Comparing the syntax of different languages or usages: a Natural Language Processing perspective

Philippe Blache & Stéphane Rauzy

Laboratoire Parole et Langage
CNRS & Aix-Marseille Université
France

Question 1: variations within a same language

- Annotated corpora / statistical modeling

$$
\begin{aligned}
X\hat{\beta} = {} & 0.95 \\
& - 1.34\{\text{c}\} + 0.53\{\text{f}\} - 3.90\{\text{p}\} + 0.96\{\text{t}\} \\
\text{(a)} \quad & + 0.99\{\text{accessibility of recipient} = \text{nongiven}\} \\
\text{(a)} \quad & - 1.1\{\text{accessibility of theme} = \text{nongiven}\} \\
\text{(b)} \quad & + 1.2\{\text{pronominality of recipient} = \text{nonpronoun}\} \\
\text{(b)} \quad & - 1.2\{\text{pronominality of theme} = \text{nonpronoun}\} \\
\text{(c)} \quad & + 0.85\{\text{definiteness of recipient} = \text{indefinite}\} \\
\text{(c)} \quad & - 1.4\{\text{definiteness of theme} = \text{indefinite}\} \\
\text{(d)} \quad & + 2.5\{\text{animacy of recipient} = \text{inanimate}\} \\
& + 0.48\{\text{person of recipient} = \text{nonlocal}\} \\
& - 0.03\{\text{number of recipient} = \text{plural}\} \\
& + 0.5\{\text{number of theme} = \text{plural}\} \\
& - 0.46\{\text{concreteness of theme} = \text{nonconcrete}\} \\
\text{(e)} \quad & - 1.1\{\text{parallelism} = 1\} - 1.2 \cdot \text{length difference (log scale)}
\end{aligned}
$$

Problem: modeling does not form a system (not a grammar)

Question 2: variations between languages

- Parallel corpora / quantification
- Treebanks /tagset, rules distribution
- Grammars / evaluation

	CT Grammar	ATIS Grammar	PT Grammar
Rules	24,456	4,592	15,039
Nonterminals	3,946	192	38
Terminals	1,032	357	47
# Test Sentences	162	98	30
Average Length	8.3	11.4	5.7
# Grammatical	150	70	30
Average # Parses	5.4	940	7.2×10^{27}

Problem: we learn few from such comparisons

Example: PP in French / Chinese

French	
Rule	*Frequency*
NP → D- Nc	5 144
NP → Nc	4 517
NP → D- Nc PP	2 907
NP → Np	1 832
NP → Nc PP	1 214
NP → Ppn	1 166
NP → D- Nc AP	1 138
NP → D- Np	664
NP → Nk	623
NP → NP Cc NP	601
NP → Nc AP	596
NP → D- AP Nc	474

Chinese	
Rule	*Frequency*
NP → NN	165 306
NP → NR	47 817
NP → NN NN	34 355
NP → PN	30 560
NP → NP NP	27 874
NP → DNP NP	22 967
NP → CP NP	14 749
NP → NT	12 510
NP → QP NP	11 710
NP → DP NP	11 008
NP → ADJP NP	10 788
NP → NN NN NN	4 814

Example: syntactic alternation

- In French: the Adj/N order

Prenominal:	NP $\rightarrow$ Det N AP
Postonominal:	NP $\rightarrow$ Det AP N

- In Chinese: dative constructions

The *gei* object construction:	VP $\rightarrow$ V NP gei NP
The *Vgei* double object construction:	VP $\rightarrow$ Vgei NP NP
The double object construction:	VP $\rightarrow$ V NP NP

Problem

- Now: what is it possible to do?
  - We can observe variations in corpora
  - We can describe variations from treebanks (rules distribution)
  - We can predict variations from data analysis (statistical modeling)
- Limits
  - Frequencies makes it possible to predict, but ...
  - ... they cannot explain
- Where does the problem come from?
  - Separate representation grammars vs. parameters

Solution

- Inclusive system, integrating grammars and parameters, implementing the notion of "importance" of the syntactic relations
- Needs:
  - Formalism for representing both information
  - Resources for acquiring the parameters
- One solution
  - Representing information in terms of constraints instead of rules
  - Developing new resources such as constraint treebanks

Outline

- Context: the existing *treebanks*
  - The Chinese Treebank, the French Trebank: main features
  - Extracting the implicit phrase-structure grammar from the treebanks
- The Chinese / French constraint-based grammar
  - The constraint-based formalism: *Property Grammars*
  - Generating the constraint grammars from the PSG
- Perspectives

Part I

Context

The source treebanks: CTB, FTB

- The Chinese Treebank:
  - 51,447 sentences
  - 1,196,329 words; 1,931,381 hanzi (Chinese characters)
  - Annnotation has Penn Treebank-style labeled brackets
- The French Treebank:
  - 20,648 sentences, comprising
  - 580,945 words
  - Annotations follow a specific constituency-based labelling

The tagsets

Chinese	
ADJP	adjective phrase
ADVP	adverbial phrase
VP	verb phrase
NP	noun phrase
PP	preposition phrase
QP	quantifier phrase
DP	determiner phrase
FRAG	fragment
PRN	parenthetical
UCP	unidentical coordination phrase
IP	simple clause headed by I (INFL)
CP	clause headed by C
CLP	classifier phrase
DNP	phrase formed by "XP + DEG"
DVP	phrase formed by "XP + DEV"
LCP	phrase formed by "XP + LC"
LST	list marker

French	
AP	Adjectival phrase
VPinf	Infinitival phrase
AdP	Adverbial phrase
Srel	Relative clause
COORD	Coordinated phrase
Ssub	Subordinated clause
NP	Noun phrase
Sint	Internal, inflected sentence
PP	Prepositional phrase
VN	Verb kernel
VPpart	Participial phrase
SENT	Independent sentence

The *French Treebank*: example

```
<SENT>
     <PP fct="MOD">Au <NP>debut</NP></PP>,
     <VN fct="SUJ">on ramassait</VN>
     <VPinf fct="OBJ">
          <PP fct="DE-OBJ">de <NP>quoi</NP></PP>
          <VN>remplir</VN>
          <NP fct="OBJ">quinze sacs_poubelle</NP>
     </VPinf>,
     <Sint>
          <VN>indique</VN>
          <NP fct="SUJ">Roger,
               <NP>ouvrier <PP>a<NP>la regie</NP></PP></NP></NP>
     </Sint>.
</SENT>
```

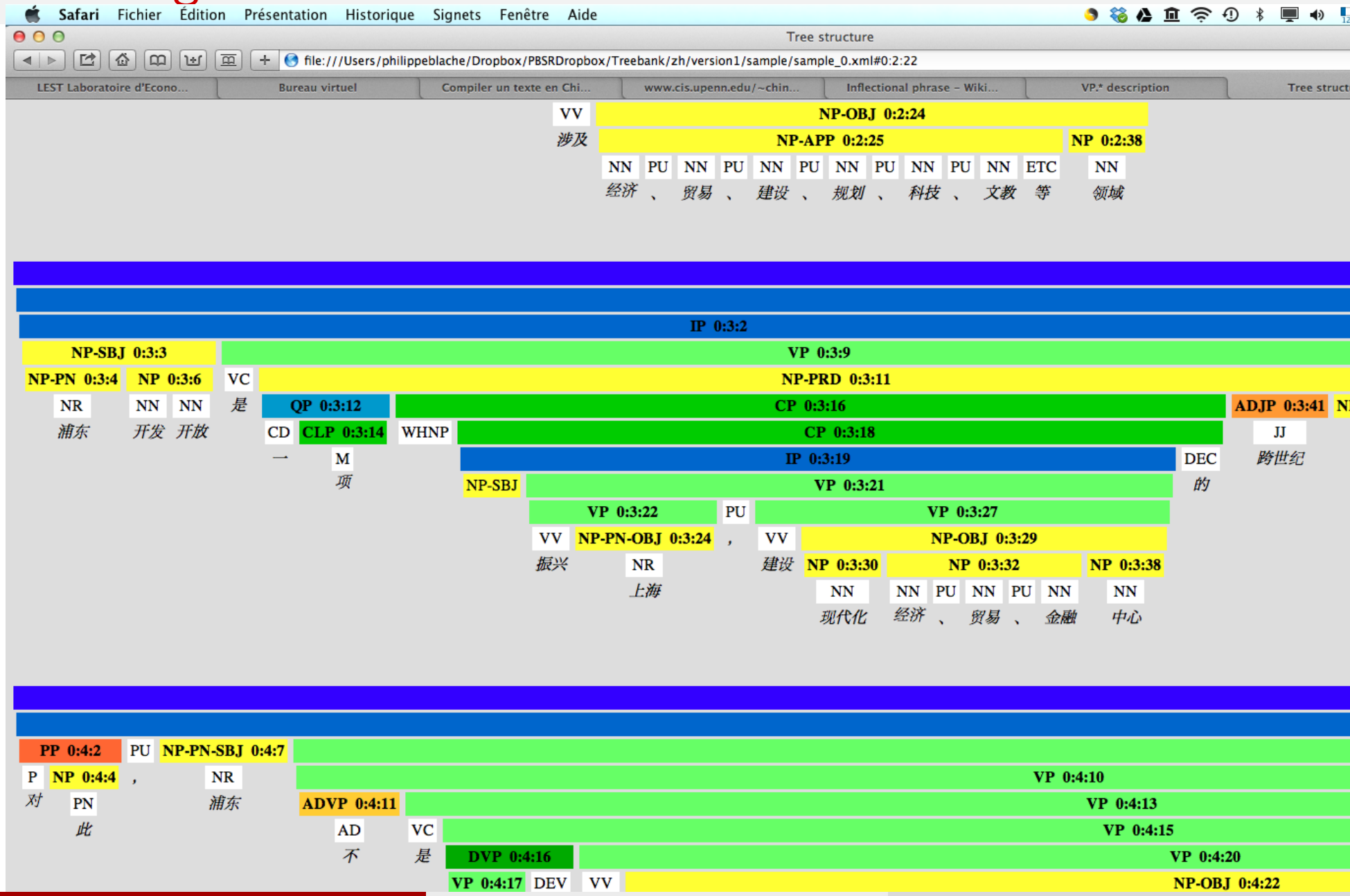
The *Chinese Treebank*: example

```
( (IP    (NP-SBJ (NN 外商)
              (NN 投资)
              (NN 企业))
         (VP  (PP-TMP (P 在)
              (LCP (IP (NP-SBJ (-NONE- *PRO*))
                   (VP (VV 改善)
                       (NP-OBJ (NP-PN (NR 中国))
                          (NP (NN 出口)
                          (NN 商品)
                          (NN 结构)))))
                   (LC 中)))
         (VP (VV 发挥)
              (AS 了)
              (NP-OBJ (ADJP (JJ 显著))
                   (NP (NN 作用)))))
          (PU 。)) )
```

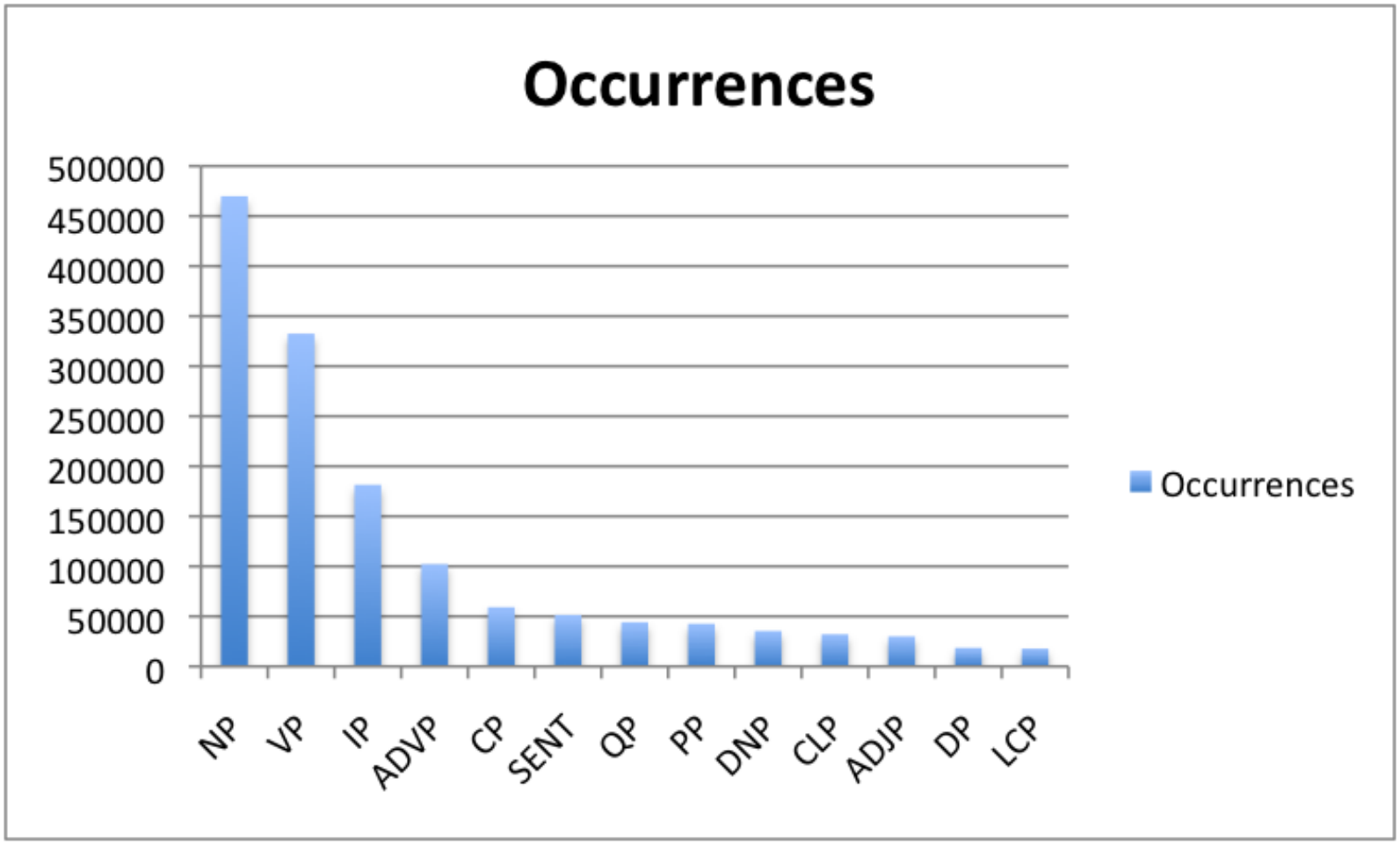
Speech: disfluencies

```
( (IP    (ADVP (AD 因此))
         (NP-SBJ (QP (CD 许多))
              (NP (NN 伊拉克人)))
         (VP (VV 认为)
         (PU ， )
         (IP-OBJ (NP-SBJ (DP (DT 这)
                         (CLP (M 份)))
                      (NP (NN 报告)))
            (VP  (ADVP (AD 并))
                 (VP (VE 没有)
                 (NP-OBJ (DNP (ADJP (ADVP (AD 多))
                                    (ADJP (JJ 大)))
                              (DEG 的))
                         (ADJP (JJ 实际))
                         (DFL (NP (PU <)
                            (NN b a c k g r o u n d))
                            (PU >))
                          (NP (NN 意义)))))))
         (PU 。)))
```

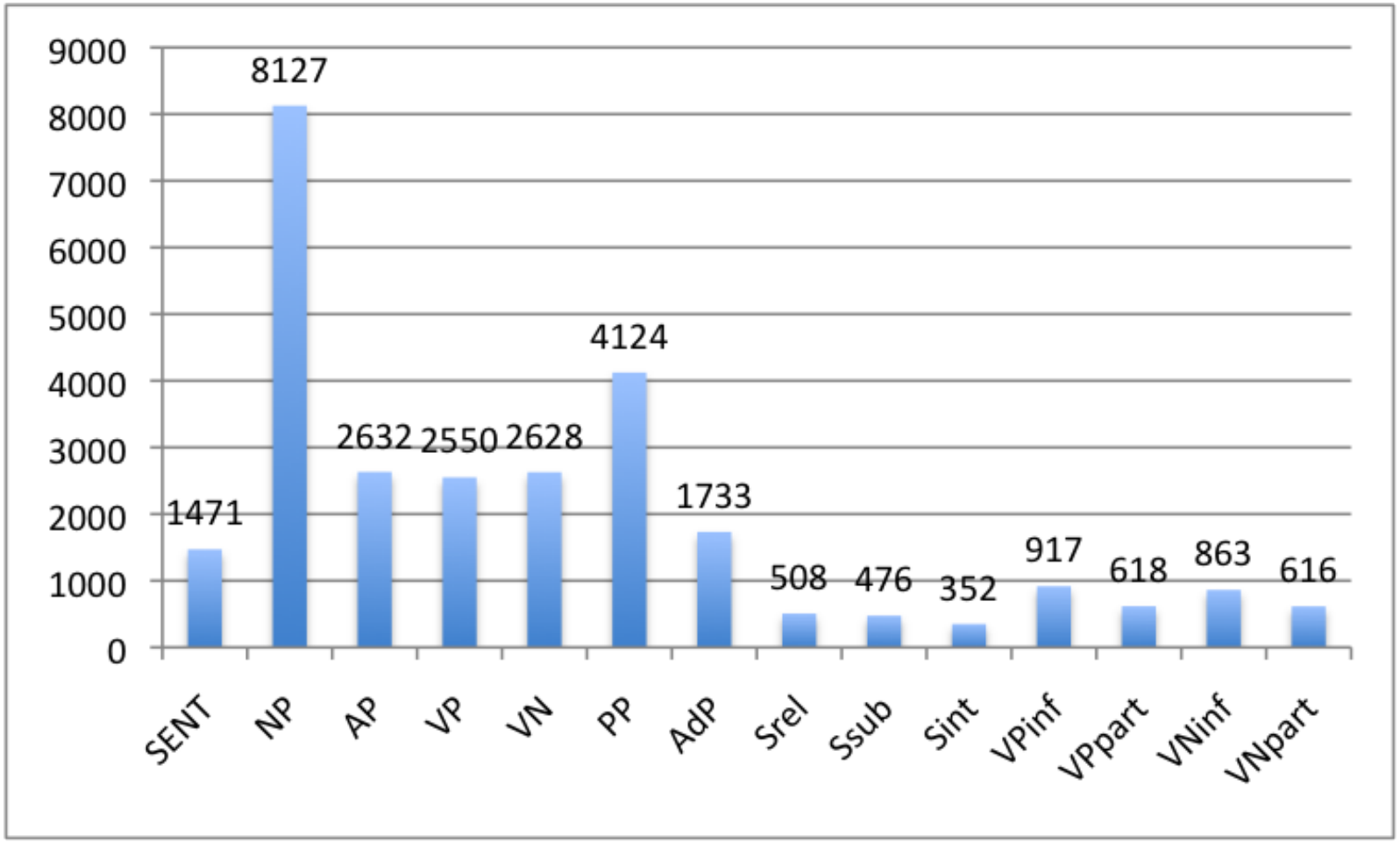
Browsing


Safari Fichier Édition Présentation Historique Signets Fenêtre Aide
Tree structure
file:///Users/philippeblache/Dropbox/PBSRDropbox/Treebank/zh/version1/sample/sample_0.xml#0:2:22
LEST Laboratoire d'Econo...
Bureau virtuel
Compiler un texte en Chi...
www.cis.upenn.edu/~chin...
Inflectional phrase - Wiki...
VP.* description
Tree struct
VV 涉及
NP-OBJ 0:2:24
NP-APP 0:2:25
NP 0:2:38
NN 经济 PU 、 NN 贸易 PU 、 NN 建设 PU 、 NN 规划 PU 、 NN 科技 PU 、 NN 文教 ETC 等 NN 领域
IP 0:3:2
NP-SBJ 0:3:3
VP 0:3:9
NP-PN 0:3:4
NP 0:3:6
VC 是
NP-PRD 0:3:11
NR 浦东
NN 开发 NN 开放
QP 0:3:12
CP 0:3:16
ADJP 0:3:41
CD 一
CLP 0:3:14
WHNP
CP 0:3:18
JJ 跨世纪
M 项
IP 0:3:19
DEC 的
NP-SBJ
VP 0:3:21
VP 0:3:22
PU ，
VP 0:3:27
VV 振兴
NP-PN-OBJ 0:3:24
VV 建设
NP-OBJ 0:3:29
NR 上海
NP 0:3:30
NP 0:3:32
NP 0:3:38
NN 现代化 NN 经济 PU 、 NN 贸易 PU 、 NN 金融 NN 中心
PP 0:4:2
PU ，
NP-PN-SBJ 0:4:7
P 对
NP 0:4:4
NR 浦东
PN 此
VP 0:4:10
ADVP 0:4:11
VP 0:4:13
AD 不
VC 是
VP 0:4:15
DVP 0:4:16
VP 0:4:20
VP 0:4:17
DEV
VV
NP-OBJ 0:4:22

Category distribution in the treebank



Category distribution for French (FTB)



Extracting the underlying Chinese CFG

Example: VP (332 516 occurrences)

Rk	Constituents	Occurrences
0	VV NP-OBJ	53 745
1	ADVP VP	45 358
2	VV	33 614
3	VV VP	15 972
4	VA	14 800
5	VV IP-OBJ	12 898
6	VP PU VP	10 754
7	VC NP-PRD	8 489
8	VV AS NP-OBJ	8 471
9	ADVP ADVP VP	7 219
10	VP VP	6 699

Rk	Constituents	Occurrences
11	VV PU IP-OBJ	5 375
12	VE NP-OBJ	5 203
13	VV NP-OBJ IP	4 359
14	VC VP	3 832
15	VV NP-PN-OBJ	3 464
16	PP-LOC VP	3 255
17	MSP VP	3 057
18	PP-DIR VP	2 985
19	NP-TMP VP	2 596
20	PP-MNR VP	2 594
21	VV QP-OBJ	2 302

The underlying CFG

IP → NP-SBJ VP	93 654	ADJP → JJ	28 136
IP → NP-SBJ VP PU	12 084	ADJP → ADVP ADJP	1 179
IP → S PU S PU	7 036		
IP → NP-PN-SBJ VP	6 056	ADVP → AD	98 773
IP → ADVP NP-SBJ VP	3 580	ADVP → CS	2 643
NP → NN	165 306	PP → P NP	22 771
NP → NR	47 817	PP → P LCP	8 499
NP → NN NN	34 355	PP → P IP	4 487
NP → PN	30 560	PP → P NP-PN	4 283
NP → DNP NP	22 224		
VP → VV NP-OBJ	53 745	QP → CD CLP	23 247
VP → ADVP VP	45 358	QP → CD	11 505
VP → VV	33 614	QP → OD CLP	2 101
VP → VV VP	15 972	QP → ADVP QP	1 598
VP → VA	14 800	QP → OD	1152

The underlying CFG (2)

CLP → M	31 872	DVP → VP DEV	1 509
DNP → NP DEG	19 654	LCP → NP LC	11 232
DNP → NP-PN DEG	4 557	LCP → IP LC	3 888
DNP → ADJP DEG	4 369	LCP → QP LC	1 509
DNP → PP DEG	2 277		
DNP → QP DEG	1 742	VRD → VV VV	3 356
DP → DT	10 113	VSB → VV VV	1 310
DP → DT CLP	55 80		
DP → DT QP	2 359	FLR → SP	2 076
		FLR → IJ	1 819

The underlying CFG: size of the grammar

Category	CF rules
IP	4 785
NP	3 593
VP	2 765
DFL	383
FRAG	328
UCP	301
QP	283
PP	224
CP	216
FLR	213
PRN	125
DNP	113
LCP	110
ADVP	81
ADJP	73

Category	CF rules
DP	55
INC	42
VRD	39
DVP	32
CLP	29
VCD	27
VSB	22
VNV	20
INTJ	19
LST	16
VPT	16
VCP	9
WHPP	4
WHNP	2

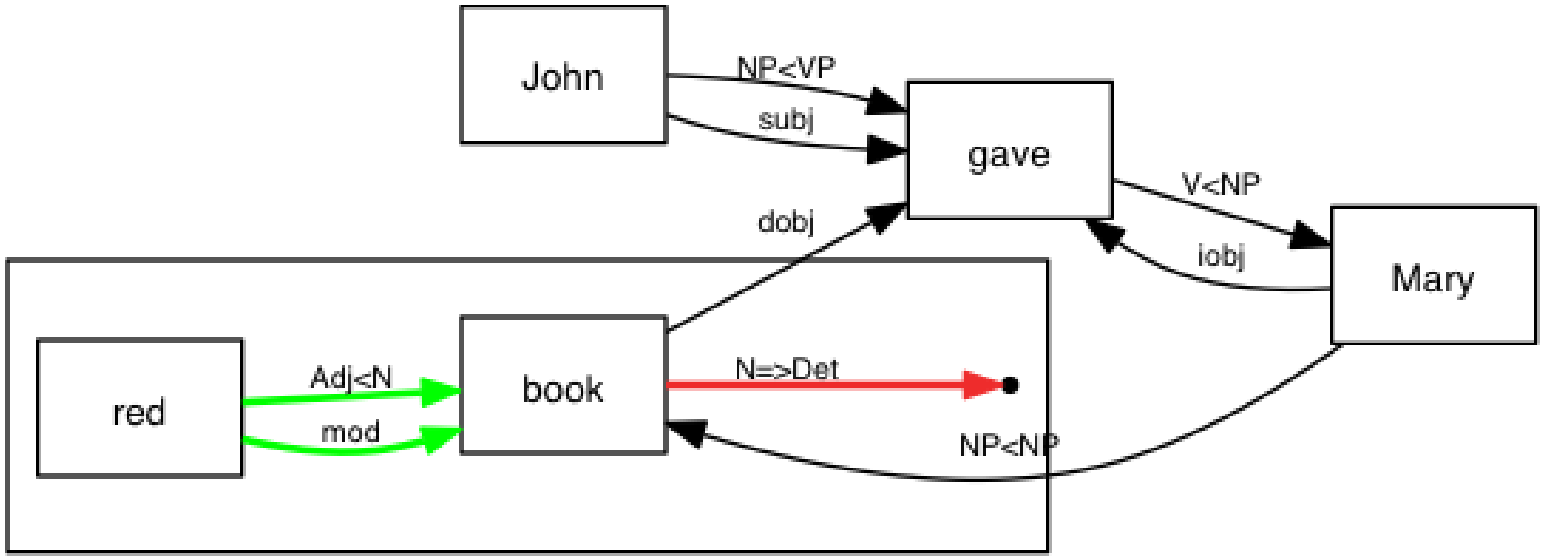
Part II

Representing information with constraints

Property Grammars in a nutshell

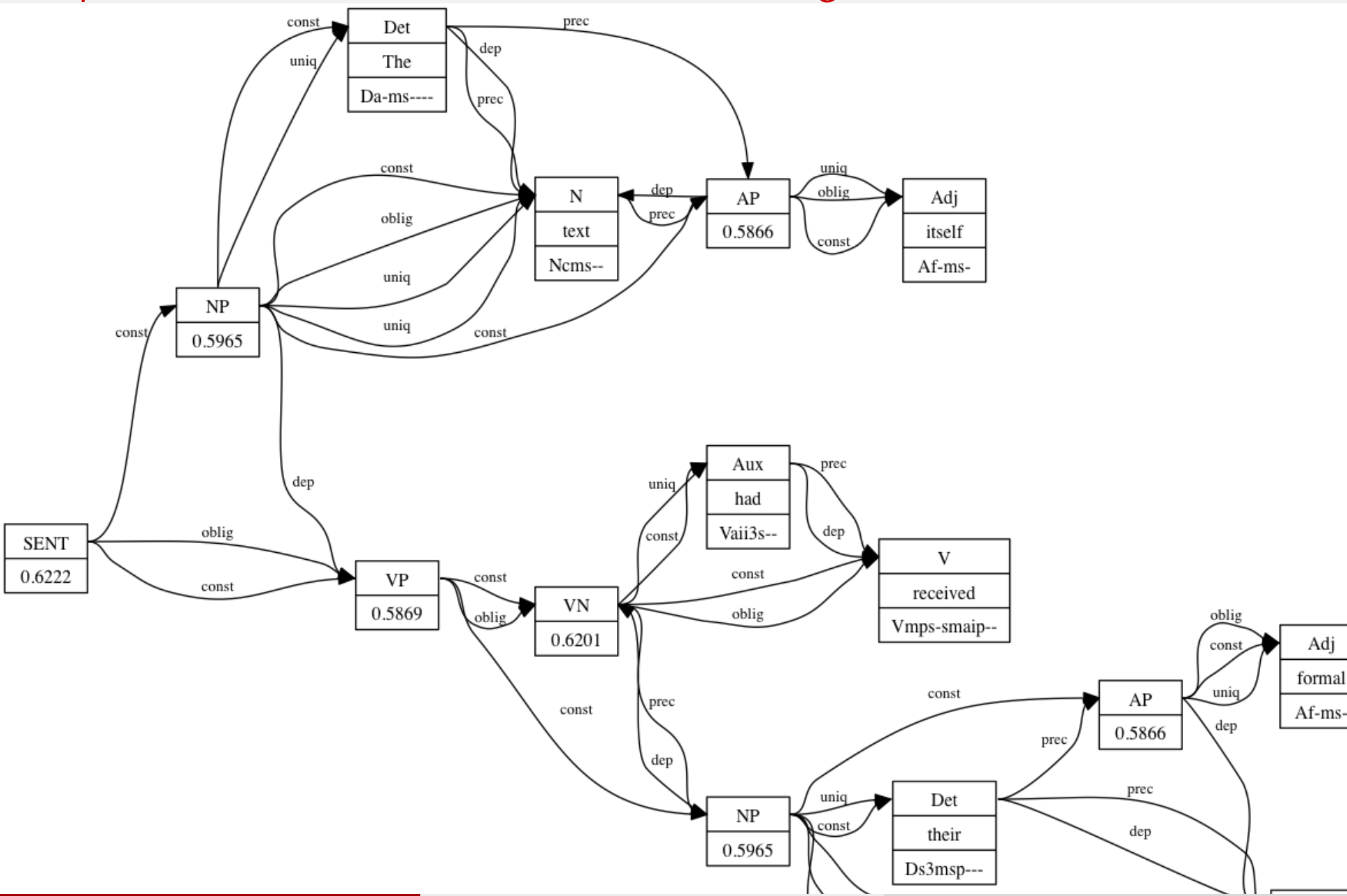
Constraint	*Description (NP in French)*
Linearity	Linear precedence relations $Det \prec N$; $Det \prec AP$; $N \prec PP$; $N \prec Sup$; $N^* \prec N$; $N \prec SRel$
Requirement	Mandatory cooccurrence between categories $N[com] \Rightarrow Det$, $Compl \Rightarrow N$
Exclusion	Cooccurrence restriction between categories $AP \not\leftrightarrow Sup$; $N \not\leftrightarrow Pro$; $Pro \not\leftrightarrow AP$; $AP[card] \not\leftrightarrow Det[ind]$
Dependency	Dependency relations $Det \rightsquigarrow N$; $AP \rightsquigarrow N$; $Sup \rightsquigarrow N$; $PP \rightsquigarrow N$; $SRel \rightsquigarrow N$
Obligation	Possible heads of a phrase $Oblig = \{N, AP, Pro\}$
Uniqueness	Constituents that cannot be repeated in a phrase $Uniq = \{Det, Sup, AP[card], AP[\neg\ card], PP, Pro, SRel\}$

Constraint relaxation



"John gave Mary red book."

Example: "*The text itself had received their formal agreement*"



The CFG rules for AP in the FTB

Constituents	Occ.	Constituents	Occ.
Af	1930	A- Ssub	1
A-	302	AdP Af Wm PP Wm PP	1
AdP Af	159	AdP Af Wm NP Wm Ssub	1
Af PP	63	Af Wm Ssub Wm PP	1
Af VPinf	19	AdP Wm AdP Af PP	1
AP Af	17	Af AdP	1
AdP Af Ssub	13	AdP Af AdP	1
AdP Af PP	8	AP Wm Cc AP Wm	1
A- PP	7	NP Af	1
Af Ssub	6	PP Af	1
AP A-	5	Af NP	1
AdP A-	4	AP AdP	1
AdP Af VPinf	3	AdP Wq Af Wq	1
Af PP:COORD	3	A- Wm A-	1
Af NP:COORD	2	AdP Af NP	1
AdP AdP Af	2	AdP Af NP PP VPinf	1
Af PP PP	2	Af Wm NP Wm PP	1

The PG constraints for AP in French

Constituency	$\{$AdP, A, VPinf, PP, Ssub, AP, NP$\}$
Linearity	A $\prec \{$VPinf, Ssub, PP, NP, AP$\}$ AdP $\prec \{$A, Ssub, PP$\}$ AP $\prec \{$A, AdP$\}$ PP $\prec \{$Ssub$\}$
Dependency	$\{$AdP, VPinf, PP, Ssub, NP$\} \rightsquigarrow$ A
Uniqueness	$\{$A, VPinf, Ssub$\}$
Obligation	$\{$A$\}$
Exclusion	VPinf $\otimes \{$PP, Ssub$\}$

Example: PP realizations in the CTB

PP → P NP	27 215	PP → P PU NP	71
PP → P LCP	8 518	PP → P CP	67
PP → P IP	4 517	PP → PP PU PP	59
PP → -NONE-	1 575	PP → P PU IP	46
PP → ADVP PP	534	PP → NP	45
PP → P QP	398	PP → P IP PU	36
PP → PP PP	167	PP → VV NP	25
PP → P DP	139	PP → P FLR LCP	21
PP → P FLR NP	87	PP → P FLR IP	20
PP → P UCP	80	PP → P PU LCP	20

Constraints describing the PP in Chinese

Constituency	{P, NP, LCP, IP, ADVP, PP, QP, DP, FLR, UCP, PU, CP, VV}
Linearity	P $\prec$ * ADVP $\prec$ PP FLR $\prec$ NP, LCP, IP PU $\prec$ NP, PP, LCP VV $\prec$ NP
Uniqueness	{NP, IP, QP, LCP, ADVP, DP}
Requirement	ADVP $\Rightarrow$ PP VV $\Rightarrow$ NP
Exclusion	NP $\otimes$ *{P FLR, PU, VV} LCP $\otimes$ *{P, FLR} IP $\otimes$ *{PU, FLR} {QP, DP, UCP, CP} $\otimes$ *{P} FLR $\otimes$ *{P, NP, LCP, IP, PU} PU $\otimes$ *{P, NP, PP, IP, LCP} VV $\otimes$ *{NP}

Distribution of the constraints in the CTB-grammar

	Linearity	Obligation	Uniqueness	Exclusion	Requirement	TOTAL
IP	39	0	32	52	19	142
NP	47	36	22	55	12	172
VP	53	0	29	58	6	146
DFL	31	39	25	46	5	146
FRAG	35	17	23	42	20	137
UCP	25	0	16	30	13	84
QP	27	0	24	37	6	94
PP	24	0	24	35	16	99
CP	22	0	26	35	11	94
FLR	18	35	23	37	0	113
PRN	16	16	22	30	11	95
DNP	28	0	26	31	7	92
LCP	22	0	30	34	12	98
ADVP	14	0	32	37	3	86
ADJP	15	19	22	29	8	93
DP	20	13	23	28	10	94

Identifying constraint weights from the grammar

- Relative importance of the constraints, according to their distribution
- Indices:
  - Number of rules in the grammar involved by the constraint
  - Cumulative frequency of these rules
- Example for the PP:

Type	Constraint	#Rules	Frequency
Linearity	P $\prec$ NP	26 rules	27 433
	P $\prec$ PRN	1 rule	2
Uniqueness	QP	5 rules	410
	DNP	2 rules	7

Some figures (in the FTB)

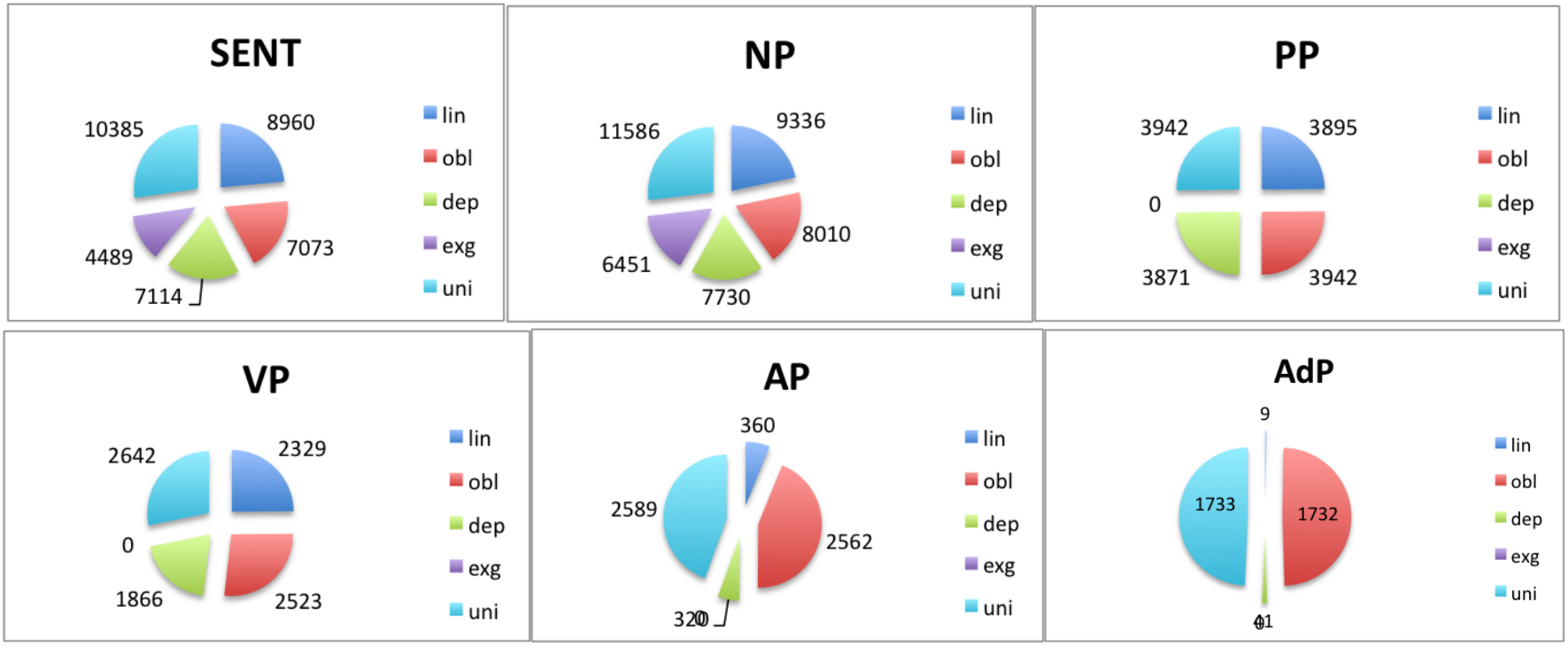
SENT	1 471
NP	8 127
AP	2 632
VP	2 550
VN	2 628
PP	4 124
AdP	1 733
Srel	508
Ssub	476
Sint	352
VPinf	917
VPpart	618
VNinf	863
VNpart	616

Number of constraints per category

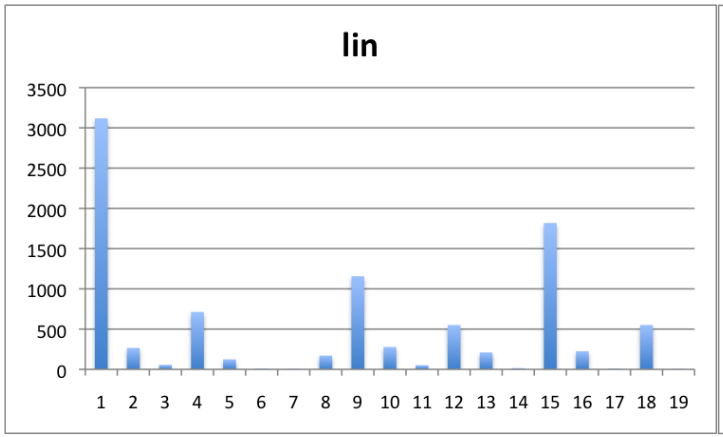
Lin	27 367
Obl	32 602
Dep	21 971
Exc	89 293
Req	11 022
Uni	38 007

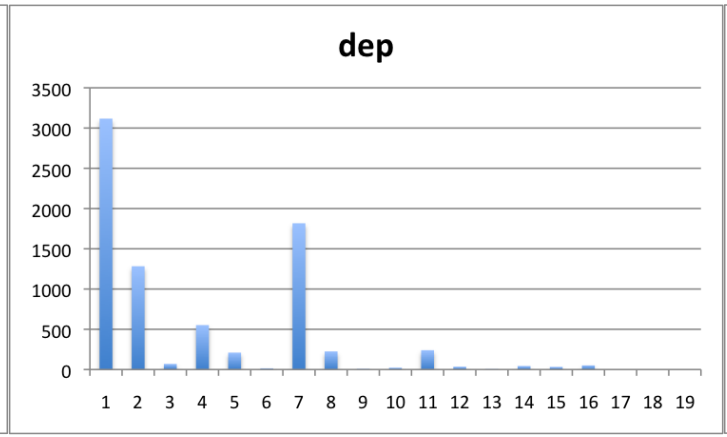
Number of constraints per type

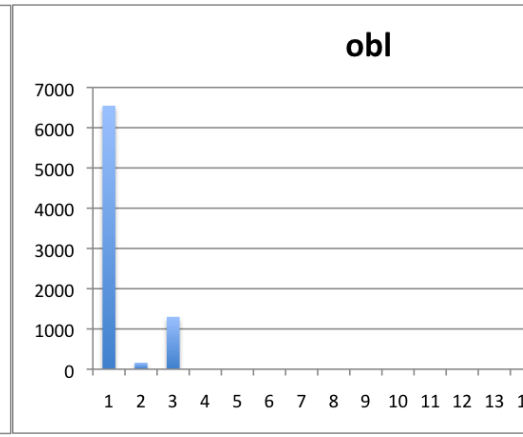
Distribution by category

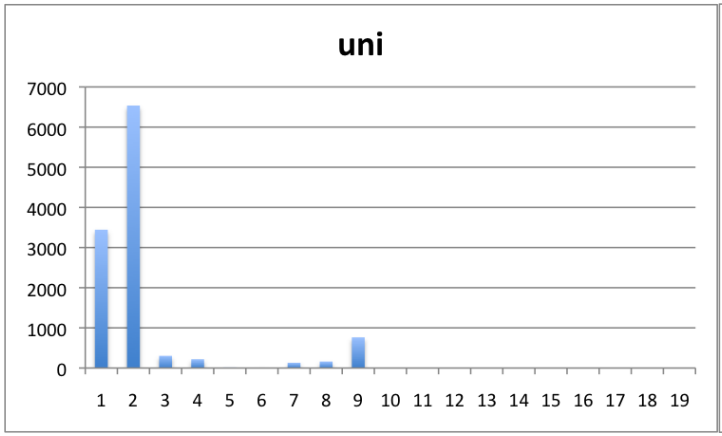


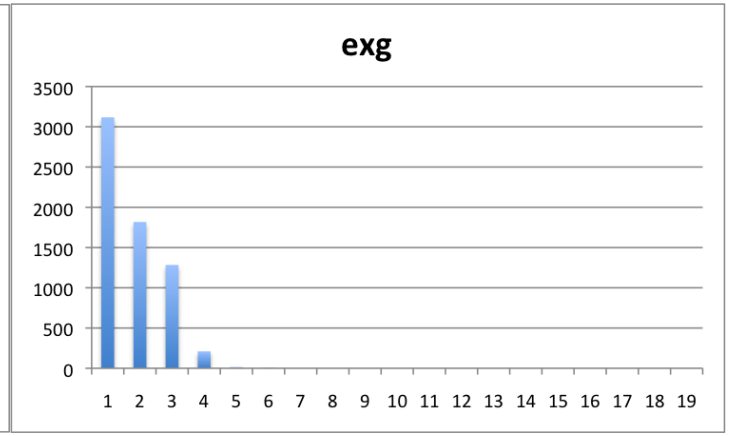
Distribution for the NP











Hard constraints for the NP in French

Type	Const. index	Property
Linearity	1	Det $\prec$ N
	4	Det $\prec$ AP
	9	Det $\prec$ PP
	15	N $\prec$ Srel
	18	N $\prec$ NP
Requirement	1	Det $\Rightarrow$ N
	2	PP $\Rightarrow$ N
	3	AP $\Rightarrow$N
Dependency	1	Det $\rightsquigarrow$ N
	2	AP $\rightsquigarrow$ N
	4	NP $\rightsquigarrow$ N
	7	PP $\rightsquigarrow$ N
Obligation	1	N
	3	Pro
Uniqueness	1	Det
	2	N

Conclusion

- A generic method for:
  - Acquiring PG grammars from constituency treebanks
  - Enriching a constituency treebank
- Applications:
  - Development of probabilistic constraint parser
  - Flexible grammatical description
  - Psycholinguistics: processing difficulty
  - Linguistic typology: comparing complexity of languages