Peut-on bien chunker avec de mauvaises étiquettes POS ?

Iris Eshkol-Taravella, Isabelle Tellier, Yoann Dupont, Ilaine Wang

Peut-on bien chunker avec de mauvaises étiquettes POS ?

Résumé. Dans cet article, nous testons deux approches distinctes pour chunker un corpus oral transcrit, en cherchant à minimiser l'étape de correction de l'étiquetage en POS. Nous ré-utilisons tout d'abord un chunker appris sur des données écrites, puis nous tentons de ré-apprendre un chunker spécifique de l'oral à partir de données annotées et corrigées manuellement. L'objectif est d'atteindre les meilleurs résultats possibles pour le chunker en se passant autant que possible de la correction manuelle des étiquettes POS, trop couteuse. La méthodologie choisie est donc guidée par un principe d'économie. Notre travail montre qu'il est possible d'apprendre un nouveau chunker performant pour l'oral à partir d'un corpus de référence annoté de petite taille, sans correction manuelle des étiquettes POS.

Abstract. In this paper, we test two distinct approaches to chunk transcribed oral data, trying to minimize the phase of POS labeling correction. First, we use an existing chunker, learnt from written texts, then we try to learn a new specific chunker from manually corrected labeled oral data. The purpose is to reach the best possible results with as few costy manual corrections of the POS labels as possible. Our methodology is thus economy oriented. Our work shows that it is possible to lean a new effective chunker for oral data from a labeled reference corpus of small size, without manual correction of POS labels

Mots-clés : chunker, étiquetage POS, apprentissage automatique, corpus oral, disfluences
Keywords: chunker, POS labeling, machine learning, oral corpus, disfluencies

1 Introduction

L'annotation du discours oral transcrit est une tâche difficile. Le langage oral est en effet différent de l'écrit, il se caractérise par des phénomènes qui lui sont propres regroupés sous l'appellation générale de *disfluences*. Nous nous intéressons dans cet article au processus de segmentation de textes en chunks. Les chunks sont des constituants continus et non-récursifs (Abney, 1991). L'objectif de la tâche de chunking est d'identifier la structure syntaxique superficielle d'un énoncé, c'est-à-dire de reconnaître ses constituants minimaux sans spécifier leur structure interne et leurs fonctions syntaxiques. Pour les transcriptions de l'oral, pour lesquelles une analyse syntaxique complète est souvent impossible, le chunking représente un degré d'analyse adapté. Il a en effet par exemple été démontré que ses constituants sont le lieu de réalisation privilégié des réparations à l'oral (BLANCHE-BENVENISTE C., 1997 : 47).

Plusieurs stratégies sont envisageables pour construire un chunker. Une des techniques efficaces avérée pour cette tâche est l'apprentissage automatique à partir de données annotées. (SHA F, PEREIRA P., 2003). Le problème qui se pose souvent avec cette technique est le besoin de disposer d'un corpus d'apprentissage de grande taille, contenant suffisamment de répétitions pour que l'apprentissage automatique en tire parti. Or, rares sont les corpus annotés de l'oral. Est-il possible d'appliquer une méthode d'apprentissage si le corpus de référence dont on dispose est petit ? C'est peut-être envisageable pour la tâche de chunking, car ce processus est surtout fondé sur les résultats de l'étiquetage des textes en POS. Et apprendre un chunker qui s'appuie sur des étiquettes POS demande a priori beaucoup moins de données annotées qu'apprendre un étiqueteur POS qui s'appuie sur des mots, parce que la variabilité des données de départ (les étiquettes POS dans un cas, les mots dans l'autre) est bien moindre.

Nous voulons donc explorer dans cet article la possibilité d'apprendre un chunker efficace pour l'oral avec un nombre limité de données annotées, mais sans passer par une étape d'apprentissage d'un étiqueteur POS spécifique de l'oral (qui requerrait plus de données annotées), et en minimisant la correction manuelle de l'étiquetage POS. Les deux principales approches que nous avons testées pour cela sont l'utilisation d'un chunker appris sur un corpus écrit et l'apprentissage d'un nouveau chunker à partir de données orales annotées (même si elles ne sont pas nombreuses) avec des chunks plus adaptés à l'oral. Dans chaque cas, plusieurs expériences ont été effectuées, mettant ou non en œuvre une phase de correction manuelle des étiquettes POS issues d'un étiquetage morphosyntaxique préalable.

L'article suit la structure suivante. Tout d'abord, nous évoquons la tâche de chunking, ses spécificités dans le cas de l'oral ainsi que les corpus à notre disposition : le corpus annoté de textes écrits (French Treebank) de Paris 7 et un extrait du corpus oral transcrit ESLO 1 qui est la cible de notre chunking final (section 2). Nous décrivons ensuite (en section 3) les chunkers qui seront utilisés : ils proviennent tous de la même technique d'apprentissage automatique, mais appliquée sur la base de données différentes. Nous exposons enfin dans la dernière partie (la section 4) les résultats des deux stratégies utilisées pour chunker les données orales transcrites, avec différents degrés de corrections manuelles requis.

2 La tâche et les données

2.1 Chunking des données orales transcrites

Les chunkers, aussi appelés « shallow parsers », sont bien adaptés aux données orales transcrites dans lesquelles les énoncés ne sont pas souvent « finalisés ». Deux problèmes majeurs se posent aux outils annotant l'oral : les disfluences, qui rompent la linéarité du discours, et le manque de ponctuation dans les transcriptions. Pour (Dister 2007), les disfluences sont les « marques typiques des énoncés en cours d'élaboration » qui « constituent un piétinement sur l'axe syntagmatique de l'énoncé et […] nécessitent d'être prises en compte par le système d'étiquetage. ».

Les disfluences typiques sont les suivantes (extraits du corpus ESLO, décrit plus loin) :

- les hésitations : *madame euh comment vous faîtes une omelette*
- les faux-départs : *il va y avoir encore des encore mais*
- les répétitions : *le le*
- les autocorrections : *juste après le la fin du premier cycle*
- les reformulations : *on fait ce que l'on appelle un carton c'est-à-dire le le ce dessin-là agrandi*
- les amorces : *vous v- vous êtes in- institutrice*
- etc.

Elles constituent un vrai problème pour l'analyse automatique de l'oral (Adda-Decker et al. 2003, Antoine et al. 2003, Benzitoun 2004, Valli et Véronis 1999, etc.) car elles réduisent considérablement les performances des outils construits pour l'écrit standard. Nos propres expériences confirmeront ce constat (cf. section 4.1). La notion de phrase, essentiellement graphique, a rapidement été abandonnée par les linguistes qui s'intéressent à l'oral ; par conséquent les transcriptions ne sont en général pas ponctuées pour éviter l'anticipation de l'interprétation (Blanche-Benveniste et Jeanjean, 1987).

Il existe des solutions spécifiques pour le chunking du français transcrit :

- (Blanc et al., 2010) ont essayé d'annoter un corpus oral français en « super-chunks » (chunks contenant les multi-mots complexes), en appliquant des cascades de transducteurs utilisant des ressources lexicales et syntaxiques. Le processus est fondé sur une étape de prétraitement des données consistant dans le reformatage et l'étiquetage des disfluences. Une approche similaire a été adoptée par (Véronis et Vailli 1999) pour l'étiquetage morphosyntaxique de l'oral.
- (Antoine et al., 2008) ont proposé une autre stratégie incluant une étape de post-correction pour traiter les erreurs liées aux disfluences.

Notre approche est différente. Suite à (Benveniste, 2005), nous considérons que les phénomènes de disfluences doivent être inclus dans l'analyse linguistique, même s'ils exigent des traitements spécifiques. Pour faire face aux données réelles et pour éviter les programmes *ad hoc* écrits à la main, nous privilégions les techniques issues de l'apprentissage automatique.

2.2 Le French TreeBank (FTB) et ses étiquettes

Le premier corpus dont nous devons tenir compte, parce qu'il a fixé les jeux d'étiquettes (aussi bien au niveau des POS qu'à celui des chunks) que nous utilisons, est FTB (French TreeBank)[1]. C'est un corpus de phrases écrites syntaxiquement analysées qui peut être facilement transformé en phrases annotées en POS et en chunks (Abeillé et al., 2003). Le jeu réduit de 30 étiquettes POS est décrit dans (Crabé, Candito, 2008). Les sept types de chunks extraits de ces données, avec les étiquettes POS correspondant à leur tête, sont les suivants :

- les groupes nominaux ou *NP* (incluant *CLO, CLR, CLS, NC, NPP, PRO, PROREL, PROWH* : notons que les pronoms sont ici considérés comme des chunks nominaux autonomes et pas inclus dans les noyaux verbaux) ;
- les groupes verbaux ou *VN*, incluant les formes interrogatives, infinitives, modales (*V, VIMP, VINF, VPP, VPR, VS*);
- les groupes prépositionnels ou *PP*, incluant les groupes nominaux introduits par une préposition (*P, P+D, P+PRO*);
- les groupes adjectivaux ou *AP*, incluant les éventuels adverbes modifieurs d'adjectifs (*ADJ, ADJWH*) ;
- les groupes adverbiaux ou *AdP*, incluant les modifieurs de phrases (*ADV, ADVWH, I*);
- les groupes de conjonction ou *CONJ* (*CC, CS*).

2.3 ESLO 1

Le deuxième corpus utilisé est un tout petit extrait du corpus oral transcrit ESLO 1 (Enquête Sociolinguistique d'Orléans)[2] (Eshkol-Taravella et al. 2012) constitué de 8093 mots correspondant à 852 tours de parole.

Les conventions de transcription dans ESLO respectent deux principes : l'adoption de l'orthographe standard et le non-recours à la ponctuation de l'écrit. Les marques typographiques comme le point, la virgule, le point d'exclamation ou encore la majuscule en début d'énoncé sont absentes. La segmentation a été faite soit sur une unité intuitive de type « groupe de souffle » repérée par le transcripteur humain, soit sur un « tour de parole », défini uniquement par le changement de locuteurs.

3 Etiqueteur et chunkers utilisés

3.1 SEM, un étiqueteur-chunker appris sur le FTB

L'outil d'annotation utilisé dans un premier temps est SEM[3] (Tellier et al. 2012), un segmenteur-étiqueteur capable d'enchaîner plusieurs annotations successives. SEM est spécialisé sur les textes écrits, puisqu'il a été appris uniquement à partir du FTB, ses étiquettes sont donc celles présentées précédemment. Il permet soit de chunker un texte déjà annoté en POS, soit d'enchaîner « annotation POS + chunks ». Nous exploiterons par la suite ces deux usages distincts.

SEM a été appris à l'aide d'un CRF (Conditional Random Fields) linéaire (Lafferty et al. 2001), implémenté dans le logiciel Wapiti[4] (Lavergne et al. 2010). Pour l'étiquetage en POS, SEM utilise une ressource extérieure : le LeFFF (Lexique des Formes Fléchies du Français) (Sagot, 2010) intégré dans les données sous la forme d'attributs booléens. Pour le chunker, le modèle CRF s'appuie à la fois sur l'étiquetage POS et sur les tokens initiaux.

Le découpage en chunks est traduit par une annotation qui suit le format standard *BIO* (*B* pour Beginning, *I* pour In, *O* pour Out). Avec SEM, chaque mot (ou token) du corpus reçoit donc, outre son étiquette POS, une étiquette qui est la concaténation du type de chunk dont il fait partie et d'une étiquette (*B* ou *I*) qui indique la position qu'il y occupe.

---

1 http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php

2 http://eslo.tge-adonis.fr/

3 http://www.lattice.cnrs.fr/sites/itellier/SEM.html

4 http://wapiti.limsi.fr

3.2 Technique d'apprentissage de nouveaux chunkers

Nous ne chercherons pas à apprendre un nouvel étiqueteur POS spécifique de l'oral mais, en revanche, nous chercherons dans certaines expériences à apprendre un nouveau chunker à partir de données orales annotées à la fois en POS et en chunks. Pour apprendre ce nouveau chunker (en fait, il y en aura plusieurs, suivant la nature des étiquettes utilisées), nous utiliserons, comme cela avait été fait pour apprendre SEM, des CRF linéaires.

Les CRF sont des modèles graphiques probabilistes non dirigés discriminants particulièrement efficaces pour la prédiction d'étiquettes. Dans le cas des modèles linéaires, ils cherchent la meilleure séquence d'étiquettes *y* à associer à la séquence de données d'entrée *x,* en maximisant une probabilité P(y|x). La probabilité P(y|x) s'exprime dans un CRF par une combinaison pondérée (les poids étant les paramètres de l'apprentissage) de fonctions caractéristiques (ou features), qui caractérisent des configurations locales de données et d'étiquettes. Pour définir l'ensemble des features de son modèle, l'utilisateur d'un programme comme Wapiti spécifie des patrons (ou templates) : sortes d'expressions régulières pouvant faire intervenir n'importe quelle propriété des données d'entrée, et une (cas des patrons unigrammes) ou deux (patrons bigrammes) étiquettes successives. Les patrons seront instanciés sur l'ensemble d'apprentissage, constitué de couples (x,y), en autant de features que de positions où ils peuvent s'appliquer.

Dans le cas de l'apprentissage d'un chunker, les données d'entrée *x* sont constituées des tokens (ou mots) du texte et des étiquettes POS associées, les étiquettes cibles de *y* sont les différents types de chunks associés à *B* ou *I*. Les patrons que nous utiliserons pour apprendre ce(s) nouveau(x) chunker(s) ont été copiés sur ceux utilisés pour l'apprentissage de SEM, et seront toujours les mêmes pour chaque expérience. Ils figurent dans la Table 1 :

Attribut	Fenêtre sur x	Type de feature sur y
token	[-2, 0]	unigramme
POS	[-2, 1]	unigramme et bigramme
Couple de POS	{-2, 0} et {-1, 0}	unigramme

TABLE 1 : spécification des patrons (templates) définissant les features des modèles CRF de chunking

4 Deux séries d'expériences

Nous décrivons dans cette section les deux séries d'expériences réalisées avec le corpus oral transcrit et les résultats obtenus. La Table 2 montre l'annotation du même exemple extrait d'ESLO 1 par différents processus, qui incluent (en gras : colonnes **III**, **IV** et **V**) ou non (colonne **II**) une phase de correction manuelle. Les différentes colonnes de ce tableau synthétique serviront soit de données d'entrée soit de données de références à nos différentes expériences. Elles seront décrites en détail au fur et à mesure que nous les utiliserons.

I	II	III	IV	V
Tokens	SEM POS	**POS corrects**	**Chunks « FTB » corrects**	**Chunks adaptés à l'oral**
euh	*DET*	***I***	***AdP-B***	***IntP-B***
l-	*DET*	***UNKNOWN***	***AdP-B***	***UNK-B***
dans	*P*	***P***	***PP-B***	***PP-B***
ma	*DET*	***DET***	***PP-I***	***PP-I***
classe	*NC*	***NC***	***PP-I***	***PP-I***

TABLE 2 : les différentes données d'entrée/de référence utilisées

Pour nos évaluations, deux chunks seront considérés comme égaux lorsqu'ils partagent exactement les mêmes frontières et le même type. Nous évaluerons les résultats du chunking avec la micro-average des F-mesures des différents chunks (moyenne des F-mesures de ces chunks pondérées par leurs effectifs) et leur macro-average (moyenne sans pondération

des F-mesures). Notons que sur le FTB, en validation croisée à 10 plis, SEM a été évalué avec une exactitude 97,33% pour l'étiquetage en POS, une micro-average de 97,53 et une macro-average de 90,4 pour le chunker.

4.1 Première approche : utilisation d'un chunker appris sur l'écrit

4.1.1 Utilisation directe de SEM

Le premier test consiste à appliquer SEM, sans aucune adaptation ni ré-apprentissage, sur les données transcrites de l'oral. SEM est utilisé sur le texte brut, et produit à la fois l'étiquetage en POS et celui correspondant au chunking. Dans la Table 2, cela correspond à prendre comme données d'entrée pour le POS la colonne **I** (les tokens), et comme données d'entrée pour le chunker les colonnes **I** et **II** (les étiquettes POS fournies par SEM sur ESLO 1).

4.1.1.1 Etiquettes de chunks utilisées

Pour évaluer la qualité d'un chunker sur l'oral, il faut constituer un corpus de référence en corrigeant l'annotation en chunks produite par SEM sur l'extrait d'ESLO 1, avec les mêmes étiquettes de chunks que SEM (c'est la colonne **IV** dans la Table 2). Découper en chunks la transcription de l'oral pose des problèmes spécifiques à cause entre autres des disfluences. Nous explicitons ici les choix que nous avons faits.

L'exemple de l'énoncé annoté dans la Table 2 (*euh l- dans ma classe*) montre bien le type de difficultés rencontrées. Les *euh* d'hésitation, ne pouvant pas être une tête de chunk, font partie de chunks adverbiaux (*AdP)*. Cela concerne également les interjections comme dans l'exemple ci-dessous :

(on/CLS)NP (peut/V)VN (commencer/VINF)VN ***(bon/I)AdP*** *(alors/I)AdP*

Les faux départs et les amorces (comme *l-* dans l'exemple de la Table 2) qui sont impossibles à interpréter font partie également de chunks adverbiaux (*AdP)*. Les autres cas où l'interprétation est possible, l'annotation se fait selon le contexte. Dans l'exemple :

(vous/PRO)NP (êtesV)VN ***(in-/NC)NP*** *(institutrice/NC)NP*

l'amorce *in-* semble correspondre exactement au début du mot suivant *institutrice*, elle est donc annotée en tant que nom commun (*NC*) et forme par conséquent un chunk nominal autonome (*NP*). Dans l'exemple suivant :

(chez/P vous/PRO)PP (chez/P ***v-/PRO)PP***

la répétition de la même préposition *chez* et l'équivalence entre l'amorce *v-* et le début du pronom *vous*, laisse supposer qu'il s'agit de la répétition du même groupe prépositionnel.

Dans le cas de répétitions du type « faits de parole », où ce phénomène fait partie des disfluences de l'oral (contrairement aux « faits de langue » où la répétition est due à la syntaxe (Henry, 2005)), les deux possibilités se présentent pour le chunking :

- Si l'élément répété est la tête du groupe syntaxique, il est nécessaire de distinguer deux chunks car un chunk ne peut pas avoir deux têtes :

  (et/CC)CONJ (et/CC)CONJ *(elle/CLS)NP (me/CLO)NP (disait/V)VN*

- Dans le cas inverse, les deux éléments appartiennent au même chunk :

  (la/DET la/DET *belle/ADJ jeune/ADJ fille/NC)NP*

4.1.1.2 Résultats

Le chunking ainsi réalisé est évalué avec une micro-précision de 77,24 et une macro-précision de 76. On a donc perdu plus de 20 points de F-mesure en moyenne (micro-average) en faisant passer un programme appris avec des textes écrits sur des données transcrites de l'oral. Ce mauvais résultat est le point de départ de différentes tentatives d'amélioration. L'objectif des expériences qui suivent est de corriger le minimum de données manuellement pour améliorer au maximum les performances du chunker.

4.1.2 Utilisation de SEM après correction de l'étiquetage POS

4.1.2.1 Méthodologie

Le chunking précédent était appliqué en cascade après un étiquetage POS du corpus qui était lui-même sans doute médiocre. La première idée pour améliorer le chunking est bien sûr de corriger manuellement l'étiquetage POS de l'oral avant de lui appliquer la phase de chunking. Ce processus a permis par la même occasion d'évaluer la qualité de l'étiquetage POS de SEM sur l'oral à 80,98%, soit 17% de moins que sur des données similaires à celles qui ont servi à apprendre. La fonction « chunker seul » de SEM peut donc ensuite s'appliquer sur le corpus corrigé avec de bonnes étiquettes POS (les colonnes **I** et **III**).

4.1.2.2 Corrections des étiquettes POS

Pour corriger les étiquettes POS, certaines conventions ont été adoptées concernant les disfluences de l'oral (voir les colonnes **II** : les étiquettes POS annotées par SEM et **III** : les étiquettes POS corrigées selon les conventions établies). Les faux départs et les amorces (comme *l-* dans l'exemple du Table 2) ont reçu une étiquette *(UNKNOWN)* qui correspond aux mots étrangers et aux néologismes dans FTB. Les marqueurs discursifs ainsi que les *euh* d'hésitation ont été étiquetés en tant qu'interjection (*I*). Cette étiquette POS est celle disponible dans SEM qui correspond le mieux à ces unités caractéristiques de l'oral.

4.1.2.3 Résultats

La correction des erreurs de l'étiquetage POS porte surtout sur les différences entre l'écrit et l'oral. Par exemple, la forme *bon* est utilisée en tant qu'adjectif dans 99% des cas dans le corpus écrit FTB, alors qu'elle est beaucoup plus fréquente dans le corpus oral en tant qu'interjection (83%).

La nouvelle micro-average du chunker est maintenant de 87,74 alors que sa nouvelle macro-average est de 88,43. Ces résultats sont en quelque sorte à mi-chemin des précédents : la moitié environ des erreurs de chunking sur l'oral peut donc être imputée à des erreurs d'étiquetage POS.

4.2 Deuxième approche : Apprentissage d'un chunker spécifique de l'oral

La deuxième approche consiste à apprendre un nouveau chunker à partir du corpus oral transcrit, en tenant plus compte des spécificités de l'oral. En effet, tant qu'à ré-apprendre un nouveau chunker, autant en profiter pour définir un jeu de chunks adapté.

Nous avons déjà noté précédemment qu'apprendre automatiquement un étiqueteur POS exige un volume important de données annotées (ou, à défaut, des ressources lexicales), car il est basé sur les mots. Le corpus de référence que nous avons constitué est trop petit pour servir de données d'apprentissage à un étiqueteur POS spécifique de l'oral. La situation est différente pour le chunking qui est fondé surtout sur les POS, et bénéficie donc d'une moindre variabilité des valeurs de ses données d'entrée. Nous allons donc essayer d'apprendre un chunker de l'oral sans pour autant passer par l'apprentissage d'un étiqueteur POS. Nous proposons trois expériences différentes pour évaluer l'influence de la qualité des étiquettes POS sur la capacité à apprendre un bon chunker de l'oral.

4.2.1 Modification des étiquettes de chunks

Pour tenir compte des spécificités de l'oral, nous avons choisi d'ajouter deux types de chunks nouveaux qui lui sont propres (voir la colonne **V** du Table 2). Ainsi, la liste des chunks a été élargie par deux nouveaux venus :

- Chunk *UNKNOWN*

L'étiquette *UNKNOWN* existe dans le jeu d'étiquettes POS du FTB, où elle est attribuée aux mots étrangers. Nous l'avons utilisée aussi pour désigner les chunks correspondant aux erreurs de transcriptions, aux faux départs ou aux amorces dont l'interprétation est impossible. Dans notre exemple, la forme *l-* est difficile à comprendre. S'agit-il d'un pronom, d'un déterminant ou d'une amorce ? L'étiquette *UNKNOWN*, déjà choisie pour cette forme au niveau POS, est donc étendue dans ce cas au chunk.

- Chunk d'interjection (*IntP*)

Nous avons pu observer le problème qui se pose avec les marqueurs discursifs et les *euh* d'hésitation qui ont été classés, faute d'une autre étiquette adaptée dans SEM, dans les chunks adverbiaux. L'ajout d'un nouveau chunk *IntP* (chunk interjection) destiné à accueillir tous les marqueurs discursifs, les *euh* d'hésitation et les interjections, résout au moins partiellement ce problème :

(des/DET idées/NC laïques/ADJ)NP (quoi/I) IntP

Cependant, lorsque les interjections se retrouvent à l'intérieur d'un groupe syntaxique, ils s'intègrent dans le chunk dont ils font partie :

- *(l'/DET école/NC **euh/I** publique/ADJ)NP*
- *(une/DET **euh/I** anglaise/NC)NP*
- *(des/DET **hm/I** inconvénients/NC)NP*

Dans les trois exemples ci-dessus, le *euh* d'hésitation et l'interjection *hm* appartiennent à un chunk nominal.

Ce nouvel étiquetage en chunks a été manuellement validé sur nos données ESLO (colonne **V** de la Table 2), et constitue la nouvelle référence grâce à laquelle nous allons à la fois apprendre et évaluer notre nouveau chunker.

4.2.2 Apprentissage et test avec les étiquettes POS corrigées

La première expérience consiste à apprendre un chunker à partir des données annotées en POS corrigées (la colonne **III** de la Table 2) et les chunks adaptés à l'oral (la colonne **V**). Un protocole de validation croisée à 10 plis a été utilisé. La nouvelle micro-average des F-mesure atteint maintenant 96,65 alors que leur macro-average vaut 96,08. Les résultats se sont donc significativement améliorés, et rejoignent ceux qui avaient été annoncés pour SEM sur FTB.

Si on observe de plus près les F-mesures des différents types de chunks, en comparaison avec les expériences précédentes, on constate une forte progression de l'annotation des chunks adverbiaux (*AdP*). Ces chunks sont très nombreux dans notre corpus au cours des premières expériences, car ils regroupent les adverbes, les marqueurs discursifs, les *euh* d'hésitation et les interjections. L'introduction d'un nouveau chunk *(IntP)* annotant ces différents phénomènes (sauf les adverbes) a considérablement réduit le nombre de chunks adverbiaux dans le corpus de référence, ce qui modifie significativement leur F-mesure. Pendant les premières expériences, la F-mesure du chunk *(AdP)* varie entre 58,14 (avec les POS non corrigées) et 71,87 (avec les POS corrigées). Désormais, la F-mesure atteint 95,76 pour le chunk *(AdP)* et 99,4 pour le chunk *(IntP)*. L'apprentissage a donc bien réussi à distinguer les deux types de chunks.

Les erreurs constatées concernent souvent des « exceptions » aux règles générales. C'est le cas des verbes qui forment d'habitude un chunk verbal *(VN)* sauf quand ils suivent une préposition. Ainsi, dans l'exemple ci-dessous, le verbe est annoté comme la tête d'un chunk verbal :

(à/P)PP (me/CLR)B-NP (marier/VINF)B-VN

alors qu'il fait partie ici d'un chunk prépositionnel (*PP*) :

(à/P me/CLR marier/VINF)PP

Les cas où les interjections et les marqueurs formant généralement un chunk *(IntP)* sont inclus dans un autre chunk posent aussi problème. Le chunker appris propose :

(l'/DET école/NC)NP ***(euh/I) IntP*** *(publique/ADJ)AP*

à la place de :

(l'/DET école/NC ***euh/I*** *publique/ADJ)****NP***

Enfin, en cas de répétition de deux étiquettes morphosyntaxiques, le chunker inclut parfois les deux mots dans le même chunk, violant ainsi la contrainte qui voudrait que chaque chunk ne devrait contenir qu'une seule tête. Il annote ainsi :

(et/CC parce_que/CS)CONJ

(ils/CLS)NP (réfléchissaient/V pensaient/V)VN (beaucoup/ADV)AdP

à la place de

(et/CC)CONJ (parce_que/CS)CONJ

(ils/CLS)NP (réfléchissaient/V)VN (pensaient/V)VN (beaucoup/ADV)AdP

Mais les très bons résultats du chunker ne sont atteints que sur des données qui ont elles-mêmes reçu un étiquetage POS correct. Or, aucun étiqueteur POS de l'oral n'ayant été appris, notre nouveau chunker risque de voir ses performances se dégrader significativement en situation réelle, avec de mauvaises étiquettes POS. Pour quantifier ce problème et essayer d'y remédier, nous avons mené deux nouvelles expériences qui ne font pas l'hypothèse de disposer d'un étiquetage POS corrigé lors de la phase d'utilisation du chunker.

4.2.3 Apprentissage avec les étiquettes POS corrigées, test sur les étiquettes non corrigées

La deuxième expérience de cette série vise à évaluer la dégradation de performance subie quand le chunker appris sur des étiquettes POS corrigées (colonnes **III** et **V**) est utilisé sur des données avec des étiquettes POS non corrigée (colonne **II**). Etant donné le faible volume de données dont nous disposons, nous avons pour cela reconduit l'expérience précédente en validation croisée à 10 plis, en prenant soin lors de chaque étape de respecter le protocole suivant :

- l'apprentissage est réalisé à l'aide des colonnes **I**, **III** et **V**
- le chunker appris est appliqué en test sur les colonnes **I** et **II**
- le résultat obtenu est comparé à la colonne de référence **V**

Nous obtenons ainsi une micro-average des F-mesure de 73,81, et une macro-average de 59,62, ce qui constitue une grosse dégradation (cf. les détails des valeurs dans la Table 3). Les performances sont particulièrement mauvaises pour le nouveau type de chunk *IntP*, car très peu d'étiquettes POS *I* sont correctement attribuées par SEM dans ESLO 1. En effet, dans FTB, les seules interjections présentes correspondent à des phrases d'un seul mot suivi d'une ponctuation. Or ESLO 1 ne contient pas de ponctuation, et cet indice n'aide donc en rien le chunker. Cela concerne surtout les mots étiquetés en tant qu'adverbes lors de l'étiquetage POS (312 fois) comme *oui, non, alors* qui sont tous annotés par SEM comme chunk adverbial à la place de chunk interjection. Un seul chunk *(IntP)* a été reconnu lors de cette expérience, et cela semble-t-il de façon quasiment « fortuite ». Les représentants du nouveau chunk *(UNKNOWN)* n'ont pas non plus été identifiés, ce qui s'explique naturellement par le fait que SEM n'a pas attribué l'étiquette POS *UNKNOWN* là où notre correction manuelle l'avait fait (sur les disfluences en particulier). Le problème mentionné dans la première série d'expériences et concernant la forme *bon* persiste également dans les résultats de ce test. Ayant une étiquette POS *ADJ*, *bon* est étiqueté en tant que chunk adjectival *(AP)* et non comme chunk interjection *(IntP)*. Dans l'exemple suivant (où la dernière colonne donne la proposition du nouveau chunker, tandis que l'avant-dernière donne la bonne étiquette), les deux unités *bon* et *alors* ne reçoivent pas une bonne étiquette de chunk :

on	*CLS*	*B-NP*	*B-NP*	
peut	*V*	*B-VN*	*B-VN*	
commencer		*VINF*	*B-VN*	*B-VN*

bon	*ADJ*	*B-IntP*	*B-AP*
alors	*ADV*	*B-IntP*	*B-AdP*

L'absence de correction des POS cause donc ici des erreurs prévisibles de chunking, surtout pour les nouveaux types de chunks qui s'appuient sur des propriétés de l'oral que SEM ne prend pas en compte. Il reste à voir si un chunker appris directement sur des étiquettes POS non corrigées se comporterait mieux.

4.2.4 Apprentissage et test avec les étiquettes POS non corrigées

La dernière expérience vise à apprendre le chunker de l'oral en se servant uniquement des étiquettes POS fournies par SEM, sans aucune correction ni en apprentissage ni en test sur ces POS. Cette fois, notre validation croisée emploie donc les colonnes **I**, **II** et **V** de la Table 2. L'objectif de cette dernière expérience est de voir s'il est possible d'apprendre un bon chunker en se fondant sur des étiquettes POS médiocres. Existe-t-il des régularités dans les erreurs au niveau morpho-syntaxique dont l'apprentissage pourrait tirer parti ? Pourrait-on donc se passer d'une correction manuelle de l'étiquetage POS (et d'un ré-apprentissage d'un étiqueteur POS de l'oral) pour obtenir tout de même *in fine* un chunker de l'oral correct ? C'est tout l'enjeu de cet ultime test.

Nous obtenons dans cette expérience une micro-average de 88,84, et une macro-average de 81,76, soit des résultats (comme on pouvait s'y attendre) intermédiaires entre les deux précédents (cf. les détails dans la Table 3). Cette fois, on constate que les chunks *(IntP)* sont très bien reconnus (plus de 93 de F-mesure), alors que SEM substitue à l'étiquette POS correcte *I* des étiquettes assez variées (typiquement ADV, ADJ, NC et V). Mais les interjections sont à la fois fréquentes et assez peu variées dans notre corpus de l'oral (*euh, hm, oui, non, etc.*) et celles présentes dans l'ensemble d'apprentissage suffisent apparemment au chunker appris (qui a aussi accès aux mots ou tokens et pas uniquement aux POS) à les identifier. Ainsi, l'exemple précédent reçoit cette fois l'étiquetage :

on	*CLS*	*B-NP*	*B-NP*	
peut	*V*	*B-VN*	*B-VN*	
commencer		*VINF*	*B-VN*	*B-VN*
bon	***ADJ***	***B-IntP***	***B-IntP***	
alors	*ADV*	*B-IntP*	*B-AdP*	

La forme *bon* est étiquetée ici correctement *(B-IntP)* malgré une erreur d'étiquetage POS où elle est reconnue en tant qu'adjectif. Dans le corpus d'apprentissage, ce mot est le plus souvent employé comme marqueur discursif, ce qui facilite sa désambiguïsation. Les unités *oui, non* très fréquentes dans le corpus (comme mentionné avant) reçoivent maintenant aussi une bonne étiquette.

Sur le chunk *(UNKNOWN)*, le nouveau chunker obtient une bonne précision (92,86%) :

vous	*DET*	*B-NP*	*B-NP*
êtes	*NC*	*B-VN*	*B-VN*
in-	***ADJ***	***B-UNKNOWN***	***B-UNKNOWN***
institutrice		*NC*	*B-NP*	*B-AdP*
n-	***ADV***	***B-UNKNOWN***	***B-UNKNOWN***
peut-être		*VINF*	*B-AdP*	*B-AdP*
non	***ADV***	***B-IntP***	***B-IntP***	

mais un mauvais rappel (18,57%). Cela vient sans doute du fait que les chunks inconnus peuvent parfois correspondre à des mots connus mais employés dans un mauvais contexte, comme dans l'exemple suivant :

euh	***V***	***B-IntP***	***B-IntP***
les	***DET***	***B-UNKNOWN***	***B-NP***
dans	*P*	*B-PP*	*B-PP*
ma	*DET*	*I-PP*	*I-PP*
classe	*NC*	*I-PP*	*I-PP*

En outre, les amorces présentent une bien plus grande variabilité que les interjections ; toutes ne peuvent pas être présentes dans l'ensemble d'apprentissage et l'accès aux tokens ne suffit donc pas à compenser le mauvais étiquetage POS. Il ne semble pas y avoir de règle évidente quant aux chunks *(UNKNOWN)* reconnus. L'hypothèse la plus probable est que SEM a reconnu uniquement les mots qu'il a déjà vus dans son ensemble d'apprentissage. On pourrait sans doute largement améliorer les capacités de notre nouveau chunker à reconnaître les amorces en lui donnant accès à certaines propriétés des tokens : dans ESLO 1, les amorces sont en effet systématiquement terminées par un tiret - : ajouter cette propriété aux attributs pris en compte dans les features devrait permettre de les identifier bien plus sûrement que pas leur contexte. Mais nous voulions utiliser le même ensemble de features (copiées sur celles utilisées pour l'apprentissage de SEM) pour toutes nos expériences, pour ne pas biaiser les comparaisons.

Les détails des résultats obtenus sur les différents types de chunks pour nos deux dernières expériences sont présentés dans la Table 3.

Type de chunk	Expérience 4			Expérience 5		
	Précision	Rappel	F-mesure	Pécision	Rappel	F-mesure
AP	50,73%	71,23%	59,26	71,76%	64,38%	67,87
AdP	55,9%	79,48%	65,64	83,78%	85,83%	84,79
CONJ	89,42%	89,42%	89,42	89,8%	91,42%	90,6
IntP	33,33%	0,12%	0,24	95,82%	91,87%	93,8
NP	81,16%	85,34%	83,2	91,93%	90,6%	91,26
PP	71,99%	81,55%	76,48	81,57%	82,41%	81,99
UNKNOWN	N/A	N/A	N/A	92,86%	18,57%	30,95
VN	78,13%	87,23%	82,43	89,75%	90,89%	90,32

TABLE 3 : Résultats des différents types de chunks dans les deux dernières expériences

La synthèse des résultats de l'ensemble de nos expériences est présentée dans la Table 4.

Expériences	Première approche : utilisation d'un chunker appris sur l'écrit (référence : la colonne **IV**)		Deuxième approche : Apprentissage d'un chunker spécifique de l'oral (référence : la colonne **V**)		
évaluation	POS non corrigées	POS corrigées	POS corrigées	Apprentissage sur POS corrigés, test sur POS non corrigés	POS non corrigés
exactitude des POS (%)	*80,98*	*100*	*100*	*80,98*	*80,98*
Micro-average	*77,24*	*87,74*	*96,65*	*73,81*	*88,84*
Macro-average	*76*	*88,43*	*96,08*	*59,62*	*81,76*

TABLE 4 : Synthèse des résultats des micro et macro-averages des F-mesures dans l'ensemble de nos expériences

5 Conclusion

Tout d'abord, notre première série d'expériences montre qu'un étiqueteur morphosyntaxique associé à un chunker appris sur un corpus écrit fait environ 17% d'erreurs supplémentaires en POS, et 20% en chunking, sur des données orales transcrites. Cet écart important justifie de trouver des stratégies d'adaptation ou de contournement pour traiter les corpus oraux. Comme l'erreur en chunking n'est pas significativement plus importante que l'erreur en POS, la solution

de corriger les POS apparaît *a priori* comme la plus « naturelle ». Cette correction manuelle des POS augmente le résultat du chunking de 10 points de F-mesure en moyenne, mais reste 10 points en dessous des performances moyennes du chunker sur l'écrit. Même avec un étiquetage POS parfait, l'écart entre l'écrit et l'oral en matière de chunking se mesure avec ces 10 points d'écart en moyenne.

L'idée de corriger directement les étiquettes de chunks est donc la suite logique de cette approche. Nous avons pour cela choisi de coller aux propriétés de l'oral plutôt que de chercher à faire entrer à tout prix les données orales dans le cadre défini pour l'écrit, d'où le choix des deux nouveaux types de chunks introduits. Ce faisant, nous n'avons pas choisi la facilité car la tâche de chunking devient plus complexe (il faut désormais discriminer parmi neuf types de chunks au lieu de sept). Pour l'apprentissage automatique d'un nouveau chunker spécifique de l'oral, le pari a été fait qu'un petit nombre de données d'apprentissage pouvait suffire.

Les trois expériences de la deuxième approche permettent de caractériser assez finement l'apport des étiquettes POS à la phase de chunking. En présence d'étiquettes POS correctes et cohérentes avec les chunks (première expérience), l'apprentissage automatique joue parfaitement son rôle, et permet d'apprendre un chunker d'aussi bonne qualité que celui qui avait été appris sur l'écrit avec beaucoup plus de données. Il n'y a donc pas de malédiction propre à l'oral en matière de chunking : même les disfluences peuvent y être bien traitées, à condition de disposer d'exemples de référence, même en quantité restreinte. En revanche, un tel chunker dépend fortement des étiquettes POS sur lesquelles il s'appuie : l'absence de correction manuelle (deuxième expérience de la série) fait chuter ses performances. Il n'est donc pas réellement exploitable en conditions réelles : en effet, tant qu'à corriger les étiquettes POS, autant ré-apprendre dans ce cas un étiqueteur POS de l'oral…

La dernière expérience est la plus prometteuse : elle montre qu'on peut apprendre un chunker spécifique de l'oral (y compris pour la reconnaissance des interjections par exemple) d'assez bonne qualité, en s'appuyant uniquement sur un petit nombre de données annotées avec des étiquettes POS médiocres (car non adaptées à l'oral). Les erreurs du POS ont bien été *compensées* par l'apprentissage du chunker, qui fait en moyenne moins d'erreurs de chunking qu'il n'y a d'erreurs d'étiquetage POS. Les mots, même en petites quantités, permettent cette compensation, et sans doute aussi le fait que les erreurs de POS sont suffisamment « régulières » pour que le chunker puisse les « rectifier ».

Cette expérience montre que l'apprentissage automatique d'un chunker spécifique de l'oral semble pouvoir assez bien se passer d'un étiquetage POS correct. Il est intéressant de constater que les résultats obtenus pour le chunker dans cette dernière expérience sont très proches de ceux de la deuxième expérience de la première approche, c'est-à-dire en appliquant SEM sur des étiquettes POS corrigées manuellement. La différence est que le nouveau chunker obtenu avec la dernière expérience est applicable sans plus de correction manuelle sur de nouvelles données orales, ce qui n'est pas le cas de ce que proposait cette précédente expérience. Ainsi, tant qu'à corriger des données, il vaut semble-t-il mieux s'attaquer aux données qui servent à l'apprentissage (les nouveaux chunks dans la dernière expérience) qu'aux données qui servent de support à un programme déjà appris (les POS dans l'expérience 2).

Il reste bien sûr à confirmer que le même genre de démarche peut être valable dans d'autres contextes, par exemple pour d'autres langues, ou en changeant la variation écrit/oral par une autre, comme un changement de domaine ou de types d'écriture (les tweets pourraient par exemple se substituer à l'oral). Le fait que l'apprentissage direct d'un nouvel étiqueteur focalisé sur une tâche cible est préférable à une séquence d'apprentissages intermédiaires avait par ailleurs déjà été constaté (Eshkol et al. 2010). Le caractère cumulatif des erreurs n'est donc pas une fatalité : il semble qu'on puisse réussir une tâche de « haut niveau » en s'appuyant sur des informations de « niveau inférieur » de qualité médiocre par apprentissage automatique, du moment que la correction des erreurs d'un niveau à un autre suive une certaine régularité.

Références

ABNEY S. (1991). Parsing by chunks. In R. Berwick, R. Abney, and C. Tenny, editors, *Principle-based Parsing*. Kluwer Academic Publisher.

ABEILLE A., CLEMENT L., et TOUSSENEL F. (2003). Building a treebank for french. In A. Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

ANTOINE J-Y., MOKRANE A., et FRIBURGER N. (2008) Automatic rich annotation of large corpus of conversational transcribed speech: the chunking task of the epac project. In Proceedings of *LREC'2008*, may 2008.

BENZITOUN C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ?, Actes de *RÉCITAL*, Maroc, Fes.

BLANC O., CONSTANT M., DISTER A. et WATRIN P. (2008). Corpus oraux et chunking. In *Journées d'étude sur la parole (JEP)*, Avignon, France.

BLANCHE-BENVENISTE C. (2005). Sémantique de l'oral, dans *Sémantique et corpus. Les aspects dynamiques de la composition sémantique de l'oral*.

BLANCHE-BENVENISTE C., JEANJEAN C. (1987). *Le français parlé, transcription et édition.* Paris, Didier Erudition.

BLANCHE-BENVENISTE C. (1997). *Approches de la langue parlée en français*. Paris, Ophrys.

BLANCHE-BENVENISTE C. (2000). Transcription de l'oral et morphologie, *Romania Una et diversa, Philologische Studien für Theodor Berchem* (Gille M. et Kiesler R. Eds). Tübingen : Gunter Narr, 61-74.

CONSTANT M., TELLIER I. (2012) Evaluating the impact of external lexical ressources unto a crf-based multiword segmenter and part-of-speech tagger. In Proceedings of *LREC 2012.*

CRABBE B, CANDITO M (2008). Expériences d'analyse syntaxique du français, Actes de *Traitement Automatique des Langues Naturelles (TALN 2008*), Avignon.

DISTER A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL*, Thèse de Doctorat, Université de Louvain.

ESHKOL I., TELLIER I, TAALAB S., BILLOT S., (2010). Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques, *10es Journées Internationales d'analyse statistique des données textuelles JADT 2010*, Rome, 9-11 juin, 2010.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C., TELLIER I., (2012) Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. in *Ressources linguistiques libres*, TAL. Volume 52, n° 3, 17-46.

HENRY S. (2005). Quelles répétitions à l'oral ? Esquisse d'une typologie, G. Williams (Éd.), *La Linguistique de corpus*, Rennes, Presses universitaires de Rennes, 81-92.

LAVERGNE T, CAPPE O, ET YVON F. (2010). Practical very large scale CRFs. In Proceedings of *ACL'2010*, 504–513. Association for Computational Linguistics, July 2010.

LAFFERTY J, MCCALLUM A, ET PEREIRA F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of *ICML 2001*, 282–289.

SHA F, PEREIRA P. (2003). Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL, 213–220.

TELLIER I., DUCHIER D., ESHKOL I., COURMET A., MARTINET M. (2012), Apprentissage automatique d'un chunker pour le français, Actes de *Traitement Automatique des Langues Naturelles (TALN 2012*), article court, Grenoble.

TELLIER I., ESHKOL I., TAALAB S., PROST J-P. (2010), POS-tagging for Oral Texts with CRF and Category Decomposition, *Research in Computer Science*, special issue *: Natural Language Processing and its Applications*, 79-90

VALLI A., VERONIS J. (1999). Etiquetage grammatical des corpus de parole : problèmes et Perspectives. L'oral spontané. *Revue Française de Linguistique Appliquée*, vol. IV-2, 113-133.