Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank

Thomas Gaillat

Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank

Thomas Gaillat[1]
(1) Université Paris-Diderot – CLILLAC-ARP (3967) & Université de Rennes 1
thomas.gaillat@univ-rennes1.fr

RÉSUMÉ

Cet article aborde la problématique de l'annotation automatique d'un corpus d'apprenants d'anglais. L'objectif est de montrer qu'il est possible d'utiliser un étiqueteur PoS pour annoter un corpus d'apprenants afin d'analyser les erreurs faites par les apprenants. Cependant, pour permettre une analyse suffisamment fine, des étiquettes fonctionnelles spécifiques aux phénomènes linguistiques à étudier sont insérées parmi celles de l'étiqueteur. Celui-ci est entraîné avec ce jeu d'étiquettes étendu sur un corpus de natifs avant d'être appliqué sur le corpus d'apprenants. Dans cette expérience, on s'intéresse aux usages erronés de *this* et *that* par les apprenants. On montre comment l'ajout d'une couche fonctionnelle sous forme de nouvelles étiquettes pour ces deux formes, permet de discriminer des usages variables chez les natifs et non-natifs et, partant, d'identifier des schémas incorrects d'utilisation. Les étiquettes fonctionnelles éclairent sur le fonctionnement discursif.

ABSTRACT

Automatic tagging of a learner corpus of English with a modified version of the Penn Treebank tagset

This article covers the issue of automatic annotation of a learner corpus of English. The objective is to show that it is possible to PoS-tag the corpus with a tagger to prepare the ground for learner error analysis. However, in order to have a fine-grain analysis, some functional tags for the study of specific linguistic points are inserted within the tagger's tagset. This tagger is trained on a native-English corpus with an extended tagset and the tagging is done on the learner corpus. This experiment focuses on the incorrect use of *this* and *that* by learners. We show how the insertion of a functional layer by way of new tags for the forms allows us to discriminate varying uses among natives and non-natives. This opens the path to the identification of incorrect patterns of use. The functional tags cast a light on the way the discourse functions.

MOTS-CLÉS : Apprentissage L2, corpus d'apprenants, analyse linguistique d'erreurs, étiquetage automatique, *this, that*

KEYWORDS : Second Language Acquisition, learner corpus, linguistic error analysis, automated tagging, *this, that*

1 Introduction

Le travail présenté ici se rapporte au domaine de l'acquisition d'une seconde langue, en l'occurrence de l'anglais. Il se fonde sur l'étude de corpus d'apprenants qui sont devenus des outils indispensables pour l'analyse des productions dans le domaine de l'acquisition d'une seconde langue (Dagneaux, Denness & Granger, 1998). Leur utilisation permet, par exemple, d'analyser les erreurs commises afin de proposer des stratégies de remédiations mises en œuvres dans des didacticiels. La constitution de corpus d'apprenants rencontre un défi de plus par rapport aux corpus de natifs du fait qu'elle s'accompagne de la mise en place d'un système d'annotation dans lequel les erreurs commises par les apprenants sont caractérisées. Dans le domaine de la morphosyntaxe, une première voie empruntée a consisté à annoter les erreurs manuellement (Granger, 1993)(Diaz-Negrillo & Fernandez-Domingez, 2006). Si ce travail offre une richesse pour la description des erreurs, il n'en reste pas moins fastidieux et coûteux. Par ailleurs, ces méthodes mélangent la caractérisation des erreurs et la description des catégories grammaticales dans une même annotation de type partie du discours (PoS), car les erreurs d'apprenants portent sur un mot dont le mauvais usage est décrit en fonction de leur position grammaticale. La deuxième voie consiste à automatiser le processus d'annotation, en distinguant les types d'annotation. L'annotation PoS sur corpus d'apprenants a fait l'objet d'expérimentations multiples (Van Rooy & Schafer, 2003)(De Haan, 2000) dont l'une des solutions a consisté à post-éditer certaines étiquettes afin d'améliorer la précision globale des étiquetages, ou afin d'y intégrer des informations relatives aux erreurs d'apprenants. Une autre approche proposée par (Diaz-Negrillo, Meurers, Valera & Wunsch, 2010) consiste à prendre en compte l'erreur dans un second temps seulement. Les auteurs développent le concept d'annotation PoS tripartite selon l'idée que la distribution, la morphologie et le lexique constituent le socle à partir duquel les erreurs peuvent être systématiquement déduites. Selon eux, il faut élaborer une annotation triple reprenant chaque catégorie sans y adjoindre d'interprétation. Nous nous inscrivons dans cette perspective en considérant que l'étiquetage automatique PoS effectué à partir d'un jeu d'étiquettes d'anglais natif permet d'apporter un premier niveau d'information concernant la distribution effective des mots reflétant un usage erroné ou non. Grâce à ces informations distributionnelles, le travail d'analyse d'erreurs peut être abordé puisque les questions de compatibilité syntaxique entre les constituants de la phrase sont au centre d'une grande part des erreurs d'apprenants. Il s'agit donc d'annoter les textes, y compris les erreurs, du point de vue de leur distribution pour permettre ensuite une caractérisation fine de ces erreurs.

L'objectif de cet article est de montrer qu'il est possible d'utiliser un étiqueteur (PoS) pour annoter automatiquement un corpus d'apprenants. Cependant, pour que l'analyse des erreurs faites par les apprenants soit suffisamment fines, des étiquettes spécifiques aux phénomènes linguistiques à étudier doivent être insérées parmi les étiquettes de l'étiqueteur (ce qui distingue cette expérience de celle de Van Rooy *et al)* et celui-ci doit être entraîné avec ce jeu d'étiquettes étendu sur un corpus natif avant d'être appliqué sur le corpus d'apprenants. Cette approche est développée dans cet article en abordant la problématique des usages de *this* et *that*. En effet, une étude antérieure (Gaillat, 2013) a montré que les apprenants éprouvent des difficultés concernant l'usage des démonstratifs. Les marqueurs avec lesquels *this* et *that* sont en concurrence dans les erreurs ne sont pas les mêmes selon leur réalisation fonctionnelle : *the* pour les emplois en déictiques, *it* pour

les emplois pronominaux. Afin de pouvoir effectuer un relevé précis et exhaustif de toutes ces formes, il est nécessaire de s'appuyer sur l'annotation PoS du corpus pour extraire les *this* et *that* selon les différentes catégories grammaticales auxquelles ils appartiennent. Ainsi, les *this* et *that* en usage pro-forme, par exemple, peuvent être isolés et mis en regard avec les pronoms *it*. Notre étude montre que l'étiquetage permet de mettre à jour des schémas incorrects d'utilisation de *this* et *that* par les apprenants. À partir de là, un travail d'analyse des erreurs peut être effectué.

L'article est organisé de la façon suivante. La section 2 aborde la méthodologie suivie pour mener notre expérience à partir de deux corpus. Après les avoir décrits, nous abordons la problématique du jeu d'étiquettes et sa modification en utilisant TreeTagger (Schmid, 1994). La section 3 présente l'annotation automatique faite avec TreeTagger. Les résultats obtenus permettent une analyse de la qualité de l'étiquetage mais aussi une analyse d'erreurs d'apprenants.

2 Méthodologie

Dans cette section, la nature des deux corpus utilisés dans cette étude est décrite. Ensuite, le problème des étiquettes utilisées par TreeTagger pour *this* et *that* est abordé et leur modification est décrite.

2.1 Les corpus

Le corpus utilisé pour la phase d'apprentissage de TreeTagger est le Penn Treebank (Marcus *et al*, 1993). Il s'agit d'un corpus d'anglais natif composé de 4,5 millions de mots en anglais américain et correspondant aux articles parus dans le *Wall Street Journal*. Ce corpus a fait l'objet d'une annotation syntaxique, c'est-à-dire un système d'arbre décrivant les dépendances syntaxiques, et d'une annotation PoS. Le jeu d'étiquettes PoS, qui comporte 36 étiquettes PoS et 12 pour la ponctuation et les symboles monétaires, a été appliqué automatiquement sur le corpus avant qu'une procédure de correction manuelle ne soit mise en œuvre pour permettre à des annotateurs humains de modifier les étiquettes erronées. Dans le cadre de notre étude, deux échantillons sont formés à partir du corpus. Le premier constitue un échantillon servant à la phase d'apprentissage qui est décrite ultérieurement dans cet article. Il comprend 1 824 168 mots et étiquettes. Le second échantillon en contient 63 092 et est utilisé pour tester la qualité de l'étiquetage. Le type de production (écrite) diffère du corpus d'apprenants (oral), mais le Penn Treebank est utilisé en raison de sa grande fiabilité (*Gold standard*).

L'échantillon utilisé pour la phase d'annotation du corpus d'apprenants provient du corpus Diderot-LONGDALE[1] et plus spécifiquement de la partie orale constituée à l'université de Paris-Diderot sous le nom de Charliphonia. Le Diderot-LONGDALE est un corpus oral d'apprenants d'anglais. Des étudiants des niveaux L1 à L3 ont été suivis sur trois années, et ont participé à des entretiens libres avec des lecteurs natifs. Les enregistrements audio recueillis ont été retranscrits par des étudiants d'anglais de niveau M2 pour ensuite être

1 *Cf.* http://www.uclouvain.be/en-cecl-longdale.html

vérifiés par des enseignants-chercheurs. Chaque enregistrement correspond à des questions portant sur des expériences personnelles et l'étudiant est invité à y répondre librement. Cet échantillon est composé de 3 243 mots ou ponctuations et d'autant d'étiquettes vérifiées manuellement.

2.2 Les problèmes des étiquettes de TreeTagger pour *this* et *that*

Les erreurs d'apprenants concernant l'usage de *this* et *that* se caractérisent par le fait qu'elles se situent sur l'axe paradigmatique plutôt que sur l'axe syntagmatique. Les difficultés ne sont en effet pas dues à des problèmes de position syntaxique mais plutôt à des mécanismes de substitution entre des formes de fonction syntaxique identique. Il existe deux branches principales de substitutions. La première touche au système déictique et au type de référence endophorique ou exophorique[2]. Deux groupes d'erreurs se distinguent. Un certain nombre d'erreurs dénote des confusions entre l'un et l'autre des types de référence. Les apprenants utilisent une forme exophorique dans un contexte endophorique. Cela se traduit par des substitutions de l'une des formes par l'autre. L'autre groupe d'erreurs se situe au sein même des processus de référence endophorique. Dans ce cas, c'est la valeur référentielle de la forme qui est mal maîtrisée et l'apprenant opère aussi une substitution entre les deux formes.

La deuxième branche de substitutions concerne des interactions du système déictique avec les deux micro-systèmes pronominal et déterminatif. *This* et *that* peuvent endosser deux fonctions syntaxiques dans la phrase en anglais. On peut les retrouver devant un syntagme nominal ou en position de syntagme nominal. Quirk *et al* (1985) les distinguent en usage déterminant ou nominal. Pour notre part, nous reprenons les termes de déterminant et de pro-forme (Lapaire et Rotgé, 1998 : 50-51) qui traduit une référence sémantique plus étendue qu'un groupe nominal isolé. Pour les apprenants, les interactions se traduisent par des difficultés de choix entre un *this / that* et un *it* ou *the*, selon la fonction syntaxique. Les erreurs liées à la fonction déterminative renvoient au statut morphosyntaxique des éléments et leur position dans la chaîne syntagmatique. Les erreurs liées à la fonction pro-forme se caractérisent du point de vue sémantique des référents qui sont mal identifiés par les apprenants. Cette distinction entre les deux types d'erreurs repose donc sur une distinction fonctionnelle qui doit apparaître dans l'étiquetage. Afin de comprendre la manière dont les étiquettes propres à *this* et *that* sont traitées par l'étiqueteur automatique TreeTagger, il est utile de s'appuyer sur les exemples suivants :

> (1) <A> would you consider pizza an Italian food </A><B> (em) yes but it's not it's not really f= it's typic but it's not (em) we can eat *that* everyday everywhere now and . but (em) my grandma does *this* by herself.
>
> (2) I don't know between New Zealand and (er) Latin America I can't choose so I think I will do both <laughs> (em) because it's so different from (er) from France (er) . I would like to discover the= all these new cultures and (er) if I had *this*

2 La référence est endophorique quand le référent se trouve dans le discours du locuteur. Elle est exophorique quand il se trouve dans la situation dans laquelle se trouve le locuteur physiquement.

opportunity to visit these countries I would live in a very typical family of the country.

En (1), le locuteur natif (marqué <A>) pose une question sur l'entité *pizza* à l'apprenant (marqué <B>). Celle-ci produit une réponse dans laquelle elle fait référence plusieurs fois à cette entité, y compris avec les pro-formes *this* et *that*. Dans les deux cas, il s'agit respectivement d'usages inattendus et erronés, non pas du point de vue de leur distribution, mais du point de vue de leur valeur référentielle. En (2) le locuteur commet une autre erreur de substitution mais elle diffère de (1) par l'élément avec lequel le *this* se substitue, c'est-à-dire *the*. Des vérifications auprès de natifs anglophones montrent que l'usage du pronom *it* aurait été privilégié en (1) et le déterminant *the* en (2). Ce qui différencie ces erreurs provient de la catégorie grammaticale des formes : soit elles jouent le rôle de déterminant en tête de syntagme nominal, soit elles jouent le rôle de pro-forme en remplaçant un syntagme nominal. Pour identifier ces types d'erreur, il faut pouvoir isoler les usages pro-formes des usages déterminants. Or, le jeu d'étiquettes de TreeTagger attribue une seule étiquette dans les deux cas. Concrètement, dans les deux exemples, l'étiquette DT (qui renvoie à déterminant) est attribuée à chaque occurrence, ce qui rend impossible une requête ultérieure pour extraire les cas de *this* en pro-forme uniquement. L'analyse de l'erreur de substitution avec *it* n'est donc pas possible. Il en va de même pour les substitutions avec *the*.

2.3 Modification des étiquettes PoS de *this* et *that* dans le Penn Treebank

Pour étiqueter, le logiciel TreeTagger nécessite un fichier formaté spécifiquement, produit lors de la phase d'apprentissage, qui se nourrit d'un corpus correctement étiqueté. L'apprentissage implique donc l'usage d'un corpus de natifs – le Penn Treebank - incluant les étiquettes modifiées. Il convient par conséquent de procéder à l'identification de toutes les occurrences de *this* et *that* dans le corpus de natifs et de les modifier. On pourra ensuite procéder au formatage des données décrit en 2.4. La première tâche consiste à repérer l'ensemble des formes selon la position syntaxique qu'elles occupent. Pour ce faire, l'outil Tregex (Levy & Andrew, 2006) est utilisé. Grâce à des expressions régulières, celui-ci permet de visualiser les arbres syntaxiques composant le corpus et d'effectuer des requêtes sur les arbres syntaxiques. Cela permet de combiner des contraintes constituées de mots et d'élément syntaxiques tels que les syntagmes verbaux ou les propositions subordonnées. De cette manière il est possible d'explorer les arbres syntaxiques.

Un inventaire des formes *this* et *that* est donc effectué. Pour ce qui concerne la distinction pro-forme / déterminant des formes, il s'agit d'abord d'identifier toutes les formes de *that* et *this* étiquetées DT. La requête suivante, exprimée en expression régulière :

/^DT$/ < that > /^NP.*/

permet d'identifier les formes *that* étiquetées DT en partie du discours et faisant partie d'un syntagme nominal. Cependant, en procédant par recoupements entre les calculs à partir des étiquettes et à partir des arbres syntaxiques, des erreurs sont trouvées. Par

exemple la requête : /^IN.*/ < that > /^NP.*/[3]

correspond aux *that* étiquetés IN et dominés directement par un syntagme nominal. Une illustration en contexte donne l'exemple suivant :

> wsj_0277.mrg-26 The Mitsubishi family company acquired that property from the government some 100 years ago [...]

Dans ce type de cas, le *that* ne peut être à la fois subordonnant et jouer le rôle de déterminant pour le nom *property*. Cela révèle donc une erreur d'étiquetage. Cependant, du fait de la position de *that* juste après un verbe, on comprend que cette configuration ait mené à l'erreur car, dans bien des cas, le verbe suivi d'un *that* introduit une subordonnée complétive. En procédant de la sorte et en diversifiant les requêtes pour identifier toutes les fonctions possibles de *that*, 871 erreurs d'étiquetage de *that* sont trouvées.

Du fait de ces erreurs, deux tâches de modification d'étiquettes sont nécessaires. Tout d'abord, il convient de corriger les erreurs d'étiquetage PoS dans le Penn Treebank. Ensuite, il s'agit de modifier les étiquettes DT en permettant la distinction déterminant et pro-forme. Dans les deux cas, l'utilisation du module TSurgeon du logiciel Tregex permet d'effectuer les changements nécessaires.

3 Annotation automatique avec TreeTagger et résultats

Cette section se focalise sur l'annotation automatique du logiciel TreeTagger (Schmid, 1994). La méthode d'annotation et le fonctionnement de TreeTagger sont présentés. Ensuite, la phase d'apprentissage, sur un corpus de natifs et avec un jeu d'étiquettes modifié, est passée en revue. Le processus d'annotation est détaillé et les résultats obtenus sont analysés du point de vue des performances dans un premier temps. Dans un second temps, l'analyse considère, d'un point de vue qualitatif, l'apport de l'annotation et des étiquettes modifiées, dans le cadre d'une analyse linguistique d'erreurs portant sur *this* et *that*.

3.1 Méthode d'annotation, fonctionnement et apprentissage

TreeTagger est un outil d'annotation basé sur une méthode probabiliste. À partir de la représentation d'un arbre décisionnel binaire, le programme estime la probabilité de trigrammes PoS. Par exemple, la probabilité d'une pro-forme, précédée d'un verbe et d'un pronom, est calculée et stockée en mémoire, et constitue une branche de l'arbre (PP, VB, TPRON). L'arbre de décision permet la classification de toutes les instances de PoS dans différentes branches lorsqu'elles sont présentées au programme dans sa phase d'apprentissage. Lors de la phase d'annotation, TreeTagger s'appuie sur l'arbre construit et

3 IN correspond à une préposition ou une conjonction de subordination et NP correspond à un syntagme nominal. Le symbole > signifie que l'argument de gauche est dominé dans l'arborescence par l'argument de droite. Le symbole < signifie l'inverse.

la probabilité d'un trigramme donné pour attribuer une étiquette à chaque mot. L'évaluation des performances de l'annotateur se fait par calcul de la précision globale (« accuracy ») pour l'ensemble des étiquettes, et du rappel et de la précision pour des étiquettes spécifiques.

Le fonctionnement de TreeTagger se fait en deux temps : un apprentissage sur un corpus étiqueté puis le processus d'attribution des étiquettes sur un corpus vierge. Il est important de noter que TreeTagger est en fait constitué de deux programmes : *train-tree-tagger* et *tree-tagger*. Le travail décrit en section2a pour objectif de préparer les données d'apprentissage traitées par le module *train-tree-tagger*. Après la modification des étiquettes, la préparation du fichier d'apprentissage nécessaire à ce module peut être opérée afin d'extraire les paires de mots ou ponctuation et d'étiquettes PoS. L'extraction des paires étiquettes / mots s'effectue depuis les fichiers bruts contenant les structures syntaxiques sous forme de jeu de parenthèses, les PoS et les mots / ponctuation et les paires sont placées sur des lignes uniques d'un fichier texte. L'illustration 1 schématise ce processus en montrant un extrait du Penn Treebank et le résultat obtenu, c'est-à-dire le fichier utilisé par le module *train-tree-tagger*.

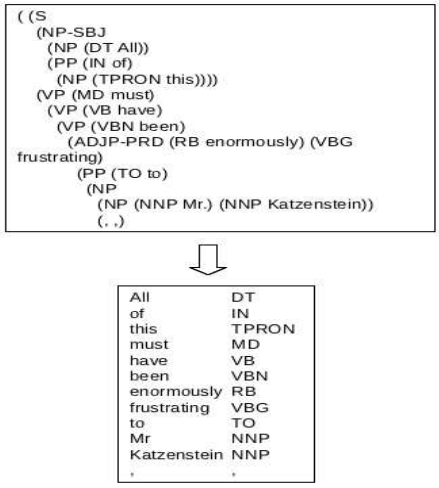

FIGURE 1 - Extraction des paires de mot / PoS depuis les arbres syntaxiques du Penn Treebank

Lors de son exécution[4], le module construit un fichier de paramétrages (*output file*) qui sera ensuite utilisé par le module d'annotation. Sa construction se fait sur la base de plusieurs fichiers dont l'un correspond au lexique extrait du corpus d'apprentissage. Celui-ci inclut notamment, les entrées correspondant à *this* et *that* et présentées ci-dessous au format attendu par TreeTagger :

that DT that TPRON that TCOM that RB that TREL that
this DT this TPRON this RB this

4 En respectant la syntaxe suivante : train-tree-tagger {options} <lexicon> <open class file> <input file> <output file>

Bien que l'article se focalise sur les pro-formes et déterminants, on voit que d'autres usages sont aussi distingués par l'étiquetage. Le jeu d'étiquettes d'origine du Penn Treebank distingue plusieurs catégories. DT et RB peuvent correspondre à des occurrences de *this* et *that,* l'étiquette DT pouvant être assignée à des pro-formes ou déterminants sans distinction. A ces étiquettes s'ajoutent WDT pour les déterminants en WH et IN pour les conjonctions de subordination et les prépositions. Le nouveau jeu d'étiquettes adopté vise à clarifier les réalisation fonctionnelles possibles de *this* et *that.* On peut retrouver les deux formes dans la composition d'un syntagme nominal comme exprimé en 2.2. Pour ce qui concerne la forme *that,* Biber *et al* (1999 : 85 ; 195) pointent les deux autres fonctions possibles dans la construction de l'hypotaxe en tant que pronom relatif et en tant que complétif. Les entrées *this* et *that* du lexique reflètent donc la modification du jeu d'étiquettes apportée ce qui donne DT pour les déterminants, TPRON pour les pro-formes, TCOM pour les complétifs, RB pour les adverbiaux et TREL pour les pronoms relatifs.

Au final, le fichier de paramétrage inclut une série d'informations telles que l'arbre décisionnel, les étiquettes possibles pour chaque mot du corpus et l'échantillon d'apprentissage corrigé et étiqueté tel que décrit en 2.3.

3.2 Processus d'annotation

Le processus d'annotation se fait avec le module *tree-tagger*. L'exécution du programme nécessite les arguments à indiquer dans l'ordre suivant : le fichier de paramétrage créé lors de l'exécution du programme *train-tree-tagger*, le fichier correspondant au corpus non étiqueté, et le nom de fichier de ce même corpus une fois étiqueté. Dans le cadre de cette étude, deux annotations sont lancées. Une première consiste à reproduire l'expérience de Schmid en annotant l'échantillon test du Penn Treebank mais avec les nouvelles étiquettes propres à *this* et *that.* La seconde est appliquée à un échantillon test du corpus Diderot-LONGDALE décrit en 2.1. Dans les deux cas d'annotation le même fichier d'apprentissage, créé à partir du Penn Treebank, est utilisé.

L'objectif de la première annotation est de pouvoir comparer la qualité de l'étiquetage et de la mettre en regard avec les résultats obtenus et décrits par Helmut Schmid. L'échantillon test du Penn Treebank, comprenant des étiquettes modifiées, permet en outre de vérifier la qualité de la prise en charge des nouvelles étiquettes introduites pour la distinction des *this* et *that.* L'objectif de la seconde annotation est de pouvoir évaluer sa qualité pour ce qui concerne l'ensemble des étiquettes, mais aussi de vérifier si les nouvelles étiquettes introduites sont correctement gérées par TreeTagger sur de l'anglais non-natif.

3.3 Résultats

Les résultats sont exprimés à partir des calculs de précision globale (« *accuracy* » en anglais) et de rappel pour ce qui concerne les *this* et *that.* Pour ce qui concerne l'annotation de l'échantillon test du Penn Treebank, la précision globale est de 95,79 %. Sur 31 546 étiquettes, 30 220 ont été correctement attribuées. Ce résultat est tout à fait

similaire aux 96 % rapportés dans l'expérience de Schmid. Les modifications d'étiquettes ne semblent donc pas avoir eu d'impact global sur la qualité de l'étiquetage. Si on prend les *this* dans leur globalité, c'est-à-dire toutes étiquettes confondues attribuées à la forme, la précision globale est de 92,10 %. Les *that* sont correctement étiquetés dans 84 % des cas. La baisse de performance concernant les *that* provient certainement de la variété plus grande des étiquettes qu'il peut recevoir, ce qui introduit plus d'incertitude dans le processus décisionnel de TreeTagger.

Si les observations globales informent sur la qualité de l'étiquetage des mots, il convient d'étudier plus en détail la situation étiquette par étiquette, pour chacune des formes, afin d'explorer la qualité de traitement des catégories grammaticales. Le comportement de TreeTagger pour les étiquettes DT et TPRON est d'autant plus intéressant qu'elles permettent une distinction nouvelle dans le Penn Treebank (*cf.* Tableau 1).

	Rappel %	Précision %	F-Score %	Nombre d'occurrences vraies attendues
This DT	100	91,04	95,31	61
This TPRON	60	100	75	15
That DT	75	78,94	76,94	20
That TPRON	55	88,23	68,18	27

TABLEAU 1 - Résultats de l'étiquetage des *this* et *that* avec les étiquettes déterminant et pro-forme pour le Penn Treebank.

En mettant en regard les différentes étiquettes, on s'aperçoit que les rappels DT sont systématiquement plus élevés que les rappels TPRON au sein d'une forme donnée. Lors du processus d'annotation, TreeTagger manque donc moins d'étiquettes DT que de TPRON. Ceci est peut-être dû au fait que DT est une étiquette fonctionnelle qui se laisse caractériser en trigramme car elle est positionnelle (pré-nominale). À l'inverse, les résultats en précision révèlent des valeurs TPRON systématiquement supérieures à DT pour une forme donnée. Cela traduit le fait que TreeTagger commet moins d'erreurs d'étiquetage lorsqu'une étiquette TPRON est attribuée. L'étiquette TPRON n'est pas uniquement tributaire de la position du mot, elle se caractérise aussi de manière plus sémantique ce qui rend sa détection plus aléatoire. Mais une fois détectée, ceci est fait avec plus d'assurance. Le détails des erreurs d'étiquetage par étiquette apparaît dans les deux matrices de confusion (*cf.* Tableaux 2 et 3).

this	DT	TPRON	RB
Tagged DT	61	6	0
Tagged TPRON	0	9	0
Tagged RB	0	0	0

TABLEAU 2 - Matrice de confusion pour *this* dans le Penn Treebank.

that	DT	TPRON	TCOM	TREL	RB
Tagged DT	15	0	3	1	0
Tagged TPRON	0	15	1	1	0
Tagged TCOM	5	11	142	8	0
Tagged TREL	0	1	13	59	0
Tagged RB	0	0	0	0	0

TABLEAU 3 - Matrice de confusion pour *that* dans le Penn Treebank.

Pour ce qui concerne l'annotation de l'échantillon du corpus d'apprenants, on obtient une précision globale de 91 %. Ce chiffre est à comparer avec les 96 % obtenus lors de l'annotation du corpus de natifs Penn Treebank. La différence semble donc traduire une difficulté de TreeTagger à gérer certaines étiquettes. Cette difficulté peut trouver deux origines. D'une part, ce corpus contient des erreurs commises par les apprenants. Ces erreurs constituent des configurations syntaxiques qui ne sont pas répertoriées dans l'arbre décisionnel créé lors de la phase d'apprentissage sur le corpus de natifs. D'autre part, il s'agit d'un corpus oral caractérisé par un grand nombre d'hésitations, de répétitions, et de phrases non terminées. Là encore, cela se traduit par des configurations syntaxiques non traitées lors de l'apprentissage. Lorsque TreeTagger parcourt le corpus à annoter, il rencontre donc des suites syntaxiques non apprises. Il a alors recours à des stratégies par défaut qui ne suffisent pas à permettre la sélection de l'étiquette correcte dans tous les cas.

Comme pour le Penn Treebank, il est intéressant d'observer l'étiquetage des *this* et *that* toutes étiquettes confondues. Pour ce qui concerne les *this,* les 22 occurrences de vrais positifs sont toutes étiquetées d'une des étiquettes possibles, et seulement 17 le sont correctement. Cela donne un rappel et une précision de 77,27 % puisque tous les *this* sont étiquetés avec une des étiquettes possibles. Pour ce qui concerne les *that,* le rappel est de 51,02 % et la précision de 50 %. Ces résultats montrent donc une différence de qualité d'annotation entre les deux corpus. Là encore, les caractéristiques oral et non-natif du corpus d'apprenants peuvent expliquer cette différence.

Afin d'affiner l'observation de ces résultats, les calculs peuvent être effectués en prenant chaque étiquette DT ou TPRON pour chaque forme (*cf*. Tableau 4). Les résultats sont mitigés. Pour ce qui concerne les *this*, les valeurs obtenues pour l'étiquette DT sont du même ordre que celles du Penn Treebank même si elles sont plus faibles (93,75% en rappel et 78,94% en précision). Les matrices de confusion renseignent sur les confusions de Treetagger (*cf.* Tableaux 4 et 5) et la distinction du *this* TPRON est problématique. Le traitement de *that* avec les étiquettes DT et TPRON est complètement erroné. Cependant, il faut noter que l'échantillon utilisé et préparé manuellement ne contient que peu d'occurrences des formes avec les étiquettes recherchées. Ce faible nombre peut donc donner une représentation extrême qui ne reflète pas nécessairement la réalité. Pour pouvoir tirer des conclusions sur l'attribution de ces étiquettes, il conviendrait d'accroître la taille de l'échantillon afin d'obtenir un plus grand nombre d'occurrences de chaque étiquette.

this	DT	TPRON	RB
Tagged DT	15	4	0
Tagged TPRON	1	2	0
Tagged RB	0	0	0

TABLEAU 4 - Matrice de confusion pour *this* dans le corpus Diderot-Longdale.

that	DT	TPRON	TCOM	TREL	RB
Tagged DT	0	1	0	0	0
Tagged TPRON	0	0	0	0	0
Tagged TCOM	2	9	21	1	3
Tagged TREL	0	1	7	4	0
Tagged RB	0	0	0	0	0

TABLEAU 5 - Matrice de confusion pour *that* dans le corpus Diderot-Longdale

3.4 Analyse d'erreurs d'apprenants avec les étiquettes modifiées

Si du point de vue des performances de traitement des étiquettes modifiées, la qualité de l'étiquetage est faible pour le Longdale, les matrices de confusion révèlent néanmoins les catégories grammaticales qui génèrent des erreurs d'étiquetage. Si ces erreurs ne permettent pas de diagnostiquer les erreurs d'apprenants, leur présence peut servir d'indice pour la signalisation de certains types d'erreur. L'apprentissage s'étant fait sur de l'anglais natif, on peut émettre l'hypothèse que les différences syntaxiques propres aux apprenants sont mal traitées par TreeTagger. C'est en ceci que les erreurs d'étiquetage peuvent être le révélateur de fonctions syntaxiques utilisées par les apprenants, qui diffèrent de l'anglais natif. En effet, si par exemple TreeTagger confond des étiquettes *that* TPRON en les étiquetant TCOM, cela pourrait signifier que des configurations syntaxiques d'apprenants dans lesquelles se trouve le *that* TPRON s'approchent de celles du *that* TCOM. L'apprentissage de TreeTagger ayant eu lieu sur de l'anglais natif, certains usages de *that* en pro-forme par les apprenants pourraient donc ressembler à des usages en complétif (verbe suivi de *that*) tel que l'exemple suivant extrait du corpus COCA[5] l'illustre : « It happens *that* the Constitution didn't create the dollar. » Un retour sur les données permet d'extraire une occurrence de *that* pro-forme étiquetée TCOM après le verbe *happen (cf.* exemple ci-dessous). Ce type d'enchaînement correspond souvent à un usage complétif en anglais natif.

> DID0199-S002 - I read it so many times that even if I don't start from the beginning I know where I am and I say <begin laughter> oh yes er er er <end laughter> I remember these times I remember yeah it will happen *that* [TCOM] and *that* [TCOM] so yes I will read it again and again

Dans cet exemple, l'apprenant ne souhaite pas utiliser *that* pour introduire une

5 *Corpus of Contemporary American English* de l'université de Brigham Young (USA).

complétive. Il s'agit en fait d'un transfert de la L1 avec une transposition de l'expression en français : « il arrivera ça et ça ». L'erreur de l'apprenant se situe au niveau du sémantisme de *happen* qui ne peut être utilisé comme le verbe *arriver* en français. Là où l'anglais place l'agent en position de sujet, le verbe *arriver* peut prendre l'agent en position de complément. Cette méconnaissance de la part de l'apprenant le pousse à recourir à l'usage d'une pro-forme pour faire référence à l'agent. Ce schéma ne se retrouve pas en anglais et par conséquent TreeTagger n'a pas rencontré cette possibilité lors de son apprentissage, d'où l'erreur d'étiquetage. Ainsi, l'erreur d'étiquetage ne permet pas de diagnostiquer l'erreur, elle permet de la signaler.

Du point de vue qualitatif, TreeTagger et son jeu d'étiquettes modifié, doivent permettre le traitement d'un certain nombre de segments tels que les exemples (1) et (2) de la section2.2avec l'attribution de deux étiquettes distinctes. En (1), le *that* et le *this* reçoivent l'étiquette TPRON. Il est alors possible d'extraire toutes les occurrences de ce type à des fins d'analyse. En (1), le *that* étiqueté TPRON reprend l'entité *pizza* en ajoutant une valeur de distanciation et de monstration peu probables dans ce contexte. D'autre part, le *this*, s'il permet une reprise de l'entité, introduit l'idée d'une information nouvelle le concernant, ce qui n'est pas le cas ici. Cet étiquetage permet donc d'explorer le corpus afin d'identifier les erreurs signalant que le processus de référence par substitution au syntagme nominal est mal assuré par les apprenants.

L'exemple (2) bénéficie aussi de l'introduction de la distinction puisque seuls les cas avérés d'utilisation en déterminant du *this* sont étiquetés DT. L'extraction des occurrences du type *this* déterminant permet une analyse ciblée des erreurs s'y rapportant. En (2), avec la détermination nominale *this opportunity,* le locuteur fait référence à l'entité discursive *visite de la Nouvelle Zélande ou Amérique Latine.* Dans ce cas, l'article défini *the* suffit, alors que *this,* en usage déterminant, ajoute une notion de monstration qui ne pourrait fonctionner qu'en cas de reprise de l'entité. Or, celle-ci est en cours de construction, ce qui rend le *this* caduque. En (1) et (2), les erreurs sont du même type et se caractérisent par une substitution sur l'axe paradigmatique. L'introduction de la distinction fonctionnelle d'étiquetage DT ou TPRON permet donc l'extraction ciblée des formes potentiellement erronées et permet un travail d'analyse d'erreurs ciblé. Cela montre l'intérêt heuristique d'une annotation fonctionnelle dans le cas où des micro-systèmes d'erreurs sont explicables sur des bases fonctionnelles.

4 Conclusion

Dans cette recherche abordant la problématique de l'annotation PoS sur un corpus d'apprenants d'anglais, nous avons posé la question de savoir s'il était possible d'étiqueter un corpus sur la base d'un corpus de natifs avec des étiquettes modifiées pour satisfaire à des besoins d'analyse d'erreurs. La méthodologie employée montre qu'il est possible d'utiliser le corpus d'anglais natif Penn Treebank pour modifier les étiquettes propres à *this* et *that* afin de servir de base à l'apprentissage de TreeTagger. Au passage, un certain nombre d'erreurs d'étiquetage des deux formes dans le Penn Treebank sont détectées et corrigées. L'apprentissage fonctionne puisque lors de la phase d'annotation, les données montrent que les deux étiquettes distinguant les usages pro-forme et déterminant des deux formes sont bien prises en compte. Sur le corpus Penn Treebank, TreeTagger attribue

les étiquettes modifiées avec une relative efficacité, notamment pour l'étiquette déterminant de *this*. L'étiquetage des *that* pose plus de problèmes du fait de la variété de ses usages syntaxiques tant au niveau de l'hypotaxe que de la détermination. Sur le corpus d'apprenants d'anglais, les résultats globaux montrent une certaine robustesse. Cependant une approche détaillée révèle que les étiquettes propres à *this* et *that* sont médiocrement prises en charge à l'exception du *this* déterminant. Afin de corroborer ou d'infirmer ces résultats, il conviendrait d'accroître la taille de l'échantillon du corpus d'apprenants. Du point de vue qualitatif, on peut dire que l'étiquetage rend possible l'analyse d'erreurs d'apprenants de deux manières. D'une part, la distinction d'étiquettes conduit vers un ciblage des occurrences pour l'analyse des erreurs pouvant s'y rapporter. D'autre part, les erreurs d'étiquetage permettent de mettre à jour des schémas incorrects d'utilisation de *this* et *that* par les apprenants, par rapport à des usages de natifs. L'étiquetage modifié permet donc de discriminer des usages variables chez les natifs et non-natifs.

L'étude montre donc que s'il est possible d'étiqueter le corpus d'apprenants d'anglais, il reste néanmoins à affiner la méthode d'apprentissage de manière à favoriser la prise en charge effective des étiquettes créées pour permettre l'analyse d'erreur. Lors de la phase d'apprentissage, il serait peut-être nécessaire de mixer le corpus en y intégrant des occurrences du corpus d'apprenants. Cela permettrait l'apprentissage de configurations syntaxiques propres aux apprenants et à la nature orale de ce corpus. Grâce à une amélioration de la distinction déterminant / pro-forme, il sera alors possible d'analyser plus en détails les difficultés éprouvées par les apprenants sur les questions de référence sous-jacentes à *this* et *that*. Le jeu d'étiquettes DT / TPRON balise les deux micro-systèmes d'erreurs possibles dans le champs des emplois référentiels de *this* et *that*. Dans le cadre d'une analyse automatique des erreurs d'apprenants, on voit l'intérêt d'un ré-étiquetage PoS plus fin qui distingue les réalisations fonctionnelles distinctes annotées de manière ambigüe du point de vue de l'analyse d'erreurs.

L'usage d'outils TAL dans le processus permet l'application du jeu d'étiquettes « à la volée » sur tout autre corpus. Il devient possible de développer des outils de requêtes s'appuyant sur une annotation identique entre corpus, les rendant ainsi interopérables. Cette interopérabilité rend envisageable un travail d'analyse d'erreurs contrastive entre plusieurs corpus d'anglais de locuteurs de langues maternelle différentes, ce qui permettrait de mieux répertorier les erreurs et de comparer les micro-systèmes d'erreurs selon les L1 des apprenants.

Remerciements

Nous remercions Detmar Meurers de l'Université de Tübingen pour ses recommandations méthodologiques précieuses. Nos remerciements s'adressent aussi à Nicolas Ballier de l'université de Paris-Diderot et Pascale Sébillot de l'IRISA pour leur relecture et suggestions. Que Camille Guinaudeau de INRIA soit remerciée pour son aide au débogage de scripts. Enfin, nous adressons nos remerciements à Helmut Schmid pour le partage de certains scripts.

5 Références

Biber, D., Johanson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided Error Analysis. *System,* (26), pages 163–174.

De Haan, P. (2000). Tagging Non-native English with the TOSCA-ICLE Tagger. *In Corpus Linguistics And Linguistic Theory*, pages 69–80.

Diaz-Negrillo, A., & Fernandez-Domingez, J. (2006). Error Tagging Systems for Learner Corpora *In Spanish Journal of Applied Linguistics (RESLA) RESLA,* (19), pages 83–102.

Diaz Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *In Language Forum*, 36(1-2), pages 139–154.

Gaillat, T. (2013). *This* and *That* in Native and Learner English: From Typology of Use to Tagset Characterisation. *In Corpora and Language in Use*. Louvain : Louvain University Press. À paraître.

Granger, S. (1993). International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Ostdijk (Eds.), *Papers from the Thirteenth International Conference on Language Research on Computerized Corpora*, pages 57–72.

Lapaire, J.-R., & Rotgé, W. (1998). *Linguistique et grammaire de l'anglais* (3e édition.). Toulouse: Presses Universitaires du Mirail.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for Querying and Manipulating Tree Data Structures. In Proceeedings of the *5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *In Computational Linguistics*, 19(2), pages 313–330.

Quirk, R., Leech, G., & Svartvik, J. (1985). *A Grammar of Contemporary English*. London, Beccles and Colchester: Longman.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In P*roceedings of the International Conference on New Methods in Language Processing*, pages 14–16.

Van Rooy, B., & Schafer, L. (2003). An Evaluation of Three POS Taggers for the Tagging of the Tswana Learner English Corpus. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conferenc*e (Vol. 16), pages 835–844.