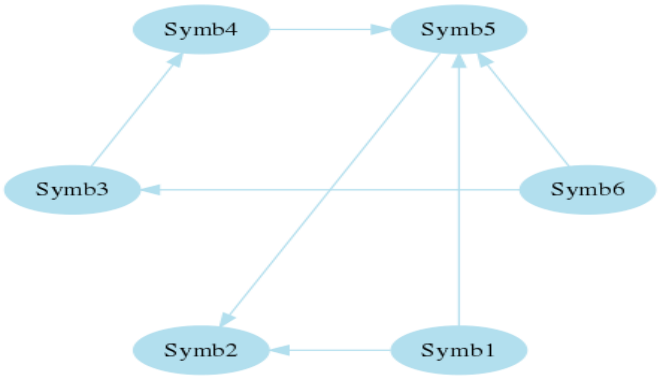
Free word order (Czech, Finnish, Hungarian) $\Rightarrow$ smaller grammar size

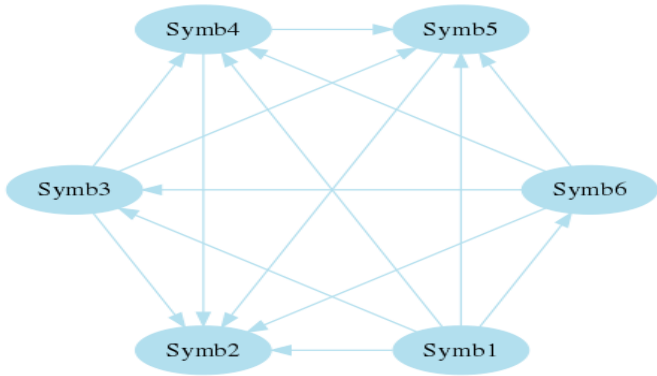
Focus: linear properties for Det in Czech and German

lang	head	prop_type	cat props	%props	occ	%(occ)	mean(w0)	mean(w1)
cs	N	precede	Det	6 15,38%	41151	7,34%	0,96	0,02
de	N	precede	Det	9 23,08%	73508	53,99%	0,96	0,16

Feature Graph: Representing Density

- *Nodes*: the arguments of the feature (e.g. dependents of a given head)
- *Edges*: relations between arguments (i.e. features such as linearity, cooccurrence, exclusion)
- *Feature graph*: given a construction (i.e. a head with its dependents) and a feature, the graph represents this feature relations between the constituents.





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

$$Density = 1$$

Density of the *linearity* feature in French and Finnish:

	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

Language classification

Common properties

Number of common properties between languages (here: Italian wrt other languages)

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

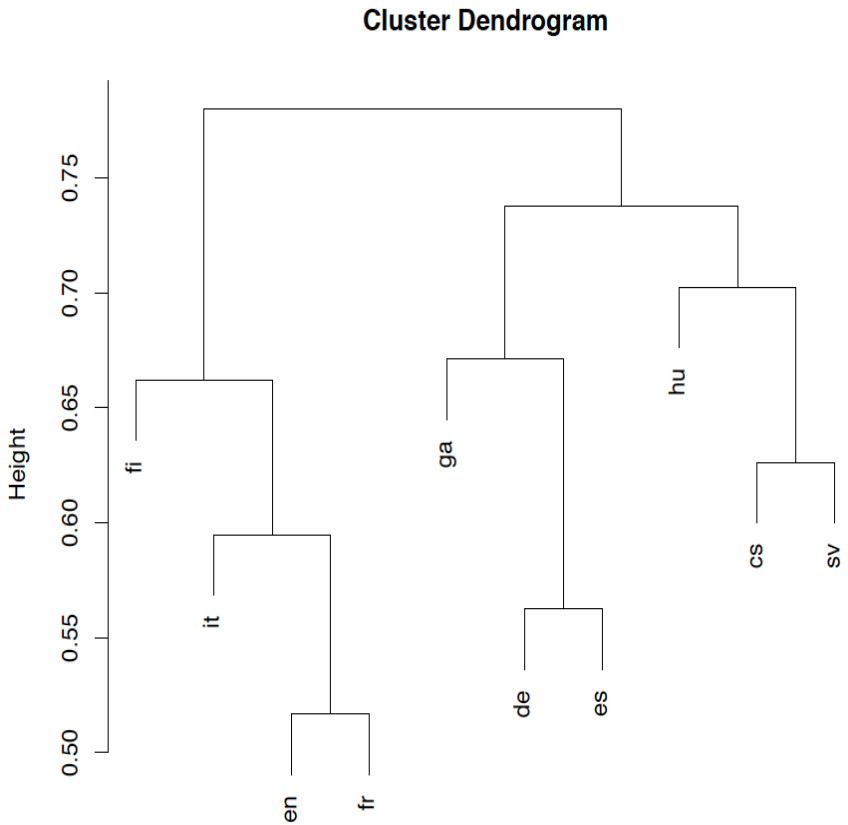
Language similarity

Ratio of common properties btw 2 languages to the total number of properties:

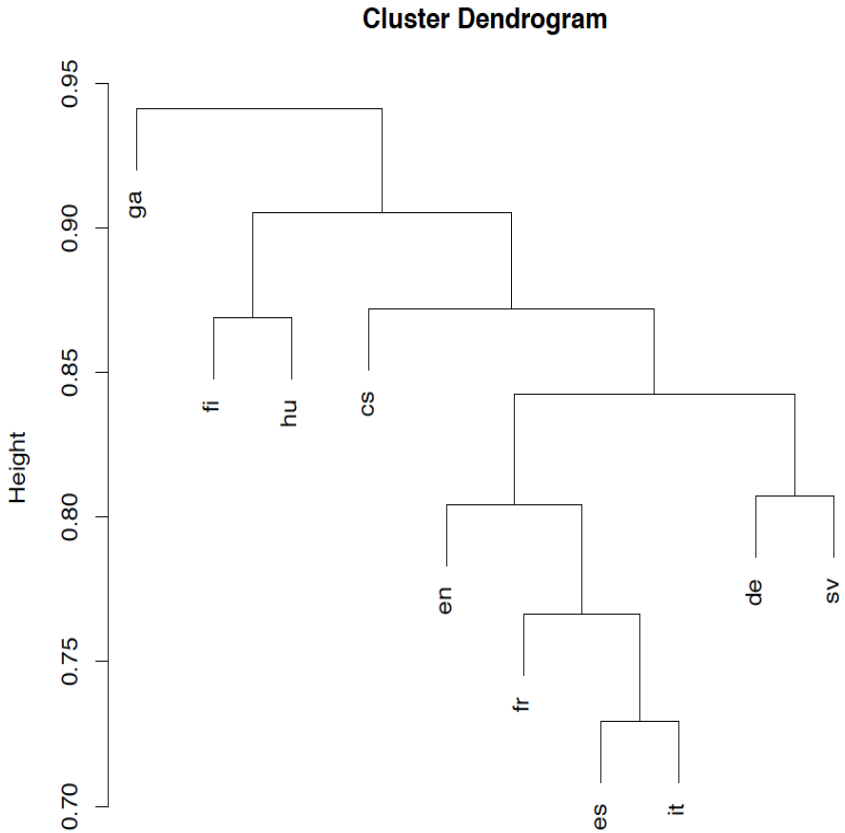
$$simil(l_1, l_2) = \frac{card(\mathcal{P}(l_1) \cap \mathcal{P}(l_2))}{card(\mathcal{P}(l_1) \cup \mathcal{P}(l_2))}$$

	#properties	cs	de	en	es	fi	fr	ga	hu	it
cs	1665									
de	1695	0,645								
en	2267	0,656	0,634							
es	2001	0,675	0,562	0,588						
fi	1299	0,724	0,756	0,655	0,701					
fr	2070	0,652	0,621	0,517	0,528	0,662				
ga	1228	0,716	0,649	0,723	0,671	0,780	0,673			
hu	969	0,702	0,707	0,759	0,738	0,732	0,741	0,726		
it	1679	0,687	0,669	0,595	0,581	0,596	0,575	0,696	0,717	
sv	1250	0,626	0,637	0,690	0,663	0,747	0,654	0,652	0,655	0,681

Hierarchical Classification



All properties



Linearity

Conclusion

- General typological features can be directly extracted from treebanks
- Different measures can be automatically calculated, rationalizing quantitative typology
- Possible extension to semantics, discourse thanks to enriched treebanks
- Complementary approach to statistical learning from annotated data