Corpus oraux et chunking

Olivier Blanc, Mathieu Constant, Anne Dister, Patrick Watrin

Corpus oraux et chunking

Olivier Blanc (1), Matthieu Constant (2), Anne Dister (3) & Patrick Watrin (4)

(1) Université de Munich, CIS, Allemagne
(2) Université Paris-Est, IGM, France
(3) Université catholique de Louvain, VALIBEL, Belgique
(4) Université catholique de Louvain, CENTAL, Belgique

ABSTRACT

This paper describes a process of partial parsing of a spontaneous spoken corpus in French. It is based on a preprocessing stage that consists in reformatting and tagging utterances that breaks the syntactic structure of the text. The chunking stage uses large-coverage and fine-grained lexical resources for general language that have been augmented with resources specific to spoken. We show that it reaches a score of 84.1% f-measure.

Keywords: corpus oraux, chunking, étiquetage

1. Introduction

Dans cet article, nous présentons une procédure de segmentation en *chunks* de corpus oraux. Cette solution s'inscrit dans un projet visant l'étiquetage moprho-syntaxique automatique de l'oral. Le *chunking* doit donc être considéré comme une première étape à partir de laquelle nous inférerons les étiquettes des différents mots du texte.

Concrètement, nous opérons en deux temps. Nous prétraitons les données transcrites de l'oral afin de les rendre compatibles avec notre *chunker*. Cette première étape nous permet ensuite d'envisager le *chunking* de la même manière que nous le faisons pour l'écrit.

Après une brève description des corpus que nous utilisons, nous décrivons successivement les modules de normalisation et de *chunking*. Nous complétons notre discussion en présentant une première évaluation de l'ensemble.

2. Le corpus

2.1. Les données

Le corpus sur lequel nous travaillons est un sous-corpus extrait de la banque de données textuelles orales Valibel. Il est composé de 60 textes, soit 443 047 mots graphiques. Ceci correspond grosso modo à 40 heures de parole. Le point commun de tous les enregistrements qui composent ce sous-corpus est qu'ils ont été effectués en Belgique francophone. Si les situations de parole sont peu variées – principalement des entretiens semi-dirigés et des conversations entre amis –, elles relèvent toutes de l'oral non planifié (par opposition à du *written to be spoken*). Une description détaillée du corpus (nombre de locuteurs, informations sociolinguistiques, situation de parole précise, *etc*) est fournie dans [7].

2.2. Conventions de transcription

Les transcriptions suivent les conventions établies au Centre de recherche Valibel [1]. Les grands principes appliqués lors de la transcription sont les suivants. Ils sont largement convergents avec ceux d'autres laboratoires travaillant sur des transcriptions de l'oral :

- utilisation de l'orthographe standard. Lorsque le transcripteur le juge nécessaire, des indications phonétiques figurent entre parenthèses ;
- abandon de la notion de phrase. Les travaux sur l'oral ont montré depuis longtemps la non-pertinence de la notion de phrase à l'oral (pour le français, [5]). Les transcriptions ne sont donc pas ponctuées. Le continuum sonore, devenu par la transcription du texte linéaire, est découpé en tours de parole (défini par le changement de locuteur), et les pauses silencieuses sont notées suivant 3 degrés : pause brève (/), pause longue (//) et silence (silence) ;
- notation de phénomènes propres à l'oral : *disfluences* (*cf.* 3.2) et chevauchements de parole.

3. Prétraitement des données

Si l'annotation de corpus oraux ne nous semble pas devoir être vue comme un problème spécifique (*cf.* [2] et [13]), dans la mesure où il n'y a pas de grammaire de l'oral par opposition à une grammaire de l'écrit [4], en l'état, ces transcriptions ne peuvent être soumises à un analyseur sans une adaptation relativement importante de celui-ci (qu'il s'agisse d'un analyseur syntaxique, ou même déjà d'un simple concordancier). En effet, le format des transcriptions, ainsi que la particularité des données liées au mode de production de l'oral, rendent l'automatisation de leur traitement difficile. Plutôt qu'une adaptation coûteuse de l'outil, nous avons choisi de prétraiter les données, de manière totalement automatisée (voir aussi l'expérience menée par [12]).

3.1. Choix de transcription

Par manque de place, nous ne traiterons dans cette section que le cas des chevauchements de parole. On pourrait y discuter également du problème posé par

[1] *cf.* `http://www.uclouvain.be/81836.html`

l'absence de ponctuation dans nos données.

Les paroles superposées. Les chevauchements de parole (*overlapping segment*) brisent la linéarité de la lecture du texte transcrit. En effet, quand le chevauchement est interne au tour de parole (c'est-à-dire qu'un locuteur B commence à parler alors qu'un autre locuteur A est déjà en train de parler, mais sans prendre définitivement la parole puisque le locuteur A poursuit son propos), la transcription prend la forme suivante :

> **ilrMF0** (aspiration) ben je préférerais pas si **|- ça ne te <ilrTC1> ça va alors -|** dérange pas [ilrTC1r]

Dans la quasi-totalité des cas, ces chevauchements internes ne brisent pas la structure syntaxique de l'énoncé en cours : le locuteur A poursuit, certes avec une intrusion de l'interlocuteur dans son discours, mais il achève néanmoins ses propos selon une structure syntaxique "classique". Le lieu du chevauchement ne constitue pas une rupture. Les chevauchements internes ont donc été extraits des tours de parole dans lesquels ils figurent, pour en faire des tours de parole à part entière. L'exemple ci-dessus devient donc, après extraction :

> **ilrMF0** (aspiration) ben je préférerais pas si |- ça ne te **@1132** -| dérange pas
> **@1132 ilrTC1** ça va alors

On a inséré le signe @ qui permet de garder la trace de l'existence du chevauchement interne. Celui-ci est noté en dessous du tour de parole dans lequel il figurait. Les tours de parole internes sont numérotés afin d'éviter tout risque de confusion, notamment lorsqu'on a plusieurs chevauchements internes dans un même tour de parole.

3.2. Particularités de l'oral

Par particularités de l'oral, nous entendons ce que la littérature appelle en général disfluences [11]. Nous avons analysé de manière approfondie certaines de ces disfluences, afin de les traiter au mieux lors de cette première phase du travail (pour le détail, voir [7]).

Les répétitions.

> **ilrMS1** (...) **sans // sans** bafouiller **sans sans sans** que **la la** langue fourche quoi

Les amorces de morphèmes. Nous appelons amorce le phénomène langagier qui consiste en "une interruption de morphèmes en cours d'énonciation" [9, p. 79].

> **ilrMS1** (...) **euh** avec la boulangerie **euh je fai/ je faisais** des pralines chez moi (...) [ilrMS1r]

Selon la typologie de [9], les amorces se répartissent en 3 catégories, qui reçoivent un prétraitement différent.

L'autocorrection immédiate (*le la fille*) : 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.

Le *euh* : appelé également pause pleine, est très fréquent à l'oral spontané. Comme l'illustre l'exemple suivant, il arrive qu'il cooccurre avec une autre disfluence, ici une répétition :

> **ilpMJ1** (...) le système démocratique dans lequel ils |- vivent <ilpGV0> mm -| / les institutions **euh** qui les qui les servent et cetera [ilpMJ1r]

Le prétraitement que nous effectuons, de manière totalement automatisé, consiste à baliser les portions de texte qui constituent, selon la terminologie de Shriberg (1994), le reparandum pour ne garder dans le texte à analyser par le chunker que le repair.

4. Le *chunker*

L'annotation en *chunks* repose sur une cascade d'automates à états finis [1]. Dans cette section, nous donnons un aperçu relativement bref de ce processus. Pour une description plus complète, nous renvoyons à [3].

La procédure de *chunking* s'articule en trois étapes successives : (1) segmentation lexicale en mots simples et composés (*i.e.* mots composés, expressions nominales et adverbiales (semi)-figées et coocurrences) ; (2) identification et étiquetage des *super-chunks* (*i.e.* chunks intégrant la notion de structure lexicale complexe et autorisant, par conséquent, certains attachements prépositionnels et adjectivaux droits) ; (3) désambiguïsation et linéarisation.

Notre système repose entièrement sur des ressources linguistiques à large couverture. Selon la nature du texte d'entrée, ces ressources peuvent être augmentées et/ou modifiées.

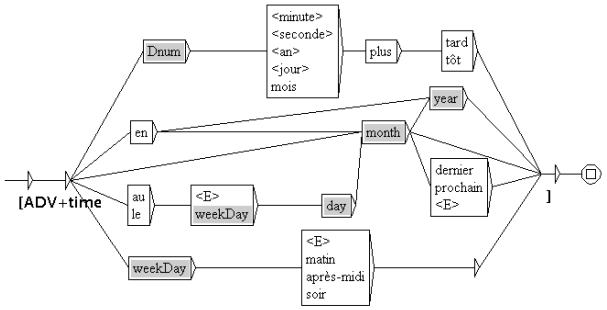

Fig. 1: Exemple de grammaire locale (adverbe de temps)

4.1. Segmentation lexicale

La segmentation lexicale s'effectue au départ d'un ensemble de ressources linguistiques (RLs) développées par des linguistes. Ces ressources se composent principalement de dictionnaires morpho-syntaxiques des formes fléchies du français. Le plus important en taille rassemble 746 198 formes simples et 249 929 formes

composées du français standard [6]. Chaque entrée est associée à un lemme, une catégorie syntaxique, des informations morphologiques (*e.g.* genre et nombre), syntaxiques (*e.g.* la structure interne des composés) et sémantiques (*e.g.* trait humain pour les noms). Les mots composés rassemblent des formes de différents types :

– noms : *pomme de terre*, *faux témoignage*
– prépositions : *au milieu de*, *à cause de*
– adverbes : *par ailleurs*, *en pratique*
– conjonctions : *bien que*, *pendant que*

Par ailleurs, nous avons construit deux dictionnaires qui prennent en compte la double particularité de notre corpus de "français parlé en Belgique francophone". L'un recense les particularités lexicales du français en Belgique, avec des formes comme *guindailleur* (*fêtard*) ou *à fond de balle* (*à fond de train*) ; l'autre tient compte de formes non recensées dans les RLs ou qui peuvent recevoir une autre catégorie grammaticale à l'oral (par exemple, *allez* qui dans certains cas doit être analysé comme une interjection et non comme une forme du verbe *aller*).

Outre les dictionnaires, les ressources linguistiques incluent également une librairie de 190 grammaires locales (*i.e.* automates lexicaux [8]). Ces grammaires visent la reconnaissance de structures nominales, prépositionnelles, déterminatives et adverbiales telles que :

– noms : noms de fonctions [*ministre anglais de l'Agriculture*]
– prépositions : prépositions locatives [*à dix kilomètres au nord de*]
– déterminants : déterminants numériques [*vingt-sept*, *des milliers de*], déterminants nominaux [*dix grammes de*]
– adverbes : adverbes de temps [*en octobre 2006*]

Ce premier module prend, en entrée, un texte segmenté en phrases et tokens. L'application des dictionnaires de mots simples et composés permet, dans un premier temps, d'associer à chaque token l'ensemble des étiquettes linguistiques possibles. Nous obtenons, en sortie, un automate représentant l'ambiguïté lexicale du texte (TFSA). Les grammaires locales sont ensuite appliquées sur cet automate qui se voit augmenté d'autant de transitions qu'il y a de séquences reconnues.

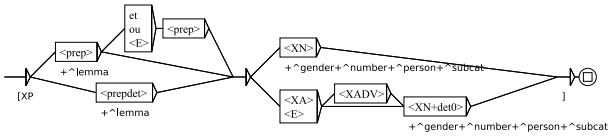

Fig. 2: *Chunk* prépositionnel

4.2. Segmentation en Super-Chunks

Afin de réaliser la segmentation en *chunks* proprement dite, nous appliquons une cascade de transducteurs sur l'automate du texte. De la même manière que pour les grammaires locales décrites au point précédent, chaque reconnaissance se traduit par l'ajout d'une transition supplémentaire au sein de l'automate du texte. La cascade est composée d'un réseau de 18 grammaires réparties en huit étapes successives :

1. *chunks* adverbiaux (`XADV`) : adverbes simples et complexes reconnus lors de la segmentation lexicale
2. *chunks* adjectivaux (`XA`) : adjectifs (précédés le cas échéant d'un adverbe)
3. *chunks* nominaux (`XN`) : noms simples, entités nommées et certaines catégories de pronoms
4. *chunks* prepositionnels (`XP`) : `XN` précédé d'une préposition
5. *chunks* verbaux (cascade de 4 transducteurs) : formes passives et actives de l'infinitif (*i.e.* `XVI(P)`), des participes passé et présent (*i.e.* `XVK(P)` et `XVG(P)`) et du verbe simple (*i.e.* `XV(P)`)

Notons que chaque *chunk* identifié peut hériter de propriétés morpho-syntaxiques de sa tête et de ses constituants.

4.3. Désambiguïsation incrémentale

La segmentation en *chunks* produit un automate du texte comprennant l'ensemble des analyses possibles. Afin de supprimer cette ambiguïté et de linéariser cet automate, le *chunker* inclut un module de désambiguïsation incrémentale opérant en trois étapes successives et optionnelles.

1. L'heuristique du plus court chemin. Cette heuristique conserve uniquement les chemins les plus courts au sein de l'automate. De cette manière, nous donnons la priorité aux analyses comprenant un ou plusieurs éléments complexes.

2. Règles manuelles. Chacune des 26 règles implémentées pour cette expérience se compose de trois parties distinctes : deux contextes (gauche et droit) formalisés sous la forme d'un automate et un partie centrale représentant l'ambiguïté (*i.e.* une liste d'analyses possibles). Si cette ambiguïté, de même que ses contextes d'apparition, sont observés au sein de l'automate du texte, nous ne conservons qu'une unique analyse (les autres sont supprimées).

3. Règles stochastiques. Nous disposons de listes fréquentielles apprises sur corpus étiquetés. Ces listes nous permettent de conserver, pour chaque cas d'ambiguïté, l'analyse la plus fréquente.

5. Expérimentations et résultats

Notre corpus de référence compte 5 336 mots (*i.e.* 2 335 chunks), annotés manuellement. Un script d'évaluation nous permet de comparer automatiquement les annotations de ce corpus à celles obtenues par application de notre chunker. Nous avons effectué une double expérience : la première utilisant les ressources linguistiques propres à l'oral et la seconde les désactivant.

Les mesures utilisées sont la *précision*, le *rappel* et la *mesure* F1, telles qu'elles sont définies dans [10]. La

précision correspond au pourcentage de chunks correctement étiquetés ; le rappel indique le pourcentage de chunks dans le corpus de référence qui ont été identifiés ; et la mesure F1 est une combinaison du rappel et de la précision.

	PRÉCISION	RAPPEL	MESURE F1
AVEC	82,8 %	85,6 %	84,1%
SANS	81,5 %	83,5 %	82,5%

Tab. 1: Résultats

Par comparaison aux résultats obtenus lors de la campagne EASY [10], nous constatons que notre chunker dépasse l'état de l'art (*cf.* Table 1). En effet, si notre précision est inférieure de 2 points au meilleur chunker de la campagne, notre rappel est nettement supérieur. Par conséquent, la mesure F1 d'appréciation globale montre un gain de 5 points par rapport à ce même chunker[2]. La qualité de ces résultats est directement imputable à la finesse de nos ressources linguistiques et, plus encore, au module de normalisation syntaxique appliqué en amont du chunking. La normalisation explique également la faible différence (moins de 2 %) que l'on observe entre les expériences *avec* et *sans* ressources propres à l'oral.

Une analyse plus avancée des résultats, nous a permis d'identifier 3 sources d'erreurs principales. (1) Certaines structures complexes sont absentes de nos ressources lexicales : la forme adverbiale *nulle part*, par exemple, est analysée XN parce qu'elle est composée d'un déterminant et d'un nom. (2) La sélection aléatoire en cas d'ambiguïté résiduelle est également responsable de nombreuses errreurs d'étiquetage : une forme telle que *de*, préposition ou déterminant, conduit à la confusion entre les analyses XN et XP. (3) Il existe finalement une catégorie d'erreurs, plus restreinte, ressortissant au module de normalisation : les répétitions et disfluences, non détectées lors du prétraitement, brisent la structure syntaxique et, par conséquent, la reconnaissance elle-même.

Bien que très brève, cette évaluation tend à prouver que le recourt à un processus de normalisation nous permet de limiter efficacement les perturbations syntaxiques de l'oral et justifie donc pleinement notre approche. Notons, cependant, qu'il nous reste encore à gérer certains cas d'agrammaticalité non pris en compte par nos grammaires (*e.g. Il m'avait voulu donner*).

6. Conclusion et perspectives

Les résultats obtenus sont extrêmement encourageants. Ils montrent que la procédure de normalisation d'un corpus de l'oral permet d'approcher la structure de l'écrit, même si certains paramètres particuliers viennent moduler notre hypothèse. Dans un futur proche, nous pensons améliorer encore ces résultats car il existe de nombreuses erreurs récurrentes que l'on peut résoudre en augmentant nos ressources lexicales ou en intégrant un module stochastique (avec chaînes de Markov cachées notamment).

[2]Précisons toutefois que notre corpus d'évaluation de même que notre définition du chunk s'éloignent sensiblement du cadre défini pour la campagne EASY.

Références

[1] Steven P. Abney. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4) :337–344, 1996.

[2] Christophe Benzitoun. L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? In *Actes de RÉCITAL (21 avril 2004, Fès)*, 2004.

[3] Olivier Blanc, Matthieu Constant, and Patrick Watrin. Segmentation in super-chunks with a finite-state approach. In *Proceedings of FSMNLP 2007 – Finite-State Methods for Natural Language Processing*, 2007.

[4] Claire Blanche-Benveniste, Mireille Bilger, Christine Rouget, and Karel van den Eynde. *Le Français parlé. Études grammaticales.* CNRS Éditions, Paris, 1990.

[5] Claire Blanche-Benveniste and Colette Jeanjean. *Le Français parlé. Transcription et édition.* Didier Érudition, Paris, 1987.

[6] Blandine Courtois. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87 :11–22, 1990.

[7] Anne Dister. *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel.* PhD thesis, Université catholique de Louvain, 2007.

[8] Maurice Gross. The construction of local grammars. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 329–352. The MIT Press, Cambridge, Mass., 1997.

[9] Berthille Pallaud. Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé*, 17 :79–101, 2002.

[10] Patrick Paroubek, Anne Vilnat, Isabelle Robba, and Christelle Ayache. Les résultats de la campagne easy d'évaluation des analyseurs syntaxiques du français. In *Actes de TALN 2007 (Toulouse)*, 2007.

[11] Elizabeth Shriberg. *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, Université de Berkeley, 1994.

[12] André Valli and Jean Véronis. étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2) :113–133, 1999.

[13] Jean Véronis. Annotation automatique de corpus : panorama et état de la technique. In Jean-Marie Pierrel, editor, *Ingénierie des langues*, pages 111–129. Hermès, Paris, 2000.