Approches syntaxiques en français parlé : vers la structuration en unités minimales du discours

Anne Dister, Liesbeth Degand, Anne Catherine Simon[1]
Université Catholique de Louvain

Abstract

L'objectif principal de notre projet est de définir les unités minimales du discours oral (dorénavant MDU, *minimal discourse units*), afin d'étudier leur contribution dans la structuration du discours. Nous avons développé une méthode pour identifier de manière non ambigüe les MDU dans des discours oraux. L'hypothèse sur laquelle se base notre recherche est que les MDU doivent être définies en fonction de deux types de critères linguistiques observables : des critères syntaxiques et des critères prosodiques. Dans cet article, nous décrivons uniquement le volet syntaxique de la recherche, en explicitant nos critères d'annotation syntaxique des données orales.

Keywords : données textuelles orales, annotation syntaxique, unité minimale du discours, unité de rection, séquence fonctionnelle

1. Introduction

En analyse du discours, il est généralement admis que le discours est hiérarchiquement structuré (Mann et Thompson 1988 ; Polany 1988 ; Roulet *et al.* 2001) : un fragment de discours est composé d'un ensemble de segments plus petits reliés les uns aux autres d'une manière cohérente. Ce qui diffère par contre d'un modèle à l'autre, c'est la manière dont on définit ces segments constitutifs des discours. Certains auteurs évitent le problème en ne définissant pas de manière stricte les éléments minimaux. D'autres se limitent à une définition pratique (parfois naïve) en vue d'une segmentation automatisée (pour une revue détaillée voir Passoneau et Litman 1997 : 105-108).

L'hypothèse sur laquelle se base notre recherche est que les MDU doivent être définies en fonction de deux types de critères linguistiques observables : des critères syntaxiques et des critères prosodiques (Degand et Simon 2005).

Notre méthodologie nous fait séparer strictement les plans syntaxique et prosodique lors d'une première analyse des données, pour confronter les résultats obtenus de manière indépendante dans un deuxième temps.

Dans cet article, nous décrivons uniquement le volet syntaxique de la recherche, en explicitant nos critères d'annotation syntaxique des données orales.

[1] (prénom.nom@uclouvain.be)

2. Des données textuelles orales

Les données sur lesquelles nous travaillons sont des données textuelles orales : il s'agit de transcriptions d'enregistrements effectués dans différentes situations de communication (cours ou conférence, journal ou interview radiophoniques, conversations spontanées entre amis, etc.).

Transcire du matériau sonore n'est pas chose aisée, et ne relève pas d'un simple travail de copiste (Blanche-Benveniste et Jeanjean 1987). Transcrire nécessite de faire une série de choix, qui bien souvent ont des répercussions sur les analyses ultérieures. Nos transcriptions suivent les conventions établies au centre de recherche Valibel (Dister *et al.* 2006). On peut les résumer comme suit : adoption de l'orthographe standard (pas de « trucage orthographique ») ; notation des phénomènes propres à la planification des énoncés oraux, souvent appelés *disfluences* (répétitions, amorces de morphèmes, *euh*, etc.) ; absence de ponctuation.

L'absence de ponctuation des transcriptions va de paire avec l'abandon de la notion de phrase à l'oral. En effet, il n'y a pas de correspondance stricte entre phénomènes prosodiques[2] et ponctuation écrite. À une petite pause ne correspond pas nécessairement une virgule à l'écrit, pas plus qu'à une pause plus grande, une ponctuation forte. Il n'y a pas de relation bi-univoque entre les deux. De plus si certains énoncés se laissent enfermer relativement facilement dans le moule de la ponctuation graphique, comment s'en sortir avec des productions dans lesquelles foisonnent répétitions de mots, inachèvements et autres phénomènes propres à l'oral qui se construit ?

Blanche-Benveniste et Jeanjean (1987 : 139) plaident pour des transcriptions non ponctuées : « La ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir. » Selon elles, en ponctuant, le transcripteur « suggèr[e] une analyse avant de l'avoir faite » (1987: 142).

C'est ainsi que la notion de phrase, dans les études de productions orales, a été abandonnée et que s'est posée la question de l'unité minimale pertinente à prendre en considération[3]. Dans le cadre de notre travail, nous avons choisi de considérer l'unité de rection.

3. Les unités syntaxiques

Nous nous inscrivons dans le cadre d'une grammaire de dépendance développée depuis les années 1970 par le groupe aixois de recherche en syntaxe (GARS) autour de Claire Blanche-Benveniste (Blanche-Benveniste *et al.* 1990). Il s'agit pour nous de segmenter le flux oral en séquences plus petites, que nous appelons *unités de rection*. Celles-ci sont ensuite elles-mêmes segmentées en séquences fonctionnelles. Les critères de découpage que nous avons développés sont reproductibles, et doivent pouvoir s'appliquer à n'importe quel énoncé de français parlé.

[2] Que ce soit les pauses (silencieuses ou pleines) ou les contours intonatifs (montants, descendants, ou complexes). Nous parlons ici d'oral non planifié et non d'oral obtenu à partir d'un texte lu.

[3] Voir l'ouvrage collectif dirigé par Berthoud et Mondada (2000) pour la position des différents auteurs sur le sujet ; voir aussi Béguelin (2000 et 2002), Simon (2001).

3.1. Les unités de rection

Une unité de rection (UR) est un composant syntaxique dans lequel des éléments régis s'articulent autour d'un élément recteur. En général, cet élément est un verbe tensé, mais il peut aussi s'agir d'un nom, d'un adjectif, etc.

Nous classons les UR parmi les 5 types suivants : les unités de rection complètes, les inachevées, les elliptiques, les averbales et les unités de rection complètes « plus ».

3.1.1. Les unités de rection complètes (URC)

Une unité de rection complète est une unité qui apparait comme achevée, syntaxiquement et sémantiquement : les compléments obligatoires sont présents, les séquences entamées sont complètes, etc. La taille de ces unités de rection peut être très variable, comme l'illustrent les deux exemples suivants.

je dors

cette journée de guerre n'a pas empêché les ministres des affaires étrangères de la ligue arabe de se réunir dans la capitale libanaise

3.1.2. Les unités de rection inachevées (URI)

Une unité de rection est inachevée soit parce qu'une des places obligatoires de la valence n'est pas instanciée, soit parce que l'un des syntagmes de l'UR (qu'il soit obligatoire ou non) est inachevé.

un jour je vais me / je suis pas bagarreur

j'aime pas cette fille que

Dans le 1er exemple, on attend un verbe à l'infinitif après le pronom *me* ; dans le 2e exemple, la place de la valence est instanciée mais laissée inachevée après le relatif.

Dans les cas d'autocorrection immédiate, on ne considère pas que l'UR est inachevée. Nous appelons *autocorrection (immédiate)* le phénomène langagier qui consiste pour un locuteur à énoncer un morphème suite à un autre morphème différent qui appartient à la même catégorie grammaticale. Ce deuxième morphème vise à corriger le premier morphème énoncé.

on il est parti

Le pronom sujet *on* est corrigé par un autre pronom sujet. On considère que l'ensemble de cette suite appartient à la même UR.

La correction peut concerner une suite de plusieurs morphèmes :

il a il a il a on a réussi

Les amorces de la construction sont intégrées dans la même UR.

Par contre, dans les exemples suivants, la suite corrigée n'appartient pas à la même catégorie que la séquence corrigée :

elle ne Marie ne va pas bien

c'est un le premier poème est un long poème

Dans ces 2 cas, on fait de la séquence amorcée une URI :

elle ne | Marie ne va pas bien

c'est un | le premier poème est un long poème

3.1.3. Les unités de rection elliptiques (URE)

Une unité de rection elliptique est une unité de rection dans laquelle l'un des éléments est omis, sans pour autant faire de cette UR une unité inachevée.

Les URE recouvrent deux grands cas de figure :

- le sujet est omis

faut pas que tu viennes

valait mieux se taire

- un parallélisme de construction fait qu'un (ou plusieurs) élément(s) n'est (ne sont) pas repris. Souvent, une coordination sépare l'URC et l'URE :

on est loin du minimalisme de la poésie | et | loin de l'hermétisme aussi → URC et URE

la circulation devrait connaitre des perturbations | et | être rétablie vers 19 heures → URC et URE

Notons que certains éléments présents dans l'énoncé bloquent une analyse en URE que l'on ferait sans la présence de ces éléments. Ainsi, la distinction que nous établissons entre les deux énoncés suivants qui seront analysés différemment à cause de la présence de la locution *à la fois*. Dans le 1er cas, on aura une URC et une URE, dans le second une seule URC. En effet, faire la même analyse que dans le 1er énoncé est impossible : sans *et fragiles*, le 1er segment est également une URE et non plus une URC, puisque la présence de *à la fois* indique une suite obligatoire à *être fort.*

nos liens sont forts et fragiles → nos liens sont forts | et | fragiles →URC et URE

*nos liens sont **à la fois** forts et fragiles* → URC

3.1.4. Les unités de rection averbales (URA)

À la différence d'une URC, le noyau d'une URA n'est pas un verbe conjugué mais un nom, un pronom, un adjectif, un adverbe, un verbe à l'infinitif ou encore une interjection.

formidable

heureusement

moi ?

Une URA peut recevoir une expansion dans laquelle on a un verbe tensé :

ce gamin qui n'arrive jamais à rien !

3.1.5. Les unités de rection complètes « plus » (urc+)

Une unité de rection complète « plus » est une unité de rection complète qui contient des marqueurs de discours ou des séquences ajointes (cf. 3.2.5.).

3.2. Les séquences fonctionnelles

Le découpage en séquences fonctionnelles se base principalement sur la segmentation décrite par Bilger et Campione (2002) pour des énoncés du français parlé. L'avantage de ce type d'annotation est qu'il est relativement aisé à faire, et qu'il ne nécessite pas de multiplier les

étiquettes. Une unité de rection est ainsi décomposée en ses différents constituants : séquence sujet, séquence objet, séquence verbe, séquence régie, séquence associée et insertions.

Toute séquence peut faire l'objet d'expansions (liste, adjectif, participe passé, relative, etc.) qui sont intégrées à la séquence. Ces expansions sont détaillées à un autre niveau d'analyse.

3.2.1. La séquence sujet (SS)

La séquence sujet est constituée d'un sujet réalisé sous une forme lexicale

le chat mange

le chat que j'avais trouvé l'année dernière dans mon jardin et qui avait été abandonnée par mes voisins mange

3.2.2. La séquence objet (SO)

La séquence objet comprend les éléments appartenant à la valence du verbe, lorsque ceux-ci sont réalisés sous une forme lexicale. Il ne s'agit donc pas du sens de « complément d'objet » de la grammaire traditionnelle.

il se rend à Esneux

Le complément locatif *à Esneux*, qui est dans la valence du verbe, est codé comme un SO.

Chaque SO entre dans une relation de proportionnalité avec une proforme.

il se rend à Esneux = il s'y rend

le chat mange la souris = le chat la mange

3.2.3. La séquence verbe (SV)

La séquence verbe comprend le verbe, le sujet s'il est pronominalisé, les clitiques, les particules négatives, les verbes modaux, les infinitifs (cf. Bilger et Campione 2002 : 120).

le chat mange la souris

il la mange

Pierre veut dormir jusqu'à 10 heures du matin

il est de la responsabilité de tout dirigeant d'avoir une juste appréhension de la situation
Patrick Couthin aime regarder les filles qui marchent sur la plage

Les infinitifs font partie du SV quand ils sont sélectionnés par le sujet. Dans l'exemple suivant, c'est le cas de *regarder*, mais pas de *marcher* dont le sujet est le pronom *les* :

Patrick Couthin aime les regarder marcher sur la plage

cet enfant aime se regarder sourire dans la glace

3.2.4. La séquence régie (SR)

Une séquence régie est une séquence qui appartient à la rection du verbe (est organisée par le verbe) mais n'entre pas dans sa valence. Pour vérifier si une séquence est régie par un verbe ou non, on applique le test de la pronominalisation : si la séquence peut être remplacée par un pronom, alors elle est régie ; dans le cas contraire, elle est associée (Blanche-Benveniste *et al.* 1990). Une séquence régie peut être soumise à l'extraction.

en revenant de la fête de son collègue papa est tombé dans les escaliers

c'est en revenant de la fête de son collègue qu'il est tombé dans les escaliers

Une SR peut être introduite par des marqueurs de rection de type *parce que*. Suivant que ces éléments régis sont à droite ou à gauche du verbe recteur, on notera

SRg (séquence régie à droite) ou *SRd* (séquence régie à gauche)

on n'a pas arrêté de me sonner tout le trajet (SRd)

il est venu parce que mon père a beaucoup insisté (SRd)

3.2.5. La séquence associée (SA)

Certaines séquences ne sont pas dans la rection (du verbe) : on les appelle des *associés* ou des *adjoints* (Blanche-Benveniste *et al.* 1990). On applique les mêmes tests que pour les séquences régies : l'impossibilité de pronominalisation ou d'extraction montrent que la séquence est associée à la construction verbale.

franchement j'ai pas envie de continuer à lire ce livre

** c'est franchement que j'ai pas envie de continuer à lire ce livre*

au fond je me demande si je l'aime encore

**c'est au fond que je me demande si je l'aime encore*

Les **dislocations** (doubles marquages) sont des séquences associées :

il est beau Philippe (SAd)

Elisa c'est un prénom magnifique pour une fille (SAg)

Les SA entrent dans la composition des URC+.

3.2.6. L'insertion (insert)

Certaines unités de rection, certaines séquences peuvent être interrompues par des insertions qui constituent une forme de parenthèse dans le discours.

Ces insertions, appelées aussi *incises*, peuvent être plus ou moins complexes :

et c'est c'est assez passionnant parce que ça se lit effectivement quelqu'un vous l'a dit d'ailleurs comme un roman

4. De la structuration syntaxique aux unités minimales du discours

Le découpage syntaxique que nous effectuons, sur la base des critères que nous venons de développer, nous permet d'identifier des unités de rection. Nous confrontons celles-ci aux segments dégagés par l'analyse prosodique, les *unités intonatives majeures* (UIM).

La correspondance des plans syntaxique et prosodique nous a permis de dégager 3 types d'unités, qui mettent en jeu des stratégies différentes de la part des locuteurs :

- *congruence* : à une unité de rection correspond une UIM ;

- *condensation* : plusieurs unités de rection sont regroupées dans une seule UIM ;

- *dislocation* : une unité de rection est répartie sur plusieurs UIM distinctes.

Notre hypothèse est que les MDU prototypiques sont celles qui correspondent à la 1re catégorie (congruence), où une unité de rection complète correspond à une unité intonative

majeure. Cette catégorie serait la catégorie non marquée, par rapport aux deux autres qui mettraient en jeu des stratégies différentes.

Références

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

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

BERTHOUD A.-Cl., MONDADA L. (Eds) (2000). Modèles du discours en confrontation. Lang. Berne.

BILGER M. et CAMPIONE E. (2002), « Propositions pour un étiquetage en "séquences fonctionnelles" », in Recherches sur le français parlé 17, Université de Provence, pp. 117-136.

BLANCHE-BENVENISTE Cl. et JEANJEAN C. (1987), *Le français parlé. Transcription et édition.* Didier Érudition. Paris.

BLANCHE-BENVENISTE Cl., BILGER M., ROUGET Chr. et van den EYNDE K. (1990). *Le français parlé : études grammaticales*. Éditions du CNRS. Paris.

DEGAND L. et SIMON A. C. (2005). « Minimal Discourse Units : Can we define them, and why should we ? », in Aurnague M., Bras M., Le Draoulec A. et Vieu L. (Éds), *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories*, Biarritz, 14-15 novembre 2005 : 65-74.

DISTER A., FRANCARD M., GERON G., GIROUL V., HAMBYE Ph., SIMON A. C., WILMET R. (2006). *Conventions de transcription régissant les corpus de la banque de données VALIBEL* (http://valibel.fltr.ucl.ac.be/, corpus oraux, conventions de transcription).

MANN W. C. et THOMPSON S. A. (1988). « Rhetorical structure theory : Toward a functional theory of text organization », in *Text* 8 : 243-281.

PASSONEAU R. J. et Litman, D. J. (1997). « Discourse Segmentation by Human and Automated Means », in *Computational Linguistics* 23: 103-139.

POLANYI L. 1988. « A formal model of the structure of discourse », in *Journal of Pragmatics* 12 : 601-638.

ROULET E., FILLIETTAZ L.et GROBET A (2001). *Un modèle et un instrument d'analyse de l'organisation du discours.* Peter Lang, Bern.

SIMON A. C. (2001). « Le rôle de la prosodie dans le repérage des unités textuelles minimales », in *Cahiers de linguistique française* 23.