View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Hal-Diderot

tu veux couper là faut dire pourquoi - Propositions pour une segmentation syntaxique du français parlé

Christophe Benzitoun, Anne Dister, Kim Gerdes, Sylvain Kahane, Paola Pietrandrea, Frédéric Sabio

To cite this version:

Christophe Benzitoun, Anne Dister, Kim Gerdes, Sylvain Kahane, Paola Pietrandrea, et al.. tu veux couper là faut dire pourquoi - Propositions pour une segmentation syntaxique du français parlé. Congrès Mondial de linguistique française, Jul 2010, New Orleans, États-Unis. CMLF, pp.2075-2090, 2010. <hal-00576854>

HAL Id: hal-00576854

https://hal.archives-ouvertes.fr/hal-00576854

Submitted on 15 Mar 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

tu veux couper là faut dire pourquoi
Propositions pour une segmentation syntaxique du français parlé

Benzitoun, Christophe

ATILF, Nancy Université & CNRS
Christophe.Benzitoun@atilf.fr

Dister, Anne

Facultés universitaires Saint-Louis et Université de Louvain
anne.dister@uclouvain.be

Gerdes, Kim

LPP, Sorbonne Nouvelle & CNRS
kim@gerdes.fr

Kahane, Sylvain

Modyco, Université Paris Ouest Nanterre & CNRS
sylvain@kahane.fr

Pietrandrea, Paola

Université Roma TRE / Lattice, ENS & CNRS
pietrand@uniroma3.it

Sabio, Frédéric

LPL, Université de Provence & CNRS
frederic.sabio@orange.fr

1 Introduction

Cet article s'intéresse à une question théorique majeure : la segmentation de transcriptions de français parlé en unités syntaxiques fondamentales. A ce jour, cette question n'a pas encore trouvé de réponse satisfaisante. Du côté de l'analyse syntaxique de l'écrit, on se tient habituellement à la ponctuation pour laquelle les pratiques sont relativement fluctuantes et n'obéissent pas à des critères stables. En effet, lorsque l'on ponctue un texte à l'écrit, on a vraisemblablement recours à des critères syntaxiques, sémantiques ou prosodiques (et d'autres encore) de façon variable selon les auteurs. Ce présupposé est donc déjà problématique pour l'analyse syntaxique de l'écrit. Mais pour l'oral spontané, la question se pose de manière encore plus aigüe en raison d'une architecture syntaxique moins bien connue et de l'absence de tradition unifiée de ponctuation, toutefois en partie compensée par les marqueurs prosodiques.

Nous avons abordé la problématique de la segmentation syntaxique de manière pratique étant donné que notre étude se situe dans le cadre d'un projet d'annotation syntaxique et prosodique de français parlé, le projet ANR Rhapsodie[1]. Dans ce papier, nous décrivons l'élaboration de critères reproductibles et opératoires utilisés pour la segmentation des transcriptions brutes du corpus Rhapsodie, critères qui ne prétendent pas régler l'ensemble des problèmes théoriques. Cette segmentation est la première étape indispensable pour une

annotation syntaxique exhaustive et pour faciliter l'analyse syntaxique automatique des textes. En définitive, nous présentons la segmentation proposée et quelques-uns des problèmes que nous avons rencontrés au cours de l'élaboration du manuel d'annotation.

Ce travail nous semble d'autant plus important qu'il permettra en plus un éclairage sur l'écrit dans lequel le lien entre segmentation graphique et unité syntaxique parait hégémonique. Or, à notre connaissance, il n'a jamais été démontré que les différentes sortes de ponctuation délimitaient le champ de la syntaxe. Au contraire, depuis quelques années, des études remettent en cause la pertinence de la notion de phrase (graphique) comme unité délimitant le champ d'investigation de la syntaxe (cf. notamment Charolles et Combettes, 1999 ; Béguelin, 2000 et 2002 ; Berrendonner, 2002 ; Blanche-Benveniste, 2002). Les auteurs de ces travaux proposent également des unités de substitution jugées plus opératoires que la phrase ainsi conçue, unités qui ont été essentiellement élaborées en s'appuyant sur des corpus oraux. L'oral est d'ailleurs un terrain particulièrement fertile pour mener ce type de réflexion, étant donné que le recours à la ponctuation pour présenter les transcriptions y est plus que problématique (pour des exemples en français, voir Blanche-Benveniste et Jeanjean, 1986). Et dès lors que l'on travaille sur le français parlé, c'est tout naturellement vers ces unités que l'on s'oriente, que l'on soit à la recherche d'unités « fonctionnelles » (Bilger et Campione, 2002) ou d'unités syntaxiques (Rossi-Gensane, 2007). Pour autant, tous les problèmes ne sont pas résolus et la segmentation systématique de corpus oraux représente encore un véritable défi. Pour l'aborder, nous avons décidé de nous inscrire dans un cadre distinguant deux modules, à savoir la micro- et la macrosyntaxe (Blanche-Benveniste et al., 1990 ; Berrendonner, 1990 ; Cresti, 2000 ; Andersen et Nølke, 2002) et de leur donner une indépendance assez forte. Il existe bien évidemment d'autres travaux portant sur la segmentation du français parlé. On peut citer, par exemple, les projets menés par le centre de recherche VALIBEL (Dister et al., 2008 ; Degand et Simon, 2005 et 2009). Mais même si la définition des unités est précisée, il n'en reste pas moins que les nombreux problèmes rencontrés au cours d'une telle tâche ne sont pas détaillés. C'est d'ailleurs une des principales originalités du présent article que de tenter d'une part de mieux délimiter ce que l'on aimerait appeler une unité syntaxique (contextuellement) maximale et, d'autre part, d'aborder les cas difficiles et la manière dont nous les avons résolus (au moins provisoirement). Ainsi, un petit détour par l'oral oblige à revoir certains réflexes et à se poser la question de savoir s'il n'est pas possible de définir une unité de traitement syntaxique plus opératoire que la phrase graphique, dont on rappelle qu'elle est utilisée sans discussion par de nombreuses théories syntaxiques et par les analyseurs automatiques de l'écrit basés sur des grammaires formelles (dont le travail d'analyse repose généralement sur une segmentation préalable en phrases). Notre travail revient donc par bien des côtés à redéfinir la ponctuation et à proposer des critères pour ponctuer de manière unifiée les corpus d'oral spontané.

Dans un premier temps (section 2), nous présentons nos deux unités d'analyse, à savoir l'unité rectionnelle (UR) et l'unité illocutoire (UI) en nous intéressant tout particulièrement aux cas où leurs frontières ne coïncident pas (cas de non-congruence). Nous étudions aussi différents types d'énoncés complexes, notamment les questions-réponses, l'instanciation, le discours rapporté, la greffe, les parallélismes inter-UI ou encore les UI discontinues (section 3).

2 Unités rectionnelles et unités illocutoires

A la différence de ce qui se passe pour l'analyse de l'écrit où l'on part généralement du découpage en phrases graphiques avant de rechercher les liens entre les éléments et d'aboutir éventuellement à la conclusion que certains éléments sont « détachés », nous recherchons d'abord les liens de dépendance, puis construisons les UR et regroupons celles qui vont ensemble. Nous adoptons ainsi une approche de bas en haut et non de haut en bas. Une UR est une unité construite autour d'une tête, qui n'est syntaxiquement dépendante d'aucun élément de rang supérieur dans un texte ou discours donné. De cette tête dépend un ensemble d'éléments. Il s'agit donc d'une unité basée sur des contraintes imposées par les catégories grammaticales. La rection se

caractérise par les contraintes imposées sur une position donnée en termes de parties du discours, de marques morphologiques et de possibilités de restructuration (commutation avec un pronom, effacement, passivation, clivage, etc.).

Il est important de souligner le fait que les UR ne sont pas définies dans l'absolu. C'est seulement dans un texte donné que l'on peut dire que certaines constructions ne dépendent d'aucune catégorie du cotexte. Ce que nous prenons comme unité de découpage d'un texte, ce sont donc les constructions qui, dans ce texte, ne dépendent d'aucune catégorie grammaticale.

En face de cette première unité, il y a l'UI dont la délimitation est liée à la reconnaissance de la force illocutoire qui peut affecter un segment dans un texte donné. Il peut paraitre étrange de trouver une telle unité dans un article sur la syntaxe, mais elle nous a semblé incontournable. Nous rejoignons en cela par exemple Creissels (2004) qui dans sa présentation des *notions de base de l'analyse syntaxique* définit la « phrase » en termes de « contenu propositionnel » et d'« opération énonciative », ce qui est très proche de la présente unité. UR et UI sont deux unités relativement autonomes qui ont leurs propres règles de formation et leurs propres combinatoires. L'UR est en fait l'unité maximale de la microsyntaxe alors que l'UI fait partie du domaine de la macrosyntaxe. Et conformément à Blanche-Benveniste et al. (1990), nous considérons que ces deux modules de l'analyse syntaxique sont complémentaires mais que la sortie de l'un ne constitue pas l'entrée de l'autre. Nous verrons que les UI sont constituées d'UR variées, allant de la simple interjection à des constructions complexes à plusieurs enchâssements. Les UI combinent généralement plusieurs UR, mais leurs frontières ne coïncident pas forcément. En effet, il arrive qu'une UR se prolonge au delà d'une UI.

2.1 Microsyntaxe et UR

La *microsyntaxe* vise à décrire des constructions syntaxiques conçues comme des ensembles rectionnels complets (voir également les clauses de Berrendonner (1990, 2002), Delais-Roussarie & Choi-Jonin (2004)). Nous parlons d'unité rectionnelle plutôt que d'ilot, car en pratique, dans les textes, un ensemble d'éléments reliés par des relations de rection ne forment pas toujours des ilots, au sens d'ensemble d'éléments adjacents dans la chaine parlée. Il y a toutes sortes d'insertions possibles : parenthèses, greffes, etc.

On définit donc les UR de manière volontairement restreinte : un élément constructeur (de catégorie variée) entouré d'unités qui dépendent de lui. A ce niveau, le principe consiste à segmenter dès lors que l'on ne peut plus effectuer de rattachement microsyntaxique à l'intérieur du texte. Par exemple, chacune des séquences suivantes peut former une unique UR :

(1) je me levais le matin [CRFP]
(2) j'étais avec des clients [CRFP]
(3) tu refais pas des chênes centenaires
(4) une affaire douteuse qui a mal fini

Dans (1), par exemple, il y a un verbe constructeur (*se lever*) qui construit deux éléments, à savoir un sujet (*je*) et un complément (*le matin*). Nous appelons *UR Verbale* (URV) toute unité de rection complète dont la tête est un verbe fini. Dans (4), nous avons affaire à une *UR Nominale*.

Les *associés* (recouvrant en partie les *compléments de phrase*) n'étant pas régis par une catégorie grammaticale, nous en faisons des UR à part entière. Ainsi dans :

(5) à mon avis c'est tout à fait l'inverse [Corpaix]

il y a deux UR, *à mon avis*, d'une part, et *c'est tout à fait l'inverse*, d'autre part. Nous séparons à l'aide de < ces deux UR (voir dans la section 2.4 la sémantique plus précise de ce délimiteur). Nous traitons de la même façon des énoncés moins canoniques :

(6) moi < ma famille < j'avais que ma mère quand j'habitais là [Corpaix]

Dans l'exemple précédent, nous considérons qu'il y a trois UR.

2.2 Unité illocutoire

On appelle *unité illocutoire* une portion de discours comportant un unique acte illocutoire, soit une assertion, soit une interrogation, soit une injonction[2]. Il y a derrière chaque UI un acte de langage que l'on peut mettre en évidence par l'introduction d'un segment recteur comme « je te dis », « je te demande », « je te conseille », « je te supplie », etc. (Récanati, 1979). Un test pour le découpage réside donc dans la possibilité d'insérer de tels segments (voir section 2.4). Le découpage en UI nous semble essentiel dans le cadre d'un projet d'annotation intono-syntaxique, car ces unités sont prosodiquement marquées (Blanche-Benveniste, 1997 ; Cresti, 2000).

Dans leur définition de la phrase, de nombreux auteurs tablent sur la correspondance entre unité illocutoire et unité syntaxique. Ainsi pour Le Goffic (1993 : 8), la phrase constitue à la fois un « acte de discours » tout en pouvant être décrite comme « le niveau supérieur de la syntaxe ». Pour Riegel et al. (1994: 26) : « Une phrase donnée est une entité structurale abstraite que l'on peut caractériser par un ensemble de règles de bonne formation phonologique, morphologique et sémantique. Elle se réalise sous la forme concrète d'énoncés. » Et on peut en dire autant de Creissels (2004). Pour ces auteurs, les règles de bonne formation des « phrases » correspondent au respect des relations de rection. Les phrases sont des UR qui se réalisent concrètement dans un énoncé (c'est-à-dire en étant pourvues d'une force illocutoire, ce qui correspond bien à notre notion d'UI). La différence fondamentale avec notre approche, c'est qu'ils décident généralement de nommer ces unités fondamentales « phrase » et privilégient dans l'analyse les UR qui sont des projections rectionnelles de verbe fini et qui, dans leurs approches, coïncident avec une UI. Pour notre part, nous considérons plutôt que le domaine où s'appliquent les règles de bonne formation de la grammaire, au sens usuel de ce terme, est l'UR dont l'extension peut différer de l'UI[3]. Ainsi, dans notre conception, UR et UI ne se correspondent pas toujours. Nous allons voir dans la suite de cette section qu'il existe des UI comprenant plusieurs UR. Nous verrons dans la section 3 qu'il existe à l'inverse des UR qui dépassent la frontière de l'UI.

Nous utilisons pour l'instant le délimiteur // pour découper le texte en unités illocutoires (voir également section 3 les délimiteurs //+ et //=). En voici quelques exemples :

(7) a. **ouais** // **voilà** // ça c'est c'est toujours le même schéma //
b. hein c'est c'est c'est c'est comme si tu mettais un poids dans la balance // **vlam** //
d'un seul coup tout tout tout il y en a un qui monte à toute vitesse et puis l'autre qui descend à toute vitesse // alors ça ça ça c'est ça c'est c'est la cata //
c. le lendemain **bombe** //

Notons qu'il existe des unités illocutoires qui ne contiennent pas d'URV, à l'image des exemples ci-dessus (partie en gras).

2.3 Entassement

Dans l'exemple (7b), nous voyons également que nous considérons que, tant qu'une UR n'est pas achevée, il s'agit de la même UI. Nous distinguons néanmoins deux cas :

- le cas où le locuteur finit par achever tous les segments amorcés et où l'on peut alors regrouper ensemble les segments qui se poursuivent de la même façon, comme dans :

  (7) b. alors { { ça | ça | ça } c'est | ça { c'est | c'est } } la cata //

- le cas où un segment n'est jamais achevé, ce que nous indiquons par le symbole & :

(7) b. d'un seul coup { { tout | tout | tout } & | il y en a un qui monte à toute vitesse }

Les délimiteurs { | } utilisés dans cet exemple encodent ce que l'on appelle un *entassement* (Gerdes et Kahane, 2009) dans le prolongement des listes paradigmatiques et de l'analyse en grille proposées dans Blanche-Benveniste et al. (1990). Ce dispositif est une extension de la notion de place rectionnelle et relie des éléments à l'intérieur d'une UR qui occupent la même position syntaxique. Ceci peut être dû à une « disfluence » (comme dans la dernière UI de (7b)), une reformulation ou bien encore une simple coordination (cf. (8d))[4]. Il est souvent difficile de distinguer entre disfluence involontaire et reformulation (8a et b) ou encore entre reformulation et coordination (8b et c) :

(8) a. elle est { infirmière | euh médecin }
b. elle est { infirmière | peut-être médecin }
c. elle est { infirmière | ^ou médecin
d. {mais | mais | mais | mais} {**les lois sociales | le droit de grève | ^et tout ça**} <
{ça s'est fait | ça s'est fait} sur {**des dizaines | ^et des dizaines**} d'années //

Dans tous ces cas, nous utilisons la même notation afin de rendre compte du phénomène d'entassement paradigmatique : les différents segments qui viennent occuper la même place syntaxique sont entourés par des accolades et séparés par des barres verticales. On appelle ces segments les *couches* de l'entassement. Les conjonctions de coordination sont précédées du symbole ^ (accent circonflexe), car elles jouent uniquement un rôle dans l'entassement et ne doivent pas être prise en compte lors du « dépliage » des entassements (Gerdes & kahane 2009, Benzitoun et al. 2009). En revanche, nous n'annotons pas les autres marqueurs d'entassements tels que les *adverbes paradigmatisants* (Nølke 1983) qui jouent également un rôle par rapport au contexte de l'entassement. Dans le dépliage des exemple (8b) et (8c), il existe une différence évidente entre la conjonction de coordination *ou* et l'adverbe paradigmatisant *peut-être* :

(9) a. elle est peut-être médecin
b. *elle est ou médecin

Rection plus entassement nous donnent une structure syntaxique connexe. Dans les exemples (8a-c), la première couche occupe la position d'attribut et la relation syntagmatique entre les deux couches induit une relation paradigmatique entre *infirmière* et *médecin*. La deuxième couche, quant à elle, hérite de cette même position syntaxique : *est* —attr→ *infirmière* | *médecin*. Notons encore que le choix d'une seule couche (non abandonnée) pour chaque entassement nous donne en général une construction grammaticale avec une structure syntaxique simple (ne contenant que des liens de rection et de modification).

La connexité de la structure syntaxique d'un énoncé est à nouveau remise en question par les phénomènes macrosyntaxiques, comme les éléments détachés, décrits dans la section suivante.

2.4 Macrosyntaxe, noyau, pré- et post-noyau

La *macrosyntaxe* vise à décrire les regroupements d'unités, et notamment d'UR, au sein d'une même UI (nous verrons en 3.1 qu'elle permet aussi de décrire les dégroupements d'UR que l'on observe dans les textes). Par exemple, les UR (1) et (2) ci-dessus sont en fait insérées dans une unique UI :

(10) je me levais le matin j'étais avec des clients [CRFP]

et (3) s'insère également dans un complexe plus large formant UI :

(11) un arbre euh si la forêt est détruite hein tu refais pas des chênes centenaires hein

Dans (10), une commerçante travaillant dans un magasin qui vend des produits oléicoles explique qu'elle est habituée depuis toute petite à être en contact avec la clientèle, car ses parents tenaient un hôtel-restaurant.

Le lien entre les deux unités rectionnelles verbales dans (10) et entre *un arbre* et *tu refais pas des chênes centenaires* dans (11) ne peut pas être décrit en termes de dépendance microsyntaxique. En effet, *je me levais le matin* n'est pas construit par le verbe *être* (et réciproquement) et un *arbre* ne dépend pas non plus microsyntaxiquement du verbe *refaire*. Pourtant, on peut reconnaitre un lien macrosyntaxique. La première UR en (10) *je me levais le matin*, aussi bien que *un arbre* en (11) ne sont pas des unités *autonomes* d'un point de vue illocutoire. Deux UR reliées au plan macrosyntaxique ne s'entassent pas, ce qu'on vérifie en remarquant la différence d'interprétation lorsque l'on réitère la conjonction dans (12) :

(12) a. je lui ai dit que je me levais le matin j'étais avec des clients
b. je lui ai dit que je me levais le matin et que j'étais avec des clients (sens différent)

La force illocutoire de (10) est portée par l'UR *j'étais avec des clients* qui permet d'interpréter l'énoncé entier comme une assertion. De même, en (11) c'est l'unité *tu refais pas des chênes centenaires,* interprétable comme une assertion même en isolation, qui porte la force illocutoire de l'énoncé. Les unités porteuses de la force illocutoire sont le plus souvent autonomisables dans le contexte précis où elles sont employées. Cela veut donc dire qu'avec une intonation identique on peut ne garder que l'unité porteuse de la force illocutoire et supprimer les éléments périphériques. Ces unités ont un statut important dans notre approche.

On appelle *composantes illocutoires* (CI) les différentes composantes d'une UI. On appelle *noyau* la CI qui porte la force illocutoire. L'une des caractéristiques du noyau est ainsi de pouvoir être nié ou interrogé. Plus précisément, c'est lui qui est le véritable destinataire d'une négation ou d'une interrogation qui serait portée par l'UI. Les paraphrases en (13) illustrent cela :

(13) ce n'est pas vrai que je me levais le matin j'étais avec des clients ≈ je me levais le matin je n'étais pas avec des clients

Les CI qui précèdent et suivent le noyau sont appelées respectivement des *pré-noyaux* et des *post-noyaux*. Nous les séparons en utilisant le délimiteur < pour les pré-noyaux et le délimiteur > pour les post-noyaux :

(14) il y a plein de trucs < tu les vois après > en fait > les défauts [C-ORAL-ROM]

Le noyau d'une UI est toujours une unité microsyntaxique, mais il ne s'agit pas toujours d'une UR. Il se peut que, en raison d'une structure communicative bien particulière, la force illocutoire soit portée par une partie seulement d'une UR :

(15) a. à ma mère <+ je ne parle plus //
b. deux euros >+ ça a couté //

Dans (15b), c'est *deux euros* qui porte la force illocutoire, raison pour laquelle nous avons employé le symbole >+ et non <+. *Deux euros* est donc bien le noyau[5]. L'ajout du symbole + indique que les CI de part et d'autre font partie de la même UR. Ainsi, cela nous permet de distinguer les exemples en (15) du suivant :

(16) la moindre contrariété < je suis angoissé //

3 Énoncés complexes

Nous allons présenter un certain nombre d'exemples qui nous ont posé problème dans le cadre de l'annotation syntaxique du corpus Rhapsodie et qui illustrent la non correspondance entre frontières d'UR et d'UI. Nous allons également étayer les principes qui nous permettent de décider où sont les frontières de ces unités.

3.1 UR au delà de l'UI

Nous avons vu jusqu'ici quelques exemples de segmentation d'une UI en plusieurs UR. Nous allons montrer ici qu'il y a des cas, traditionnellement nommés *épexégèses* ou *complément différés*, pour lesquels on peut

considérer que c'est au contraire une UR qui est segmentée en plusieurs UI. Considérons les deux exemples suivants (L1 et L2 indiquent les tours de parole de deux locuteurs et ? signalent le fait qu'il s'agit d'une question) :

(17) a. L1 il a jeté le livre
L2 dans la poubelle ?
b. il parle anglais et bien

Dans ces deux exemples, il y a deux actes illocutoires : en (17a) c'est évident puisqu'il y a deux tours de parole et que l'un est une assertion et l'autre une interrogation ; en (17b) il y a deux assertions et seul cela peut expliquer l'usage du *et* entre deux syntagmes de catégories si différentes. Dans les deux cas, il n'y a pas autonomie microsyntaxique de la construction à la base du deuxième acte illocutoire. Celle-ci peut en effet enchaîner avec la construction précédente pour former une construction microsyntaxique canonique du type verbe + ajout. On peut ainsi paraphraser les exemples précédents par :

(18) a. L1 il a jeté le livre
L2 il l'a jeté dans la poubelle ?
b. il parle anglais et (en plus) il parle bien anglais

Dans un tel cas, lorsque l'UI n'a pas d'autonomie microsyntaxique et qu'elle peut (voir ci-dessus) s'adosser à une autre UI, nous considérons qu'elle appartient à la même UR. Comme précédemment, nous indiquons que la frontière illocutoire n'est pas une frontière d'UR en ajoutant un +. Entre deux morceaux séparés par //+, on a donc toujours une relation syntaxique, rection ou entassement. Voici l'annotation que nous proposons pour les deux exemples ci-dessus :

(17) a. L1 il a jeté le livre //+
L2 dans la poubelle //
b. il parle anglais //+ et bien //

Dans l'exemple suivant (emprunté à Debaisieux, 2007), une UI à valeur interrogative vient s'entasser sur l'UI précédente de son interlocuteur et la réponse à la question poursuit encore l'UR :

(19) L1: ils avaient honte par rapport {aux Marseillais |} //+
L2: {| aux Marseillais } //+
L1: parce qu'ils parlaient pas le même provençal qu'eux // [Corpus Debaisieux]

Nous proposons une telle analyse pour les structures du type question-réponse. Alors que pour des exemples tels que (20a) on envisage traditionnellement que la réponse est elliptique, nous considérons pour notre part qu'il n'y a pas d'ellipse, ni de fragment autonome mais que la réponse forme avec elle une seule et même UR. Plus précisément, la réponse contient un segment qui s'entasse sur le pronom interrogatif et éventuellement d'autres éléments qui complètent l'UR. On peut ainsi extraire de cette UR complexe un chemin équivalent à (20b).

(20) a. L1 : {comment |} il s'appelle //+
L2 : {| Coluche} je crois //
b. L2: il s'appelle Coluche je crois

Les questions-réponses ne sont pas les seuls phénomènes où une UI vient s'entasser sur un segment de la précédente UI et l'instancier. On observe également cela avec des groupes indéfinis :

(21) vous avez donné {quelque chose de plus|} à la femme //+ {|des armes de persuasions} // [Rhapsodie]

Considérer qu'une construction microsyntaxique s'arrête nécessairement avec l'arrivée d'un nouveau tour de parole ou après une rupture intonative forte, comme on le fait généralement, est un choix axiomatique qui n'est pas justifié empiriquement dans l'état de nos connaissances. Nous adoptons un axiome différent qui ne

suppose pas de congruence a priori entre UR et UI, la charge de la preuve étant à ceux qui imposent cette contrainte supplémentaire sur les composantes de la description syntaxique.

3.2 Constructions introduites par des conjonctions : la question de l'intégration syntaxique

L'analyse des séquences introduites par une conjonction de subordination pose problème depuis longtemps, à la fois parce que les données orales non planifiées présentent des fonctionnements syntaxiques et discursifs qui n'ont pas été clairement dégagés dans le cadre phrastique des grammaires traditionnelles, et parce que le recours généralisé à la notion classique de « subordination » tend à unifier de façon artificielle plusieurs types de configurations syntaxiques qu'il conviendrait au contraire de distinguer soigneusement (voir Blanche-Benveniste, 2002 ; Benzitoun, 2006 ; Debaisieux, 2007 pour un examen de certaines difficultés). Pour notre entreprise d'annotation de l'oral, il nous paraît indispensable de distinguer entre les subordonnées « classiques », qui sont sous la dépendance grammaticale d'un verbe, et celles qui présentent un fonctionnement syntaxique et pragmatique différent.

Les subordonnées régies

Dans des exemples comme :

(22) nous avons vu un crépuscule *alors que nous étions au sommet de la mosquée* //
(23) tu aimais la poésie *parce que justement tu pouvais la choisir* //
(24) on le reprendrait *s'il conjuguait mal ses verbes* //

La séquence conjonctionnelle (en italiques) constitue une « subordonnée » au sens tout à fait classique du terme : elle a le statut d'un élément régi par le verbe constructeur (fonction d'ajout) ; elle est intégrée à l'UI en cours et appartient au noyau macrosyntaxique. A ce titre, elle ne fait l'objet d'aucune annotation particulière.

Parmi les indices qui conduisent à analyser ces séquences comme régies par le verbe, on peut mentionner (à la suite de Blanche-Benveniste et al., 1990) :

- la sensibilité aux modalités portées par le verbe (ex : on *ne* le reprenait *que* s'il conjuguait mal ses verbes) ;
- la reformulation dans d'autres dispositifs de la rection [clivage, « si-dispositif » pour les causales] (ex. : *c'est* alors que nous étions au sommet de la mosquée *que* nous avons vu un crépuscule, *si* tu aimais la poésie *c'est* parce que justement tu pouvais la choisir) ;
- la possibilité de listage paradigmatique (ex. : tu aimais la poésie parce que tu pouvais la choisir *et pas parce que tu aimais réciter les textes*) ;
- la possibilité de faire précéder la conjonction par un adverbe à effet paradigmatisant (on le reprenait *uniquement* s'il conjuguait mal ses verbes) ;
- la proportionnalité avec une proforme (ex. : nous avons vu un crépuscule *à ce moment-là*).

Notons que les séquences conjonctionnelles régies peuvent avoir deux valeurs macrosyntaxiques distinctes :

- Si elles forment un élément pré-noyau, on indiquera leur statut macrosyntaxique par « < » et leur statut d'élément régi par « + » :

(25) *si ça va pas* <+ je retourne chez mes parents // [Corpaix]

- Si elles constituent le noyau à elles seules, on note « > » pour donner leur statut macrosyntaxique, et « + » pour signaler l'existence d'un lien de rection (cf. l'exemple 15.b pour un même type d'annotation) :

(26) L1 vous allez aller chercher des champignons //
L2 *seulement s'il fait beau* >+ on ira //

Les subordonnées non régies

Par définition, celles-ci ne passent pas les tests de la rection rappelés plus haut. On différencie deux types :

- les subordonnées non régies qui apparaissent à l'intérieur d'une UI : elles ne portent aucune valeur illocutoire propre. Elles sont soit antéposées à la construction, soit aisément antéposables. On les note soit par < soit par > selon leur statut de pré- ou post-noyau. Par exemple :

(27) a. *comme il est bientôt 8 heures* < il faudrait se dépêcher //
b. *si tu as soif* < il y a de la bière dans le frigo //
c. *il y a de la bière dans le frigo* > si tu as soif //
d. *il a dû pleuvoir* > puisque la chaussée est mouillée //

Le second type de subordonnée non régie est plus délicat à repérer : il s'agit des cas où une conjonction n'instaure aucun lien de rection et semble constituer avec la séquence qui suit une unité illocutoire autonome. En voici quelques exemples dans les extraits suivants :

(28) a. ^et ^puis là < alors < on a pu y rester tant qu'on a voulu // il faisait pas chaud // ^et ^puis il y avait moins de monde aussi // *^parce ^que l'Acropole < pour voir une colonne < "tu vois" il faut écraser les gens* // poussez-vous un peu s'il vous plaît // c'est abominable // on est obligé de faire la queue // nous < la queue < on l'a pas fait // [Debaisieux, 2006]
b. ce film n'a pas du tout fonctionné >+ en France tout du moins // *^parce ^que en Amérique < beaucoup de gens sont allés le regarder* // [Japon]
c. généralement < les mâles sont aussi plus beaux et plus colorés dans la plupart des espèces // *^bien ^que chez les poissons comme les Trichogaster leeri < ils sont exactement pareils* // [aquario]
d. on se fait régulièrement reprendre quand on fait une faute de grammaire ou une faute de syntaxe en français // *^alors ^que en en sicilien en tout cas (puisque je parle plus facilement le sicilien que l'italien) < "euh" on peut on peut se permettre de faire des fautes de grammaire* // [Corpaix]

Il est évident qu'au plan sémantique, le morphème *parce que* de l'exemple (28a) n'a pas pour fonction d'indiquer un quelconque lien causal entre l'état de chose représenté dans la construction qui précède (*il y avait moins de monde*) et l'état de chose représenté dans la construction qui suit (*pour voir une colonne il faut écraser les gens*). Au plan grammatical, la séquence conjonctionnelle de ces exemples n'est pas sous la dépendance syntaxique d'un verbe recteur. En revanche, elle porte sa valeur illocutoire propre. On le voit notamment par la possibilité de gloser la séquence conjonctionnelle avec un verbe de parole, ce qui montre qu'on a une énonciation autonome :

(29) a. si je dis ça c'est parce qu'en Amérique, beaucoup de gens sont allés le regarder
b. je dis ça bien que chez ces poissons ils sont exactement pareils

Ces UI à introducteur conjonctionnel (que nous annotons par un ^ comme pour les conjonctions de coordination) constituent des sortes d'assertions secondes, forcément branchées sur une assertion qui précède ; c'est pourquoi il n'est jamais possible de les trouver en tête avec une valeur de pré-noyau :

(30) * parce que en Amérique beaucoup de gens sont allés le regarder < ce film n'a pas du tout fonctionné en France //

S'il est clair que ces « fausses subordonnées » forment des UI indépendantes et ne sont pas régies par un verbe au même titre que les subordonnées en (22) et suivantes, on peut néanmoins considérer qu'elles entretiennent une certaine dépendance syntaxique avec l'UI qui précède. Elles ne portent pas sur le contenu de l'assertion qui précède mais sur l'assertion elle-même, comme le montre les paraphrases en (29). Nous avons décidé de ne pas encoder cette forme particulière de dépendance. Elle peut être récupérée dans notre corpus par l'étude des UI qui sont introduites par une « conjonction de subordination ».

3.3 Coordination d'UR

Lorsqu'on a des coordinations d'UR verbales, notamment dans des séquences narratives, on peut hésiter à considérer qu'il s'agit de plusieurs UI, plutôt que d'une grande UI. Considérons un exemple :

(31) j'étais à l'école juste à côté puis un peu plus vers rue Saint-Dominique donc c'est les quartiers la la partie que je connais le mieux et puis on va dire que euh après à part à part euh le le quand je sors je vais surtout vers euh vers les endroits où où j'ai où j'ai mes racines

Dans un tel cas, nous décidons de faire de chaque conjoint une UI séparée, dans la mesure où chacun pourrait faire l'objet d'une assertion séparée et que chacun porte ainsi sa propre force illocutoire. Voici donc la segmentation proposée :

(32) j'étais à l'école { juste à côté |} //+
^puis {| un peu plus vers rue Saint-Dominique } //
^donc c'est { les quartiers | { la | la } partie } que je connais le mieux //
^et ^puis on va dire que ["euh" après { à part | à part } "euh" {{le | le} & | quand je sors } < je vais surtout {vers "euh" | vers} les endroits {{où | où} j'ai | où j'ai} mes racines] //

On notera que le deuxième segment, bien que formant une UI, n'est pas autonome syntaxiquement : nous considérons qu'il s'entasse sur le complément du premier (*juste à côté*) et appartient à la même UR. On notera également dans cet exemple un pré-noyau qui se trouve dans une complétive, c'est-à-dire dans un enchâssement microsyntaxique. Nous indiquons de tels enchâssements par l'utilisation des crochets [ ]. On a ainsi une CI[6] marquée comme pré-noyau et délimitée par [et <. Par ailleurs, comme nous l'avons vu dans la section précédente, nous marquons par ^ les *introducteurs* d'UI. Il s'agit de conjonctions qui occupent la première position d'une UI et qui dans leur fonctionnement s'apparentent davantage à des marqueurs de pile qu'à des pré-noyaux.

Dans certains cas néanmoins, une coordination de deux UR ne peut pas être segmentée en deux UI :

(33) et puis sinon s'il est pas content < comme comme disait euh Chevènement < un ministre hein < ça ça obéit ou ça ferme & enfin ça ferme sa gueule // et puis ça obéit //

En effet, en raison de la portée large de *s'il est pas content*, on ne peut pas paraphraser par deux interventions successives de *je te dis* : *je te dis « s'il est pas content un ministre ça obéit » je te dis « ou ça ferme (sa gueule) »*. En revanche, les reformulations qui suivent, bien que toujours dans la portée de *s'il est pas content*, sont considérées comme des énoncés indépendants.

Les corrélatives comme (34) présentent un cas particulier intéressant d'entassement d'UR.

(34) plus tu manges plus tu grossis

Contrairement aux coordinations qui précèdent, il s'agit bien d'une unique UI, car aucune des deux UR ne peut être énoncée sans l'autre et aucune des deux UR n'est porteuse à elle seule de la force illocutoire. Enfin les éléments introducteurs *plus ... plus ...* s'apparentent aux autres marqueurs d'entassement que sont les

conjonctions de coordination par leur placement obligatoire en tête du conjoint (*ni…ni…, ou…ou*). Bien qu'on ne puisse pas dire qu'il s'agit de deux segments qui commutent dans une même position régie[7] comme dans les cas prototypiques d'entassement, nous considérons néanmoins qu'il s'agit bien d'un entassement par analogie constructionnelle. Nous proposons donc l'encodage suivant :

(35) {^plus tu manges | ^plus tu grossis } //

3.4 Discours direct

Le discours direct présente une difficulté particulière en raison d'un enchâssement d'actes illocutoires. Considérons l'exemple suivant :

(36) a. il a dit [casse-toi > pauvre con //] //
b. Marcel Achard écrivait [elle est très jolie // elle est même belle // elle est élégante //] //

Le discours rapporté dans ces exemples possède sa propre force illocutoire et [marque le début d'une UI. Néanmoins, le segment qui précède (*il a dit* ou *Marcel Achard écrivait*) ne forme pas un acte illocutoire autonome ni une UR complète. Ainsi, il y a enchâssement du discours rapporté à l'intérieur de l'énoncé complet. Dans la mesure où cette UI occupe une position microsyntaxique, nous utilisons comme pour l'exemple (32) les crochets [ ] pour marquer cette enchâssement microsyntaxique. Mais à la différence de (32), le segment enchâssé forme bel et bien une UI avec sa propre force illocutoire, ce que nous indiquons par le délimiteur d'UI // à la fin de ce segment.

3.5 Greffe

La greffe est un cas un peu limite d'UI. Il s'agit du procédé qui consiste à remplir une position syntaxique à l'aide d'une autre catégorie que celle attendue (Deulofeu, 1999). On est donc face à une rupture de sous-catégorisation. En général, cette rupture consiste en une sorte de commentaire périphrastique venant combler ou commenter un choix lexical :

(37) a. vous suivez la ligne du tram qui passe vers la [je crois que c'est une ancienne caserne //] je sais pas voilà // [Rhapsodie]
b. vous avez dit que euh [disons ma carrière pour simplifier //] témoigne de ma bonne conduite // [Rhapsodie]
c. tu as pas un emploi du temps avec euh [tel jour je fais ça // tel autre jour je fais ça //] // [Corpus Debaisieux]

C'est un peu comme dans le discours rapporté où une UI vient occuper une position régie à l'intérieur d'une UR. Nous le notons de la même manière que le discours direct, même si la greffe n'est pas un phénomène lexicalisé[8] et ne se produit pas dans des positions où elle est attendue, à la différence des discours directs qui se produisent dans des positions régies bien spécifique (comme l'objet du verbe *dire*). On peut observer des phénomènes particuliers où le locuteur poursuit simultanément son énoncé matrice et sa greffe en les entrelaçant (voir section 3.7).

3.6 Parallélismes entre UI

Notre approche de bas en haut nous a permis d'observer des regroupements d'UI assez fréquents et réguliers dans les transcriptions d'oral que nous avons eues à annoter. Ce seront les unités les plus complexes de notre proposition de segmentation. Nous utilisons le délimiteur //= pour noter ces regroupements d'UI. Nous en avons distingué 5 types :

(i) Répétition à l'identique de l'UI complète. Cette répétition a une fonction sémantique d'intensification du contenu propositionnel de la première UI:

(38) a. et puis voilà //= puis voilà //
b. ouais // **bah oui** //= **bah oui** //= **bah oui** //= **bah oui** //
c. faut faut faire ses preuves //= faut faire ses preuves //

(ii) Regroupements à valeur de confirmation composés par deux UI construites autour d'un même élément constructeur qui régit dans la première UI un élément atténué (angl. *hedged*) par une particule discursive (*quoi,* en (39a) ; *pratiquement* en (39b)) et, dans la seconde UI, le même élément présenté sans atténuation (et par conséquent confirmé) :

(39) a. ben oui // alors c'est c'est bon ben **c'est comme ça** quoi //= **c'est comme ça** //
b. dans les dix quinze ans < on va pas s'en apercevoir // mais en dix quinze ans < clac // ça ça va cogner // et **c'est irréversible** pratiquement //= à moins de ramener des chercheurs étrangers < **c'est irréversible** parce qu'on ne forme pas un chercheur en en cinq ans // [Rhapsodie]

L'exemple (39b) montre, comme dans les cas de parallélismes entre UI, que la répétition lexicale intéresse le noyau et qu'elle ne touche pas le vocabulaire du pré-noyau (cf. Blanche-Benveniste, 1997 : 128).

(iii) Regroupements d'UI parallèles à valeur d'approximation, construits autour d'un même élément constructeur qui régit dans les deux UI des éléments en relation de co-hyponymie entre eux. En (40), par exemple, deux UI sont construites autour de *ça s'est fait.* La première est modifiée par *sur des dizaines et dizaines d'années*, la seconde par le co-hyponyme *sur cinquante cent ans* :

(40) mais mais mais mais les lois sociales le droit de grève et tout ça < ça s'est fait **ça s'est fait sur des dizaines et des dizaines d'années** //= **ça s'est fait sur** presque enfin **cinquante cent ans** [Rhapsodie]

(iv) Regroupements d'UI synonymiques à valeur confirmative. Il s'agit d'un regroupement de deux UI construites autour d'un même élément constructeur affirmé dans la première UI et nié dans la seconde, qui régit dans les deux UI des éléments en relation d'opposition sémantique entre eux. Dans l'exemple (41), la première UI est construite autour de *c'est,* qui a comme attribut *comme ça* ; la seconde UI est construite autour de *c'est pas*, négation de *c'est* qui a comme attribut *autrement*, opposé de *comme ça*.

(41) puis on lui dit //+ **c'est comme ça** //= **c'est pas autrement** //

(v) Regroupements à valeur contrastive d'UI syntaxiquement parallèles. L'élément constructeur de la première UI a une relation sémantique (d'opposition, de synonymie, de co-hyponymie) avec l'élément constructeur de la seconde UI. Les éléments régis présentent également une relation sémantique entre eux. Des connecteurs adversatifs peuvent aider à renforcer la relation de contraste entre les deux UI. Dans (42) par exemple, le parallélisme concerne deux UI composées d'un pré-noyau et d'un noyau. Le pré-noyau de la première UI est en opposition sémantique avec le pré-noyau de la seconde (*recherche appliquée, recherche fondamentale*), les noyaux expriment des jugements opposés (*oui, moins*), le *mais* sert à renforcer le contraste entre les deux énoncés.

(42) en tout cas la recherche fondamentale < elle elle reste libre heureusement //= la recherche appliquée < moins //= mais la recherche fondamentale < oui // [Rhapsodie]

En outre, notre conception, exposée ci-dessus, entre en conflit avec la tradition grammaticale qui reconnait une coordination de propositions. Or, les notions de « proposition » et de « coordination » sont confuses car elles ne permettent pas de distinguer, en pratique, des entassements intra-UR et des successions d'UR ou d'UI

(voir également Béguelin 2000 pour une position proche). C'est la raison pour laquelle nous évitons autant que possible de recourir à ces notions.

3.7 UI discontinues

Une UI peut interrompre le déroulement d'une autre UI :

(43) je me levais tu vas rire le matin à cinq heures

Dans un tel cas, on considère que *tu vas rire* forme une UI insérée dans l'UI *je me levais le matin à cinq heures*. Nous proposons deux manières équivalentes de noter cette insertion. Soit en mettant l'UI inséré entre parenthèses :

(44) je me levais (tu vas rire) le matin à cinq heures //

soit en indiquant par le symbole # que l'UI se poursuit plus loin et reprend à la prochaine occurrence de # :

(45) je me levais # // tu vas rire // # le matin à cinq heures //

Les deux notations sont rigoureusement équivalentes — (=> #// et) => //# —, mais le symbole # permet également d'encoder des cas plus complexes comme le suivant, où une greffe donne lieu à un ajout ultérieur qui fonctionne ensuite comme une insertion, bien qu'en un sens il appartienne toujours à la même UI :

(46) on avait critiqué le le journal de [je crois que c'était le Provençal #] on l'avait critiqué par rapport à (# ou le Méridional //) par rapport à la mort de [comment il s'appelle //+ pas Coluche //+ l'autre //] // [Corpaix] (Blanche-Benveniste et al., 1990)

L'interaction entre locuteurs donne lieu aussi à de nombreuses discontinuités. Dans l'exemple suivant, le locuteur L1 est interrompu à trois reprises par son interlocuteur. Cela ne l'empêche pas de poursuivre un énoncé assez complexe, tout en interagissant avec son interlocuteur par des *ouais* qui ponctuent les interventions de L2. La séquence de délimiteurs //#+ indique que l'UI se termine (//) mais que l'UR continue plus loin (#+).

(47) L1: mais mais sinon euh bon & // en tout cas la recherche fondamentale < elle elle reste libre //#+
L2: ouais ouais //
L1: # heureusement //#+
L2: donc ouais // en XXX //
L1: ouais //
L2: faut voir après //
L1: ouais // # la recherche appliquée < moins //= mais la recherche fondamentale < oui //

4 Conclusion

La délimitation des unités maximales de la syntaxe a constitué la tâche la plus difficile dans les choix d'annotation syntaxique dans le cadre du projet Rhapsodie. Le travail pratique de découpage nous a amenés à considérer deux types d'unités — les unités rectionnelles et les unités illocutoires. Ces deux types d'unités obéissent à des principes d'organisation très différents, pris en charge respectivement par la micro- et la macrosyntaxe. Ces deux dimensions possèdent une relative indépendance et aucun des deux niveaux n'englobe l'autre.

Remarquons encore que la définition de l'unité rectionnelle nécessite de prendre en compte l'entassement. En effet, certains segments, comme les reformulations ou les réponses à une question partielle ne se rattachent ni par la rection, ni par une relation macrosyntaxique et n'ont pourtant aucune autonomie syntaxique.

L'un des objectifs du projet Rhapsodie est de livrer une annotation syntaxique complète. La segmentation présentée ici n'est qu'une première étape. La délimitation sert de prétraitement avant l'analyse automatique par des analyseurs développés pour l'écrit (cf. Benzitoun et al. 2009). Il nous reste également à faire le travail de recherche de corrélation entre ces délimiteurs et les marques prosodiques.

Nous espérons, à travers une tâche pratique d'annotation syntaxique, avoir contribué à faire avancer des questions théoriques fondamentales concernant les différents niveaux d'organisation syntaxique du discours et les différents modules syntaxiques en jeu.

Références bibliographiques

Andersen, H.L., Nølke, H. (Eds) (2002). Macro-syntaxe et macro-sémantique*. Actes du colloque international d'Århus*, 17-19 mai 2001. Bern: Peter Lang.

Béguelin, M.-J. (Dir.) (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*. Bruxelles : De Boeck & Larcier.

Béguelin, M.-J. (2002). Clause, période ou autre ? La phrase graphique et la question des niveaux d'analyse. M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase ?, Verbum 24 (1-2),* 85-107.

Benzitoun, Chr. (2006), *Description morphosyntaxique du mot* quand *en français contemporain*. Thèse non publiée, Université de Provence.

Benzitoun, Chr., Sabio F. (2009). Où finit la phrase ? Où commence le texte ? L'exemple des regroupements de constructions verbales. Journées d'étude *Ce que le texte fait à la phrase*, Caen, 3-4 décembre.

Benzitoun, Chr., Dister, A., Gerdes, K., Kahane, S., Marlet, R. (2009). annoter du des textes tu te demandes si c'est syntaxique tu vois. *Arena Romanistica, 4*, 16-27.

Benzitoun Chr., Debaisieux J.-M., Deulofeu J., Dister A., Gerdes K., Kahane S., Lefeuvre Fl., Pietrandrea P., Rossi-Gensane N., Sabio F., Victorri B. (en développement), *Guide d'annotation syntaxique Rhapsodie*, http://rhapsodie.ilpga.fr/wiki.

Berrendonner, A. (1990). Pour une macro-syntaxe. *Travaux de linguistique 21*, 25-31.

Berrendonner, A. (2002). Les deux syntaxes. M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase ?*, *Verbum*, *24 (1-2)*, 23-36.

Bilger, M., Campione, E. (2002). Propositions pour un étiquetage en "séquences fonctionnelles". *Recherches sur le français parlé, 17*, 117-136

Blanche-Benveniste Cl., (1997). *Approches de la langue parlée en français*. Paris: Ophrys

Blanche-Benveniste, Cl., Bilger, M., Rouget, Ch., van den Eynde, K. (1990). *Le français parlé. Etudes grammaticales*. Paris: Editions du CNRS

Blanche-Benveniste, Cl., (2002). Phrase et construction verbale. M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase?, Verbum 24 (1-2)*, 7-22.

Blanche-Benveniste, Cl., Jeanjean, C. (1986). *Le Français parlé. Transcription et édition*. Paris : Didier Érudition.

Bonvino, E., Masini, Fr., Pietrandrea P. (2009), List Constructions: a semantic network. *Troisième Conférence Internationale de l'AFLiCo*, Nanterre. Accessible à http://francescamasini.caissa.it/Presentations_files/parigi_draft.pdf.

Charolles M.), Combettes (B.). 1999. Contribution pour une histoire récente de l'analyse du discours. *Langue française, 121*, 76-116.

Creissels D. (2004). *Cours de syntaxe générale*. *Chapitre 1*, Hermès. Accessible à http://lesla.univ-lyon2.fr/sites/lesla/IMG/pdf/doc-346.pdf.

Cresti, E. (2000). *Corpus di italiano parlato*. Florence : Accademia della Crusca.

Debaisieux, J.-M. (2007). La distinction entre dépendance grammaticale et dépendance macrosyntaxique comme moyen de résoudre les paradoxes de la subordination. *Faits de Langue, 28,* 119-132.

Degand, L., Simon, A.C. (2005). Minimal Discourse Units: Can we define them, and why should we? *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories*, Biarritz, 65-74.

Degand, L., Simon, A. C. (2009). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours 4* (http://discours.revues.org/index.html).

Delais-Roussarie, E., Choi-Jonin I. (2004). Existe-t-il des indices intonatifs de segmentation en unités macro-syntaxiques ? *Actes de JEP-TALN*, Fès, Maroc.

Deulofeu, J. (1999). *Recherches sur les formes de la prédication dans les énoncés assertifs en français contemporain (le cas des énoncés introduits par le morphème* que*)*. Thèse d'état, Université Paris 3.

Dister, A. (2008). *Guide de codage pour le projet MDU. Partie syntaxique: découpage en unités de rection et en séquences fonctionnelles*, http://valibel.fltr.ucl.ac.be.

Dister, A., Degand, L., Simon, A. C. (2008). Approches syntaxiques en français parlé: vers la structuration en unités minimales du discours. *Proceedings of the 27th Conference on Lexis and Grammar*, L'Aquila, 10-13 September 2008, 27-34.

Gerdes, K., Kahane, S. (2009) Speaking in Piles. Paradigmatic Annotation of a Spoken French Corpus. *Proceedings of the fifth* C*orpus Linguistics Conference,* Liverpool.

Guénot Marie-Laure (2006). La coordination considérée comme un entassement paradigmatique : description, représentation et intégration. P. Mertens, C. Fairon, A. Dister et P. Watrin (Éds.), *Verbum ex machina*, *Actes de TALN 2006*, *Cahiers du Cental* 2 (2), Presses universitaires de Louvain, Louvain-la-Neuve.

Le Goffic, P. (1993). *Grammaire de la phrase française*. Paris : Hachette.

Nølke, H. (1983). *Les adverbes paradigmatisants : fonction et analyse*. Copenhague : Akademisk Forlag.

Récanati, F. (1979). *La transparence et l'énonciation : pour introduire à la pragmatique.* Paris : Seuil.

Riegel, M., Pellat, J.-C., Rioul, R. (1994). *Grammaire méthodique du français.* Paris : PUF.

Rossi-Gensane N. (2007). Quelles unités syntaxiques pour l'oral ? *PFC : enjeux descriptifs, théoriques et didactiques*, *Bulletin PFC, 7.*

Sabio Fr. (2006) L'antéposition des compléments dans le français contemporain : l'exemple des objets directs. *Lingvisticae Investigationes, 29 :1*. Fascicule spécial : *Ordre des mots et topologie de la phrase française*. K. Gerdes et C. Muller (éds.), 173-182.

---

[1] Projet financé par l'Agence Nationale de la Recherche (contrat ANR Rhapsodie 07 Corp-030-01, Corpus prosodique de référence du français parlé, dirigé par Anne Lacheret). Le guide d'annotation syntaxique et les corpus déjà annotés sont en ligne (Benzitoun et al., en développement), avant la livraison du corpus complet avec son système de requête.

[2] Nous ne faisons pas d'hypothèses sur une classification des actes illocutoires dans cet article. Nous considérons ici les trois modalités d'énonciation que reprennent la plupart des grammaires (cf. par ex. Riegel et al. (1994). L'assertion se caractérise par la possibilité de se voir attribuer une valeur de vérité. Cela inclut les exclamations (*Que ce livre est beau !*).

[3] Les UI obéissent également à des règles de bonne formation, mais d'une part celles-ci sont d'une autre nature et d'autre part elles n'ont pas fait l'objet d'une description aussi systématique, ni d'une réelle modélisation. Cet article peut être vu comme un pas dans cette direction.

[4] Cet exemple illustre néanmoins des valeurs sémantiques particulières de la coordination. S'il s'agit bien d'une coordination, la valeur de {les lois sociales | le droit de grève | ^et tout ça} ou de {des dizaines | ^et des dizaines} n'est pas au niveau sémantique une simple addition des conjoints, mais plutôt la construction d'un hyperonyme. Nous prévoyons dans le cadre du projet Rhapsodie un typage des entassements et notamment des coordinations, que nous ne présenterons pas ici. Cf. Guénot (2006), Gerdes & Kahane (2009) et Bonvino et al (2009) pour une ébauche de classification.

[5] Cf. Sabio (2006) pour l'analyse des exemples du type (15b).

[6] Par contre, le statut macrosyntaxique du segment qui régit le segment enchâssé (ici *on va dire que*) n'est pas encore clair pour nous. S'agit-il d'une CI ? Un des objectifs de la constitution du corpus annoté est de récupérer suffisamment de segments de ce type pour en dresser une typologie.

[7] Certaine constructions figées posent le même problème :

(i) a. il travaille *jour et nuit*
b. il s'est plongé *corps et âme* dans son travail

Dans ces exemples, aucun des deux conjoints ne peut occuper seul la place (*il travaille jour ; *il s'est plongé corps).

[8] Bien que certaines constructions comme *il a emprunté **je ne sais quel** parcours machiavélique* ou *il est parti **il y a** deux heures* sont probablement des greffes lexicalisées.