Linguistique de corpus 2

Analyse syntaxique automatique

Patrick Paroubek

LIMSI-CNRS
Dépt. CHM - Groupe LIR
Bât. 508 Université Paris XI, 91403 Orsay Cedex
pap@limsi.fr

mercredi 15 mars 2017 / L3 - Cours 6

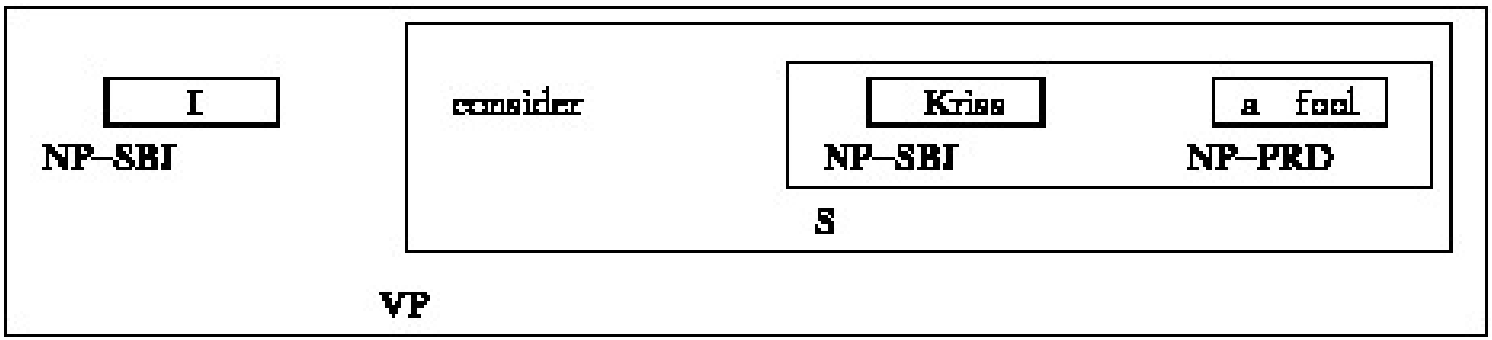
Annotations syntaxiques
les principales représentations

Le but de l'analyse syntaxique automatique est de fournir une analyse complète ou partielle de la structure d'un énoncé en termes de :

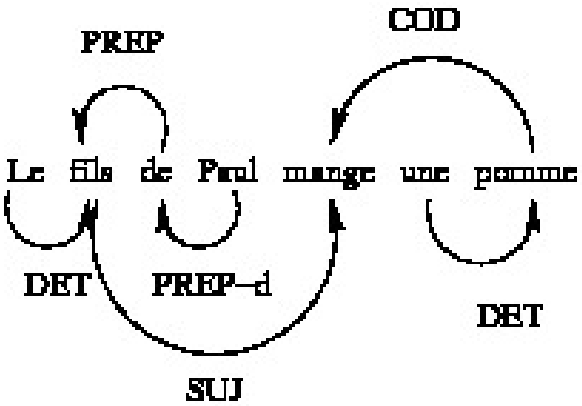
- "**chunks**" (lit. morceaux), des séquences de mots ayant une signification/fonction syntaxique commune,
- **constituants**, des séquences de mots qui représentent une unité fonctionnelle au sein d'une structure hierearchique,
- **dépendances**, des relations liant un mot donné (appelé "**la tête**") et un de ses dépendants,
- ou tout autre représentation propre à la théorie syntaxique utilisée...

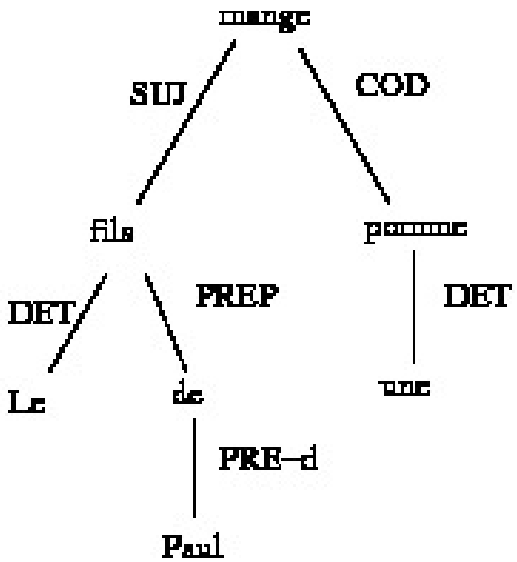
Un exemple d'analyse en constituants extrait du corpus arboré (corpus avec des annotations représentant des arbres syntaxiques) PennTreebank pour l'anglais, pour l'énoncé : "*I consider Kriss a fool*".

- S : clause déclarative simple (pas introduite par une conjonction de subordination ni par un pronom interrogatif et qui n'a pas d'inversion sujet-verbe)
- NP-SBJ : constituant nominal sujet de surface (sujet syntaxique),
- VP : constituant verbal
- NP-PRD : constituant nominal prédicatif

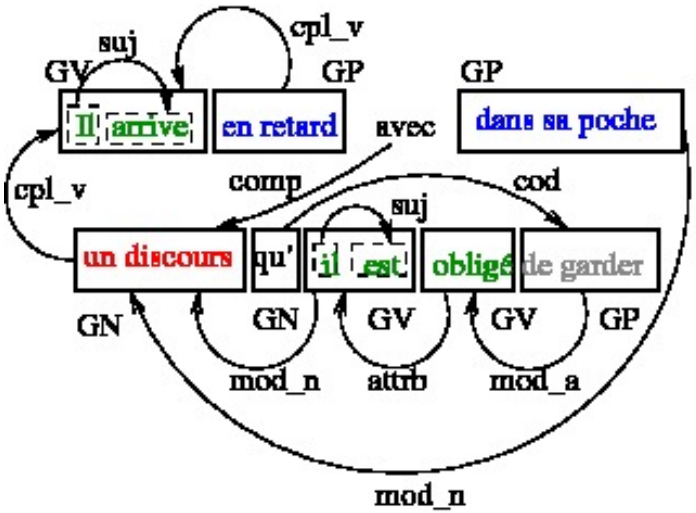


Un exemple d'analyse en dépendances avec les annotations de l'analyseur SYNTEX (D. Bourrigault) :
"Le fils de Paul mange une pomme".





Le formalisme de la campagne d'évaluation PASSAGE des analyseurs syntaxiques du français (2006-2009) associe dans sa représentation un niveau "chunks" (groupes syntaxiques) et des relations grammaticales (triplets ou quadruplet de la forme relation-arg1-arg2 ou relation-arg1-arg2-arg3).



Groupes Syntaxiques

GV = groupe verbal
GP = groupe prepositionnel
GN = groupe nominal

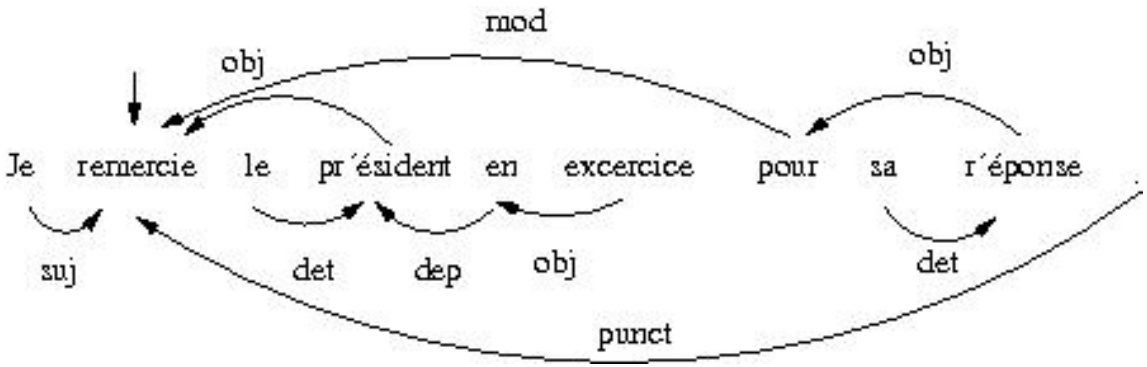
Relations syntaxiques

suj = sujet–>verbe
cpl_v = complement –> verbe
comp = complementeur –> complemente'
mod_n = modifieur –> nom
attrb = attribut –> verbe
mod_a = modifieur –> adjectif
cod = cod –> verbe

ConLL est un format de donnée extensible initialement développé pour des campagnes d'évaluation d'analyse syntaxique en dépendances et qui est utilisé par une large communauté.
Il permet de représenter les mots, d'un énoncé, les informations morpho-syntaxiques (parties du discours, lemme, etc.) et des dépendances syntaxiques. Il utilise une représentation matricielle dont la première colonne contient les formes de l'énoncé, puis les autre colonnes suivantes leurs étiquettes morpho-syntaxiques, et ensuite les dépendances syntaxiques.
C'est un format extensible ; auquel on peut ajouter de nouvelle couches d'analyse par simple ajout de colonnes à la matrice de représentation. Les dépendances sont représentées aux moyen de deux colonnes, l'une pour le type de la dépendance, l'autre pour l'adresse de sa cible, qui référence une ligne de la matrice, la source de la dépendance étant la forme courante.

1	Je	cln	CL	CLS	s=suj	2	suj	2	suj
2	remercie	remercier	V	V	m=ind|n=s|p=3	0	root	0	root
3	le	le	D	DET	g=m|n=s|s=def	4	det	4	det
4	président	président	N	NC	g=m|n=s|s=c	2	obj	2	obj
5	en	en	P	P	p=3	4	dep	4	dep
6	exercice	exercice	N	NC	g=m|n=s|s=c	5	obj	5	obj
7	pour	pour	P	P	_	2	mod	2	mod
8	sa	son	D	DET	g=f|n=s|s=poss	9	det	9	det
9	réponse	réponse	N	NC	g=f|n=s|s=c	7	obj	7	obj
10	.	.	PONCT	PONCT	s=s	2	ponct	2	ponct

Extrait d'annotation ConLL issu du corpus Sequoia v4.0.



Représentation graphique des dépendances de l'extrait d'annotation ConLL issu du corpus Sequoia v4.0.

Bien entendu, ce formalisme général permet de représenter une grande variété de formalisme à base de dépendances synatxiques, dont la sémantique doit être précisée dans un guide d'annotation. Dans le cas qui nous intéresse, il s'agit du Guide d'annotation ConLL FTB : http ://alpage.inria.fr/statgram/frdep/Publications/FTB-GuideDepSurface.pdf. Notons que nous aurons essentiellement deux types d'information, des étiquettes (morpho-syntaxique, sémantique, se rapportant à un système de classification particulier, etc.) associées à une forme particulière, et des dépendances étiquetées qui vont relier des formes ensembles, avec la condition qu'une forme ne peut engendrer qu'une dépendance, mais par contre plusieurs dependances peuvent aboutir sur une même forme.

Le guide d'annotation Bonsai recense 12 dépendances pour les gouverneurs verbaux : *suj (Sujet), obj (objet), de_obj (SP argumental en de, non locatif), a_obj (SP argumental en à, non locatif), p_obj (autre SP argumental), ats (Attribut du sujet), ato (Attribut de l'objet), mod (Modifieur), aux_tps (auxiliaires de temps), aux_pass (auxiliaires du passif), aux_caus (verbe causatif, en cas de complexe causatif + inf), aff (clitiques figés)*

et 8 dépendances pour gouverneurs non verbaux : *mod (Modifieurs repérés structuralement, comme par exemple les adjectifs épithètes, autres que les relatives), mod_rel (Relatives adnominales), coord (Relation portée par un coordonnant, avec comme gouverneur le coordonné immédiatement précédent), arg (utilisé dans le cas de prépositions « liées », ex. « Charybde en Scylla », dep_coord (Relation portée par un coordonné différent du premier, avec comme gouverneur le coordonnant immédiatement précédent), det (Relation portee par les déterminants), ponct (Relation portée par tout dépendant typographique, sauf pour les virgules jouant le rôle de coordonnant), dep (Relation sous-spécifiée, pour les dépendants prépositionnels (pas de gestion de la distinction argument / ajout pour les gouverneurs non verbaux).*

Il s'agit là des représentations utilisées pour l'annotations automatique. L'annotation manuelle proposée dans [**?**] ajoute les 8 dépendances suivantes : *mod_loc (Modifieurs sémantiquement locatifs, au propre ou au figuré), mod_cleft (Pour la subordonnée dans le cas d'une clivée), p_obj_agt (Pour les compléments d'agent, passif ou causatif), p_obj_loc (Dépendants argumentaux locatifs, source, destination, ou localisation), suj_impers (Pour le sujet explétif il), aff_moyen (Pour le clitique se en cas de moyen), arg_comp (Utilisé pour relier une comparative et son gouverneur), et finalement arg_cons (Utilisé pour relier une consécutive et son gouverneur adverbial.*

On distingue d'une part

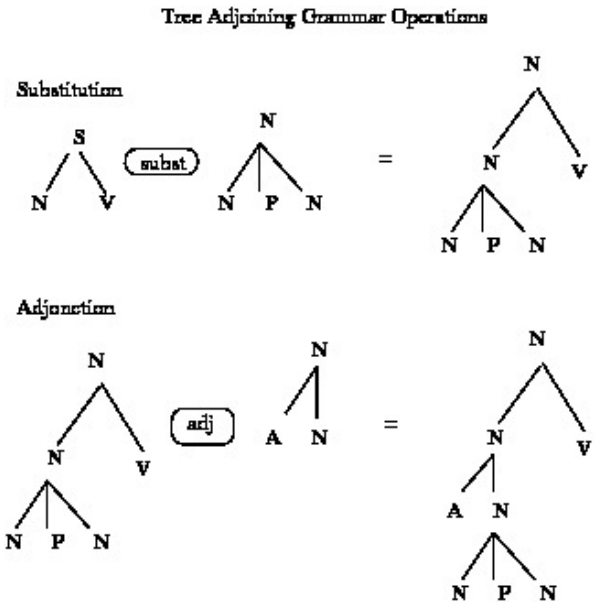
- les analyses de surface (*shallow parses*), se limitant à un niveau de groupes ou de relations
- les analyses profondes (*deep parses*), construisant des structures syntaxiques sur plusieurs niveaux (recursivité)

et d'autre part

- les analyses complètes, l'énoncé est analysé entièrement (la structure syntaxique couvre tous les mots de l'énoncé et forme une structure monolithique)
- les analyses partielles, certains mots de l'énoncé sont dépourvus d'annotations et les annotations d'un même énoncé peuvent être séparées en plusieurs îlots disjoints.

En analyse automatique, les représentation syntaxiques combinent des éléments atomiques du formalisme syntaxique au moyen d'une ensemble d'opérateurs ou règles sur les structures syntaxiques pour construire l'arbre syntaxique qui correspond à un énoncé (approche de la grammaire générative). Par exemple le formalisme des grammaires d'arbres adjoints (Tree Adjoining Grammars) de A.K. Joshi, combine des arbres au moyens de 2 opérations :

- la substitution (remplacement d'une feuille par un arbre dont la racine porte la même étiquette que la feuille qu'il remplace)
- l'adjonction remplacement d'un noeud interne de l'arbre par un arbre dont la racine porte la même étiquette que la feuille qu'il remplace et dont l'une des feuille porte la même étiquette que la racine de l'arbre.


Tree Adjoining Grammar Operations
Substitution
S
subst
N
N
V
N
P
N
=
N
N
V
N
P
N
Adjonction
N
N
V
N
P
N
adj
N
A
N
=
N
N
V
A
N
N
P
N

En TAG la construction d'une analyse produit 2 résultats :

1. l'arbre dérivé, c'est-à-dire l'arbre syntaxique couvrant l'énoncé analysé
2. l'arbre de dérivation, l'arbre représentant les différentes opérations de substitution (traits pointillés) et d'adjonction (trait plein) qui ont servi à produire l'arbre dérivé à partir des arbres contenus dans la grammaire définie pour le language analysé.

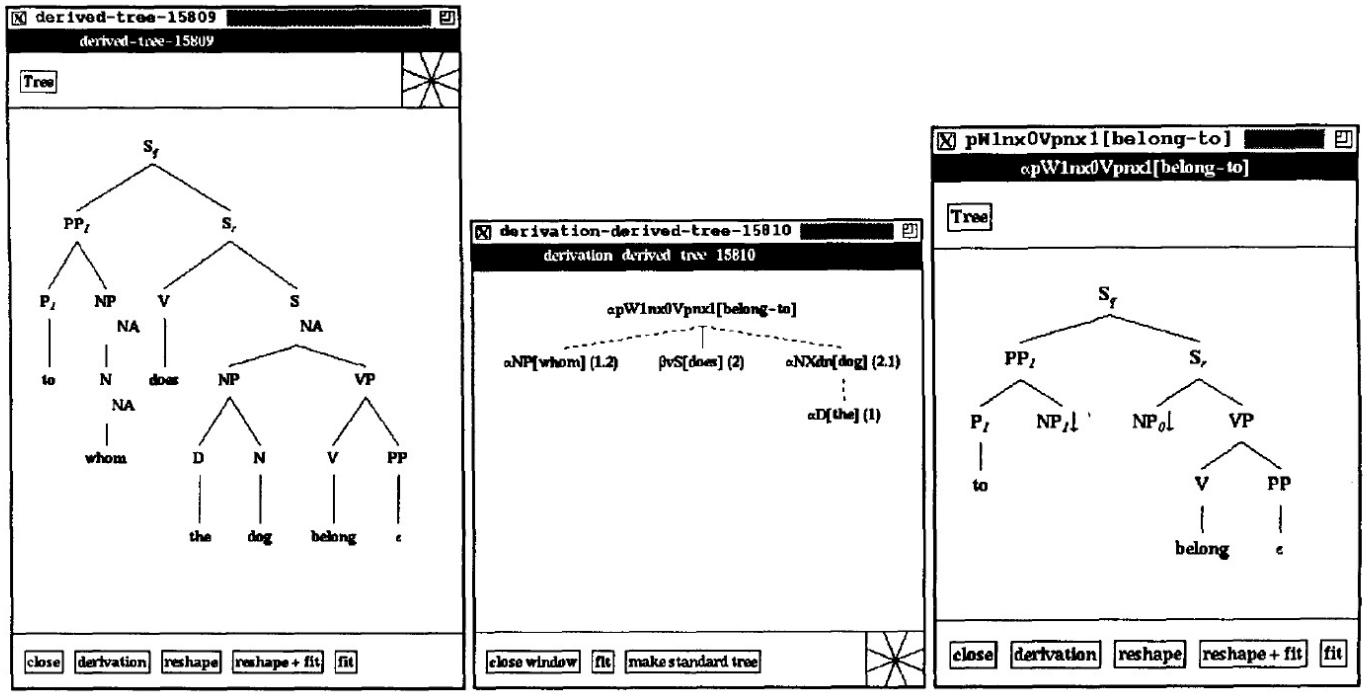


Figure 4: *left*, a derived tree, *middle*, its derivation, *right*, an elementary tree participating in the derivation.

1. La campagne EASY

1. Annotations pour l’analyse syntaxique

1. Les données

2. Les résultats

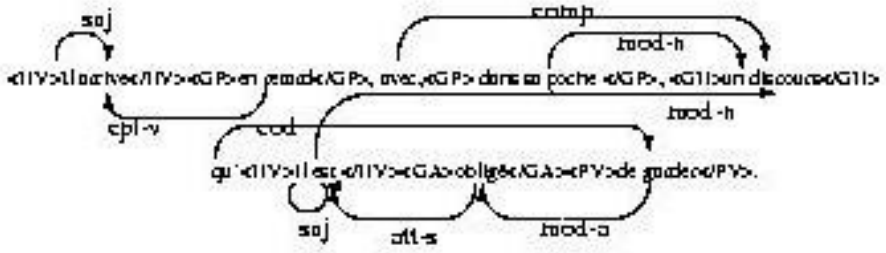
Objectif: évaluation d'analyse syntaxique

5 fournis. corpus, 12 participants, 15 systèmes évalués

- France Telcom R&D
- INRIA (ATOLL 1,2)
- LATL
- LIC2M
- LIRMM
- LORIA
- XEROX
- LPL (1,2 & 3)
- PERTIMM
- SYNAPSE
- ERSS
- TAGMATICA

Les fournisseurs de corpus :

- ATILF (littéraire)
- DELIC (oral transcrit, emails)
- ELDA (oral ESTER, MLCC, sénat, questions TREC traduites, questions Amaryllis, web)
- LLF (Le Monde)
- STIM (médical)



Il arrive en retard, avec, dans sa poche, un discours qu'il est obligé

de garder.

Guide d'annotation (A. Vilnat) :
http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html

5 types de constituants

1. GN groupe nominal
2. GP groupe prépositionnel
3. NV noyau verbal
4. GA groupe adjectival
5. GR groupe adverbial

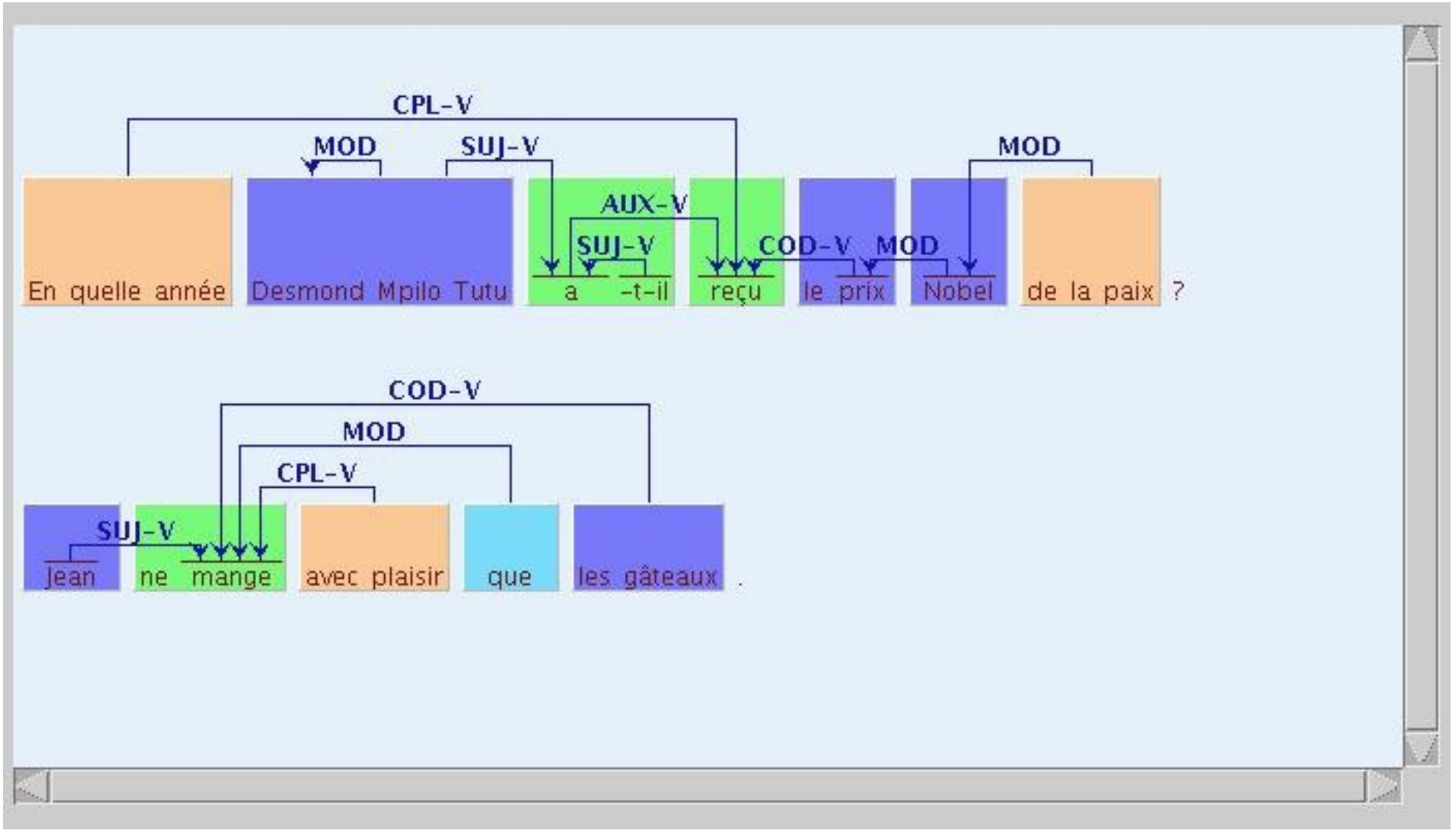
14 types de relations

1. Sujet - Verbe
2. Auxiliaire - Verbe
3. Objet direct - Verbe
4. Complément - Verbe
5. Modifieur – Verbe
6. Complémenteur
7. Attribut -Sujet/Objet
8. Modifieur - Nom
9. Modifieur - Adjectif

1. Modifieur – Adverbe
2. Modifieur – Préposition
3. Coordination
4. Apposition
5. Juxtaposition

Outils de validation :

•éditeur graphique (E. Giguet)



<DOCUMENT fichier="oral_delic_1.xml">
<E ID="E1">
<F ID="E1F1">fref-f-c3</F>
</E>
<E ID="E2">
<F ID="E2F1">voilÃ </F>
</E>
<E ID="E3">
<F ID="E3F1">ben</F>
<F ID="E3F2">je</F>
<F ID="E3F3">travaille</F>
<F ID="E3F4">dans</F>
<F ID="E3F5">un</F>
<F ID="E3F6">pressing</F>
</E>

<DOCUMENT fichier="oral_elda_1.xml">
<E ID="E1">
<F ID="E1F1">14</F>
<F ID="E1F2">heures</F>
<F ID="E1F3">À </F>
<F ID="E1F4">Paris</F>
<F ID="E1F5">,</F>
<F ID="E1F6">midi</F>
<F ID="E1F7">en</F>
<F ID="E1F8">temps</F>
<F ID="E1F9">universel</F>
<F ID="E1F10">,</F>
<F ID="E1F11">l'</F>
<F ID="E1F12">information</F>
<F ID="E1F13">continue</F>
<F ID="E1F14">sur</F>
<F ID="E1F15">RFI</F>
<F ID="E1F16">.</F>
</E>

<?xml version="1.0" encoding="UTF-8"?>
<DOCUMENT fichier="\Oral Elda\oral_elda_1EASY.UTF8.xml" xmlns:xlink="http://www.w3.org/1999/xlink">
<E id="E1">
<constituants>
<Groupe type="GN" id="E1G1">
<F id="E1F1">14</F>
<F id="E1F2">heures</F>
</Groupe>
<Groupe type="GP" id="E1G2">
<F id="E1F3">Ã </F>
<F id="E1F4">Paris</F>
</Groupe>
<F id="E1F5">,</F>
<Groupe type="GN" id="E1G3">
<F id="E1F6">midi</F>
</Groupe>
<Groupe type="GP" id="E1G4">
<F id="E1F7">en</F>
<F id="E1F8">temps</F>
</Groupe>
<Groupe type="GA" id="E1G5">
<F id="E1F9">universel</F>
</Groupe>
<F id="E1F10">,</F>
<Groupe type="GN" id="E1G6">
<F id="E1F11">l'</F>
<F id="E1F12">information</F>
</Groupe>
<Groupe type="NV" id="E1G7">
<F id="E1F13">continue</F>
</Groupe>

ANNOTATIONS EN CONSTITUANTS

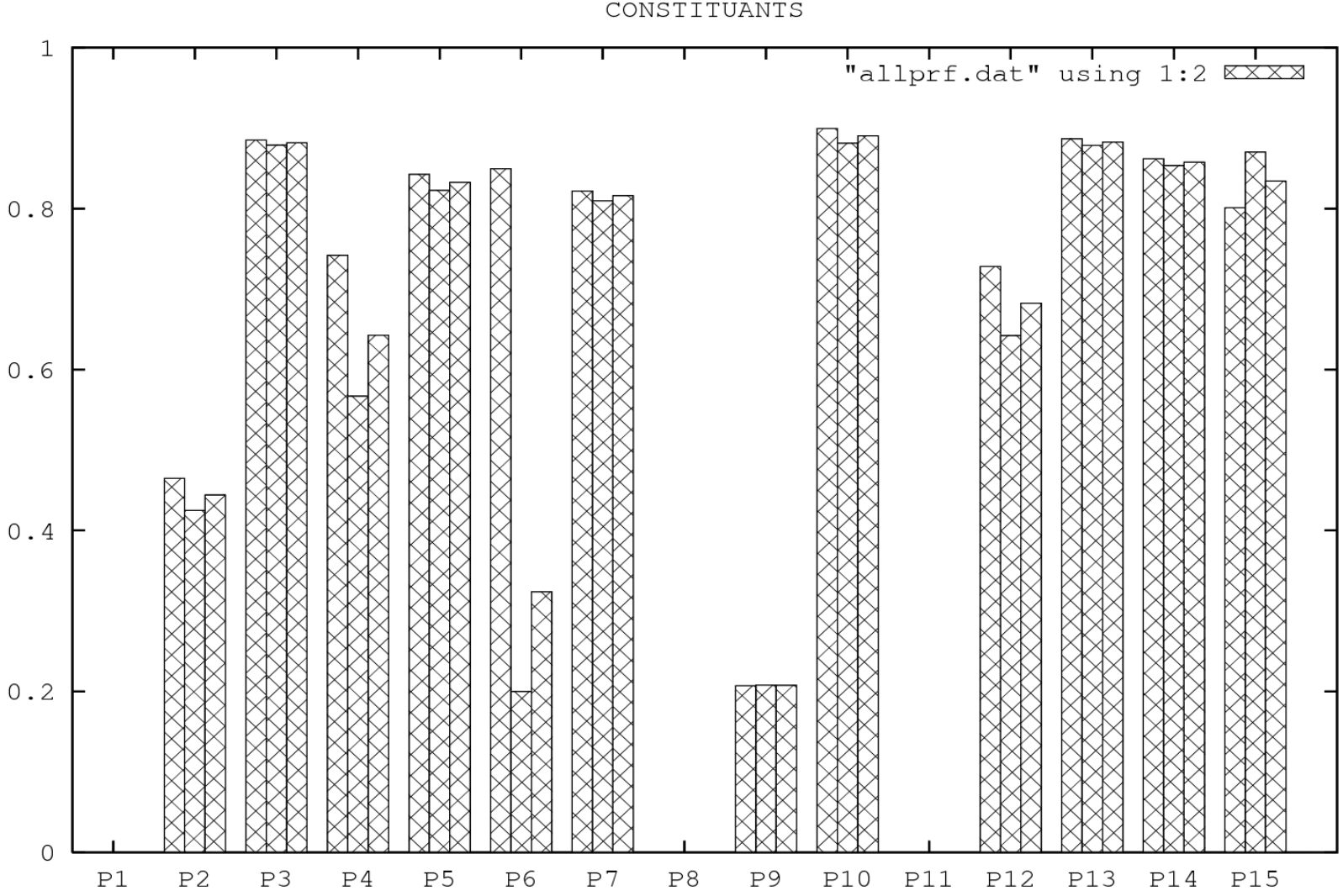
<Groupe type="GP" id="E1G8">
<F id="E1F14">sur</F>
<F id="E1F15">RFI</F>
</Groupe>
<F id="E1F16">.</F>
<F id="E1F17">Â§</F>
</constituants>
<relations>
<relation xlink:type="extended" type="MOD-N" id="E1R2">
<modifieur xlink:type="locator" xlink:href="E1G4"/>
<nom xlink:type="locator" xlink:href="E1F6"/>
<a-propager booleen="faux"/>
</relation>
<relation xlink:type="extended" type="SUJ-V" id="E1R3">
<sujet xlink:type="locator" xlink:href="E1G6"/>
<verbe xlink:type="locator" xlink:href="E1G7"/>
</relation>
<relation xlink:type="extended" type="CPL-V" id="E1R4">
<verbe xlink:type="locator" xlink:href="E1G7"/>
<complement xlink:type="locator" xlink:href="E1G8"/>
</relation>
<relation xlink:type="extended" type="MOD-N" id="E1R5">
<modifieur xlink:type="locator" xlink:href="E1G5"/>
<nom xlink:type="locator" xlink:href="E1F8"/>
<a-propager booleen="faux"/>
</relation>
<relation xlink:type="extended" type="MOD-N" id="E1R6">
<modifieur xlink:type="locator" xlink:href="E1F1"/>
<nom xlink:type="locator" xlink:href="E1F2"/>
<a-propager booleen="faux"/>
</relation>
</relations>
</E>

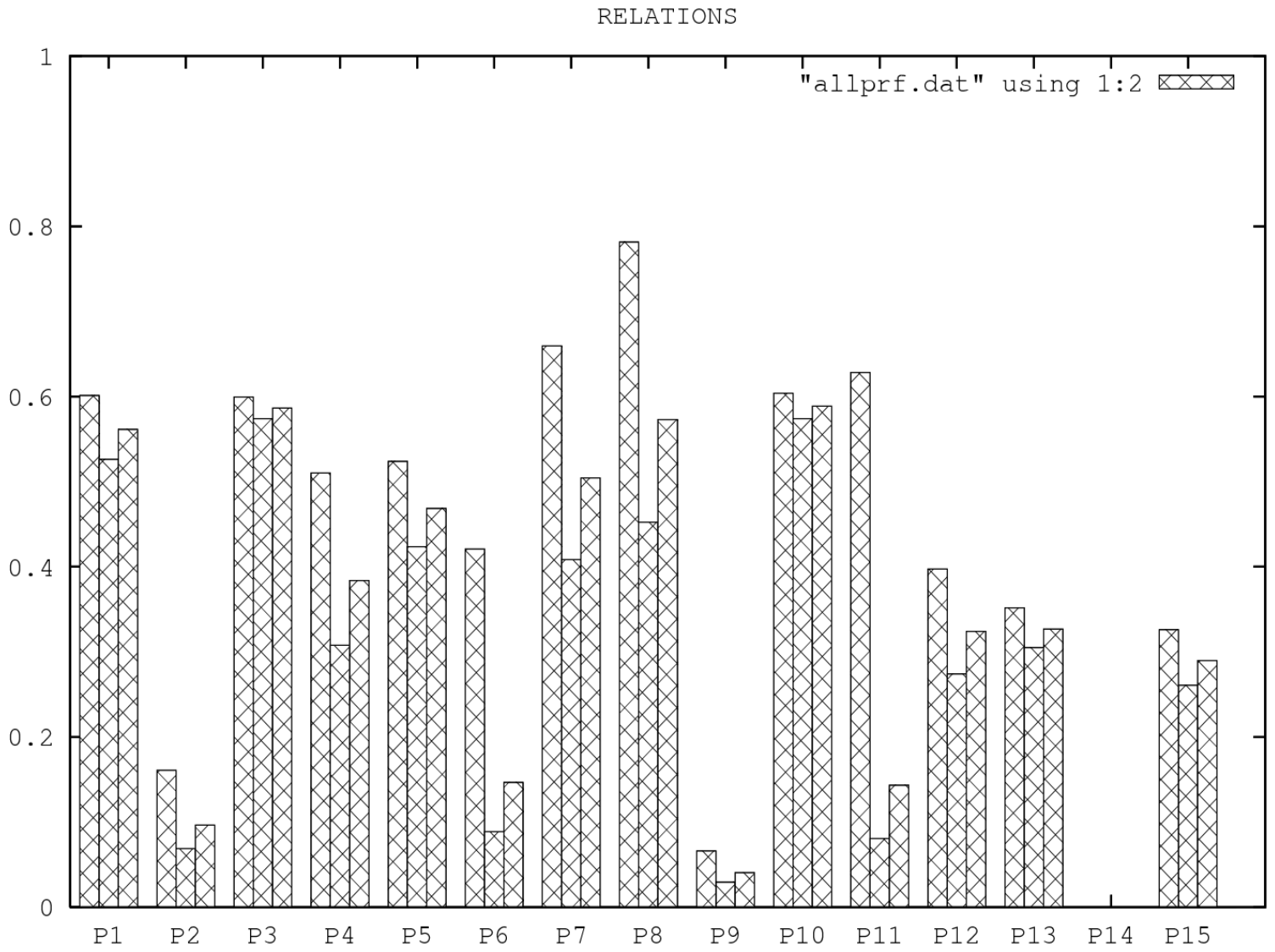
ANNOTATIONS EN RELATIONS

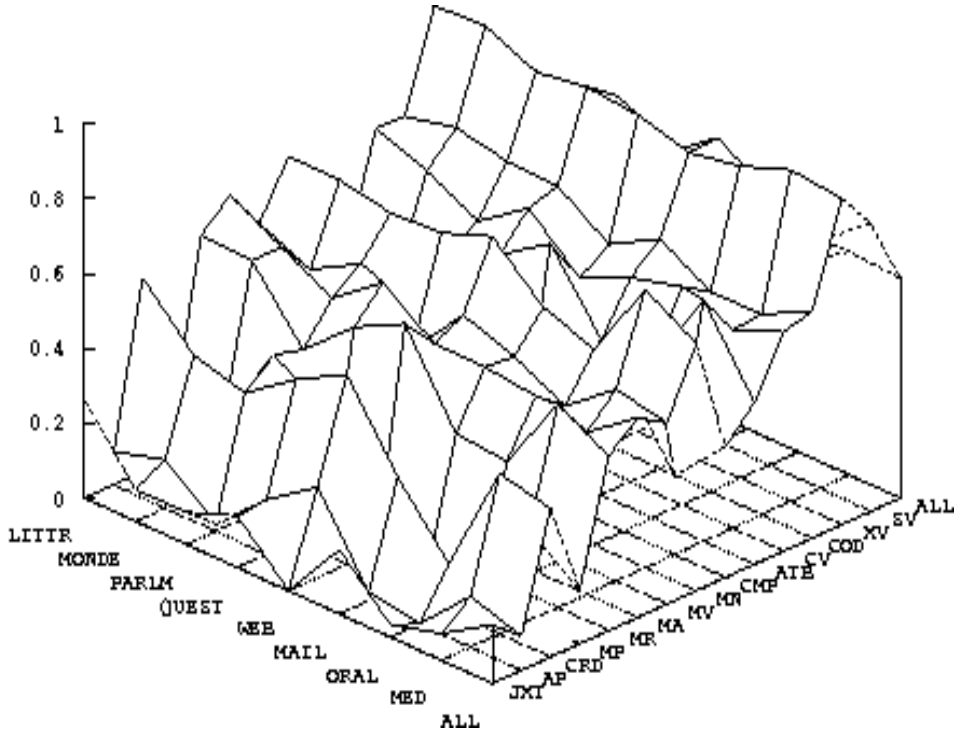
Mesures de précision et rappel :

- par participant,
- type de constituant,
- par type de corpus.

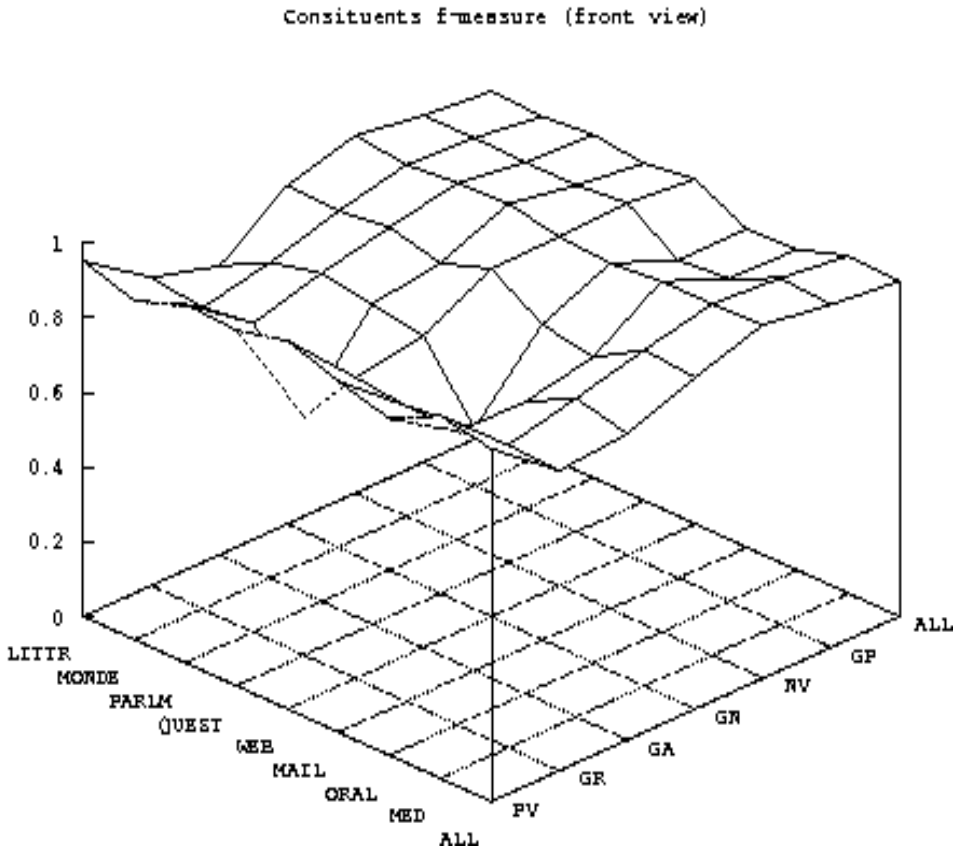
Mesures strictes (égalité stricte des adresses)
et relachement de contrainte sur les adresses
de début et de fin de groupes


CONSTITUANTS
"allprf.dat" using 1:2
1
0.8
0.6
0.4
0.2
0
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15


RELATIONS
"allprf.dat" using 1:2
1
0.8
0.6
0.4
0.2
0
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15



Resultat d'un participant a EASY en f-mesure pour les relation pour tous les types de corpus et pour toutes les annotations.



Resultat d'un participant a EASY en f-mesure pour les constituants pour tous les types de corpus et pour toutes les annotations.

Écrire un automate qui reconnaisse les noyaux verbaux selon les consignes d'annotation syntaxique PASSAGE et qui les entoure avec des balise XML <NV> </NV>.