Vers un treebank du français parlé

Anne Abeillé[1, 2] Benoit Crabbé[1, 3]
(1) LLF, CNRS-Université Paris Diderot, 75013 Paris, PRES Sorbonne Paris Cité, IUF
(2) Alpage, INRIA, Université Paris Diderot, 75013 Paris, PRES Sorbonne Paris Cité
abeille@univ-paris-diderot.fr, bcrabbe@univ-paris-diderot.fr

RÉSUMÉ

Nous présentons les premiers résultats d'un corpus arboré pour le français parlé. Il a été réalisé dans le cadre du projet ANR Etape (resp. G. Gravier) en 2011 et 2012. Contrairement à d'autres langues comme l'anglais (voir le Switchboard treebank de (Meteer, 1995)), il n'existe pas de grand corpus oral du francais annoté et validé pour les constituants et les fonctions syntaxiques. Nous souhaitons construire une ressource comparable, qui serait une extension naturelle du Corpus arboré de Paris 7 (FTB : (Abeillé *et al.*, 2003))) basé sur des textes du journal Le Monde. Nous serons ainsi en mesure de comparer, avec des annotations comparables, l'écrit et l'oral. Les premiers résultats, qui consistent à réutiliser l'analyseur de (Petrov *et al.*, 2006) entraîné sur l'écrit, avec une phase de correction manuelle, sont encourageants.

ABSTRACT

Towards a treebank of spoken French

We present the first results of an attempt to build a spoken treebank for French. It has been conducted as part of the ANR project Etape (resp. G. Gravier). Contrary to other languages such as English (see the Switchboard treebank (Meteer, 1995)), there is no sizable spoken corpus for French annotated for syntactic constituents and grammatical functions. Our project is to build such a resource which will be a natural extension of the Paris 7 treebank (FTB : (Abeillé *et al.*, 2003))) for written French, in order to be able to compare with similar annotations written and spoken French. We have reused and adapted the parser (Petrov *et al.*, 2006) which has been trained on the written treebank, with manual correction and validation. The first results are promising.

MOTS-CLÉS : Corpus arboré, français parlé, corpus oral, analyse syntaxique automatique.

KEYWORDS: Treebank, spoken French, spoken corpus, parsing.

1 Introduction

Nous présentons les premiers résultats d'un corpus arboré pour le français parlé. Il a été réalisé dans le cadre du projet ANR Etape (resp. G. Gravier) entre 2010 et 2012. Les corpus arborés (Treebank) pour les autres langues ont une partie écrite et une partie orale : Penn Treebank (Switchboard (Meteer, 1995)), Verbmobil pour l'allemand, Prague Dependency Treebank pour le tchèque (Mikulova, 2008). A notre connaissance, il n'existe pas de grand corpus oral du français

annoté et validé pour les constituants et les fonctions syntaxiques. Les corpus oraux annotés existants pour le français suivent des schémas spécifiques : annotation en micro et macro syntaxe pour le corpus Rhapsodie (cite Deulofeu 2011), annotation en dépendances de (Cerisara *et al.*, 2010), annotation en chunks du corpus Otim (Blache *et al.*, 2010) Nous souhaitons construire une ressource qui soit une extension naturelle du Corpus arboré de Paris 7 (FTB (Abeillé *et al.*, 2003)) basé sur des textes du journal *Le Monde*. Nous serons ainsi en mesure de comparer, avec des annotations comparables, l'écrit et l'oral. Nous procédons en trois temps : une phase de prétraitement avec ponctuation et balisage des dysfluences, une phase d'analyse automatique, une phase de correction manuelle. Pour la seconde phase, nous avons adapté le parseur de (Petrov *et al.*, 2006) entraîné sur le FTB ; pour la troisième phase, nous avons adapté et enrichi les consignes du Corpus arboré de Paris 7 (Abeille *et al.*, 2013).

2 De l'écrit à l'oral

Contrairement à d'autres langues comme l'anglais (Switchboard (Meteer, 1995)) il n'existe pas de grand corpus oral du français annoté et validé pour les constituants et les fonctions syntaxiques. Nous souhaitons construire une ressource comparable, qui serait une extension naturelle du Corpus arboré de Paris 7 (FTB (Abeillé *et al.*, 2003)) basé sur des textes du journal *Le Monde*. Une extension à l'oral devrait permettre à terme de mener des études comparatives sur des données comparables de la syntaxe du français écrit et du français oral.

Le corpus écrit est annoté lexicalement (lemme, catégories et sous-catégories lexicales, morphologie flexionnelle, mots composés), en constituants et en fonctions et a été validé manuellement. Il est distribué depuis 2001 et est accompagné d'un guide d'annotation (135pp). Le jeu d'étiquettes morphologiques est relativement riche (218 catégories) alors qu'on compte 12 étiquettes de syntagmes et 8 étiquettes de fonctions. Les choix généraux d'annotation reposent sur un schéma surfaciste d'annotation de constituants majeurs qui se veut compatible avec plusieurs théories syntaxiques. Contrairement au Penn Treebank (Marcus *et al.*, 1993) le corpus français ne comporte pas de catégories vides ni de constituants discontinus.

Contrairement à d'autres initiatives d'annotation pour le français (Deulofeu *et al.*, 2010), et suivant en cela les initiatives pour d'autres langues (Meteer, 1995; Mikulova, 2008) la représentation de données orales proposée ici repose sur l'hypothèse que la syntaxe de la phrase orale ne nécessite pas un réaménagement en profondeur du schéma d'annotation de l'écrit, même si des aménagements légers sont nécessaires. Ce choix a pour conséquence de rendre disponible l'outillage déjà existant (analyseurs, outils d'édition de treebank) pour faciliter et accélérer le travail d'annotation.

Plusieurs versions du French Treebank sont actuellement utilisées (Schluter et van Genabith, 2007; Blache et Rauzy, 2012). Nous nous appuyons sur la représentation simplifiée décrite notamment par (Crabbé et Candito, 2008) qui permet l'analyse automatique avec les algorithmes d'analyse en constituants à l'état de l'art. En particulier nous nous appuyons sur un jeu de catégories lexicales réduit (28 catégories) et une liste de mots composés réduite aux mots composés grammaticaux. Cette version réduite a l'avantage de se convertir de manière déterministe vers une représentation en dépendances syntaxiques projectives (Candito *et al.*, 2009) qui est de plus en plus utilisée. Annoter en constituants permet donc de bénéficier des deux types de représentations.

3 Les données orales

Les données orales que nous utilisons sont des données du corpus ESTER 3 issues du projet ETAPE (Gravier *et al.*, 2012) dédié à l'évaluation de systèmes de reconnaissance automatique de la parole. Les données sont constituées d'extraits de débats de télévision et de radio françaises.

Les données annotées ici constituent un sous-ensemble de ce corpus constitué des émissions radiophoniques de *France Inter* de l'année 2010 : cinq émissions de *un temps de pauchon* et une émission du *Masque et la plume*, ce qui représente près d'une heure trente de temps de parole. Dans le premier cas il s'agit d'interviews non préparées donnant la parole à des inconnus. Dans le second, il s'agit d'un débat public très animé avec au moins dix journalistes sur le plateau, plus des commentaires de spectateurs. Nous avons également un extrait du corpus français de CORAL-ROM (Cresti *et al.*, 2004). L'extrait annoté est *L'allumage* (Poitiers 2001). CORAL-ROM présente un type de conversation informel et spontané entre deux amies : qui représente 14 minutes de parole. Les données de référence ESTER 3 sont transcrites orthographiquement, ponctuées et enrichies avec un balisage des diffluences, selon le format *transcriber* (Barras *et al.*, 1998). De manière à uniformiser nos données de travail, nous avons également reformaté les données CORAL-ROM dans ce même format. Au vu de l' extrait donné en Figure 1, on constate que les données de départ sont déjà structurées, en particulier on observe que l'on a un balisage pour la musique `<Event desc="musique" type="noise" extent="begin"/>` et les bruits parasites, un balisage pour les disfluences comme pour les marqueurs de discours `<Event desc="dm" type="lexical" ... />` mais aussi les répétitions, les révisions `<Event desc="rev" type="lexical" ... >` et les hésitations ainsi qu'une segmentation en tours de parole. On distingue trois types de caractéristiques des données orales qui touchent à la segmentation, la présence de chevauchements et à la présence de disfluences.

Segmentation Nous partons ici d'une transcription enrichie, c'est-à-dire avec des ponctuations fortes, mais avec peu de ponctuations faibles, et pas de mots composés. On voit sur l'exemple qu'un tour de parole ESTER peut comporter plusieurs phrases ou aucune. On a également observé que certaines phrases recouvrent plusieurs tours de parole. On note finalement que la ponctuation renseignée dans les transcriptions de départ n'a pas un statut clair : les annotateurs la renseignent plutôt pour indiquer des pauses dans le flux de parole que comme marque syntaxique. C'est pourquoi nous avons revu la segmentation manuellement avant l'analyse automatique.

Les chevauchements On trouve en particulier dans les transcriptions du *Masque et la plume* un nombre non négligeable de chevauchements. Ceux-ci sont annotés dans le format ESTER en suivant un schéma comme illustré en Figure 1 : où la balise XML `<Overlap>` encode la portée d'un chevauchement. L'attribut `type` indique le locuteur qui domine l'échange par la valeur `primary` et celui que l'on entend moins est renseigné par la valeur `backchannel`.

Les disfluences Outre les questions de segmentation et de chevauchements, les disfluences sont typiques de l'oral. La transcription ESTER les renseigne sous forme de balises XML, on recense ainsi quatre types de disfluences :

– Hésitations : *euh*
– Répétitions qui concernent la répétition à l'identique : *qui a retardé un peu <nos> nos commentaires, qui avait été sérieusement amoché <au> au masque et la Plume, a été bluffé par le jeu <de> de Morgan Freeman. . .)*
– Révisions qui concernent des révisions de forme : *<le>la grandiloquence, beaucoup <de> d'auditeurs, autre chose <qu'un> qu'une guerre*

```
<Turn speaker="spk2" startTime="428.447" endTime="430.539">
   <Sync time="428.447"/>
   <Event desc="rev" type="lexical" extent="begin"/>
       si
   <Event desc="rev" type="lexical" extent="end"/>
       il y avait pas une route
   <Sync time="429.187"/>
   <Overlap type="primary" extent="begin"/>
     qui desservait ce terrain
   <Event desc="dm" type="lexical" extent="begin"/>
     quoi
   <Event desc="dm" type="lexical" extent="end"/>
     ?
   <Overlap type="primary" extent="end"/>
   <Overlap type="backchannel" extent="begin" speaker="spk3" subtype="out-field"/>
     non il y avait pas une route .
   <Overlap type="backchannel" extent="end"/>
</Turn>
```

FIGURE 1 – Extrait d'un fichier Le Masque et la plume au format Transcriber

– Marqueurs de discours qui sont des mots ou des locutions qui ont une valeur illocutoire sans avoir de fonction syntaxique dans l'énoncé comme par exemple *ah,bref, mais bon voilà, non non non, na na na...*

L'annotation des marqueurs de discours n'étant pas toujours cohérente, nous l'avons reprise, avec une liste de 115 marqueurs (simples ou composés). En particulier les connecteurs, les conjonctions de coordination en début de phrase, ou les pronoms disloqués, ne sont pas traités comme des marqueurs de discours. De façon générale, nous traitons les balises de diffluences comme des étiquettes de syntagmes, qui peuvent avoir une structure interne.

4 Le schéma d'annotation

Nous indiquons dans cette section les lignes directrices et les conventions adoptées pour l'annotation en syntaxe des données de l'oral. Le schéma d'annotation est dérivé du schéma d'annotation pour le treebank écrit (Abeillé *et al.*, 2003).

On supprime les informations ayant trait au bruit et à la musique considérées comme extralinguistiques. Par contre on préserve les balises de synchronisation avec la piste sonore, notées `<Sync>` dans ESTER 3 (Figure 1) encodées par des sous-arbres de racine `Sync` attachés avec les mêmes conventions que les disfluences. Nous présentons plus en détails dans la suite de cette section les choix quant à la segmentation et à la gestion des dysfluences.

4.1 Linéarisation et segmentation des données orales

Comme pour l'écrit, une des premières décisions à prendre lorsqu'on annote un corpus en syntaxe porte sur la segmentation en mots. Contrairement au corpus écrit, la segmentation pour le corpus oral minimise le nombre de mots composés. Nous nous sommes pour cela appuyés sur les travaux

antérieurs de (Crabbé et Candito, 2008) en ne retenant qu'un nombre minimal de mots composés, en particulier des mots composés grammaticaux comme des conjonctions de subordination, de coordinations, des déterminants, prépositions ... et quelques mots composés propres à l'oral *n'est-ce pas, s'il vous plaît, tant pis...* qui ont un impact sur la syntaxe et l'analyse de la phrase. La liste exacte des mots composés est définie et documentée dans (Abeille *et al.*, 2013).

Nous nous appuyons également sur une segmentation en phrases, même si le choix de tel ou tel découpage ne va pas de soi. Plusieurs notions sont possibles : une notion phonétique ou la phrase est délimitée par la durée des pauses, ce qui est le cas de la transcription ESTER 3, une notion dialogique où la phrase correspond à un tour de parole, une notion discursive où la phrase correspond à un acte de langage, et une notion syntaxique où la phrase correspond à une plus grande unité syntaxique complète (avec enchâssement possible). Ici nous avons considéré qu'un tour de parole non constitué uniquement de bruit ou de musique correspond au moins à une phrase, même fragmentaire. Un tour de parole peut lui-même être découpé en plusieurs phrases racines. Nous nous appuyons pour cela sur des critères syntaxiques, discursifs et prosodiques. Une séquence autonome associée à un acte de langage forme une phrase racine. En revanche, nous ne considérons pas qu'une phrase recouvre des tours de paroles différents, c'est-à-dire qu'une même phrase commencée par un locuteur soit terminée par un autre locuteur [1]. En cas d'interruption et pour repérer les syntagmes inachevés nous utilisons plutôt une annotation d'inachèvement (-INA) comme étiquette supplémentaire sur les noeuds racine des syntagmes jugés inachevés.

Ces critères étant donnés, voyons comment sont traités les cas de chevauchements. Les structures à chevauchements ESTER 3 suivent un schéma tel qu'illustré en figure 2 à gauche (où le balisage XML est simplifié). Pour gérer les cas de chevauchement dans l'annotation syntaxique, le principe a été de fusionner les parties en `backchannel` associées à un locuteur X au tour de parole suivant (resp. précédent selon les cas) de ce locuteur X dans les données transcrites, ce qui permet d'éviter de découper artificiellement une phrase complète énoncée par ce locuteur X. Par contre, pour préserver l'information, nous avons également introduit des marques dans les arbres sous forme de noeuds feuilles pour indiquer la portée du chevauchement suivant le schéma donné en figure 2. Chacun des quatre noeuds feuilles ainsi introduit dans les arbres est

```
<Turn speaker="Y">
  w[y,1] ... w[y,b-1]
<Overlap>
   w[y,b+1] ... w[y,e-1]
</Overlap>
   w[y,e+1] ... w[y,n]
<Backchannel speaker="X">
   w[x,1] ... w[x,e-1]
</Backchannel>
</Turn>
<Turn speaker = "X">
   w[x,e+1] ... w[x,n]
</Turn>
```

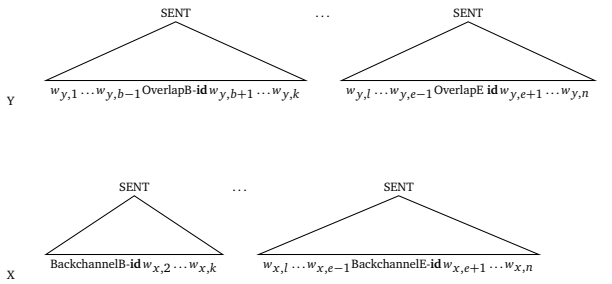

FIGURE 2 – Encodage des chevauchements dans les arbres

1. Les annotations ESTER 3 comportent parfois plusieurs tours de paroles consécutifs pour un même locuteur. Nous avons refusionné ces séquences de manière à éviter qu'une phrase prononcée par un même locuteur ne soit artificiellement découpée.

de plus annoté par un identifiant unique (noté **id** dans le schéma) permettant d'identifier à quel chevauchement ce noeud fait référence. Ce qui permet de gérer des chevauchements multiples dans un même document et dans un même tour de parole. Notons que coder le chevauchement sous forme d'un noeud non terminal dans les arbres ne serait pas suffisamment général, car cela empêche de coder des chevauchements qui portent sur plusieurs phrases ou des chevauchements qui présentent des structures à croisement [2]

4.2 La gestion des disfluences

Les disfluences sont annotées dans les données ETAPE par des balises XML qui groupent une séquence de mots comme étant disfluente. Schématiquement pour une phrase $w_1 \dots w_n$, une disfluence à la forme suivante : $w_1 \dots w_{b-1} \langle D \rangle w_b \dots w_{e-1} \langle /D \rangle w_e \dots w_n$. Où D représente un code XML pour hésitation, révision, répétition ou marqueur de discours. Les disfluences sont intra-phrastiques, peuvent avoir une structure interne (dans le cas de répétions ou de révisions par exemple) mais ne présentent pas de schémas de croisement non projectifs. Nous les représentons comme des noeuds syntagmatiques dans les arbres, comme illustré en Figure 3.

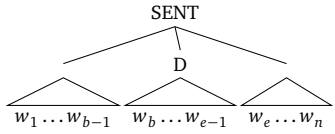

FIGURE 3 – Disfluences

L'attachement des disfluences dans les arbres de constituants n'étant pas naturellement déterministe, nous choisissons d'attacher les répétitions au premier syntagme qui contient le matériel répété, et les révisions au premier syntagme qui contient le matériel révisé. En cas d'hésitation sur le noeud auquel attacher la disfluence, on tranche pour l'attachement au noeud le plus haut dans l'arbre.

4.3 Les catégories utilisées

Catégories syntagmatiques	AdP, AP, COORD, NP, PP, VN, VPinf, VPpart
	Sint (parenthétique ou incise), Srel (relative), Ssub (subordonnée), SENT (racine)
Catégories lexicales	ADJ, ADJINT (adjectif interrogatif), ADV, ADVINT (adverbe interrogatif),
	ADVEX (adverbe exclamatif), (V (indicatif qui inclut conditionnel), VINF (infinitif)
	VIMP (impératif), VPP (part passé), VPR (part présent), VS (subjonctif)
	NC (nom commun), NPP (nom propre), CC (conj coord), CS (conj sub)
	CLS (clitique sujet), CLO (clitique objet ou complément), CLR (clitique réfléchi)
	P (préposition), P+D (au, du, des), P+PRO (auquel, duquel, desquels) PRO
	PROINT (pronom interrogatif), PROREL (pronom relatif)
	DET, DETINT (déterminant interrogatif), DETEX (déterminant exclamatif),
	ET (mot étranger), I (interjection), UK (mots inachevés/non reconnus)
	HES, REP, REV,DM
Symboles Fonctionnels	SUJ,OBJ,A-OBJ,DE-OBJ,P-OBJ,MOD,ATS,ATO,DIS,VOC
Marque d'inachèvement	INA

TABLE 1 – Jeu d'étiquettes utilisé dans le treebank oral

Le schéma d'annotation est un format en constituants et en fonctions dont les arbres sont annotées par un jeu d'étiquette utilisé par (Crabbé et Candito, 2008) et qui simplifie le jeu d'étiquette

2. Formellement, les balises de chevauchement n'encodent pas nécessairement des structures d'arbres projectifs.

du treebank écrit quant aux jeux de symboles préterminaux (étiquettes morphosyntaxiques). On ajoute à ce jeu d'étiquettes les symboles non terminaux HES, REV, REP, DM qui encodent respectivement les disfluences (hésitation, révision, répétition, marqueur de discours), et des symboles supplémentaires SYNC, OVERLAPB, OVERLAPE,BACKCHANNELB,BACKCHANNELE qui encodent dans les arbres les annotations de synchronisation son et de chevauchement extraites du format des annotations ETAPE.

De plus, certains noeuds comportent des annotations structurées par plus d'un attribut. Ainsi en plus de la catégorie syntaxique, on renseigne pour les noeuds arguments du verbe, c'est-à-dire les noeuds frères du noeud VN et les clitiques arguments leur fonction syntaxique prise parmi le jeu décrit par (Abeillé et Barrier, 2004) auquel on ajoute deux nouvelles fonctions de vocatif et de disloquées (notées VOC, DIS). Un troisième attribut booléen (noté INA) peut être renseigné sur un noeud non terminal pour indiquer qu'il encode un syntagme inachevé.

4.4 Quelques observations

Statistiques descriptives Suivant ce schéma d'annotation nous avons annoté 2118 phrases des corpus ESTER 3 et CORAL-ROM. En détaillant les différents sous-corpus, le treebank annoté se résume par la table suivante :

	Le masque et la plume	Un temps de Pauchon	CORAL-ROM(L'allumage)	Total
Occurrences	15260	11932	5050	32242
Phrases	795	882	441	2118
Lg. moy. phrases	19.1	13.5	11.5	15.2

TABLE 2 – Statistiques descriptives

Observations qualitatives On observe un certain nombre de particularités déjà mentionnées pour l'oral (Blanche-Benveniste, 1997). On observe une abondance de discours rapporté et d'incises (incise notée Sint :MOD en 1), un nombre important de syntagmes inachevés et d'énoncés fragmentaires. Un nombre important de phrases commencent par un marqueur discursif (2) ou une conjonction de coordination (phrase annotée comme COORD en 4) :

(1) (VN ils faisaient) (NP :OBJ (REV des :Det) (Sint :MOD je sais pas moi) des :Det trucs) (c-oral-rom)

(2) (DM bon-A alors-ADV) (VN raconte-moi) (NP :OBJ ton week-end)) (c-oral-rom)

On observe aussi de nombreuses juxtapositions (comme en (3) ou on duplique la fonction ATS) et on peut parfois hésiter entre une annotation comme disfluence (révision ou répétition) ou comme juxtaposition. A partir du moment où les disfluences ont la même structure interne que les autres syntagmes, comme la répétition en (4) qui inclut deux syntagmes, un utilisateur qui serait en désaccord peut choisir d'ignorer certaines balises de disfluences. Les répétitions intensives (5) ne sont pas notées comme des disfluences. De même les mots annotés comme marqueurs de discours ont leur étiquette habituelle (par exemple A, V ou ADV) dominée par la balise DM, comme en (2), qui peut aussi être ignorée en cas de besoin :

(3) C'est (NP :ATS un grand couteau), (NP :ATS une sorte de hachoir) (un temps de pauchon)

(4) (COORD mais (REP (VN il y a) (NP :OBJ mêlée)), (VN il y a) (NP :OBJ mêlée) (masque et la plume)

(5) Ils sont (AP :ATS très-ADV très-ADV laids-A) (masque et la plume)

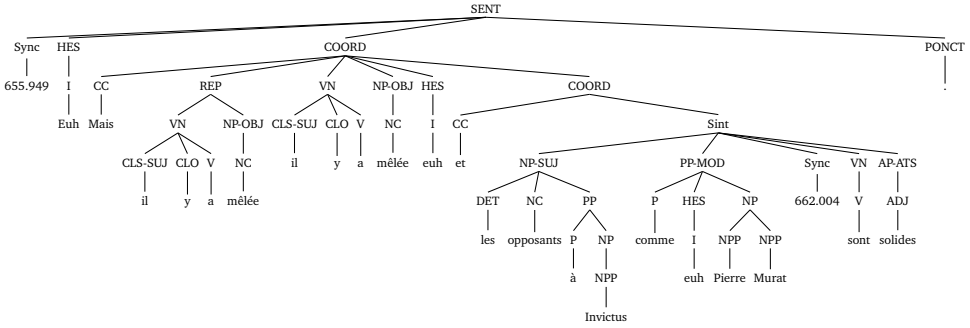

FIGURE 4 – Exemple d'arbre du Masque et la plume (après correction)

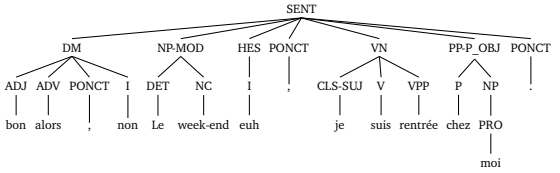

FIGURE 5 – Exemple d'arbre de CORAL-ROM (après correction)

5 Procédé d'annotation

Dans cette section nous décrivons plus précisément la méthode d'annotation qui a été déployée. Celle-ci se divise en trois étapes séquentielles.

Segmentation et linéarisation des données Lors de cette première étape, nous avons segmenté semi-automatiquement les données en phrases en nous appuyant sur la ponctuation donnée par les données au format ESTER 3. La segmentation en phrases a été systématiquement validée manuellement. Lors de cette étape le travail d'annotation a consisté tout d'abord à corriger la ponctuation ESTER 3. Celle-ci ayant été réalisée principalement sur critères phonétiques, elle a été corrigée pour refléter davantage une ponctuation grammaticale.

Lors de cette étape nous avons parfois rélinéarisé les données. En effet l'annotation ESTER 3 n'impose pas de contrainte stricte quant à l'ordre de la transcription lorsque plusieurs locuteurs parlent simultanément. Nous avons identifié quelques cas de structures syntaxiques bien formées qui étaient interrompues par le tour de parole d'un autre locuteur. Pour ces quelques cas, nous nous sommes permis de réordonner l'annotation pour restituer une cohérence quant à la structuration du texte en phrases.

Finalement, nous avons normalisé la segmentation en mots. La segmentation en mots a été réalisée de manière à minimiser la quantité de mots composés en nous basant sur la liste établie par (Crabbé et Candito, 2008). Sont retenus en priorité comme mots composés les mots composés grammaticaux (notoirement les déterminants, conjonctions de subordination et de coordination). La liste de (Crabbé et Candito, 2008) a été mise à jour et est documentée dans (Abeille *et al.*, 2013).

L'annotation syntaxique automatique la méthode d'analyse syntaxique repose sur l'hypothèse que la structure des phrases à l'oral n'est pas fondamentalement différente de celles de l'écrit. C'est plutôt la distribution de probabilité de la grammaire qui varie. Les données étant segmentées, l'étape d'analyse syntaxique couvre les tâches traditionnelles d'étiquetage morphosyntaxique, de parsing et d'étiquetage fonctionnel. Nous n'avons pas utilisé explicitement d'étiqueteur morphosyntaxique dans la mesure où l'analyseur syntaxique utilisé (Petrov *et al.*, 2006) est un modèle conjoint qui réalise déjà le tagging.

Plus spécifiquement, la méthode d'analyse utilisée tire parti des annotations en disfluences données par ESTER 3. L'analyse en constituants proprement dite est précédée d'un prétraitement qui supprime de l'entrée les disfluences, les marques de chevauchement et les balises de synchronisation avec la bande son. Celles-ci sont réintégrées dans les analyses en post-traitement. Les arbres de dysfluences sont créés de manière heuristique : la racine est la catégorie donnée par ESTER 3, celle-ci domine systématiquement les noeud préterminaux (tags) étiquetés par un 2-CRF linéaire (modèle appris sur le treebank écrit).

Les arbres sont finalement annotés en fonctions par un 2-CRF linéaire appris sur les données écrites suivant la description donnée dans (Candito *et al.*, 2009) : seuls les noeuds arguments du verbe reçoivent une étiquette fonctionnelle : il s'agit des noeuds en position frère des noeuds VN et des noeuds clitiques (en position frères du noeud V).

La correction manuelle L'étape de correction manuelle a consisté à corriger les annotations en constituant et en fonctions. Notons que nous avons travaillé avec des représentations type Penn Treebank ce qui a permis de réutiliser les interfaces graphique WordFreak destinée à l'édition d'arbres en constituants et Tregex (Levy et Andrew, 2006) pour la visualisation et la recherche de motifs, ce qui facilite considérablement le travail d'annotation. Cette partie du processus a consisté en une première étape d'annotation suivie d'une étape de discussion/adjudication entre annotateurs.

Concernant les disfluences, la correction concerne leur structure interne pour les révisions ou les répétitions comme en (6) ou leur rattachement. En (7) on a une phrase en discours rapporté (complément du verbe faire) réduite à un marqueur discursif. :

(6) moi, (REP (NP :SUJ ça-PRO) (VN-INA me-CLO)) ça me dit rien, moi (c-oral-rom)

(7) Il me fait : (Sint :OBJ (DM ben si)).

Pour les catégories lexicales, on observe le même type de corrections que pour l'écrit, concernant le mauvaise étiquetage de mots grammaticaux fréquents et ambigus comme pour *de* (préposition au lieu de déterminant) ou *que* (conjonction de subordination au lieu de pronom relatif). Les autres erreurs concernent les mots non appris sur l'écrit comme les interjections, ou plus rares comme les interrogatifs et les impératifs. Les formes verbales syncrétiques, fréquentes avec les verbes du premier groupe au présent, sont ainsi systématiquement étiquetées indicatif alors qu'il faut les corriger en impératif voire subjonctif. Pour les constituants aussi, on observe le même type de corrections que pour l'écrit concernant les mauvais rattachements de syntagmes prépositionnels ou de relative. Les autres corrections concernent l'ajout de l' étiquette INA quand le syntagme inachevé est mal formé et le rattachement des disfluences (REP, REV). Pour les fonctions, les corrections spécifiques concernent l'ajout des fonctions vocatif (8) et disloqué (9), et la réduplication des fonctions pour les juxtapositions (10). Une partie des corrections est la même que pour l'écrit concernant les sujets inversés ou la distinction entre complément et ajout

pour les syntagmes prépositionnels.

(8) (DM Allez-V) (NP-VOC Catherine) (NP encore un tour) ! (le masque et la plume)

(9) (NP :DIS moi-PRO), (NP :DIS ce qui me frappe), (VN c'-CLS-SUJ est) (NP-ATS la fin). (le masque et la plume)

(10) (NP-SUJ des chanteurs) ,(NP-SUJ des musiciens) (VN sont passés) dans ce théâtre (un temps de pauchon)

On compte 8 à 10 heures pour 100 phrases environ (en double correction). Au total, pour la correction des transcriptions et la segmentation (avant parsing) et la correction des analyses (après parsing), nous avons employé 4 annotateurs pour un total de 12 hommes-mois : 3 étudiants de Paris 7 en linguistique (M2) ou en linguistique informatique (M1), et une ancienne étudiante, spécialiste du FTB (Vanessa Combet).

6 Évaluation

Cette section propose une évaluation et une mise en perspective de la méthode de préannotation syntaxique (étape 2 du processus d'annotation), qui est l'étape clé du processus. La question que l'on se pose lorsqu'on veut annoter un corpus hors domaine consiste à déterminer la meilleure manière d'amorcer la préannotation des données de manière à faciliter la tâche des annotateurs sachant qu'on dipose d'un modèle d'analyse pour le domaine source.

En termes d'analyse syntaxique, l'annotation d'un corpus oral tombe dans la classe des problèmes d'adaptation de domaine. Celui-ci comporte deux aspects. Premièrement il s'agit d'adapter la structure : en effet nous avons vu que le schéma d'annotation de l'oral introduit de nouvelles structures et de nouvelles catégories liées aux disfluences. En second lieu il faut adapter la distribution de probabilité de la grammaire. Il s'agit du problème classique d'adapter la distribution de probabilité d'un modèle probabiliste entrainé sur un échantillon de données biaisé (un corpus écrit) à un échantillon possédant des propriétés différentes (corpus oral).

De manière à apporter une première idée de la correction de méthodes d'adaptation simples, nous comparons ici quatre méthodes qui tirent parti des données à la fois écrites et orales pour faciliter le processus de préannotation :

– **Utilisation des données écrites uniquement (E) :** Cette méthode de base consiste à analyser les données orales en utilisant uniquement un modèle d'analyse appris sur l'intégralité des données écrites (21268 phrases). Utiliser cette méthode de base ne permet pas d'envisager analyser correctement les structures propres à l'oral (dysfluences). Il s'agira essentiellement de notre *baseline*.
– **Approche par transformation/détransformation des données (T/D) :** Cette méthode consiste à prétraiter les données orales en supprimant les disfluences (balisées dans les données ESTER 3) de l'entrée donnée à l'analyseur syntaxique. Ce dernier, entraîné sur l'ensemble des données écrites (21268 phrases), doit alors prédire pour l'oral des structures qui ressemblent à celles de l'écrit. Une étape de post traitement réinsère finalement dans les arbres d'analyse les dysfluences supprimées en prétraitement.
  Chaque disfluence de k mots ainsi réinsérée est un arbre dont la racine est la catégorie de la dysfluence (donnée par ESTER 3). La racine domine immédiatement une séquence de k-tags étiquetées par un 2CRF linéaire appris sur le treebank écrit, chacun de ces $k-$tags domine le mot correpondant.

– **Approche par utilisation exclusive des données orales (O)** Dans ce troisième scénario, on suppose qu'on dispose d'un fragment de données déjà annotées pour le domaine cible. Le modèle d'analyse est entrainé uniquement sur ce fragment de données orales et n'utilise pas les données écrites.
– **Approche par utilisation combinée des données écrites et des données orales (O/E) :** Dans ce dernier scénario, le modèle est appris sur l'intégralité des données écrites et sur un fragment des données orales.

Comparaison des différentes méthodes Dans ce qui suit nous évaluons chacune de ces méthodes en fonction de la quantité de données orales utilisées pour entrainer le modèle. Dans le cas des méthodes (E) et (T/D), le fragment de données orales de référence disponible n'est pas utilisé pour l'entrainement. Les méthodes (E) (T/D) et (E/O) utilisent systématiquement l'intégralité des données écrites pour l'entrainement. Les fragments de données orales utilisés à l'entrainement du modèle par les méthodes (O) et (E/O) sont issus de données de référence déjà validées par les annotateurs.

L'analyseur utilisé est l'analyseur de Berkeley (Petrov *et al.*, 2006) tel que distribué à ce jour. Cet algorithme faiblement lexicalisé est connu pour être relativement robuste au changement de domaine. L'ensemble des tests réalisés repose sur la comparaison des prédictions de cet analyseur sur un corpus de test comportant 528 phrases. Le calcul du F-Score est réalisé avec le logiciel `evalb` (paramétrage standard, phrase de moins de 40 mots).

Nous avons évalué la correction de chacune des quatre méthodes en fonction de la taille du fragment de données orales utilisées à l'entrainement. Les résultats sont résumés dans la table 3 (**P**récision,**R**appel, **F**-score, **Tag**ging accuracy) [3]. Les lignes représentent chacune des quatre méthodes d'analyse. Les colonnes représentent la taille des données orales (en nombre de phrases) utilisées par l'analyseur lors de l'entrainement. Les chiffres indiquent le F-score de l'analyseur sur le jeu de test de 528 phrases.

Méthode	530				1060				1590			
	P	R	F	Tag	P	R	F	Tag	P	R	F	Tag
Ecrit (E)	62.2	66.4	64.3	61.7	62.2	66.4	64.3	61.7	62.2	66.4	64.3	61.7
Ecrit (T/D)	72.8	79.6	76.0	62.4	72.8	79.6	76.0	62.4	72.8	79.6	76.0	62.4
Oral (O)	64.8	64.8	64.8	66.4	68.9	69.2	69.0	70.7	70.6	71.3	71.0	72.3
Oral+Ecrit (O/E)	63.6	66.1	64.9	62.0	63.6	67.0	65.3	64.8	67.4	70.9	69.11	67.0

TABLE 3 – Evaluation des méthodes d'adaptation

Vu que les deux premières lignes représentent des protocoles qui ignorent totalement les données orales à l'entrainement, le score d'évaluation est constant. En première observation, on constate que la méthode de transformation/détransformation des données est celle qui donne les meilleurs résultats. L'explication la plus vraisemblable pour expliquer ce meilleur résultat tient probablement à (1) les données à prédire correspondent structurellement aux données apprises et (2) une partie de la solution est simplement déjà donnée : les dysfluences sont en effet copiées de l'entrée vers la sortie sans possibilité de se tromper dans leurs prédictions.

On constate également que le modèle mixte (O/E) fonctionne comparativement moins bien qu'un modèle entraîné uniquement sur les données orales (O). La raison est certainement à chercher dans le fait que les proportions de données orales et écrites de ce modèle sont inégales : 21268

3. Notons toutefois que les performances de l'analyseur varient d'un type de corpus à l'autre : ainsi on obtient un F-score de 69.5 sur les données CORAL-ROM et de 61.8 sur les données France Inter avec le modèle (E).

phrases pour l'écrit contre $k * 530$ phrases pour l'oral ($1 \leq k \leq 3$). Autrement dit, ce modèle reste fondamentalement semblable à un modèle de l'écrit.

Exploration du comportement des modèles mixtes (O/E) De manière à vérifier plus en détail si un modèle de type (O/E) permet d'obtenir un modèle satisfaisant en assurant une pondération plus appropriée des deux groupes de données (oral,écrit) nous avons procédé à une seconde expérience par rééchantillonage contrôlé des données. Dans cette seconde expérience nous avons testé dans quelle mesure un la méthode de type (O/E) se comporte en fonction de deux paramètres : (1) la proportion de données écrites dans le corpus d'entrainement et (2) la taille des données d'entrainement.

Le protocole de quantification des résultats est identique au cas précédent, nous utilisons systématiquement le même corpus de test. Ce qui change c'est la création du corpus d'entraînement. Ainsi pour chaque mesure réalisée, on a créé un corpus d'entrainement par échantillonage avec remise dans les données (angl. *bootstrapping with replacement*). Les groupes de données source (dans lesquelles on tire) sont un échantillon écrit E constitué des 21268 phrases du French treebank écrit, et d'un échantillon O constitué de 1530 phrases annotées pour l'oral. Notons k la proportion de texte écrit souhaitée dans le corpus généré. Chaque phrase c_i du corpus bootstrappé $C = c_1 \dots c_n$ est tirée avec une probabilité k dans le groupe E et $(1-k)$ dans le groupe O. Le tirage dans un groupe (E ou O) est fait de manière uniforme et avec remplacement (on peut tirer plusieurs fois le même exemple). C'est ce corpus généré aléatoirement C qui sert comme données d'apprentissage du modèle d'analyse syntaxique. Il est donc possible que certaines phrases de E ou de O ne soient pas représentées dans C échantillonné et que certaines phrase de E ou de O y soient représentées plusieurs fois.

Notons que le processus de bootstrapping nous permet de créer des corpus de tailles queclonques. Ainsi nous avons croisé chaque valeur retenue pour k (0,0.25,0.5,0.75) avec une taille de corpus n variant de 1000 à 7000 phrases. Les résultats d'anlyse sur les 528 phrases de test sont reportées dans le tableau 4. Les résultats montrent globalement qu'une pondération plus appropriée des

Données d'entrainement	1000	2000	3000	4000	5000	6000	7000
Mix(Oral,Ecrit,$k=0$)	65.57	68.09	69.15	68.2	69.1	67.4	68.0
Mix(Oral,Ecrit,$k=0.25$)	68.2	69.6	71.0	72.1	70.9	71.9	72.1
Mix(Oral,Ecrit,$k=0.5$)	67.8	69.6	71.0	72.0	69.1	67.4	67.9
Mix(Oral,Ecrit,$k=0.75$)	65.7	69.9	70.7	71.2	71.7	72.0	72.4

TABLE 4 – Evaluation par bootstrapping

deux groupes de données permet d'améliorer subtantiellement les performances de l'analyseur. Ainsi on atteint un F-Score de 72.4 pour un corpus d'entrainement de 7000 phrases comportant 75% de données écrites à comparer avec 69.1 obtenu par le mélange naïf de la première expérience. Toutefois, l'observation la plus étonnante reste la comparaison avec la méthode artisanale (T/D) F-score= 76% qui reste très nettement meilleure que la méthodes de mélange (O/E) même en contrôlant les proportions pour cette dernière. Pour confirmer la pertinence de notre méthode artisanale, il faudrait également la comparer à des méthodes d'adaptation de domaine plus élaborées que le bootstrapping, comme par exemple des méthodes d'active learning qui visent à pondérer d'avantage les exemples clés pour l'apprentissage ou encore à des méthodes d'apprentissage semi-supervisées. Il serait intéressant également de reformuler notre méthode artisanale sous forme d'analyse syntaxique de graphes acycliques orientés (DAG) où les dysfluences sont données en entrée à l'analyseur comme segments préparenthésés. Il faut toutefois noter que cette approche n'est pas parfaitement équivalente à la méthode (T/D) dans

la mesure où les segments préparenthésés seraient étiquetés par des symboles de dysfluences qui sont absents de la grammaire de l'écrit. Il faut toutefois rappeler que la méthode (T/D) s'applique à un scénario d'annotation dans lequel les disfluences sont déjà annotées. Les bonnes performances de cette méthode semblent en effet provenir du fait qu'une partie du parenthésage à prédire est donné. Dans un scénario d'analyse syntaxique de l'oral – à partir d'une source brute – déployer cette méthode demanderait en particulier de réaliser un tagger en disfluences pour l'oral dont les résultats sont supposés parfaits. Or l'étiquetage automatique de disfluences comme les répétitions ou les révisions ne représente apparemment pas un problème trivial.

7 Conclusion

Nous avons validé sur deux heures de transcription de débats radiophoniques et de dialoque informel, une méthode d'analyse syntaxique du français parlé, en constituants et en fonctions, inspiré de ce qui se fait pour d'autres langues, et qui est une extension naturelle du FTB pour le français journalistique. Nous avons enrichi le guide d'annotation du FTB, adapté et réentraîné l'analyseur de (Crabbé et Candito, 2008) et adapté une plate-forme d'annotation pour la validation manuelle. Les premiers résultats sont encourageants, à la fois en ce qui concerne les performances du parseur et les temps de correction. Les corpus radiophoniques annotés (une heure trente de temps de parole, environ 27 000 mots) seront distribués dans le cadre du consortium du projet Etape. Les annotations du dialogue c-oral-rom (Cresti *et al.*, 2004) sont disponibles et le corpus distribué par Elra. La suite du travail consistera à annoter des corpus oraux librement accessibles comme le corpus CID (Bertrand *et al.*, 2008) ou CFPP (Branca-Rosoff *et al.*, 2012).

Remerciements

Les auteurs tiennent à remercier les annotateurs qui ont contribué à corriger les annotations : Vanessa Combet, Floriane Guida, Antoine Lacambre et Mathilde Marié. Ceux-ci ont été financés par le projet ANR Etape (resp. G. gravier). Ce projet a aussi bénéficié du financement du Peps Syfrap (reps. C. Gardent) (CNRS INSHS INSII). Nous remercions Elisabeth Delais-Roussarie qui a corrigé certaines transcriptions, Mathilde Dargnat avec qui nous avons établi la liste des marqueurs de discours, Djamé Seddah pour l'aide au déploiement des outils de correction ainsi que Claire Gardent et Christophe Cerisara pour les discussions permettant de comparer annotations en constituants et annotations en dépendances.

Références

Abeille, A., Combet, V. et Crabbé, B. (2013). Conventions pour annotation syntaxique du français parlé. Rapport technique, Université Paris 7.

Abeillé, A. et Barrier, N. (2004). Enriching a french treebank. *In Proceedings of LREC*.

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In* Abeillé, A., éditeur : *Treebanks*. Kluwer, Dordrecht.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. *In First International Conference on Language Resources and Evaluation (LREC)*.

BERTRAND, R., BLACHE, P., ESPESSER, R., FERRÉ, G., MEUNIER, C., PRIEGO-VALVERDE, B. et RAUZY, S. (2008). Le cid - corpus of interactional data - annotation et exploitation multimodale de parole conversationnelle. *TAL*, 49(3).

BLACHE, P., BERTRAND, R., GUARDIOLA, M., GUÉNOT, M.-L., C.MEUNIER, NESTERENKO, I., PALLAUD, B., PRÉVÔT, L., PRIEGO-VALVERDE, B. et RAUZY, S. (2010). The OTIM formal annotation model : a preliminary step before annotation scheme. *In Proceedings LREC*.

BLACHE, P. et RAUZY, S. (2012). Enrichissement du ftb : un treebank hybride constituants/propriétés. *In Actes TALN*, Grenoble.

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

BRANCA-ROSOFF, S., FLEURY, S., LEFEUVRE, F. et PIRES, M. (2012). Discours sur la ville. corpus de français parlé parisien des années 2000. Rapport technique, Université Paris 3.

CANDITO, M., CRABBÉ, B., DENIS, P. et GUÉRIN, F. (2009). Analyse syntaxique du français : des constituants aux dépendances. *In TALN*.

CERISARA, C., GARDENT, C. et ANDERSON, C. (2010). Building and exploiting a dependency treebank for french radio broadcast. *In Proc. TLT9*, Tartu, Estonia.

CRABBÉ, B. et CANDITO, M. (2008). Expériences d'analyse syntaxique du français. *In TALN*.

CRESTI, E., do NASCIMENTO, F. B., MORENO-SANDOVAL, A., VÉRONIS, J., MARTIN, P. et CHOUKRI, K. (2004). The c-oral-rom corpus. a multilingual resource of spontaneous speech for romance languages. *In LREC*.

DEULOFEU, J., DUFORT, L., GERDES, K., KAHANE, S. et PIETRANDREA, P. (2010). Depends on what the french say : Spoken corpus annotation with and beyond syntactic function. *In Linguistic Annotation Workshop (LAW IV)*.

GRAVIER, G., ADDA, G., PAULSSON, N., CARRÉ, M., GIRAUDEL, A. et GALIBERT, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. *In Proc LREC*.

HOEKSTRA, A., MOORTGAT, M., SCHUURMAN, I. et van der WOUDEN, A. (2000). Syntactic annotation for the spoken dutch corpus project (cgn). *In Computational Linguistics in the Netherlands (CLIN)*.

LEVY, R. et ANDREW, G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. *In Proc. LREC*.

MARCUS, M. P., SANTORINI, B. et MARCINKIEWICZ, M. A. (1993). Building a large annotated corpus of english : The penn treebank. *Computational Linguistics*, 19(2):313–330.

METEER, M. (1995). Dysfluency annotation stylebook for the switchboard corpus. Rapport technique, Upenn.

MIKULOVA, M. (2008). Rekonstrukce standardizovaného textu z mluvené řeči v pražsksém závislostním korpusu mluvené češtiny. manuál pro anotátory. Rapport technique 38, UFAL.

PETROV, S., BARRETT, L., THIBAUX, R. et KLEIN, D. (2006). Learning accurate, compact, and interpretable tree annotation. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

SCHLUTER, N. et van GENABITH, J. (2007). Preparing, restructuring and augmenting a french treebank : lexicalized parsers or coherent treebanks ? *In Proceedings Pacling 2007*, Melbourne.

