Problèmes d'annotation d'un corpus oral dans le cadre de la campagne EASY

Christophe Benzitoun, Jean Véronis

Equipe DELIC – Université de Provence
29, Av. Robert Schuman 13100 Aix-en-Provence
{Christophe.Benzitoun, Jean.Veronis}@up.univ-aix.fr

Mots-clés : corpus oral, annotation syntaxique
Keywords : spoken corpus, syntactic annotation

Résumé

Nous présentons, dans cet article, les problèmes que nous avons rencontrés et les solutions que nous avons adoptées pour l'élaboration du corpus oral de référence dans le cadre de la campagne EASY.

Abstract

In this paper, we present some problems and their solutions to annotate the gold standard spoken corpus of the EASY project.

1 Introduction

A la suite de plusieurs mois de réflexion et d'expérimentation liées à la constitution du corpus oral de référence pour le projet d'évaluation des analyseurs syntaxiques EASY (cf. aussi Benzitoun et al. 2004), il a fallu trouver un formalisme permettant de coder l'intégralité des données transcrites, « spécifiques » à l'oral (pauses, intonation, répétitions, amorces, inachèvements…), qui soit en adéquation avec celui proposé dans le cadre du projet. Cette contrainte répond à l'objectif que nous nous sommes fixés de reproduire le plus fidèlement possible les énoncés produits et ainsi de garder toutes les informations, quelle qu'en soit l'origine (prosodie, travail de formulation…), à travers les divers niveaux de l'annotation. En effet, celles-ci sont potentiellement utiles pour l'analyse automatique ultérieure des corpus oraux, et notamment leur utilisation pour l'amélioration des technologies vocales.

Pour cela, nous avons élaboré deux versions du corpus. Une première version, de travail, contient toutes les informations spécifiques à l'oral, sous forme de balises supplémentaires. La seconde, qui respecte scrupuleusement le guide d'annotation fourni par les organisateurs de la campagne EASY, est générée automatiquement à partir de la première (par suppression ou transformation d'informations). Le corpus de travail pourra être utile non seulement aux participants dont l'analyseur utilise des informations autres que textuelles ou qui veulent supprimer automatiquement les « disfluences », mais aussi pour évaluer les programmes détectant les disfluences, les pauses ou l'intonation ou, plus généralement, pour l'évolution

des technologies de la parole, dont les « modèles de langage » sont souvent mis au point à partir de textes écrits reflétant assez mal le langage parlé (par exemple le journal *Le Monde*).

Le corpus lui-même est composé de dix extraits d'environ cinq minutes chacun du *Corpus de Référence du Français Parlé* (DELIC, 2004) spécialement choisis pour leur caractère monologique et leur hétérogénéité situationnelle. La transcription orthographique a été effectuée entièrement à la main par des experts avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées (DELIC, 2004) ne contiennent aucun trucage orthographique (du type *p'tit, y'a,* etc.) ni aucune ponctuation, suivant la tradition de notre équipe, qui a clairement montré que la ponctuation de l'écrit était parfaitement inadéquate à la transcription de l'oral (cf. Blanche-Benveniste & Jeanjean, 1986). Par contre, sont notés avec soin les répétitions, les amorces, les *euh* d'hésitation, les allongements, les pauses (avec leur durée exacte) et les accents et mouvements intonatifs majeurs. Ceux-ci ont tous fait l'objet d'une balise particulière (voir la transcription du corpus dans Campione (2001, vol 2.)).

2 L'unité maximale (UM)

Les organisateurs de la campagne d'évaluation devaient fournir aux participants un texte segmenté en « phrases ». Or, il a été largement montré que cette notion est une notion purement graphique et qu'elle ne se retrouve pas, ni de loin ni de près, dans les productions orales (cf. Berrendonner, 2002 ; Blanche-Benveniste, 2002). On note d'ailleurs que même à l'écrit, phrases et unités linguistiques sont souvent non concordantes (Benzitoun, 2004).

Il a donc fallu avoir recours à une méthode de segmentation spécifique à l'oral. Dans une approche analogue à celle Blanche-Benveniste (2002), nous avons donc choisi comme unité de segmentation une *unité maximale* (UM), composée d'un *constructeur* (le plus souvent verbal, mais éventuellement aussi nominal, adjectival ou adverbial), et de tous ses *dépendants* et *associés* (au sens de Blanche-Benveniste et al. (1990)). Pour ce faire, il faut absolument se défaire des présupposés théoriques entourant le marquage des relations syntaxiques. Dans l'extrait suivant, le *parce que* ouvre une nouvelle UM qui n'entretient aucun rapport syntaxique avec la précédente.

1) *enfin dans une zo*ne touristique j'ai été remarque hein↗ ++ donc c'est pas bien difficile euh de bosser* ***parce qu'****en fait bon: effectivement mes économies arrivaient à: leurs fins↗ ++*

En outre, les cas d'UM parenthétiques venant couper une autre UM sont fréquents. Le guide ne disposant pas d'une étiquette pour les énoncés parenthétiques (bien qu'ils soient loin d'être inexistants à l'écrit), ces unités restent « flottantes » dans la version sous-spécifiée du corpus, et seules les relations internes sont marquées.

3 Les constituants

Les marqueurs *euh, hein, bon, ben, quoi, disons, je veux dire*, etc., extrêmement fréquents à l'oral, mais qui n'ont pas de catégorie grammaticale traditionnelle claire (ce ne sont ni des interjections, ni des adverbes) ont été annotés avec la balise « insert » (cf. Biber et al., 1999). Ont aussi été annotés les répétitions, les mots ou constituants inachevés et les segments inaudibles. Les éléments répétés sont inclus dans un groupe, quand ils en font partie, ou sont laissés à l'extérieur du groupe. Les constituants inachevés sont marqués avec la catégorie

qu'ils auraient s'ils étaient achevés et sont reliés à leur constructeur lorsqu'un autre élément n'occupe pas déjà la position syntaxique. Dans l'exemple suivant, *plusieurs* a été annoté GN et il a été relié au verbe *rester* par l'intermédiaire de la relation « modifieur de verbe ».

2) *je suis restée euh je sais pas qu- qu-* ***plusieurs*** *euh*

Seuls les inachèvements au niveau du mot ou du constituant sont notés et pas les inachèvements relationnels. Par exemple, dans 3), *une fille qui part seule en stop* semble en attente d'un verbe constructeur. Malgré cela, nous ne marquons pas d'inachèvement car la question est souvent complexe, l'inachèvement apparent pouvant parfois être complété par des phénomènes extra-linguistiques (geste, mimique, soupir…).

3) *enfin bon là euh ça a été dur dans la famille quand même hein parce que: ++* ***une fille qui part seule en stop*** *euh en Espagne c'était pas à côté*

Les relations entre inachèvement et répétition peuvent parfois être problématiques. Nous avons donc pris la décision de ne marquer les répétitions que dans les cas où celles-ci sont contiguës et où le mot est répété de manière exacte. Dans l'exemple suivant, le premier *de* sera marqué comme étant une répétition alors que le second sera marqué comme faisant partie d'un constituant inachevé car il est suivi par *des* (Adda et al. (2005) suivant les recommandations du LDC parlent de « révision »).

4) *j'ai fait l'expérience* ***de de: des*** *métiers de la restauration*

De même, *en hiver euh il y a plus* dans 5) ne sera pas marqué répétition car il n'est pas contigu.

5) *et puis* ***en hiver euh il y a plus*** *euh bon comme en France sans doute dans les coins touristiques hein / ++ en hiver il y a plus le boulot*

Les mots inachevés sont transcrits comme ils ont été prononcés. Certaines formes n'apparaissent donc pas dans les ressources dictionnairiques d'un analyseur. Dans l'exemple suivant, l'ensemble forme un GN et à l'intérieur on met une balise « fragment » pour signaler que *bou-* est un mot inachevé.

6) *des petites bou- + fioles*

4 Les relations

Il n'a pas été utile de créer des balises spécifiques pour les relations ce qui étaye, une fois de plus, l'hypothèse de notre équipe selon laquelle il n'y a pas de relations syntaxiques propres à l'oral (mais seulement des différences de fréquences). Comme à l'écrit, les relations peuvent être à une distance tout à fait remarquable, ce qui permet de faire des hypothèses concernant nos capacités d'encodage et de décodage (voir à ce sujet l'exemple présenté dans Benzitoun et al. (2004)). Dans ce cas, un élément est généralement répété pour permettre d'effectuer le raccordement.

7) *ce qui fait que j'ai amené des affaires d'hiver des affaires euh d'été* ***plus*** *euh à cette époque j'avais j'étais en maîtrise il me restait le mémoire à faire* ***plus*** *euh donc euh les livres tout ce qu'il me fallait pour faire mon mémoire là-bas*

Un autre phénomène remarquable est celui du double marquage (Blasco-Dulbecco, 1999).

Celui-ci n'a pas fait l'objet d'une mention particulière dans le guide d'annotation ce qui nous a obligé à proposer un traitement spécial en accord avec les organisateurs. Nous avons donc opté pour le marquage de deux relations identiques. Dans l'exemple 8), il y aura donc deux relations sujet malgré l'absence d'accord morphologique du verbe avec *nous*.

8) ***nous on** est là*

A propos de l'absence d'accord morphologique, on peut signaler ce cas très intéressant dans lequel *l'enfant, le bébé, les vieux* ont dû être reliés par la relation de « juxtaposition » sur le modèle *l'enfant, le bébé, les vieux peuvent halluciner*. L'accord morphologique ne se ferait que dans le cas où les éléments juxtaposés se trouvent avant le verbe.

9) *l'enfant peut halluciner le bébé↗ +++ ben écoutez **les vieux aussi** ↘*

Le dernier point abordé est celui des éléments non dépendants qui sont néanmoins enchâssés dans un énoncé. Vu qu'il n'y avait pas d'étiquette pour les non dépendants, nous avons hésité entre « modifieur » et « complément » pour ce type d'éléments.

10) ***chaque voyage** ↗ il y a: il y a une re*mise en question au niveau euh++ physique*

5 Bilan & conclusion

Les phénomènes « spécifiques » de l'oral (en particulier les « disfluences ») sont extrêmement fréquents (de l'ordre de 10% des corpus). Leur repérage et leur traitement sont un enjeu très important pour le traitement automatique de la parole (dialogue homme-machine, reconnaissance vocale). Les phénomènes qui sont décrits ici, qui n'apparaissaient pas dans le guide d'annotation de la campagne EASY ou dont le traitement a demandé une interprétation particulière des consignes, nous amène aussi à reconsidérer notre regard sur l'écrit.

Références

ADDA G. et al. (2005), Disfluences et traitement automatique : l'heure de vérité, Journée d'étude de l'ATALA, 2 avril 2005.
BENZITOUN C. (2004), L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ?, Actes de *RECITAL*, pp. 13-22.
BENZITOUN C. et al. (2004), L'analyse syntaxique de l'oral : problèmes et méthode, Journée d'étude de l'ATALA, 15 mai 2004.
BERRENDONNER A. (2002), Les deux syntaxes, *Verbum*, Vol. XXIV, n° 1-2, pp.23-35.
BIBER D. et al. (1999), *Longman grammar of spoken and written English*, Essex, Longman.
BLANCHE-BENVENISTE CL. (2002), Phrase et construction verbale, *Verbum*, XXIV, n°1-2, pp. 7-22.
BLANCHE-BENVENISTE CL. ET AL. (1990), *Le français parlé. Etudes grammaticales*, Paris, CNRS Editions.
BLANCHE-BENVENISTE CL., JEANJEAN C. (1986), *Le français parlé. Edition et transcription*, Paris, Didier-Erudition.
BLASCO-DULBECCO M. (1999), *Les dislocations en français contemporain. Etude syntaxique*, Paris, Champion.
CAMPIONE E. (2001), *Etiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*, Thèse de doctorat, Aix-en-Provence: Université de Provence.
DELIC (2004), Présentation du *Corpus de référence du français parlé*, *Recherches sur le français parlé*, 18, pp. 11-42.