TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe

Christophe Benzitoun[1] Karën Fort[2,3] Benoît Sagot[4]
(1) ATILF, Nancy Université & CNRS, 44, avenue de la Libération, BP 30687, 54063 Nancy cedex
(2) INIST-CNRS, 2 allée de Brabois, 54500 Vandoeuvre-lès-Nancy
(3) LIPN, Université Paris 13 & CNRS, 99 av. J.B. Clément, 93430 Villetaneuse
(4) Alpage, INRIA Paris–Rocquencourt & Université Paris 7, Rocquencourt, France
christophe.benzitoun@atilf.fr, karen.fort@inist.fr, benoit.sagot@inria.fr

RÉSUMÉ

Nous présentons dans cet article un travail portant sur la création d'un corpus de français parlé spontané annoté en morphosyntaxe. Nous détaillons la méthodologie suivie afin d'assurer le contrôle de la qualité de la ressource finale. Ce corpus est d'ores et déjà librement diffusé pour la recherche et peut servir aussi bien de corpus d'apprentissage pour des logiciels que de base pour des descriptions linguistiques. Nous présentons également les résultats obtenus par deux étiqueteurs morphosyntaxiques entrainés sur ce corpus.

ABSTRACT

TCOF-POS : A Freely Available POS-Tagged Corpus of Spoken French

This article details the creation of TCOF-POS, the first freely available corpus of spontaneous spoken French. We present here the methodology that was followed in order to obtain the best possible quality in the final resource. This corpus already is freely available and can be used as a training/validation corpus for NLP tools, as well as a study corpus for linguistic research. We also present the results obtained by two POS-taggers trained on the corpus.

MOTS-CLÉS : Etiquetage morpho-syntaxique, français parlé, ressources langagières.

KEYWORDS: POS tagging, French, speech, language resources.

1 Introduction

L'annotation automatique du français parlé est généralement réalisée par le biais de prétraitements de corpus ou d'adaptation d'outils existant pour le texte (Dister, 2007; Blanc *et al.*, 2008). Une autre solution peut consister à masquer certains phénomènes tels que les "disfluences" (répétitions, amorces de mots, etc.) (Valli et Véronis, 1999). Pourtant, l'utilisation d'étiqueteurs automatiques élaborés pour et à partir de données écrites n'est pas une solution optimale étant données les particularités des corpus oraux par rapport à l'écrit. Même si l'étiquetage de corpus oraux ne représente pas un problème spécifique (Benzitoun, 2004), l'utilisation de modèles entraînés sur des données écrites donne des résultats médiocres. Ainsi, nous avons testé TreeTagger (Schmid, 1997), avec son modèle standard pour le français, sur un échantillon de 3 007 tokens extraits du corpus de référence décrit dans cet article et nous avons obtenu une précision de seulement 83,1 %.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 99–112,
Grenoble, 4 au 8 juin 2012.

Un corpus du français parlé annoté en morphosyntaxe librement disponible serait donc utile, non seulement pour les logiciels d'annotation en morphosyntaxe, mais également pour améliorer les systèmes de transcription automatique (Huet *et al.*, 2006) ou d'autres outils. Cependant, il n'existe pas encore, à notre connaissance, de corpus de français parlé spontané annoté en morphosyntaxe (parties du discours et/ou lemmes) qui soit diffusé librement. Parmi les corpus annotés mais non diffusés librement, on peut citer les projets elicop (Mertens, 2002), C-ORAL-ROM (Campione *et al.*, 2005), Valibel (Dister, 2007), Corpus de Français Parlé Parisien (Branca-Rosoff *et al.*, 2010) ou bien encore ESLO (Eshkol *et al.*, 2010).

Notre objectif est donc de développer et diffuser librement à l'ensemble de la communauté scientifique un corpus pré-annoté automatiquement puis corrigé manuellement, dont la qualité aura été précisément évaluée. Il pourra servir notamment de corpus d'apprentissage spécifique au français parlé et plus largement de corpus exploitable pour des recherches en linguistique ou en Traitement Automatique des Langues (TAL).

Nous présentons tout d'abord le corpus de français parlé TCOF (Traitement des Corpus Oraux du Français), puis la méthodologie utilisée pour l'annotation manuelle, les différentes évaluations réalisées pendant la campagne et enfin les résultats obtenus par les étiqueteurs morphosyntaxiques entrainés sur une partie du corpus annoté TCOF-POS.

2 Présentation du corpus

Le corpus d'origine que nous avons annoté est celui du projet TCOF (André et Canut, 2010), librement disponible sur le site du CNRTL [1]. Ce corpus est constitué de transcriptions de données orales recueillies dans des contextes aussi naturels que possible. Il comporte une partie d'interactions entre adultes et une autre entre adultes et enfants. En ce qui concerne la partie adulte (la seule que nous ayons exploitée jusqu'à présent), elle est composée :

– d'interactions sollicitées, dans lesquelles au moins deux locuteurs sont engagés dans des récits de vie, d'événements ou d'expériences, ou dans des explications sur un savoir-faire professionnel ou technique ;
– de conversations à bâtons rompus ou portant sur des thématiques spécifiques ;
– de données non sollicitées dans des situations publiques ou professionnelles : réunions publiques, activités professionnelles diverses.

De ce corpus, nous avons extrait un échantillon de 22 240 tokens [2], soit 11 transcriptions différentes. Cet échantillon contient des conversations, des réunions professionnelles, ainsi que des extraits d'une Assemblée Générale à l'Université. L'intégralité des paroles prononcées a été scrupuleusement retranscrite en orthographe standard, sans artifice ou aménagement orthographique (donc sans ponctuation), suivant en cela les recommandations de (Blanche-Benveniste et Jeanjean, 1987) [3] largement diffusées et utilisées. Elles sont au format généré par le logiciel Transcriber (trs - XML).

1. http ://cnrtl.fr/corpus/tcof/
2. Notre conception de la notion de token est assez élémentaire (aucune insertion possible).
3. Les conventions de transcription sont disponibles sur le site suivant : http ://cnrtl.fr/corpus/tcof/TCOFConventions.pdf

Ces transcriptions sont automatiquement converties en texte brut à l'aide d'une feuille de style XSLT (qui élimine l'intégralité des balises XML), puis d'une série d'expressions régulières qui supprime les informations non désirées, telles que les pauses. Le texte final contient les mentions des locuteurs (L1, L2, etc.), l'intégralité des paroles prononcées, ainsi que les multi-transcriptions. Il s'agit donc de transcriptions brutes non retouchées, dont voici un exemple :

> *L1 et puis je crois que c'est en je crois je crois même que c'est en zone industrielle*
> *L2 ouais ouais je pense aussi ça doit pas être en ville*
> *L1 oui mais*
> *L2 en Belgique aussi il y a des trucs euh un genre de grand tr- enfin un genre de grande galerie en Belgique et puis c'est que des magasins de fringues aussi*

Les transcriptions ont été faites dans le cadre d'un cours de deuxième année de Sciences du Langage à l'Université Nancy 2, puis revues par des enseignants de Sciences du Langage. L'anonymisation, quant à elle, a été réalisée manuellement par des étudiants-vacataires. A la lecture, ils devaient repérer les toponymes, anthroponymes, etc. puis les remplacer par un symbole et insérer un son dans la portion de signal sonore correspondante.

3 Méthodologie

L'annotation totalement manuelle de corpus étant très coûteuse, nous avons procédé, comme décrit dans (Marcus *et al.*, 1993), à une correction manuelle de corpus pré-annotés automatiquement. La nature du pré-annotateur, ainsi que les modalités de la correction manuelle diffèrent selon les étapes du processus, comme nous allons le voir dans cette section. Toutefois, toutes les pré-annotations ont été produites par différentes instances du système TreeTagger (Schmid, 1997), qui fournit pour chaque token d'entrée une étiquette morphosyntaxique et un lemme.

Comme indiqué en introduction, l'utilisation comme pré-annotateur pour un corpus de parole spontanée transcrite, d'un étiqueteur morphosyntaxique entraîné sur un corpus écrit n'est pas adaptée. Parmi les phénomènes qui posent problème, lesquels ne sont pas totalement absents des corpus écrits mais y sont bien plus rares (Benzitoun, 2004), on peut citer :
– les répétitions de mots ou de groupes de mots (*ça ça redevient ça redevient le bordel comme ça*),
– les reformulations (*peut-être séparer complètement euh junior euh homme enfin euh adulte*),
– les ruptures de construction (*ouais ouais que de la gueule que de la*),
– les amorces de mots (*moi j'aurais p- j'aurais pas mis de pantalon*),
– les incises (*euh on considérait que former les hommes et c'est toujours euh en en en vigueur ça hein former les les en- les les enfants d'aujourd'hui c'est aussi former les hommes de demain*),
– les formes non conventionnelles (*tu sais genre trop vénère ; avoir du matos en entrée de mag*),
– les particules discursives (*hein, eh ben*, etc.)...
Pour ne prendre que deux exemples, la version standard de TreeTagger pour le français considère que *bon* est systématiquement un adjectif et *quoi* un pronom, alors qu'ils sont majoritairement des particules discursives. De plus, nos transcriptions ne sont pas segmentées en « phrases » (Blanche-Benveniste et Jeanjean, 1987), ce qui peut également poser des problèmes aux outils. Par exemple, l'étiquette SENT (pour *sentence*) indiquant une frontière de phrase doit obligatoirement être présente dans le lexique servant pour l'apprentissage de TreeTagger (même s'il ne s'en sert pas par la suite). En conséquence, nous avons procédé, dès que possible, à l'entraînement de versions de TreeTagger à partir des annotations déjà obtenues sur notre corpus.

La méthodologie retenue, décrite en détail dans cette partie, peut être résumée comme suit :

1. Définition de critères de tokenisation et d'identification des composés, puis tokenisation automatique ;
2. Définition d'un jeu d'étiquettes adapté à la parole spontanée transcrite ;
3. Création d'un corpus de référence C_{ref} de 22 240 tokens par correction d'une pré-annotation automatique, effectuée par deux experts linguistes :
   – Les 10 000 premiers tokens de C_{ref} ont été pré-annotés avec la version standard de TreeTagger ;
   – Les 12 240 tokens suivants de C_{ref} ont été pré-annotés avec une version de TreeTagger entraînée sur les 10 000 premiers ;
4. Ré-annotation par deux étudiantes d'environ 7 500 tokens du corpus de référence C_{ref} (suivie d'une phase d'adjudication), afin d'évaluer la qualité des annotations dans deux configurations distinctes :
   – environ 6 000 tokens ont été pré-annotés par la version standard de TreeTagger ;
   – environ 1 500 tokens ont été pré-annotés avec une version de TreeTagger entraînée sur les 16 312 premiers tokens de C_{ref} ;

   L'objectif était ici de mesurer l'impact de la différence de qualité entre pré-annotateurs en termes de vitesse d'annotation et de précision du résultat de l'étape manuelle ;
5. Application de cette méthodologie à un plus grand nombre d'étudiants pour en valider la robustesse ;
6. Annotation par deux étudiantes d'un corpus additionnel C_{add} de 80 000 nouveaux tokens, pré-annotés avec la version de TreeTagger entraînée sur la totalité de C_{ref}.

Nous avons appliqué pour cette campagne les bonnes pratiques actuelles en annotation manuelle de corpus, qui consistent à évaluer le plus tôt possible l'accord inter-annotateurs et de mettre à jour le guide d'annotation (Bonneau-Maynard *et al.*, 2005). La répétition régulière de ce processus conduit à ce qu'on appelle maintenant l'annotation agile (Voormann et Gut, 2008).

3.1 Tokenisation et gestion des composés

Le corpus ayant fait l'objet d'une pré-annotation (voir section 3.3), nous avons pris comme base la tokenisation par défaut de TreeTagger, qui repose notamment sur un fichier de composés. Mais ce dernier s'est avéré insuffisant (par exemple, *parce que* reste découpé en deux tokens distincts mais *puisqu'ils* en un seul token). Nous l'avons donc complété au fur et à mesure, en respectant le critère suivant : toute séquence dans laquelle il est possible d'insérer un élément est découpée en plusieurs tokens, afin d'exclure les unités discontinues. Ainsi, *un peu* est découpé en deux tokens (car on peut trouver *un tout petit peu*).

3.2 Un jeu d'étiquettes adapté à la parole spontanée transcrite

Afin de bénéficier au mieux des ressources développées pour l'écrit et de limiter le travail de correction, tout en prenant en considération les phénomènes spécifiques à la parole spontanée cités ci-dessus, nous avons décidé d'utiliser un jeu d'étiquettes basé au départ sur les étiquettes par défaut fournies par TreeTagger. Nous l'avons complété à l'aide de (Abeillé et Clément, 2006).

Les répétitions, reformulations, etc. n'ont pas fait l'objet de traitements spécifiques, chacun des tokens a la catégorie qu'il a habituellement (ex : le[DET] le[DET] le[DET] chat). Au final, même si les identifiants des étiquettes sont différents, les catégories retenues sont quasiment identiques à (Abeillé et Clément, 2006), avec toutefois un peu moins de sous-catégories (notamment aucune pour les adverbes et les adjectifs) et l'ajout de la catégorie « auxiliaire » ainsi que de trois étiquettes spécifiques à l'oral : *MLT* (multi-transcription), *TRC* (amorce de mot) et *LOC* (locuteur) (cf. tableau 1). Afin de nous aider dans la rédaction du manuel d'annotation, nous nous sommes d'ailleurs inspirés de (Abeillé et Clément, 2006). Notre jeu d'étiquettes comprend 62 étiquettes.

En outre, il a été affiné durant la phase de constitution du corpus servant de référence. En effet, nous voulions que les étiquettes soient apposées de manière aussi systématique que possible pour que nos choix soient réversibles et que les modifications soient automatisables, autant que faire se peut. De ce fait, même si cela peut paraître discutable d'un point de vue théorique, nous avons privilégié les choix qui potentiellement génèrent le moins de fluctuations entre annotateurs. Par exemple, la distinction entre participe passé et adjectif n'est pas aisée et plutôt que d'obtenir une annotation de qualité moindre, nous avons préféré neutraliser celle-ci. Ainsi, chaque fois que la forme verbale existe (sauf cas de changement notoire de sens), nous avons annoté « verbe ». Dans le cas contraire, nous avons annoté « adjectif ».

Nous avons également décidé d'essayer de limiter les cas de transferts d'une catégorie vers une autre (trans-catégorisation). En effet, ceux-ci auraient artificiellement été limités aux cas rencontrés dans le corpus à annoter, sans possibilité d'avoir une vision globale du phénomène. De plus, cela aurait complexifié la tâche de correction. Ainsi, dans *rouler tranquille*, *tranquille* est considéré comme un adjectif et non comme un adverbe (ce qui, de toute façon, est discutable d'un point de vue théorique). Enfin, il n'a pas été possible d'exclure totalement les cas d'étiquettes limitées à un mot unique. Ainsi, l'étiquette « particule interrogative » ne s'utilise que pour *est-ce qu-e/i* et « prédéterminant » uniquement pour *tous*.

3.3 Création du sous-corpus de référence

Comme indiqué ci-dessus, la création de la première tranche de 10 000 tokens du corpus de référence C_{ref} de 22 240 tokens a été réalisée en utilisant comme pré-annotateur la version standard de TreeTagger, entraînée sur un corpus écrit. Nous (L. Bérard et C. Benzitoun) avons ensuite corrigé ces pré-annotations en plusieurs passes. Nous avons tout d'abord effectué des remplacements automatiques, lorsque les modifications étaient systématiques ou que l'étiquette majoritaire n'était pas celle apposée par défaut par le logiciel (ce qui est le cas pour *bon* (ADJ/INT) et *quoi* (PRO :int/INT), par exemple). Ensuite, nous nous sommes répartis les données à corriger et, après les avoir intégralement traitées, nous nous les sommes échangées pour révision. Nous avons ensuite discuté des cas où nous n'étions pas en accord jusqu'à trouver des solutions. Nous avons effectué ces étapes plusieurs fois, jusqu'à obtenir des annotations fiables. Le guide d'annotation était mis à jour à chaque étape.

Nous avons par ailleurs généré automatiquement des fichiers de fréquences, afin de faciliter le repérage des erreurs. A ainsi été calculée la fréquence de chaque étiquette pour un même lemme ou un même token, ce qui nous a permis d'identifier et de corriger quelques erreurs supplémentaires. Par exemple, *C.E.* ayant été annoté 1 fois *NAM* et 3 fois *NOM :sg* (pour un même lemme *C.E.*), cette dernière étiquette a été attribuée aux 4 occurrences. De même, cela nous a permis de corriger deux occurrences de *du*, indûment annotées *DET :ind*.

ADJ	adjectif	NUM	numéral
ADV	adverbe	PRO :clo	clitique objet
AUX :cond	auxiliaire au conditionnel	PRO :cls	clitique sujet
AUX :futu	auxiliaire au futur	PRO :clsi	clitique sujet impersonnel
AUX :impe	auxiliaire à l'impératif	PRO :dem	pronom démonstratif
AUX :impf	auxiliaire à l'imparfait	PRO :ind	pronom indéfini
AUX :infi	auxiliaire à l'infinitif	PRO :int	pronom interrogatif
AUX :pper	auxiliaire au participe passé	PRO :pos	pronom possessif
AUX :ppre	auxiliaire au participe présent	PRO :rel	pronom relatif
AUX :pres	auxiliaire au présent	PRO :ton	pronom tonique
AUX :simp	auxiliaire au passé simple	PRP	préposition
AUX :subi	auxiliaire au subjonctif imparfait	PRP :det	préposition/déterminant
AUX :subp	auxiliaire au subjonctif présent	PRT :int	particule interrogative (est-ce que)
DET :def	déterminant défini	SYM	symbole
DET :dem	déterminant démonstratif	TRC	amorces de mots
DET :ind	déterminant indéfini	VER	verbe sans flexion (voilà)
DET :int	déterminant interrogatif	VER :cond	verbe au conditionnel
DET :par	déterminant partitif (du)	VER :futu	verbe au futur
DET :pos	déterminant possessif	VER :impe	verbe à l'impératif
DET :pre	pré-déterminant (tout (le))	VER :impf	verbe à l'imparfait
EPE	épenthétique	VER :infi	verbe à l'infinitif
ETR	mots étrangers	VER :pper	verbe au participe passé
FNO	forme noyau (oui, non, d'accord, etc.)	VER :ppre	verbe au participe présent
INT	interjection et particules discursives	VER :pres	verbe au présent
KON	conjonction	VER :simp	verbe au passé simple
LOC	locuteur	VER :subi	verbe au subjonctif imparfait
MLT	multi-transcription	VER :subp	verbe au subjonctif présent
NAM	nom propre	NOM :trc	nom commun tronqué
NOM	nom commun	NAM :trc	nom propre tronqué
NOM :sig	sigle	VER :trc	verbe tronqué
NAM :sig	sigle	ADJ :trc	adjectif tronqué

TABLE 1 – Jeu d'étiquettes du corpus TCOF-POS

Nous avons ensuite appliqué les résultats obtenus par Fort et Sagot (2010) sur l'intérêt d'une pré-annotation avec un outil de qualité moyenne. Ainsi, une fois la première tranche de 10 000 tokens annotés, nous avons ré-entraîné TreeTagger sur ce sous-corpus (mais sans utiliser de lexique externe) et avons pré-annoté les transcriptions suivantes de C_{ref} (12 240 tokens) avec ce nouvel outil. La même méthodologie que celle utilisée pour corriger les 10 000 premiers tokens nous a permis de finaliser le corpus de référence C_{ref} de 22 240 tokens.

3.4 Création du sous-corpus additionnel

Le corpus diffusé est composé pour une part du sous-corpus de référence C_{ref} et pour une autre part d'un autre sous-corpus additionnel C_{add} corrigé par deux étudiantes (de L3 et M2 de Sciences du Langage) recrutées spécifiquement pour cette tâche. Dans un premier temps, afin d'évaluer *a priori* la méthodologie prévue pour l'annotation de C_{add}, nous avons mené une campagne de tests en nous servant de C_{ref} comme référence. Pour ce faire, les deux étudiantes ont eu 15 fichiers extraits de C_{ref} d'environ 500 tokens chacun [4] à corriger dans un ordre contraint. Les 12 premiers fichiers avaient été pré-annotés par la version standard de TreeTagger. Afin de mesurer l'impact de la qualité du pré-annotateur, les 3 derniers fichiers avaient été pré-annotés par une version de TreeTagger ré-entraînée à partir d'un extrait de 16 312 tokens de la référence C_{ref} (et sans lexique externe). Naturellement, ces tokens forment un sous-ensemble de C_{ref} disjoint des 15 fichiers à ré-annoter. Les étudiantes avaient l'interdiction d'échanger des informations durant la phase d'annotation.

La correction a été effectuée dans un tableur, les cellules contenant les étiquettes étant munies d'une liste déroulante se limitant au jeu d'étiquettes défini ci-dessus. La saisie était donc contrainte. Une fois la correction terminée, les fichiers annotés en parallèle ont été comparés automatiquement. Les cas de divergence entre les deux annotateurs ont ainsi été repérés automatiquement et corrigés par un expert [5].

Dans le cadre de ce travail, les mesures suivantes ont été effectuées :
– le temps mis par les étudiantes pour annoter chaque fichier ;
– la précision de chaque fichier par rapport à la référence ;
– l'accord inter-annotateurs des étudiantes (Kappa de Cohen (Cohen, 1960)) ;
– la précision après fusion et adjudication.
L'évaluation de leurs annotations sur ces 15 fichiers est reproduite ci-dessous (figures 1 et 2 et tableau 2). Elle tient compte de la lemmatisation et des parties du discours.

1e	2e	3e	4e	5e	6e	7e	8e	9e	10e	11e	12e	13e	14e	15e
107	71	80	67	60	60	57	65	50	50	52	47	32	32	31

TABLE 2 – Temps d'annotation (en minutes)

Entre le 12^e et le 13^e fichier, la différence de temps est vraisemblablement imputable au changement de pré-annotation par TreeTagger.

4. Cette taille a été retenue car nous avons observé qu'elle permet une correction rapide et une attention soutenue sans être obligé de s'interrompre en cours d'annotation.

5. Pour des raisons pratiques, il n'a pas été possible de confier cette phase à un expert externe. La personne qui l'a réalisée a également collaboré à la réalisation du corpus servant de référence, ce qui peut représenter un biais méthodologique.

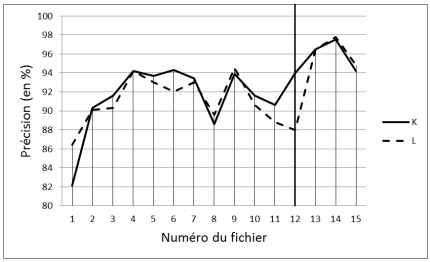

FIGURE 1 – Évolution de la précision des deux étudiantes

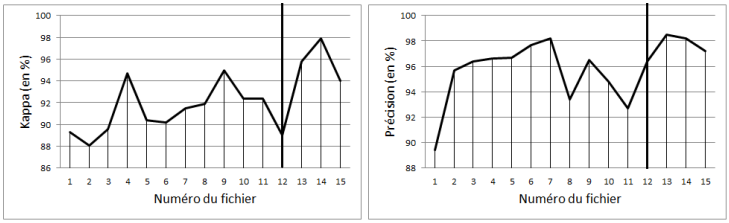

FIGURE 2 – Évolution du Kappa (à gauche) des 2 étudiantes et de la précision après adjudication (à droite)

La qualité des corrections après ré-entraînement et pré-étiquetage ainsi que le faible temps de correction (pour les 3 derniers fichiers, donc) nous ont paru suffisants pour valider notre méthodologie et ainsi poursuivre l'élaboration du corpus. Sur les 3 derniers fichiers, la précision moyenne est de 98,03 % en ne tenant compte que des étiquettes. Nous sommes donc passés à l'annotation par les deux étudiantes du corpus C_{add}. Elles ont ainsi reçu le même jeu de 160 nouveaux fichiers de 500 tokens chacun, pré-annotés par la version ré-entraînée de TreeTagger. En 60 heures, elles ont corrigé 80 000 tokens chacune, ce qui fait une moyenne d'un peu plus de 21 minutes par fichier de 500 tokens. Sur l'ensemble, l'accord inter-annotateurs (Kappa de Cohen (Cohen, 1960)) est en moyenne de 96,5 % et le temps moyen consacré à l'adjudication de 2 min. 45 sec par fichier.

4 Élargissement de l'évaluation

Afin d'évaluer le caractère robuste de notre méthodologie, nous avons élargi l'évaluation à plus d'étudiants. En effet, nous comptons augmenter de manière importante la quantité de fichiers corrigés dans les années à venir et nous voulons vérifier si notre méthodologie donne des résultats comparables quels que soient les correcteurs. Pour ce faire, nous avons adopté

la même méthodologie que celle décrite ci-dessus, à savoir une double-annotation de chaque fichier puis une adjudication pour les cas de divergence uniquement. Cette évaluation a porté sur les corrections fournies par 10 étudiants en Sciences du Langage à l'Université Nancy 2 (L3 et M2) dans le cadre de deux enseignements. A chaque binôme, nous avons donné 6 fichiers (4 fichiers pré-annotés avec le TreeTagger de base et 2 fichiers avec le TreeTagger ré-entraîné) à corriger dans un ordre contraint. Dans cette expérience, comme dans la précédente, les étudiants devaient corriger les lemmes en plus des étiquettes. Dans la suite de ce travail, les mesures que nous présentons tiennent compte à la fois des lemmes et étiquettes (sauf précision contraire).

4.1 Temps d'annotation et accord inter-annotateurs

En ce qui concerne le temps d'annotation, nous avons observé une diminution systématique avec une nette différence entre les 4 premiers fichiers et les deux derniers (voir tableau 3).

1e annot.	2e annot.	3e annot.	4e annot.	5e annot.	6e annot.
110,1	101,8	79,2	72,5	41,3	39,7

TABLE 3 – Temps d'annotation (en minutes)

Au-delà de la diminution du temps de correction inhérente à une meilleure maîtrise des étudiants, il paraît difficile d'expliquer la diminution du temps entre le quatrième et le cinquième fichier par un autre facteur que le basculement entre le TreeTagger standard et la version ré-entraînée. Le même phénomène peut être observé concernant l'accord inter-annotateurs (*cf.* figure 3). Le coefficient d'accord inter-annotateurs présenté ici est, comme précédemment, le κ de Cohen (Cohen, 1960).

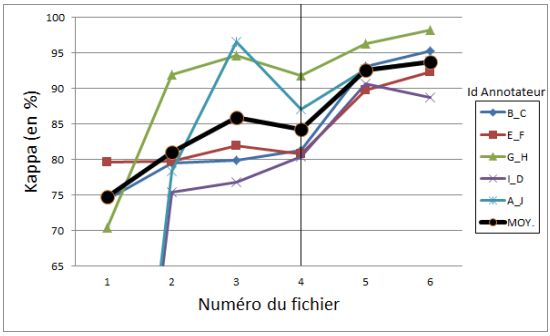

FIGURE 3 – Évolution de l'accord inter-annotateurs (kappa) des étudiants

Dans la figure 3, la courbe noire représente l'évolution de la moyenne des accords inter-annotateurs, de même que dans les graphiques suivants.

4.2 Précision

Outre une diminution significative du temps d'annotation et une augmentation de l'accord inter-annotateurs, nous avons également constaté une importante augmentation de la précision en moyenne pour chaque étudiant (*cf.* figure 4). On observe encore une fois une nette augmentation entre le quatrième et le cinquième fichier, et ce chez tous les étudiants. La figure 5 indique la précision de chaque fichier après fusion et adjudication.

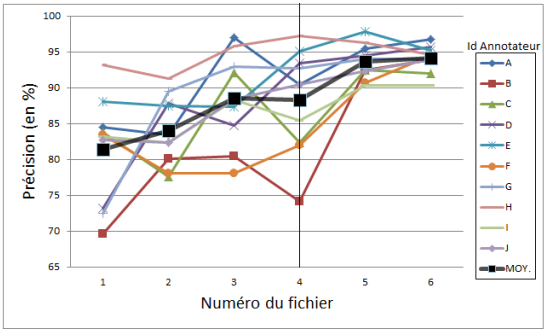

FIGURE 4 – Evolution de la précision de chaque étudiant

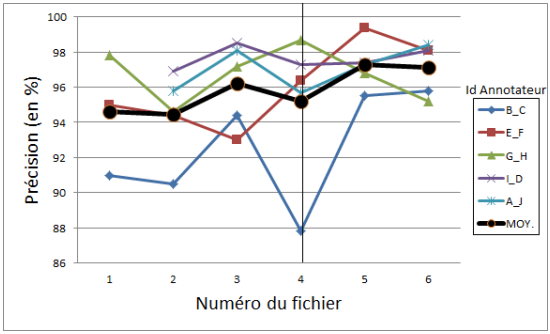

FIGURE 5 – Évolution de la précision de chaque fichier corrigé après fusion et adjudication

Finalement, sur les deux derniers fichiers annotés, le taux de précision moyen est respectivement de 97,28 % et 97,12 % en évaluant les erreurs portant à la fois sur les lemmes et les étiquettes. Si l'on prend en compte seulement les étiquettes, le taux de précision moyen est alors respectivement de 97,42 % et de 97,8 % pour ces mêmes fichiers, ce qui est légèrement inférieur à ce qui a

été relevé précédemment pour les deux étudiantes. Cependant, nous estimons que cela permet d'affirmer que les principes adoptés permettent d'obtenir des corpus annotés de qualité proche, et ce quelle que soit la personne qui corrige.

5 Premiers résultats de l'annotation automatique

Pour effectuer les premiers tests concernant l'apprentissage automatique, notre choix s'est porté sur deux étiqueteurs morphosyntaxiques : MElt (Denis et Sagot, 2009, 2012), étiqueteur état de l'art pour le français, et TreeTagger, utilisé comme pré-annotateur pour constituer le corpus, et largement utilisé bien qu'il ne soit pas celui qui donne les meilleurs résultats à l'heure actuelle (Denis et Sagot, 2009; Eshkol *et al.*, 2010). Tous deux sont librement disponibles et multi-plateformes.

Notre objectif était d'étudier la courbe d'apprentissage, et ce sous plusieurs angles : la précision de l'étiqueteur entraîné augmente-t-elle avec la taille du corpus d'entraînement ? L'utilisation d'un lexique externe augmente-t-elle de façon significative la précision de l'étiqueteur ? Avec quelle taille de corpus d'entraînement obtient-on le meilleur étiqueteur ? Quelle est sa précision ? À partir de quelle taille de corpus d'entraînement l'étiqueteur obtenu peut-il être utilisé comme pré-annotateur dans une campagne d'annotation manuelle qui consiste en la correction manuelle de l'annotation automatique ? Les deux systèmes d'étiquetage, MElt et TreeTagger, conduisent-ils à des résultats similaires concernant les questions précédentes ?

Nous avons donc procédé à l'entraînement de MElt et de TreeTagger sur 10 sous-corpus successifs du corpus de référence C_{ref}, dont la taille croît de 2 000 à 20 000 tokens. Nous avions préalablement mis de côté trois tranches de 500 tokens afin de servir d'échantillons de test. Pour rendre nos résultats comparables avec ceux présentés dans d'autres campagnes, l'évaluation de la précision s'est limitée aux seules étiquettes.

Pour l'entraînement de TreeTagger, nous avons utilisé comme lexique externe le lexique Morphalou 2.0 (Romary *et al.*, 2004). Nous avons dû convertir Morphalou au format attendu par TreeTagger puis le fusionner, pour chacun des 10 sous-corpus d'apprentissage, avec le lexique qui en est extrait. Nous avons également effectué des tests sans Morphalou, uniquement avec un lexique endogène. Pour l'entraînement de MElt, nous avons utilisé la version du lexique L*efff* (Sagot, 2010) utilisée pour l'entraînement de la version standard de l'étiqueteur MElt pour le français. Le lexique externe étant utilisé par MElt comme une source de traits pour le modèle d'étiquetage, nous avons pu conserver le jeu de catégories du lexique externe bien qu'il soit différent des catégories du corpus d'entraînement. Pour MElt, le lexique externe reste distinct du lexique extrait du corpus d'entraînement.

Premier constat : à l'exception du premier étiqueteur entraîné sur 2 000 tokens, les étiqueteurs obtenus avec MElt sont systématiquement meilleurs que ceux obtenus avec TreeTagger. Les meilleurs scores, obtenus à partir du corpus de 20 000 tokens, sont respectivement 96,9 % avec MElt et 94,9 % avec TreeTagger. Comparés aux 85–90 % annoncés par (Eshkol *et al.*, 2010) et aux 80 % obtenus par A. Dister (c.p. du 24 janvier 2008 [6]), nos résultats constituent donc une amélioration significative. Mais cela masque des variations d'un échantillon à un autre, ainsi qu'au niveau de la moyenne (voir figure 6), mais surtout des différences entre jeux d'étiquettes.

6. Diaporama disponible à l'adresse : http ://rhapsodie.risc.cnrs.fr/docs/Dister_Syntaxe_240108.pdf

Deuxième constat, sans surprise : l'utilisation du lexique externe améliore la précision de l'étiqueteur. Par exemple, sur le corpus de 2 000 tokens, TreeTagger n'atteint que 78,4 % de précision sans lexique externe contre 90,9 % avec Morphalou. Pour ce même corpus, la différence est moins importante avec MElt, mais elle est significative : la précision passe de 84,9 % à 88,1 %. Avec 20 000 tokens, l'utilisation du lexique externe permet à la précision de l'étiqueteur entraîné par MElt de passer de 95,5 % à 96,9 %. On note que MElt, sans lexique externe, donne des résultats supérieurs à TreeTagger avec lexique externe dès que le corpus d'entraînement fait plus de 12 000 tokens.

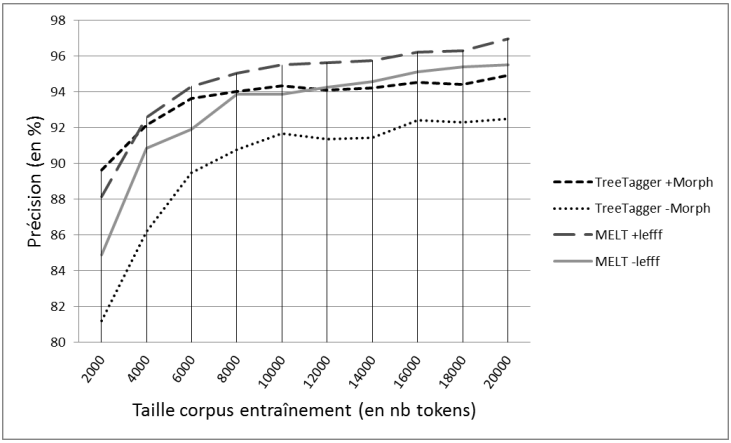

FIGURE 6 – Évolution de la précision de l'annotation automatique par tranche de 2 000 tokens

La figure 6 montre qu'à partir de 6 000 tokens, les résultats commencent à progresser de manière moins marquée, que ce soit avec MElt ou TreeTagger. Les précisions obtenues avec cette taille de corpus (94,3% avec MElt) sont suffisantes pour lancer une campagne de correction telle que nous la décrivons ci-dessus, après ré-entraînement. Il n'est pas indispensable que le corpus d'apprentissage soit plus volumineux. En tout cas, au vu de nos résultats, il n'est pas utile d'aller au-delà de 10 000 tokens si l'on utilise TreeTagger. L'utilisation de MElt semble toutefois préférable, avec des résultats qui continuent à croître jusqu'à 20 000 tokens.

6 Conclusion et perspectives

Le corpus TCOF-POS ($C_{\text{ref}} + C_{\text{add}}$) est disponible sur le site du CNRTL[7] sous licence Creative Commons BY-NC-SA 2.0[8], héritée du corpus TCOF. Il contient un peu plus de 100 000 tokens, dont un peu plus de 20 000 tokens de référence et 80 000 tokens obtenus grâce à la double-annotation (par les deux étudiantes recrutées) puis adjudication par un expert linguiste (C.

7. http ://cnrtl.fr/corpus/perceo/
8. http ://creativecommons.org/licenses/by-nc-sa/2.0/fr/

Benzitoun). Les meilleurs modèles d'étiquetage pour TreeTagger et pour MElt, qui ont une précision respectivement de 94,9 % et 96,9 % à ce stade de développement du corpus, seront également mis à disposition sous peu sur ce site (pour l'instant, seul le fichier paramètre pour TreeTagger est téléchargeable). Le lexique fusionné avec Morphalou est également disponible à cette même adresse. La version ré-entraînée de TreeTagger a déjà été utilisée par les concepteurs du Corpus de Français Parlé Parisien [9] pour annoter leurs données.

Dans le cadre d'une campagne de correction d'une pré-annotation automatique, nous avons mis en évidence le seuil de 6 000 tokens comme base de départ minimale pour ré-entraîner le logiciel. A ce stade, on obtient de bons résultats (94,3 % pour MElt et 93,6 % pour TreeTagger) et la précision progresse de manière moins marquée. Mais cette recommandation est valable lorsque le logiciel d'étiquetage est couplé à un lexique externe. Or, dans le cadre de notre campagne d'évaluation des corrections manuelles, nous n'avons pas utilisé de lexique externe pour ré-entraîner TreeTagger. Il faudra donc tester si les résultats sont de meilleure qualité lorsque l'on ajoute ce paramètre, travail que nous effectuons à l'heure actuelle.

Remerciements

Nous tenons à remercier les 12 étudiants ayant collaboré à ce projet et plus particulièrement M. Salcedo et M. Paquot, recrutées spécifiquement pour faire l'annotation. De même, L. Bérard, E. Jacquey, V. Meslard, S. Ollinger et E. Petitjean ont apporté une contribution significative à ce projet. Nous souhaitons également remercier l'ATILF pour son soutien financier dans le cadre d'un projet interne. La participation de K. Fort a été financée dans le cadre du programme Quæro [10], financé par OSEO, agence nationale de valorisation de la recherche. Celle de B. Sagot entre dans le cadre du projet ANR EDyLex (ANR-09-CORD-008).

Références

Abeillé, A. et Clément, L. (2006). *Annotation morpho-syntaxique. Les mots simples - Les mots composés Corpus Le Monde*.

André, V. et Canut, E. (2010). Mise à disposition de corpus oraux interactifs : le projet TCOF (traitement de corpus oraux en francais). *Pratiques*, 147/148:35–51.

Benzitoun, C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? *In Actes de la conférence RECITAL*, pages 13–22, Fès, Maroc.

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. et Jeanjean, C. (1987). *Le Francais parlé. Transcription et édition*. Didier Érudition, Paris, France.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A. et Mostefa, D. (2005). Semantic Annotation of the French Media Dialog Corpus. *In InterSpeech*, Lisbonne, Portugal.

9. http ://cfpp2000.univ-paris3.fr/search-transcription-tt/
10. `http://www.quaero.org`

BRANCA-ROSOFF, S., FLEURY, S., LEFEUVRE, F. et PIRES, M. (2010). Discours sur la ville. corpus de francais parlé parisien des années 2000 (CFPP2000). Rapport technique.

CAMPIONE, E., VÉRONIS, J. et DEULOFEU, J. (2005). *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*, édité par E. Cresti et M. Moneglia, chapitre 3. The French corpus, pages 111–133. John Benjamins, Amsterdam, Hollande.

COHEN, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC 2009*, Hong Kong, Chine.

DENIS, P. et SAGOT, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*. À paraître.

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, Belgique.

ESHKOL, I., TELLIER, I., TAALAB, S. et BILLOT, S. (2010). étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. *In 10th International Conference on statistical analysis of textual data (JADT 2010)*, Rome, Italie.

FORT, K. et SAGOT, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. *In Proc. of the Fourth ACL Linguistic Annotation Workshop*, Uppsala, Suède.

HUET, S., GRAVIER, G. et SÉBILLOT, P. (2006). Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique. *In Actes des 26èmes Journées d'Études sur la Parole (JEP)*, Dinard, France.

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

MERTENS, P. (2002). Les corpus de francais parlé ELICOP : consultation et exploitation. *In* BINON, J., DESMET, P., ELEN, J., MERTENS, P. et SERCU, L., éditeurs : *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*, pages 101–116. Universitaire Pers, Leuven, Belgique.

ROMARY, L., SALMON-ALT, S. et FRANCOPOULO, G. (2004). Standards going concrete : from LMF to Morphalou. *In Workshop on Electronic DictionariesWorkshop on Electronic Dictionaries, Coling 2004*, Genève, Suisse.

SAGOT, B. (2010). The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for french. *In 7th international conference on Language Resources and Evaluation (LREC 2010)*, La Vallette, Malte.

SCHMID, H. (1997). *New Methods in Language Processing, Studies in Computational Linguistics*, édité par D. Jones et H. Somers, chapitre Probabilistic part-of-speech tagging using decision trees, pages 154–164. UCL Press, Londres.

VALLI, A. et VÉRONIS, J. (1999). étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue Francaise de Linguistique Appliquée*, IV(2):113–133.

VOORMANN, H. et GUT, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.