Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé

Anne Lacheret[(1)], Sylvain Kahane[(1)], Julie Beliao[(1)], Anne Dister[(2)], Kim Gerdes[(3)], Jean-Philippe Goldman[(4)], Nicolas Obin[(5)], Paola Pietrandrea[(6)], Atanas Tchobanov [(1)]

(1) Modyco, Université Paris Ouest Nanterre & CNRS
(2) Université Saint-Louis - Bruxelles
(3) LPP, Université Paris Sorbonne Nouvelle & CNRS
(4) Université de Genève
(5) IRCAM, UMR STMS UPMC-CNRS, Paris
(6) LLL, Université François Rabelais & CNRS

INTRODUCTION

Bilan

- Journées PFC 2011, Paris : projet en cours
  - *Annotation prosodique de corpus oraux : le projet Rhapsodie :* http://rhapsodie.risc.cnrs.fr/fr/

    - Contexte : corpus annotés (en France) en syntaxe et en prosodie

    - Enjeux théoriques
      - Interface prosodie/syntaxe/discours → corpus annotés

    - Enjeux technologiques : vers des grands corpus annotés
      → corpus d'apprentissage

    - Zoom sur annotation prosodique

Aujourd'hui

- Projet achevé
  - Ressources disponibles sur http://www.projet-rhapsodie.fr/ (cretaive Common, noncommercial, etc) : 57 échantillons annotés, 3h, 30.000 mots, 89 locuteurs (h,f), différentes situations de discours

    - Échantillons sonores wave/mp3
    - Analyse acoustique : F0 corrigé, stylisation automatique (format pitch)
    - Transcription orthographique (txt)
    - Annotation macrosyntaxique (txt)
    - Annotation microsyntaxique (format tabulaire)
    - Annotation prosodique (xml, textgrid)
    - Metadonnées (xml, html)
    - Browser metadonnées
    - Outil requêtage prosodie: Rhapsodie QL

TREEBANK RHAPSODIE

Version 0.8 18.12.2012

Corpus de français parlé annoté pour la prosodie et la syntaxe

Propriété intellectuelle

VISUALISER DES EXEMPLES

- arbre de dépendance
- arbre de constituants macrosyntaxiques
- arbre de constituants prosodiques

INTERROGER LE TREEBANK

Langage de requêtes Rhapsodie QL, pour l'interrogation de la prosodie

ATTENTION!

Le moteur de requête supporte les browsers Firefox, Chrome, mais pas Internet Explorer

TELECHARGER ET ECOUTER LE TREEBANK

- Recherche dans les transcriptions et écoute synchronisée
- Téléchargement des fichiers du corpus
- Téléchargement des fichiers de codage microsyntaxique version béta 10/13 (zip)

TUTORIELS

- Tutoriel métadonnées (pdf)
- Protocole codage prosodique (wiki + pdf)
- Tutoriel requêtes (wiki)
- Tutoriel Easy Align enrichi (wiki + pdf)
- Annotation tonale (pdf)
- Contours globaux (wiki + pdf)
- Tutoriel codage macrosyntaxique (pdf)
- Tutoriel codage microsyntaxique (pdf) **Nouveau: Version béta 10/13** (pdf)
- Tutoriel transcription orthographique (wiki + pdf)
- Auteurs des tutoriels

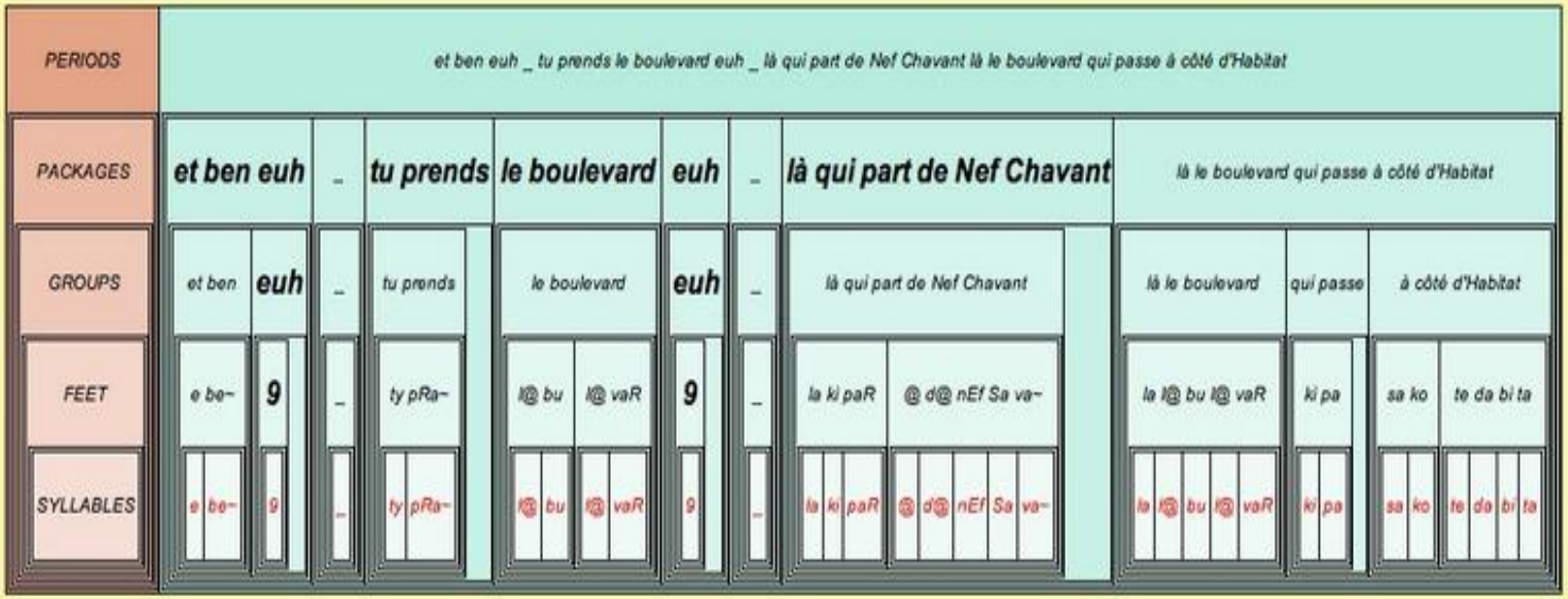
EN SAVOIR PLUS

- Présentation du projet Rhapsodie
- Développements récents
- Publications
- Journées d'étude
- Participants
  - Laboratoires
  - Membres
- Outils
- Signaler un problème

Visualiser des exemples

Arbre de constituants prosodiques

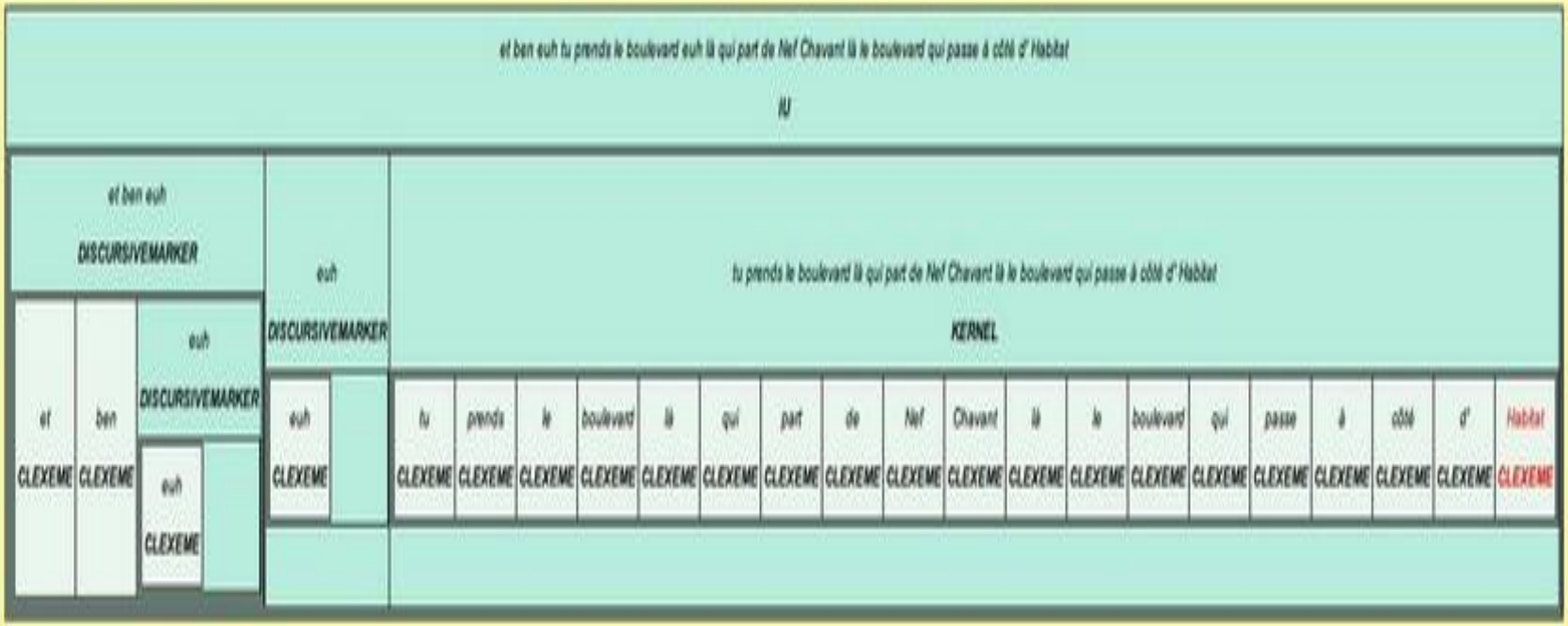
Prosodic structure



Download the structures

Arbre de constituants macrosyntaxiques

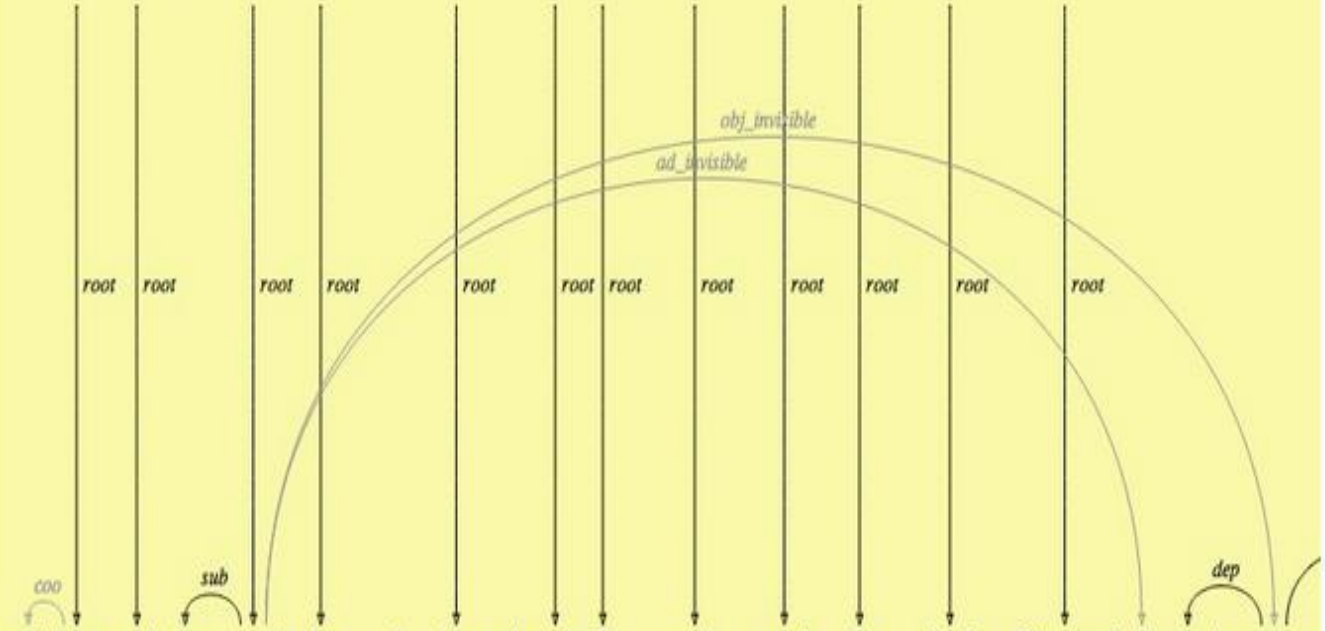
Topological Structure



Download the structures

Arbre de dépendance

Dependency structure



et	ben	euh	tu	prends	le	boulevard	euh	là	qui	part	de	Nef	Chavant	là	le	boulevard	qui	passe	à	côté	d'	Habitat
J	I	I	CL	V	UNKNOWN	UNKNOWN	I	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	ADV	D	N	QU	V	PRE	N	PRE	N

Download the structures

Télécharger

Recherche ||| Tutoriel

Informations sur le téléchargement/Download info

Recherche mots: SEARCH

Sex: Indiff. Age: Indiff.

Genre: Indiff. SubGenre: Indiff. Corpus: Indiff.

Interactivity: Indiff.

Social Context: Indiff. Event Structure: Indiff.

Channel: Indiff. Planning Type: Indiff. Quality: Indiff.

SEARCH

Accès échantillon: Indiff.

SHOW

Mon Panier View

Edited [View] [Export]
Ready [View] [Export]

Nouveau panier:

CREATE

Archives prosodie TextGrids [ZIP]
Archives prosodie XML [ZIP]
Archives pitch nettoyé [ZIP]
Archives pitch lissé [ZIP]
Archives metadonnées [ZIP]
Archives Transcription TXT [ZIP]
Archives Syntaxe TXT [ZIP]
Archives MP3 [ZIP (74 MB)]
Archives WAV [ZIP (677 MB)]

Results

Search results will display here.

Very important! Don't use the browser back button as it will bring you back to the login page. Use the search button to return to search results.

Le Treebank

Corpus design

- Représentativité typologique (typologie textuelle)
  Mais pas de corpus de référence du français
- Eviter idiosyncrasies individuelles

→ Beaucoup d'échantillons courts > peu d'échantillons longs.

→Sources externes (dont PFC)

Quid de la propriété intellectuelle ?

Interlocuteurs au CNRS, à l'ANR : néant

→Tout à construire

Propriété intellectuelle

Les données

Les échantillons du Treebank Rhapsodie soit sont issus de données primaires préexistantes, *i.e.* 32 enregistrements déjà constitués pour des projets antérieurs, en accord avec les concepteurs initiaux : **sources externes**, soit ont été collectés en interne (25 échantillons) : **source interne.**

Aussi, pour assurer la traçabilité des sources et, ce dans le respect de la propriété intellectuelle : pour toute utilisation dans une communication orale et ou une publication d'échantillon du treebank Rhapsodie, nous donnons ci-dessous les instructions pour la citation des sources.

Les sources externes : 32 échantillons

Nom de la source	Référence bibliographique	Nombre d'échantillons	Nom des échantillons dans Rhapsodie
CFPP2000	Branca-Rosoff S., Fleury S., Lefeuvre Fl., Pires M (2012), *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)* http://cfpp2000.univ-paris3.fr/	4	Rhap-D0001,CFPP2000 Rhap-D0002,CFPP2000 Rhap-D0004,CFPP2000 Rhap-D0006,CFPP2000
Corpus Avanzi	Avanzi M. (2012), *L'interface prosodie/syntaxe en français Dislocations, incises et asyndètes*, Bruxelles, Peter Lang	19	Rhap-D0007,Avanzi Rhap-D0008,Avanzi Rhap-D0017,Avanzi Rhap-D0020,Avanzi Rhap-M0001,Avanzi Rhap-M0003,Avanzi Rhap-M0004,Avanzi Rhap-M0005,Avanzi Rhap-M0006,Avanzi Rhap-M0007,Avanzi Rhap-M0008,Avanzi

Comment utiliser et citer les sources

1. Présentation de l'échantillon au fil du texte (article, exemplier, diaporama)

1.1. Source externe

- occurrence de l'exemple suivie a) du nom de l'échantillon Rhapsodie précédé du préfixe Rhap-, b) du nom du projet source

  - *j'accorde une puissance énorme à l'acte d'écrire* [Rhap-D2009, corpus Mertens]
  - c'était ils préféraient rigoler que de travailler [Rhap-D0002, CFPP2000]
  - *je suis heureux de me retrouver ce soir parmi vous* [Rhap-M2001, C-PROM]
  - et puis finalement bah on a choisi de rester [Rhap-D0003, PFC]

1.2.Source interne

- occurrence de l'exemple suivie a) du nom de l'échantillon Rhapsodie précédé du préfixe Rhap-, b) du nom du projet Rhapsodie

  - ex. *les réformes plus vite et plus fort* [Rhap-D2013, Rhapsodie]

2. Référence bibliographique à citer

2.1. Source externe : 2 références à indiquer

- La référence bibliographique correspondant à la source indiquée dans le tableau ci-dessus
- Lacheret A., Kahane S., Pietrandrea P. (2014 ed.), *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, col Studies in Corpus Linguistics, Amsterdam, Benjamins

2.2.Source interne

- Lacheret A., Kahane S., Pietrandrea P. (2014 ed.), *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, col Studies in Corpus Linguistics, Amsterdam, Benjamins

Diversité des genres de discours

Traits situationnels → Marqueurs formels → Type de genre

- Biber, D. and Conrad, S. (2009). *Register, Genre and Style*. Cambridge, CUP.

- KOCH, P. and W. OESTERREICHER (2001). Langage parlé et langage écrit, in *Lexicon der Romanistischen Linguistik*, T1-2, Tübingen, Max Niemeyer Verlag, 584-627.

Type de parole	Monologues, dialogues	
	Situation de communication	Parole privée ou publique
	Planning	Spontané, semi-spontané, planifié
	Interactivité	Interactif, semi-interactif, non-interactif
	Canal de communication	Face à face vs. Conférence, émission radiophonique ou télévisuelle
	Type de séquence	Argumentative, descriptive, procédurale, oratoire

Ecodage des métadonnées

- IMDI-CMDI format, Max Planck Institute for Psycholinguistics, Nijmegen

  - Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). *Standardizing a component metadata infrastructure*. In N. Calzolari (Ed.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, May 23rd-25th, 2012 (pp. 1387-1390). European Language Resources Association (ELRA)

Accès échantillon: D0009

SHOW

Results

METADATA:

- Access
- WrittenResource
- WrittenResource
- Source

Id Corpus PFC http://www.projet-pfc.net/pfc-recherche.html

Format

Quality Unspecified

CounterPosition

Start Unspecified

End Unspecified

TimePosition

Start Unspecified

End Unspecified

- Access

Description

Corpus source - PFC, Le projet international PFC, codirigé par Marie-Hélène Côté Université d'Ottawa, Jacques Durand ERSS, Université de Toulouse-Le Mirail, Bernard Laks MoDyCo, Université de Paris X et Chantal Lyche Universités d'Oslo et de Tromsø, s'adresse à un triple public, susceptible de s'intéresser au français oral dans ses usages attestés et dans sa variation au sein de l'espace francophone : chercheurs, enseignants/apprenants de français et grand public. http://www.projet-pfc.net/

refid: D0009

Schémas d'annotation

Transcription orthographique (Dister)

Alignement texte-son (easyalign, Goldman, J.-Ph. (2011): "Easyalign: an automatic phonetic alignment tool under praat", In INTERSPEECH-2011, 3233-3236. http://latlntic.unige.ch/phonetique.

tires mots, phonèmes, syllabes

→ Annotations modulaires

Annotation syntaxique

- Combine le modèle syntaxique développé par l'Ecole d'Aix-en-Provence (Blanche-Benvéniste)), et le modèle pragmatique élaboré dans le cadre du projet Lablita (Cresti
- Deux niveaux de cohésion syntaxique ont été fixés :
  - niveau macrosyntaxique pilote la cohésion illocutoire à l'intérieur de l'énoncé (Berrendonner) :
    - l'ensemble des relations formelles qui se déploient entre segments pour former un acte illocutoire,
  - niveau microsyntaxique pilote la cohésion syntaxique
    - Annotation des catégories, fonctions et dépendances entre unités grammaticales.

Annotation syntaxique

Disfluences et entassements paradigmatiques annotés

- Disfluence : ni relation de dépendance avec d'autres éléments, ni syntaxe interne (J. Deulofeu)
- Entassements paradigmatiques (piles)

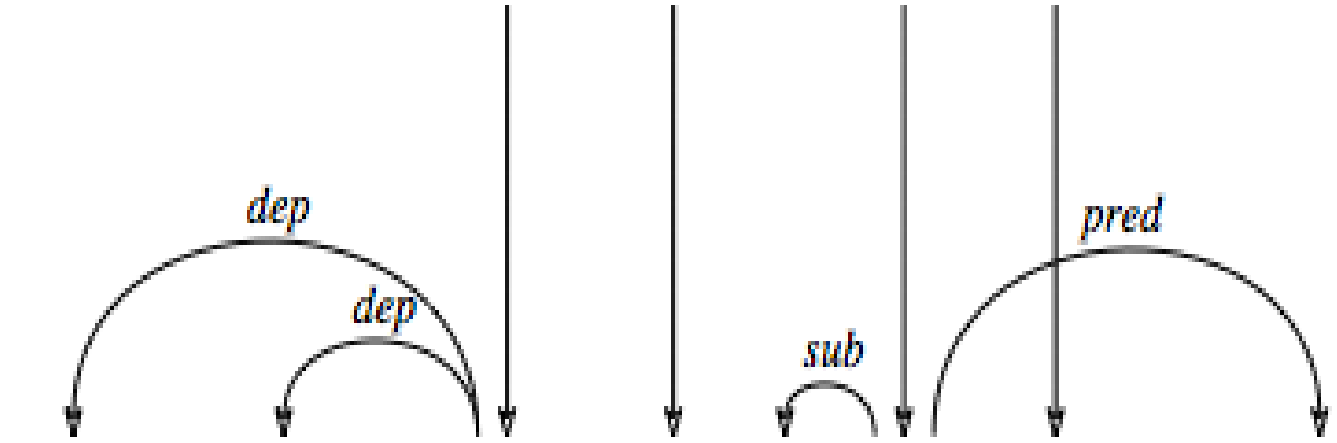
Il y a un cou~ *certainement*

un courant

Annotation microsyntaxique

- Analyse syntaxique de surface comme dans les autres treebanks (Abeillé A., Crabbé B. (2013) Vers un treebank du français parlé, Actes de TALN 2013)

- Rection syntaxique (dépendances modélisant les relations de rection : recteur et régi)



mon premier choix " euh " < c' était " euh " psychologie //

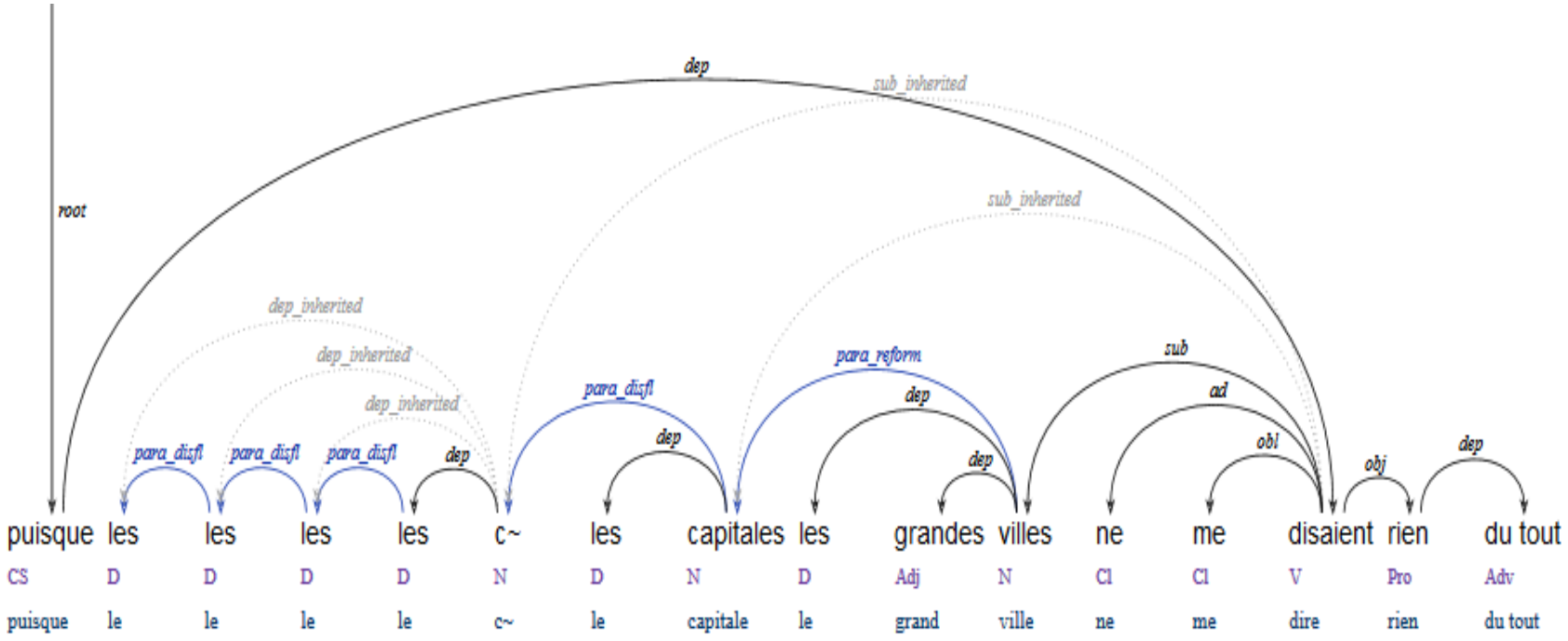
POS

- 15 étiquettes syntaxiques pour le tagging

	N	V	Cl	D	Pre	Adv	I	Adj	J	Qu	CS	Pro	Pre+D	X	Pre+Qu	total
POS	6249	5969	4177	4081	3457	2784	1978	1610	1141	800	726	718	484	198	3	34375

Entassements, piles

- Plusieurs éléments viennent occuper la même position régie



^ *puisque { { { les | les | les | les } c~ | les capitales } | les grandes villes } ne me disaient rien du tout //*

Macrosyntaxe

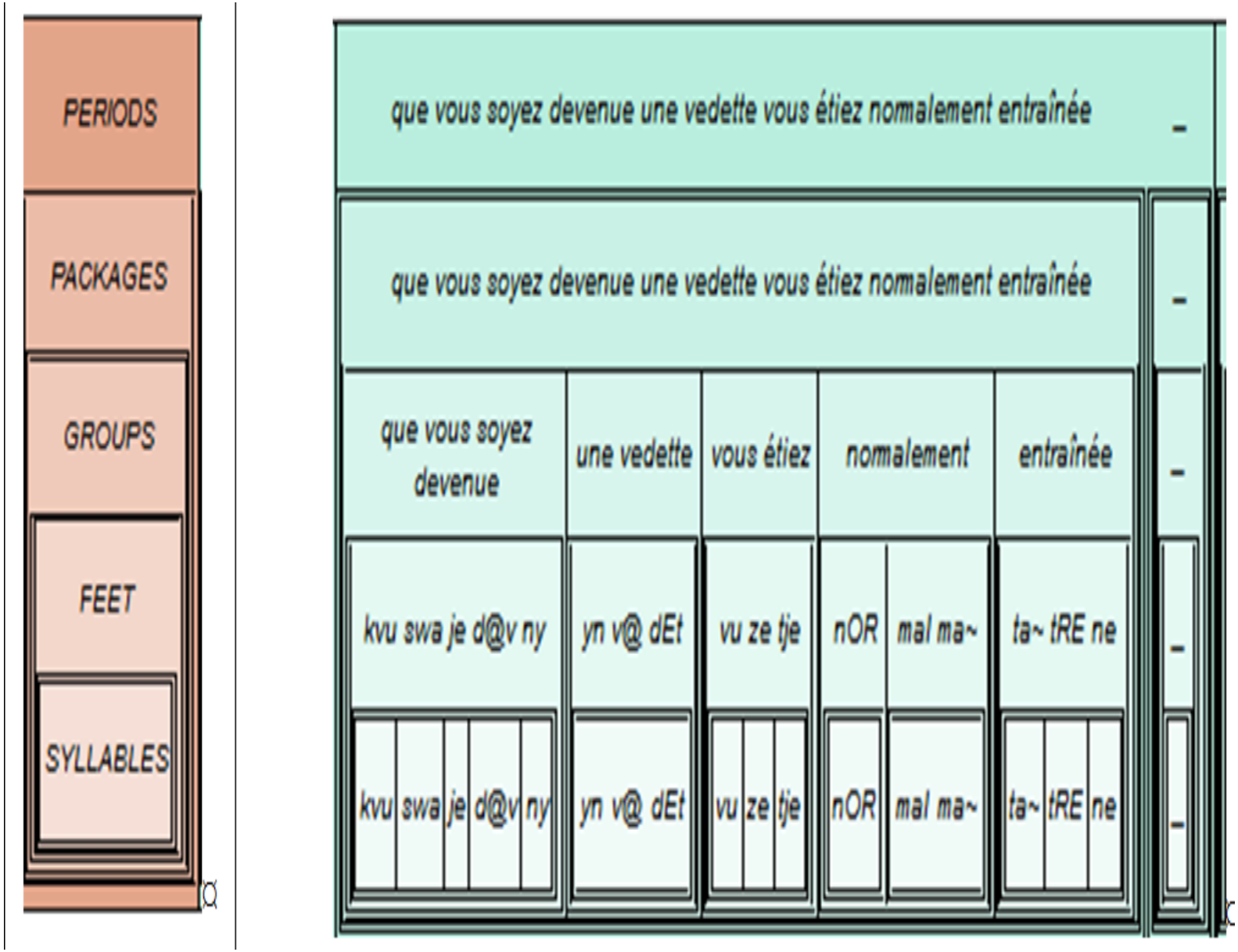
- Unités illocutoires
  - Chaque échantillon est segmenté en une succession d'unités illocutoires (UI);
  - Des composantes d'unité illocutoire :
    - un noyau obligatoire,
    - un ou plusieurs pré-noyaux optionnel(s),
    - un ou plusieurs post-noyaux, optionnel(s).
    - Etc.
- *alors < là < la psychiatrie <* ***c'est autre chose*** *// (Rhap-D0006, CFPP2000)*

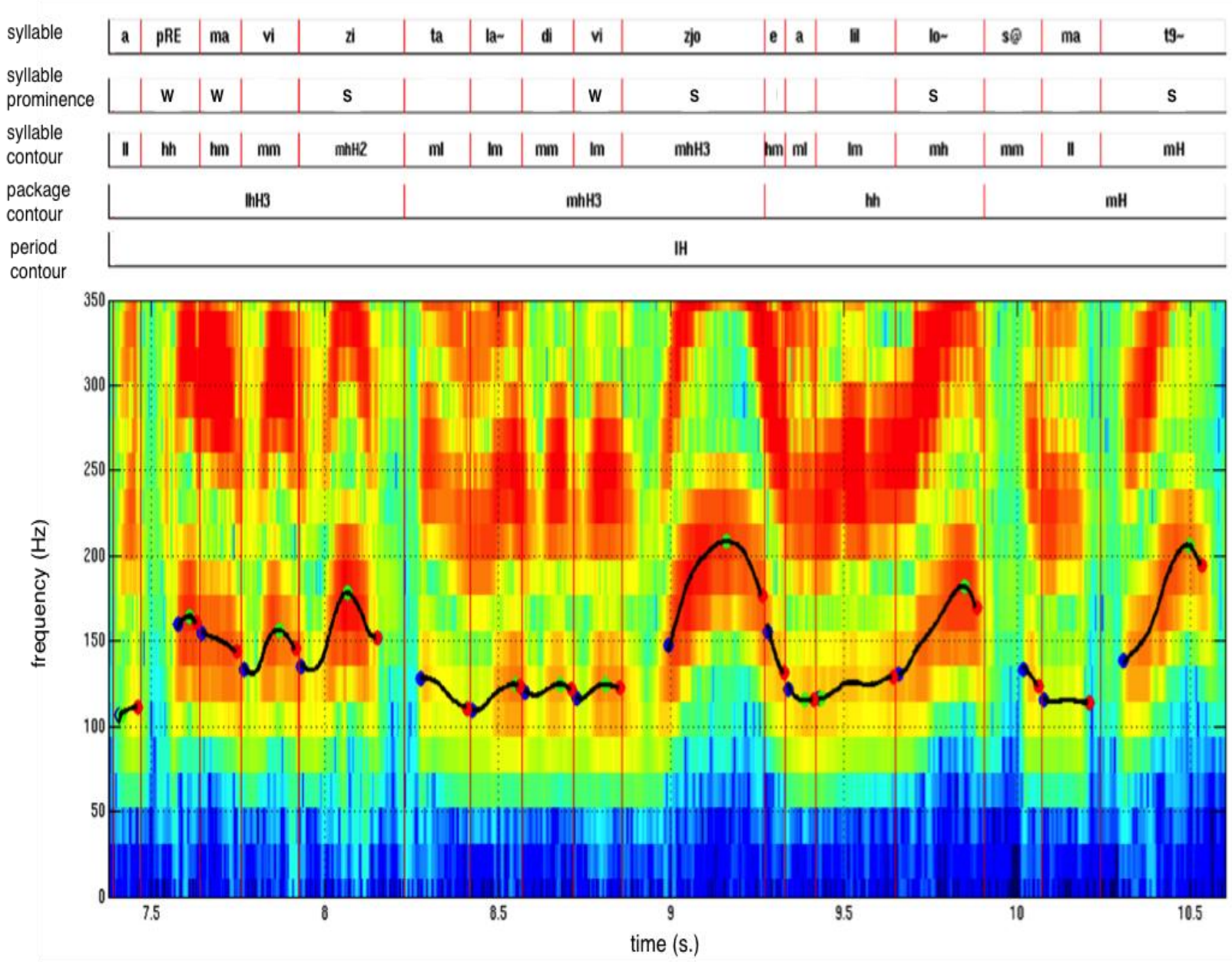
UI vs UR : notion de syntaxe coopérative

- Une UR peut traverser plusieurs UI
  - $L2 "eh bien" je crois que je ne me suis pas conduit d'une façon conforme à ce qu'on attend "euh" { d'une jeune fille d'abord | ^et d'une femme ensuite //+
  - $L1 | d'une jeune $- $L1 bourgeoise //+ [D2001, Corpus Mertens]

Prosodie (voir ppt 2011, site PFC)

- Annotation manuelle des proéminences et disfluences phonétiques
- Structure prosodique dérivée automatiquement
- Annotation tonale automatique flexible
  - Outil distribué en ligne courant 2014


PERIODS
PACKAGES
GROUPS
FEET
SYLLABLES
que vous soyez devenue une vedette vous étiez normalement entraînée
que vous soyez devenue une vedette vous étiez normalement entraînée
que vous soyez devenue
une vedette
vous étiez
normalement
entraînée
kvu swa je d@v ny
yn v@ dEt
vu ze tje
nOR
mal ma~
ta~ tRE ne
kvu swa je d@v ny
yn v@ dEt
vu ze tje
nOR
mal ma~
ta~ tRE ne


syllable
a
pRE
ma
vi
zi
ta
la~
di
vi
zjo
e
a
lil
lo~
s@
ma
t9~
syllable prominence
W
W
S
W
S
S
S
syllable contour
ll
hh
hm
mm
mhH2
ml
lm
mm
lm
mhH3
hm
ml
lm
mh
mm
ll
mH
package contour
lhH3
mhH3
hh
mH
period contour
lH
frequency (Hz)
350
300
250
200
150
100
50
0
7.5
8
8.5
9
9.5
10
10.5
time (s.)

C’est fini !