

# Chunker différents types de discours oraux : défis pour l'apprentissage automatique

Iris Eshkol-Taravella<sup>1,2</sup> Mariame Maarouf<sup>2,3</sup> Marie Skrovec<sup>2</sup> Flora Badin<sup>2</sup>

(1) MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France

(2) LLL UMR7270, 10 Rue de Tours, 45065 Orléans, France

(3) Lattice UMR8094, 1 rue Maurice Arnoux, 92120 Montrouge, France

ieshkolt@parisnanterre.fr, maarouf.mariame@gmail.com, marie.skrovec@univ-orleans.fr, flora.badin@univ-orleans.fr

## RÉSUMÉ

---

Le travail décrit le développement d'un chunker pour l'oral par apprentissage supervisé avec les CRFs, à partir d'un corpus de référence de petite taille et composé de productions de nature différente : monologue préparé vs discussion spontanée. La méthodologie respecte les spécificités des données traitées. L'apprentissage tient compte des résultats proposés par différents étiqueteurs morpho-syntaxiques disponibles sans correction manuelle de leurs résultats. Les expériences montrent que le genre de discours (monologue vs discussion), la nature de discours (spontané vs préparé) et la taille du corpus peuvent influencer les résultats de l'apprentissage, ce qui confirme que la nature des données traitées est à prendre en considération dans l'interprétation des résultats.

## ABSTRACT

---

### **Chunking different spoken speech types : challenges for machine learning**

This paper describes the development of a chunker for spoken data by supervised machine learning using the CRFs, based on a small reference corpus composed of two discourse types: prepared monologue vs spontaneous talk in interaction. The methodology respects the specific character of the processed data. The machine learning considers the results of several available taggers, without manual correction of their results. Experiments show that the discourse type (monologue vs free talk), the speech nature (spontaneous vs prepared) and the corpus size can influence the results of the machine learning process. The type of data should therefore be considered in interpreting the results.

---

**MOTS-CLÉS :** chunking, apprentissage supervisé, CRF, segmentation automatique de l'oral, corpus oral, variation discursive, genre

**KEYWORDS:** chunking, machine learning, CRF, automatic segmentation of oral data, oral corpus, discursive variation, genre

---

# 1 Introduction

La notion de phrase étant généralement considérée comme peu pertinente pour l'analyse et le traitement de l'oral (Blanche-Benveniste et al, 1990 ; Groupe de Fribourg, 2012), différents types d'unités de segmentation de l'oral ont été proposés par des recherches menées dans le cadre de projets comme Rhapsodie ou Orfeo. Le projet SegCor<sup>1</sup> porte aussi sur la segmentation des corpus oraux et propose une segmentation multiniveau. Son premier niveau est une segmentation en unités minimales syntaxiques en termes de constituance, appelées chunks.

Les chunks sont des constituants continus et non récursifs (Abney 1991). Le chunking identifie la structure syntaxique superficielle d'un énoncé et peut être effectué automatiquement. Il est fondé sur un étiquetage morphosyntaxique préalablement réalisé. Ce type d'annotation est particulièrement pertinent pour l'oral car le discours oral est parfois composé d'énoncés non « finalisés » ce qui rend une analyse syntaxique complète difficile. Blanche-Benveniste (1997) a démontré que ces constituants sont le lieu de réalisation privilégié des réparations à l'oral.

Plusieurs stratégies sont possibles pour développer un chunker. Les méthodes symboliques ont été testées dans le cadre des travaux de (Blanc et al, 2008, 2010, Antoine et al, 2008) où des cascades de transducteurs développées chunkent des transcriptions de l'oral en se fondant sur des ressources lexicales et syntaxiques. L'apprentissage automatique supervisé semble être particulièrement performant sur cette tâche comme montrent les recherches de (Sha et Pereira, 2003, Tellier et al, 2012, 2014, Tsuruoka et al, 2009). Dans la suite du travail de (Tellier et al, 2014), la recherche présentée ici utilise la méthode de l'apprentissage supervisé. Les productions orales se caractérisent par une grande variété discursive : variété situationnelle (conversation privée, débat public...), tâches langagières (expliquer, raconter, décrire...), genres (récits de voyage, interview...) ou registre de langue (courant, familier, soutenu). La nature des données traitées influence et guide le processus d'apprentissage. Dans le travail de (Tellier et al, 2014), le corpus de référence était composé d'entretiens sociolinguistiques ; dans celui-ci nous nous fondons sur deux autres situations de communication : une conférence et une discussion spontanée entre plusieurs personnes lors d'un repas.

## 2 Constitution du corpus de référence

Les données traitées proviennent de deux grands corpus de français parlé contemporain : ESLO<sup>2</sup> et CLAPI<sup>3</sup>. Deux types de discours sont sélectionnés : une conférence donnée par un locuteur, un monologue préparé (10 minutes, 2120 tokens) dans le corpus ESLO2 (M) et une discussion entre trois personnes, une interaction spontanée, se déroulant dans un contexte d'ordre privé (10 minutes, 2461 tokens) dans le corpus CLAPI (R).

---

1 Un projet franco-allemand, financé par l'Agence Nationale de Recherche (ANR-15-FRAL-0004)

2 Enquêtes Sociolinguistiques à Orléans, <http://eslo.huma-num.fr/>

3 Corpus de LAngue Parlée et Interaction, <http://clapi.ish-lyon.cnrs.fr/>

## 2.1 Prétraitement

Les deux fichiers utilisés pour ce travail sont prétraités en termes de segmentation et d'annotation. Les tokens, les annotations et le signal sonore sont d'abord alignés semi-automatiquement<sup>4</sup>. Les unités polylexicales (*comme ça, plein de, ciné club*) sont repérées ensuite grâce à la ressource Leff (Sagot et al., 2010). Le résultat du prétraitement est montré dans la Figure 1.

à	l'	école	primaire	on	avait	un	ciné	club	spk2[ortho] (2500)
à	l'	école	primaire	on	avait	un	ciné club		spk2[POSok] (2378)
PRP	D	NOM	ADJ	PR O:P	VER:impf	DET:A RT	NOM		spk2[pos] (482/2378)

Figure 1 : Résultat et visualisation du prétraitement sous Praat<sup>5</sup>

## 2.2 Typologie des chunks

La typologie de chunks est fondée sur celle présentée dans Tellier *et al.* (2014) et complétée par deux nouvelles étiquettes (FNO et ARTIC). Elle contient neuf catégories :

- adjectival phrase (AP) : chunk adjectival - l'adjectif tête placé après le verbe (*elle est trop jolie*) ;
- adverbial phrase (AdP) : chunk adverbial - un syntagme dont la tête est un adverbe (*peut-être*) ;
- nominal phrase (NP) : chunk nominal - les syntagmes nominaux intégrant les adjectifs placés avant et après le nom et les pronoms non clitiques (*tes belles chaussures*) ;
- prepositional phrase (PP) : chunk prépositionnel - les syntagmes introduits par une préposition (*de loin*) ;
- verbal phrase ou verbal nucleus (VP) : chunk verbal – les syntagmes organisés autour d'une tête verbale, associée à ses clitiques (*on nous entend*), fléchie ou non ;
- ponctuation (SENT) : les transcriptions ne contiennent pas de marques typographiques, sauf des points d'interrogation conservés pour plus de lisibilité ;
- articulateur (ARTIC) : une catégorie qui regroupe des éléments non autonomes reliant des unités de différents niveaux, qu'il y ait dépendance syntaxique ou non, comme les pronoms relatifs, les conjonctions, les marqueurs discursifs (*et, que, lequel, enfin, mais, du coup, etc.*) ;
- forme noyau (FNO) : inspirée des travaux de Benzitoun *et al.* (2012), cette catégorie regroupe des éléments autonomes, non périphériques, constituant à eux seuls une unité illocutoire (*oui, ouf, merde, d'accord, voilà, bonjour, salut, mince, santé, etc.*) ;
- inconnu (UNKNOWN) : une catégorie regroupant les chunks non identifiés, comme les amorces de mots, les mots mal orthographiés, etc.

4 Découpage en unités polylexicales et annotation en POS : Treetagger (Schmid, 1994), Dismo (Christodoulides et al., 2014) et réalignement manuel sur le signal sonore

5 Praat est un outil de transcription et d'annotation manuelle de l'oral (<http://www.fon.hum.uva.nl/paul/praat.html>).

## 2.3 Annotation manuelle

Les deux corpus prétraités sont d’abord annotés par deux chercheurs selon la typologie établie. L’accord inter-annotateur calculé en appliquant le kappa de Cohen (Cohen, 1960) est de 88%. La troisième annotation de consensus est effectuée par la suite sur le même corpus, elle sert de corpus de référence et d’évaluation pour l’apprentissage automatique. L’annotation est réalisée à l’aide du logiciel Praat (Boersma et Van Heuven, 2001) et en utilisant le format BILOU<sup>6</sup> (Ratinov et Roth, 2009) permettant de délimiter une unité mais aussi de déterminer la place de chaque terme au sein de cette unité. Grâce à Praat, les annotateurs ont accès à l’écoute des enregistrements pour mieux comprendre certaines situations ce qui n’a pas été effectué dans le cadre du travail de (Tellier *et al.* 2014). Le corpus ainsi annoté contient 1069 chunks dans M et 1455 chunks dans R répartis de manière hétérogène dans les deux corpus (la présence importante de PP 30% dans M vs 11% dans R contrairement au VP représentant 40% dans R vs 23% dans M, etc.).

## 3 Apprentissage automatique

L’apprentissage automatique vise à indiquer des frontières de chaque chunk, mais aussi à déterminer son type. Le corpus de référence ayant une petite taille, nous optons pour le modèle des CRFs (*Conditional Random Fields*) linéaires (Lafferty *et al.*, 2001) qui a déjà fait preuve d’une bonne performance pour cette tâche (Sha et Pereira, 2003, Tellier *et al.*, 2012, 2014, Tsuruoka *et al.*, 2009). Le chunking s’applique sur le corpus étiqueté en POS. Tellier *et al.* (2014) ont montré qu’on peut apprendre un chunker propre à l’oral avec des POS non corrigées et avec un corpus de référence de petite taille. Les auteurs arrivent à 88% de micro-*average*. Nous poursuivons la même démarche mais avec une méthodologie redéfinie en fonction de la spécificité des données orales : (1) les données traitées sont plus hétérogènes car elles comprennent deux types de discours oral ; (2) les annotateurs humains ont systématiquement recours à l’écoute du son pour déterminer les choix d’annotation ; (3) le jeu d’étiquettes est retravaillé (l’ajout de deux nouvelles étiquettes *ARTIC* et *FNO*); (4) les résultats de plusieurs étiquetages morpho-syntaxiques sont ajoutés dans les traits intégrés au modèle CRF ce qui permet de vérifier si une série d’étiquettes POS non corrigées proposées par différents étiqueteurs pour le même mot améliore les résultats du chunking et quels outils parmi ceux testés sont les plus pertinents pour le corpus oral traité.

Quatre étiqueteurs sont testés : TreeTagger (Schmidt, 1994) ; SEM (Tellier *et al.*, 2012) exploité par (Tellier *et al.*, 2014) et utilisant les étiquettes morpho-syntaxiques de (Crabbé *et al.*, 2008) ; parseur en dépendance syntaxique (Kahane *et al.*, 2017) développé dans le cadre du projet Orfeo, d’où nous extrayons uniquement les POS et les POS du gouverneur syntaxique du token courant ; Perceo (Benzitoun *et al.*, 2012), étiqueteur adapté pour l’oral qui a la particularité de posséder une étiquette FNO, étiquette aussi présente dans notre typologie de chunks.

---

6 B pour Begin, premier token du chunk; I pour In, un élément à l’intérieur d’un chunk ; L pour Last, dernier élément du chunk ; O pour Out, un élément extérieur, absent dans le corpus car tous les tokens font partie d’un chunk ; U pour Unit, un chunk composé d’un seul token.

Les expériences sont effectuées sur trois corpus : ESLO2 (M), CLAPI (R), ESLO2+CLAPI (M+R). L'objectif est de vérifier si le genre de discours (monologue/discussion entre 3 personnes), la nature de discours (spontané/préparé) et la taille du corpus peuvent influencer les résultats d'apprentissage. De nombreuses configurations sont testées afin d'obtenir des résultats exhaustifs en combinant et variant le nombre de patrons [token + POS] pris en compte<sup>7</sup>. Pour le parseur d'Orfeo, deux combinaisons supplémentaires sont testées (1) en prenant en compte uniquement l'étiquette POS du token, (2) l'étiquette POS du token et de son gouverneur. En premier lieu, pour chacune des combinaisons la prise en compte du token de la ligne courante est testée. Ensuite, les trois combinaisons donnant les meilleurs résultats pour chaque corpus sont sélectionnées pour les tests en incluant token+1 et token-1. Après isolation du meilleur résultat, d'autres colonnes sont ajoutées comme par exemple le lemme, récupéré depuis l'annotation TreeTagger, pour tester une possible amélioration du score. La Figure 2 montre les meilleures combinaisons de patrons pour chaque corpus.

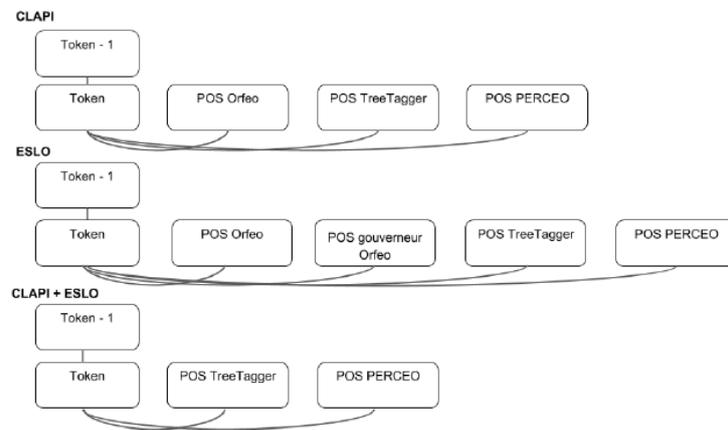


Figure 2 : Meilleures combinaisons de patrons pour chaque corpus

## 4 Résultats et évaluation

L'évaluation des résultats est effectuée en validation croisée à 10 plis<sup>8</sup> et sur trois corpus (M, R, M+R) séparément. Trois mesures différentes sont utilisées pour évaluer les performances de l'annotation : la précision, le rappel et la F-mesure<sup>9</sup>. Ces mesures sont calculées pour chaque type de chunks et à partir de ces résultats, on obtient la micro-averaging<sup>10</sup>. Cette pondération permet

7 token+SEM, token+SEM+TTG, token+SEM+TTG+Orfeo, token+SEM+TTG+Orfeo+Perceo, token+Orfeo, token+Orfeo+TTG..., token+1 et token-1.

8 En réalisant un apprentissage sur 9/10 des exemples, on se prive de peu de données tout en s'assurant de fournir une évaluation peu « biaisée » car elle est une moyenne de plusieurs expériences.

9 la moyenne harmonique du rappel et de la précision

10 la moyenne pondérée des résultats obtenus des différents types de chunks

d'attribuer plus ou moins de poids aux résultats en fonction de leur taux de présence dans le corpus. Ainsi, plus une catégorie de chunks est présente dans le corpus, plus son score aura d'importance dans le calcul de la micro-average et inversement.

	M	R	M+R
	85,8%	83,2%	85,7%
POS TreeTagger	x	x	x
POS Perceo	x	x	x
POS Orfeo	x	x	
Gouv Orfeo	x		
Tok_courant Tok_precedent	x	x	x

Tableau 1 : Tableau de meilleurs résultats obtenus en termes de micro-average

Pour le corpus R, les meilleurs résultats sont produits par la combinaison qui regroupe les outils TreeTagger, PERCEO et le parseur d'Orfeo auxquels s'ajoutent le token courant et le token précédent (83,2 %). Les meilleurs scores obtenus pour le corpus M proviennent de l'application d'une combinaison similaire à celle de R sauf qu'en plus de l'étiquette POS récupérée par le parsing en dépendance du parseur d'Orfeo, l'étiquette du gouverneur est aussi présente (85,8 %). Le monologue, extrait d'une conférence, est composé d'énoncés plus longs et ne contient pas d'interaction, les liens de dépendances y sont plus présents. Dans le cas du corpus M+R, les meilleurs résultats sont obtenus en n'utilisant que les POS de TreeTagger et de Perceo, en plus du token précédent au token courant (85,7 %). Ces résultats montrent que l'utilisation des étiquettes proposées par deux outils de l'oral, PERCEO et le parseur d'Orfeo, dans les patrons est tout à fait pertinente. La taille du corpus est un critère important aussi car le corpus plus long n'a pas besoin de beaucoup d'étiquettes POS et peut se contenter des résultats de deux outils. Il est étonnant que TreeTagger semble être plus pertinent dans ce cas que le parseur d'Orfeo. Le corpus M, un monologue préparé, se prête mieux à l'apprentissage que le corpus R, la discussion spontanée.

L'évaluation du chunker par étiquette montre que l'étiquette FNO obtient les moins bons résultats (23,52% de F-mesure). En effet, certains tokens sont ambigus, comme par exemple *oui*, *ouais*, *non*, tantôt FNO, tantôt articulateurs (marqueurs discursifs). Ainsi, un *ouais* en réponse à une question sera considéré comme prédicat autonome (un « mot-phrase ») et donc annoté (FNO), comme ici :

ELI je [VP B] vous [VP I] sers [VP L] ?  
BEA ouais [FNO U]

En revanche cette même forme peut être considérée comme élément périphérique au prédicat. Il s'agira alors d'un articulateur discursif non autonome, comme dans l'exemple ci-dessous, où *ouais* opère comme balise de clôture du tour de ELI :

ELI non [ARTIC U] mais [ARTIC U] tu [VP B] sais [VP L]  
 tu [VP B] en [VP I] mets [VP L] pas [AdP B] beaucoup [AdP L]  
 tu [VP B] en [VP I] mets [VP L] un [NP B] fond [NP L] ouais [ARTIC U]

On relève quelques autres erreurs courantes. Ainsi, de nombreux chunks NP sont annotés comme PP à cause de l’ambiguïté entre la préposition *de* suivi du déterminant défini et l’article partitif (*du, de la, etc.*), tous les deux ayant la même forme. Par ailleurs, un quart des AP sont considérés comme des VP car souvent un token de type AP suit un chunk de type VP. D’une manière générale, les frontières de chunks (les étiquettes B, L, U) sont mieux annotées (Tableau 2).

	B	I	L	U
R	0,94	0,86	0,91	0,94
M	0,92	0,87	0,93	0,9
M+R	0,93	0,86	0,92	0,93

Tableau 2 : Résultats de F-mesure pour les étiquettes BILU

## 5 Conclusion

Les productions orales se caractérisent par une grande variété discursive. L’article décrit le développement d’un chunker par apprentissage automatique avec les CRFs en utilisant un corpus de référence de petite taille comprenant les données orales de nature différente : monologue dans le cadre d’une conférence vs discussion spontanée entre 3 personnes lors d’un repas. Un genre et un type de discours peuvent influencer les résultats d’apprentissage. Ainsi, les résultats du parsing en dépendance sont plus pertinents à intégrer au modèle CRF pour le monologue où les énoncés longs se prêtent plus à ce type d’analyse. Les FNO obtiennent de meilleurs scores dans une discussion car ils y sont plus nombreux. La nature des données traitées est donc à prendre en considération dans l’interprétation des résultats. Plusieurs perspectives sont envisagées : (1) d’ajouter certaines informations issues des enregistrements comme la prosodie ; (2) de laisser les deux options dans les cas où les annotateurs humains hésitent entre différentes étiquettes possibles ce qui améliorera les résultats du chunker ; (3) d’ajouter des règles d’annotation pour certains phénomènes récurrents et systématiques comme la précision qu’un tour de parole commence toujours par une frontière B ou U ; (4) d’intégrer dans le corpus d’apprentissage le maximum de situations de communication pour généraliser le développement d’un chunker pour l’oral.

## Remerciements

Ce travail a été effectué dans le cadre du stage de Mariame Maarouf, co-encadré par Isabelle Tellier, qui nous a quittés le 1 juin 2018. Nous tenons ici à lui rendre un hommage affectueux et à lui témoigner notre gratitude pour son enthousiasme, ses idées et ses conseils avisés, sans lesquels cet article n’aurait pu voir le jour.

# Références

- Abney S. (1991). *Parsing by chunks*. In R. Berwick, R. Abney, and C. Tenny, editors, Principle-based Parsing. Kluwer Academic Publisher.
- Antoine J.-Y., Mokrane A., Friburger N. (2008). Automatic rich annotation of large corpus of conversational transcribed corpus. Actes de *LREC 2008*.
- Benzitoun C., Fort K., Sagot B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. Actes de *JEP-TALN 2012*, 99-112.
- Blanc O., Constant M., Dister A., Watrin P. (2008). Corpus oraux et chunking. Actes de *Journées d'étude sur la parole (JEP)*.
- Blanc O., Constant M., Dister A., Watrin P. (2010). Partial parsing of spontaneous spoken french. Actes de *LREC'10*.
- Blanche-Benveniste C., Bilger M., Rouget C., Van Den Eynde K. (1990). *Le français parlé*. Études grammaticales, Paris, CNRS Éditions.
- Blanche-Benveniste C. (1997). *Approches de la langue parlée en français*. Paris, Ophrys.
- Boersma P., Van Heuven V. (2001). Speak and unSpeak with Praat. *Glott International*, 5(9/10), 341-347.
- Christodoulides G., Avanzi M., Goldman J-P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. Actes de *LREC'14*.
- Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Groupe de Fribourg, (2012), *Grammaire de la période*, Berne, Peter Lang.
- Kahane S., Deulofeu J., Gerdes K., Nasr A., Valli A. (2017). Annotation micro et macrosyntaxique manuelles et automatique de français parlé. *Journée Floral*, mars 2017, Orléans.
- Lafferty J., McCallum A., Pereira F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Actes de *ICML 2001*, 282-289.
- Ratinov L., Roth D. (2009). Design challenges and misconceptions in named entity recognition. Actes de *CoNLL*.

Sagot B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. Actes de *LREC 2010*.

Schmidt H. (1994). Probabilistic part-of-speech tagging using decisions trees. Actes d'*International Conference on New Methods in Language Processing*, 44-49.

Sha F., Pereira F. (2003). Shallow parsing with conditional random fields. Actes de *HLT-NAACL 2003*, 213-220.

Tellier I., Duchier D., Eshkol I., Courmet A., Martinet M. (2012). Apprentissage automatique d'un chunker pour le français, Actes de *TALN 2012*.

Tellier I., Eshkol-Taravella I., Dupont Y., Wang I. (2014). Peut-on bien chunker avec de mauvaises étiquettes pos ? Actes de *TALN2014*.

Tsuruoka Y., Tsujii J., Ananiadou S. (2009). Fast full parsing by linear-chain conditional random fields. Actes de *EACL 2009*.

