annoter du des textes tu te demandes si c'est syntaxique tu vois

Christophe Benzitoun [a], Anne Dister [b], Kim Gerdes [c], Sylvain Kahane [d], Renaud Marlet [e]

Résumé

Rhapsodie est un projet de constitution de corpus du français parlé, annoté en prosodie et en syntaxe. Nous présentons ici les choix d'annotations syntaxiques spécifiques à l'oral ainsi que la stratégie d'annotation et les différents niveaux de représentation.

1. Introduction

Cette communication présente les choix d'annotation faits actuellement pour le volet syntaxique du projet Rhapsodie. Ce projet[1] vise à étudier l'interface syntaxe-prosodie en français parlé. Une part importante du projet est consacrée à l'élaboration[2] d'un corpus distribué librement, échantillonné en différents genres discursifs, et doté d'annotations prosodiques et syntaxiques exploitables pour l'analyse du statut de la prosodie dans le discours, de ses relations avec la syntaxe et la structure informationnelle.

Depuis le début des années 1980 ont été lancés des projets de grande envergure pour la constitution de corpus oraux dans les langues de grande diffusion. Conjointement, des consortiums internationaux de coordination de projets se sont constitués (p. ex. Clarin http://www.mpi.nl/clarin/). Dans ce contexte foisonnant, les francophones ont pris conscience du retard accumulé dans la constitution et l'exploitation des corpus oraux. C'est sans doute la raison pour laquelle de nombreux projets visant au développement de larges corpus de français parlé ont vu le jour au cours de ces 20 dernières années. Plus récemment, des structures de mutualisation et d'échanges de ressources ont été développées en France (voir la mise en place du Centre de Ressources pour la Description de l'Oral). Malgré cet effort, et contrairement au monde anglo-

[a] ATILF, Nancy Université & CNRS [b] Centre de recherche VALIBEL, UCLouvain [c] LPP, Sorbonne Nouvelle & CNRS [d] Modyco, Université Paris Ouest Nanterre & CNRS [e] LaBRI, INRIA Bordeaux Sud-Ouest.

[1] Projet ANR Rhapsodie 07 Corp-030-01 : corpus prosodique de référence du français parlé.

[2] Une partie du corpus repose sur des enregistrements existants, déjà présents dans d'autres corpus. Cependant, les transcriptions et annotations, prosodiques comme syntaxiques, seront spécifiques à Rhapsodie.

saxon, il n'existe pas aujourd'hui de corpus de français parlé réellement annoté en syntaxe et distribué librement. Le CID (Corpus of Interactional Data) est ce qui s'en rapproche le plus (Bertrand et al. 2009), mais le corpus n'est pas libre et les informations syntaxiques sont superficielles (chunks typés) ; des annotations de plus haut niveau ont néanmoins été expérimentées (Blache 2009). La seule banque d'arbres (angl. treebank) pour le français que nous connaissions est le Corpus de Paris 7 (Abeillé et al. 2003), mais il s'agit d'un corpus de l'écrit, extrait du journal *Le Monde*, et qui n'est pas totalement libre (il est distribué gratuitement à des fins de recherche universitaire, mais n'est pas librement redistribuable ni accessible en ligne).

L'originalité de Rhapsodie repose sur une annotation parallèle en prosodie et en syntaxe, l'information syntaxique incluant des annotations très complètes couvrant à la fois des phénomènes de micro- et de macrosyntaxe. Nous voulons annoter en syntaxe la totalité du texte, y compris les disfluences, alors que celles-ci sont généralement gommées avant l'annotation syntaxique ; c'est le cas des principaux corpus de référence, comme le Switchboard pour l'anglais (http://groups.inf.ed.ac.uk/switchboard) ou le CGN pour le néerlandais (http://lands.let.ru.nl/cgn). Notre traitement des disfluences repose sur les travaux menés depuis plus de trente ans sur le français parlé par l'école d'Aix, autour notamment de Claire Blanche-Benveniste et José Deulofeu, et qui tend à traiter les disflences comme des cas particuliers d'entassements paradigmatiques, au même titre que la coordination et d'autres phénomènes que nous présenterons ici.

Cette communication est centrée sur les principales difficultés que pose, de notre point de vue, l'annotation syntaxique d'un corpus oral. Un des problèmes principaux est la segmentation en unités syntaxiques maximales, segmentation qui est rendue difficile par divers phénomènes, comme les greffes et les parenthèses, ainsi que toute une série de faits que nous regroupons sous le terme générique d'*entassements paradigmatiques* (Blanche-Benveniste 1990, Guénot 2006, Gerdes et Kahane 2009).

Dans cette communication, nous ne présenterons pas l'annotation des phénomènes microsyntaxiques, c'est-à-dire tout ce qui relève de l'hypotaxe et qui sera encodé en termes de dépendances syntaxiques étiquetées par des fonctions syntaxiques, à la façon de nombreux autres corpus comme le Prague Dependency Treebank (http://ufal.mff.cuni.cz/pdt) ou l'Alpino Dependency Treebank (http://www.let.rug.nl/vannoord/trees/).

Nous allons présenter l'annotation retenue à travers la présentation de divers phénomènes, puis nous présenterons la procédure que nous souhaitons suivre pour annoter le corpus en syntaxe. La plupart de nos exemples sont des exemples attestés.

2. Annotation et phénomènes traités

Il existe aujourd'hui d'assez bons analyseurs syntaxiques pour l'écrit. Nous avons à notre disposition plusieurs de ces analyseurs et nous allons les utiliser pour l'annotation en dépendance syntaxique du corpus. Ces analyseurs sont paramétrés pour l'écrit (ils utilisent par exemple de manière consistante la ponctuation) et ont des résultats médiocres sur l'oral du fait de la présence de phénomènes d'ordre macrosyntaxique. Pour contourner ce problème, nous allons effectuer une annotation manuelle préalable du corpus avec différentes balises liées à la macrosyntaxe et limiter la tâche des analyseurs à la seule détermination de relations microsyntaxiques. C'est ce travail que nous présentons ici.

2.1 Segmentation principale

Notre unité maximale est le *Segment Rectionnel Elargi (SRE)*. Un SRE est composé d'un élément recteur principal et de tous les éléments qu'il régit ou qui lui sont associés. Nous n'allons pas tenter ici de redéfinir ce qu'est le segment, ni ce que sont les éléments associés. On se contentera de dire que les éléments associés sont des éléments qui ont l'apparence d'unités dépendantes du verbe mais qui n'en ont pas les propriétés : ils ne sont pas équivalents à une proforme, pas clivables, etc. Les SRE correspondent schématiquement à la notion intuitive de phrase.

On notera que le découpage en SRE ne tient pas compte des tours de paroles et qu'il est courant que deux locuteurs construisent ensemble une SRE. Par exemple :

Loc1 : quatre-cent-soixante-mille euros
Loc2 : attends ça fait c~
Loc1 : dix-huit millions et une chique

(Le symbole « ~ » marque une amorce de morphème.) Les éléments associés à gauche et à droite sont séparés du noyau par les marqueurs « < » et « > » :

mon pote < il veut rien faire

Certains éléments qui ont l'apparence de propositions sont traités comme des associés lorsqu'ils commutent avec un groupe nominal :

j'ai un pote < il veut rien faire (mon pote < il veut rien faire)
je me levais le matin < j'étais avec des clients (chaque matin < j'étais avec des clients)

Les marqueurs de discours (Teston-Bonnard 2006) ne sont pas intégrés dans la structure syntaxique de l'énoncé et sont isolés par des doubles guillemets :

on avait un trou entre "euh" "enfin" "tu vois" midi et cinq heures

Nous ne faisons pas de coordinations de phrases. Celles-ci sont systématiquement découpées en deux SRE. Dans le même esprit, nous découpons en deux SRE certaines propositions introduites par *parce que*, car il peut s'agir de noyaux séparés.

2.2 Entassements paradigmatiques

Les productions orales comportent un grand nombre de reformulations dues au fait que le locuteur construit son message en même temps qu'il le produit et qu'il est souvent à la recherche de la bonne formulation :

euh à Poitiers nous a~ nous sommes vingt-deux permanents hein vingt-deux agents per~ vingt-deux fonctionnaires permanents

On parle de disfluences lorsque ces reformulations sont involontaires et qu'elles produisent des segments grammaticalement incomplets comme dans l'exemple précédent. Comme nous l'avons dit dans l'introduction, dans les travaux d'annotations de l'oral que nous connaissons, les disfluences sont « nettoyées » avant l'annotation ; les disfluences et amorces de syntagmes sont marquées (p. ex. entre parenthèses) et ne sont pas intégrées à l'analyse syntaxique :

(euh) à Poitiers (nous a~) nous sommes (vingt-deux permanents hein vingt-deux agents per~) vingt-deux fonctionnaires permanents
➞ *segment analysé :* à Poitiers nous sommes vingt-deux fonctionnaires permanents

Les travaux de l'école d'Aix ont largement montré que cette entreprise de nettoyage, à supposer qu'elle soit légitime, s'avère dans de nombreux cas impossible. De plus, il est souvent difficile de séparer ce qui relève de la volonté du locuteur (cf. les « effets de style ») d'une reformulation involontaire (et disfluente, donc).

A la suite de l'école d'Aix, nous considérons que les reformulations, comme la coordination, constituent un même groupe de phénomènes, que nous nommons les *entassements paradigmatiques,* que nous encodons de la même manière, mais que nous pouvons distinguer par leur type (bien que nous ne l'explicitions pas ici ; Gerdes et Kahane 2009) :

"euh" à Poitiers < {nous a~ | nous sommes} {vingt-deux permanents | "hein" vingt-deux agents per~ | vingt-deux fonctionnaires permanents}

La balise « | » indique le *point de jonction* (appelé encore *point d'interruption* pour les disfluences ; Shriberg 1994). La balise « { » indique le *point de retour* : on peut, à la suite de Heeman et al. (2006), considérer que lors d'un entassement paradigmatique, le locuteur effectue un retour à la position initiale :

"euh" à Poitiers < { nous a~
| nous sommes } { vingt-deux permanents
| "hein" vingt-deux agents per~
| vingt-deux fonctionnaires permanents }

Cette interprétation des balises est équivalente à l'analyse en grille (Blanche-Benveniste et al. 1979). Le rôle de la balise « } » est secondaire et il nous a semblé souvent difficile de décider où la mettre. Du point de vue de la prosodie, seule la balise « | » correspond clairement à un contour prosodique marqué, le point de retour n'étant a priori pas marqué (notamment dans les nombreux cas où le locuteur n'a pas anticipé qu'il allait effectuer un entassement).

On peut définir l'entassement ainsi : le segment Y s'entasse sur un segment X qui le précède si Y occupe la même position syntaxique que X. Les conjonctions de coordinations sont considérées comme des marqueurs d'entassement et sont notés à l'aide du symbole « ^ » :

c'est {réexpliquer mais {tout le temps | tout le temps | tout le temps} | ^et & | ^et & | ^et ^alors "euh" leur demander "tu vois" qu'ils te rappellent "euh" {certains points | certaines règles | etcetera}} "quoi"

L'exemple précédent illustre un cas d'*intensification* par répétition d'un segment. Le symbole « & » marque un syntagme inachevé.

Certains entassements peuvent être non contigus. Dans ce cas, on utilise les balises « |} » et « {| » pour marquer le point d'interruption et la reprise de l'entassement. Les *instanciations* sont un exemple typique, qu'il s'agisse de l'instanciation d'un pronom interrogatif ou d'un indéfini :

Loc1: "hé" "euh" il est né {en quelle année |} > lui > François
Loc2: {| en quatre-vingt-quatre}

Françoise Giroud vous avez donné {quelque chose de plus |} à la femme {| des armes de persuasion}

Ce dernier cas a été décrit par Cl. Blanche-Benveniste sous le terme d'*effet deux-points*, en référence à la façon dont il serait ponctué à l'écrit.

Notons encore que les reformulations peuvent donner des effets sémantiques similaires à l'instanciation :

> on a fait la photo {dans & | dans le jardin | dans la petite pelouse}

2.3 Autres phénomènes

La greffe est le procédé qui consiste à remplir des positions syntaxiques à l'aide d'une autre catégorie que celle attendue (Deulofeu 1999). Il y a une rupture de sous-catégorisation. Il s'agit généralement d'une construction verbale occupant la position d'un groupe nominal. Les greffes sont notées entre crochets. Lorsque l'élément central de la greffe est présent, c'est-à-dire celui qui apparaitrait directement dans cette position syntaxique, nous le délimitons à l'aide de « * * ».

> j'ai rencontré [je crois que c'était *Paul*] à l'école hier

> vous suivez la ligne du tram qui passe vers {la & | [je crois que c'est *une ancienne caserne* "je sais pas"]} "voilà"

Enfin, nous notons entre parenthèses les *insertions*, c'est-à-dire, les unités qui viennent interrompre une SRE en cours de construction. La SRE en construction a l'obligation de reprendre et de se terminer ultérieurement pour que l'on puisse parler d'insertion ; sinon il y a simplement succession de SRE.

> les universités qui se préoccupent de la concurrence internationale (et nous pouvons très bien admettre que certaines n'aient pas à le faire) ne pourront se contenter très longtemps du système et du niveau actuels de financement en Communauté française de Belgique

3. La stratégie d'annotation et les niveaux de représentation

Notre stratégie d'annotation repose sur le fait qu'il existe aujourd'hui d'assez bons outils développés pour l'écrit. Adapter ces outils à l'oral constituerait un projet en soi, plus ambitieux que notre projet d'annotation (même si nous pensons qu'un projet tel que Rhapsodie est une étape essentielle dans le développement d'analyseurs pour l'oral et qu'une des finalités de Rhapsodie est de pouvoir contribuer à l'entraînement de tels analyseurs). Autrement dit, nous voulons utiliser des outils développés pour l'écrit sans les modifier de façon substantielle. Pour ce faire, nous réaliserons « à la main », comme nous l'avons déjà dit, un prétraitement des textes de manière à fournir aux analyseurs syntaxiques de l'écrit des segments sur lesquels on sait qu'ils

feront peu d'erreurs. Nous allons présenter plus en détail les différentes étapes de l'annotation, qui alternent plusieurs fois analyse automatique et traitement manuel. Nous sommes également en train de développer pour les traitements manuels des outils qui permettent une bonne visualisation des résultats automatiques et une correction ergonomique.

Niveau 000 : transcription brute

L'annotation syntaxique prend en entrée une transcription orthographique de la bande sonore qui peut contenir différentes marques (indiquant des changements de locuteur ainsi que des éléments paraverbaux tels que les rires, etc.). Nous appelons cette transcription la transcription brute, elle constitue le niveau 000 d'annotation. (Le nommage chiffré actuel des niveaux n'est pas définitif ; son étrangeté a des raisons « historiques ».)

Niveau 00 : prétraitement automatique simple

Un premier traitement automatique (Dister *et al.* 2009) permet d'annoter les disfluences triviales (comme les doublement de mots) et de repérer des marqueurs de discours potentiels (*hein, ben, quoi, tu vois,* …), ainsi que les marques telles que les *euh*. Le prétraitement peut faire des erreurs. Elles seront corrigées manuellement au niveau 0 qui suit immédiatement. L'objectif du prétraitement est d'accélérer le travail manuel d'annotation de ce niveau 0.

Niveau 0 : balisage des unités macrosyntaxiques maximales

L'encodage de niveau 0 est l'encodage obtenu après l'introduction des balises présentées à la section précédente. Il s'agit de la transcription de niveau 000 enrichie de balises marquant les frontières des entassements, incises, segments rectionnels élargis, etc. Ces balises sont ajoutées à la main au-delà du niveau 00. Nous développons actuellement un outil libre, Pilepilot, qui nous permet de faire ce balisage sur une présentation du type analyse en grille, c'est-à-dire où les différentes couches d'un entassement sont les unes au-dessus des autres. L'outil, qui tourne sur un serveur et ne nécessite donc pas d'installation de logiciel particulier autre qu'un navigateur, permet l'écoute directe des parties en cours d'analyse, l'annotation ergonomique grâce à une interface graphique (javascript) et la comparaison des analyses des différents annotateurs. (Voir la figure 1 ci-dessous.)

Figure 1. Outil d'annotation PilePilot

Niveau 0+ : entrée des analyseurs syntaxiques

Comme nous l'avons dit, les analyseurs syntaxiques pour le français dont nous disposons ont été développés pour l'écrit et nous n'envisageons pas, dans le cadre du projet Rhapsodie, d'effectuer de modifications substantielles de ces outils. Ils ne sont pas paramétrés pour traiter de l'oral non balisé, ni pour traiter les balises que nous avons introduites au niveau 0. Mais grâce à ces balises, nous pouvons effectuer un découpage automatique du texte et fournir à l'analyseur des morceaux qu'il est capable d'analyser. Donnons un exemple :

est-ce que vous songez {à d'autres communicat~ | "euh" à d'autres fonctions}

donnera deux segments :

(1) est-ce que vous songez à d'autres communicat~
(2) est-ce que vous songez à d'autres fonctions

Certains fragments de texte seront ainsi dupliqués et analysés plusieurs fois. Ces analyses seront cependant fusionnées automatiquement dans les niveaux ultérieurs si elles sont identiques. Si elles diffèrent, un traitement manuel sera nécessaire.

Une autre stratégie consiste à ne pas déplier les entassements et voir un énoncé avec entassement comme un graphe dirigé acyclique (DAG), c'est-à-dire comme un graphe dont les arcs sont étiquetés par les mots du texte et qui intègre tous les chemins possibles dans une structure avec entassements. Un parseur comme SxLFG (Boulier et Sagot 2005) sait gérer un DAG en entrée, mais pour l'instant il est paramétré pour choisir le meilleur chemin dans le DAG et pas pour analyser l'ensemble du DAG.

Niveau 1- : sortie des analyseurs syntaxiques

Les analyseurs nous fournissent une analyse syntaxique que l'on peut présenter sous forme d'arbre de dépendance. Il nous reste deux choses à faire : 1) traduire automatiquement ces analyses pour qu'elles correspondent exactement au jeu d'étiquettes que nous souhaitons (nous pensons notamment à la liste de fonctions syntaxiques qu'il nous faut arrêter) ; 2) appliquer sur le texte de départ (celui du niveau 0) les annotations syntaxiques calculées pour les différents segments, en fusionnant en outre les annotations syntaxiques des fragments dupliqués.

Niveau 1 : annotation syntaxique corrigée

La dernière étape est une correction manuelle du niveau 1-, absolument nécessaire, car les analyseurs syntaxiques font encore beaucoup d'erreurs (nous estimons qu'environ 30 % des dépendances devront être corrigées) et qu'ils n'utilisent pas le même jeu d'étiquettes que le nôtre. L'encodage de niveau 1 est donc une analyse syntaxique complète du texte, comprenant aussi bien la microsyntaxe (les dépendances fonctionnelles) que la macrosyntaxe.

Niveau topologique

L'analyse syntaxique de niveau 1 permet de calculer les constituants topologiques, c'est-à-dire les regroupements de mots imposés par la structure fonctionnelle (Gerdes et Kahane 2006). Nous avons postulé que les constituants prosodiques découlent de la structure topologique et nous espérons confirmer cette hypothèse en corrélant les frontières de constituants topologiques avec les frontières prosodiques.

Niveaux supérieurs

Une bonne analyse des marqueurs prosodiques nécessiterait davantage qu'une annotation syntaxique. Aussi envisageons-nous, dans une étape ultérieure, d'enrichir le corpus avec un marquage discursif comprenant un marquage des chaînes de coréférences et un marquage de relations de discours. Cela construirait une structure connexe sur l'ensemble du texte.

Conclusion

Contrairement à d'autres corpus oraux, nous souhaitons analyser l'ensemble du texte, en profondeur et sans perdre les spécificités de l'oral, comme les segments disfluents ou les greffes, d'ordinaire « corrigés ». Une des particularités de notre approche est de combiner une analyse en grille avec une analyse détaillée des relations syntaxiques. Nous avons montré avec quel

dispositif il était possible d'obtenir une analyse syntaxique complète de productions orales spontanées.

Remerciements

Nous remercions l'ensemble des participants du projet Rhapsodie, mené par l'énergique Anne Lacheret, et tout particulièrement les autres membres du groupe syntaxe, à commencer par Jeanne-Marie Debaisieux, Frédéric Sabio et Bernard Victorri. Nous remercions également Claire Blanche-Benveniste et José Deulofeu pour les passionnants échanges que nous avons eus.

Références

Abeillé, A., L. Clément, F. Toussenel (2003). Building a Treebank for French. A. Abeillé (ed) *Treebanks*, Kluwer, Dordrecht.

Blache, P. (2009). Des relations d'alignement pour décrire l'interaction des domaines linguistiques : vers des Grammaires Multimodales, in *16e conférence sur le Traitement Automatique des Langues Naturelles (TALN),* Senlis, France.

Blanche-Benveniste, C., M. Bilger, C. Rouget, K. van den Eynde (1990). *Le Français parlé. Études grammaticales*, Paris, CNRS Éditions.

Blanche-Benveniste, C., B. Borel, J. Deulofeu, J. Durand, A. Giacomi, C. Loufrani, B. Meziane, N. Pazery (1979). Des grilles pour le français parlé. *Recherches sur le français parlé*, *2*, 163-205.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, S. Rauzy (2009). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues* 49(3), 1-30.

Boullier, P. et B. Sagot (2005). Analyse syntaxique profonde à grande échelle: SxLfg. *Traitement Automatique des Langues*, 46(2), pp. 65-89.

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., M. Constant et G. Purnelle (2009). Normalizing speech transcriptions for Natural Language Processing. *Actes du colloque international Spoken Communication*, Univ. de Naples.

Gerdes, K., et S. Kahane (2006). L'amas verbal au cœur d'une modélisation topologique de l'ordre des mots. K. Gerdes et C. Muller (Eds), *Ordre des mots et topologie de la phrase française, Linguisticae Investigationes,* 29:1, 75-89.

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

Guénot, M.-L. (2006). La coordination considérée comme un entassement paradigmatique : description, formalisation et intégration, *Actes de TALN*, Louvain, Belgique, 178-187.

Heeman, P., A. McMillin, J. S. Yaruss (2006). An annotation scheme for complex disfluencies. *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*, PhD Thesis, Berkeley University.

Teston-Bonnard, S. (2006). *Propriétés topologiques et distributionnelles des constituants non régis. Application à une description syntaxique des particules discursives*, Thèse de l'Université de Provence.