Analyse morpho-syntaxique en présence d'alternance codique

José Carlos Rosales Núñez Guillaume Wisniewski
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
prénom.nom@limsi.fr

RÉSUMÉ

L'alternance codique est le phénomène qui consiste à alterner les langues au cours d'une même conversation ou d'une même phrase. Avec l'augmentation du volume généré par les utilisateurs, ce phénomène essentiellement oral, se retrouve de plus en plus dans les textes écrits, nécessitant d'adapter les tâches et modèles de traitement automatique de la langue à ce nouveau type d'énoncés. Ce travail présente la collecte et l'annotation en partie du discours d'un corpus d'énoncés comportant des alternances codiques et évalue leur impact sur la tâche d'analyse morpho-syntaxique.

ABSTRACT

PoS tagging of Code Switching

Code switching (CS) is a phenomenon consisting in alternating languages during a conversation or within a sentence. Due to the increasing volume of User Generated Content, code switching, that used to be mainly an oral phenomenon, is becoming more and more present in written texts, creating the need to adapt NLP tasks and models to this new type of content. This work presents the collection and annotation of a corpus containing CS sentences and assesses the impact of code switching on PoS tagging.

MOTS-CLÉS : Erreur d'annotation, analyse morpho-syntaxique, adaptation au domaine.

KEYWORDS: Annotation error, PoS-tagging, domain adaptation.

1 Introduction

Le *code-switching* (CS) ou alternance codique est le phénomène qui consiste à alterner les langues au cours d'une même conversation (Isurin *et al.*, 2009; Myers-Scotton, 1997). C'est un phénomène fréquent chez les locuteurs des communautés bilingues et multilingues qui ont l'habitude de passer d'une langue à l'autre au cours d'une conversation et parfois même à l'intérieur d'une phrase (Auer, 1998). La table 1 donne plusieurs exemples d'énoncés produits par des locuteurs anglais-espagnol illustrant ce phénomène.

Le code-switching est un phénomène oral que l'on ne retrouve quasiment pas à l'écrit : la quasi totalité des corpus existants (comme, par exemple, (Özlem Çetinoğlu, 2016; Ramanarayanan & Suendermann-Oeft, 2017)) est constituée de transcriptions. Mais avec l'augmentation du volume de contenu généré par les utilisateurs *(user generated content)* notamment sur les différents média sociaux (Facebook, Twitter, ...) ou les forums, de plus en plus de textes écrits comportent des énoncés écrits en plusieurs langues. En effet, par de nombreux aspects, les contenus générés par les utilisateurs ont des caractéristiques qui se rapprochent de ceux de la langue parlée. La collecte d'énoncés présentant des alternances codiques se retrouve donc simplifiée (il n'est plus nécessaire d'enregistrer et de

Conversation	◇ mi entonces ahorayou want to speak Spanish ! ◇ and we 're like " are ... I 'm sure this is like como unos chinitos ahí trabajan " ◇ no mentiradothat was a day one five dollars .
Twitter	◇ I used to think his name was Toño 😐 until they told me it was Jonny ☺ I was like pos Como Se Llame /.- 😜 ♡ ◇ The fact that Jonny already knew me I yo no 😀 It 's like Baby Porke Nunca me hablabas 😀 😀

TABLE 1: Exemples d'énoncés prononcés par des locuteurs anglais-espagnol comportant une alternance codique. Les mots anglais sont en bleu, les mots espagnols en rouge, les ponctuations, entités nommées et autres symboles en noir. Les données sont issues des deux corpus décrits à la section 2.

transcrire des dialogues), ouvrant la possibilité de nouvelles études. Mais, ce développement nécessite également l'adaptation des méthodes et des tâches existantes à ce nouveau type de données.

Ce travail comporte deux contributions : nous décrivons, dans un premier temps (§2), la collecte d'un nouveau corpus d'énoncés comportant des alternances codiques et leur annotation en partie du discours. Nous évaluerons ensuite l'impact de ce phénomène sur l'analyse morpho-sytaxique (§3).

2 Collecte et annotation des corpus

Nous allons considérer, dans nos expériences, deux corpus d'énoncés produits par des locuteurs bilingues espagnol-anglais correspondant aux deux types d'énoncés CS mentionnés dans l'introduction : la langue parlée et les contenus générés par l'utilisateur.

Le premier corpus que nous utilisons repose sur le corpus `Miami Bangor`[1], l'un des plus gros corpus de transcription contenant des alternances codiques : il est constitué des transcriptions de plus de 35h d'entretiens avec 84 locuteurs bilingues de la région de Miami. Les phrases de ce corpus ont été segmentées en mots automatiquement. Des annotateurs humains ont ensuite annoté chaque mot du corpus pour indiquer quelle était sa langue et son étiquette morpho-syntaxique, en suivant le guide d'annotation du projet UD (Nivre *et al.*, 2017). Une description complète de cette campagne d'annotation est faite dans (Soto & Hirschberg, 2017). Dans nos expériences, seules les phrases contenant un changement de langues ont été conservées. Dans la suite de cet article nous appellerons ce corpus `Conversation`.

Le second corpus est issu de la campagne d'évaluation organisée dans le cadre du second atelier *Computational Approaches to Linguistic Code Switching* (Molina *et al.*, 2016). Cette campagne avait pour objectif d'identifier la langue dont chaque mot d'un twee était issu. Comme le `Miami Bangor`, ce corpus comporte des énoncés mélangeant anglais et espagnol. Pour collecter ceux-ci, les organisateurs de la campagne ont ciblé les comptes Twitter d'utilisateurs habitant des régions dans lesquelles sont présents de nombreux locuteurs bilingues (en pratique, New-York et Miami) et qui suivent le compte Twitter de radios espagnoles. Les tweets collectés ont été segmentés et étiquetés semi-manuellement pour indiquer à quelle langue chaque mot appartenait.

À partir de cette information de langue, nous avons étiqueté automatiquement les corpus en utilisant

1. `http://bangortalk.org.uk/speakers.php?c=miami`

	n. phrases	n. mots	longueur phrase	% mots anglais	% mots espagnol	% symboles
Conversation	2 980	36 677	12 mots	39,0%	46,1%	14,9%
Twitter	1 002	15 474	14 mots	50,6%	28,8%	20,5%

TABLE 2: Principales caractéristiques des corpus utilisés dans ce travail. Les symboles correspondent à tous les mots dont il est impossible d'identifier la langue (noms propres, ponctuation, émoticône, ...).

des dictionnaires extraits de Wiktionary et des corpus anglais et espagnol du projet UD. Le guide d'annotation du projet UD a été étendu pour ajouter deux étiquettes correspondant aux hastags et aux émoticône. Deux annotateurs [2] ont ensuite vérifié et corrigé manuellement l'ensemble des étiquettes. Ce corpus sera appelé Twitter dans le reste de cet article. [3]

La table 2 résume les principales caractéristiques de ces deux corpus. Ces statistiques montrent que le corpus Twitter présente une alternance codique plus faible : la majorité des phrases ne comporte que quelques mots en espagnol et une large majorité de mots en anglais. En pratique, sur les deux corpus, environ 45% des phrases, il n'y a qu'un seul mot qui n'est pas exprimé dans la langue majoritaire, ce qui suggère que les deux corpus comportent de nombreux cas d'*emprunt lexical* et n'est pas constitué uniquement d'alternance codique à proprement parler (Myers-Scotton, 1997).

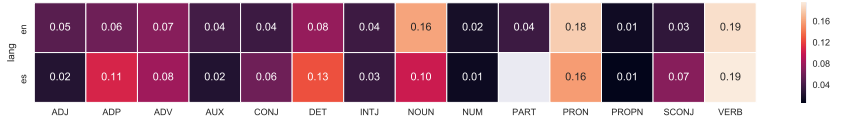
Pour caractériser les phénomènes d'alternance codique, nous avons considéré la distribution des étiquettes morpho-syntaxiques par langue à l'intérieur de chaque corpus analysé, résultat présenté dans la Figure 1. En comparant les distributions, il apparait clairement que, comme on pouvait s'y attendre, les corpus présentent des données de nature très différente (p. ex. la proportion d'adjectif varie considérablement d'un corpus à un autre). La langue majoritaire, c'est-à-dire, celle qui est la plus utilisée dans le corpus (espagnol pour Conversation et anglais pour Twitter) ne semble, par contre, ne pas avoir d'impact sur la nature des mots prononcés dans une langue ou dans l'autre.

3 Analyse morpho-syntaxique en présence d'alternance codique

Analyseur morpho-syntaxique à base d'historique Nous utilisons un analyseur morpho-syntaxique à base d'historique (Black *et al.*, 1992; Tsuruoka *et al.*, 2011). Dans ces approches, la prédiction d'une séquence d'étiquettes morpho-syntaxiques est modélisée sous la forme d'une suite de problèmes de décision, consistant chacun à prédire l'étiquette d'une observation. Chaque décision est prise par un classifieur multi-classe utilisant comme descripteurs des informations extraites de la structure d'entrée, ainsi que les décisions prises antérieurement. Nous utilisons, dans toutes nos expériences, un perceptron moyenné comme classifieur multi-classe (Collins, 2002). Nous utilisons des caractéristiques simples que l'on retrouve, à notre connaissance, dans tous les étiqueteurs morpho-syntaxique : mots courants, mots dans une fenêtre de ± 2, étiquettes des deux mots précédents (et leur conjonction), conjonction du mot courant et de l'étiquette précédente, ... [4] Une description détaillée

2. Un des annotateurs est un locuteur natif de l'espagnol ; les deux annotateurs parlent couramment l'anglais.
3. Ce corpus est téléchargeable librement à partir des pages personnelles des deux auteurs.
4. Les entrées sont également transformées : tous les nombres, les URL, les émoticônes et les mentions sont remplacées par un même token

(a) UD



(b) Conversation

(c) Twitter

FIGURE 1: Distribution des étiquettes morpho-syntaxiques selon les langues sur les corpus `UD` (haut), `Conversation` (milieu) et `Twitter` (bas).

de ce modèle se trouve dans (Wisniewski *et al.*, 2014b,a).

Les performances de ce modèle sont légèrement inférieures aux performances d'un modèle d'analyse morpho-syntaxique neuronal tout en étant beaucoup plus rapide à entrainer (notamment à cause du nombre réduit d'hyper-paramètres) : par exemple, sur les corpus anglais et espagnol du projet *Universal Dependencies*, notre modèle obtient, respectivement, une précision de 93,5% et 95,0% alors que le modèle UDPIPE (Straka *et al.*, 2016) obtient 93,5% et 95,5%.

Adaptation du modèle pour l'alternance codique Nous proposons dans cette section une modification très simple du modèle que nous venons de présenter visant à prendre en compte l'alternance de langues dans une phrase. Le principal objectif de ce modèle est de permettre de caractériser et de quantifier les problèmes soulevés par la présence d'alternance codique dans des phrases.

La méthode proposée repose sur une spécialisation du classifieur utilisé dans notre analyseur morpho-syntaxique. Elle consiste simplement à identifier la langue de chaque mot et à utiliser deux classifieurs différents, chacun adapté à une des langues en présence, pour prendre les décisions successives lors de l'inférence.

Plus précisément, nous apprenons, indépendamment, deux analyseurs morpho-syntaxiques : le premier sur un corpus anglais étiqueté avec des informations morpho-syntaxique, le second sur un corpus similaires en espagnol. Ces corpus sont identiques à ceux utilisés pour l'apprentissage d'un analyseur « classique ». Lors de l'inférence, en fonction de la langue du mot dont on cherche à prédire l'étiquette, l'un ou l'autre des modèles est utilisé pour réaliser la prédiction. Bien que les étiquettes soient prédites par des modèles indépendants (au sens où aucun information n'est partagée entre les langues au moment de l'apprentissage), l'historique est partagé. Par souci de simplification, l'identification de la langue d'un mot est réalisée de manière indépendante.

Protocole expérimental Le modèle introduit dans le paragraphe précédent a été testé sur les deux corpus introduits dans la section 2. Les modèles monolingues sont appris sur les corpus `UD_English` et `UD_Spanish` du projet UD. [5]

Identification de la langue Les langues de chaque mot du corpus de test sont prédites à l'aide de l'outil `langid.py` avec ses modèles pré-entrainés (Lui & Baldwin, 2012). Cet outil repose sur un classifieur bayésien naïf et est capable d'identifier la langue d'un *document* de manière très précise la langue d'un document, mais sa capacité à prédire la langue d'un mot unique n'a, à notre connaissance, jamais été évaluée. Nous utilisons également avec un modèle à réseaux de neurones (2 couches cachées composé de 128 et 64 neurones et une couche de sortie 'sotfmax') construit spécifiquement pour prédire la langue d'un mot sans connaissance du contexte dans lequel il a été utilisé. Ce réseau considère en entrée une représentation « one-hot » des 4-grams de lettres du mot ou le mot entier si sa longeur est plus petite que 4. Ce modèle est entrainé sur les deux corpus UD considérés.

Lorsque les mots sont pris isolément (c.-à-d. sans considérer leur context), L'outil `langid.py` est capable de prédire correctement la langue d'un mot issus d'un corpus présentant des alternances codiques dans 56,8% des cas, tandis que notre modèle atteint un taux de reconnaissance de 93,3% de reconnaissance. Cette grosse différence de performances est très certainement lié au fait que `langid.py` a besoin de plusieurs mots de la même langue regroupés dans une phrase pour identifier avec certitude la langue d'une phrase et ne peut donc être utilisé à un niveau sous-phrastique. Le modèle à base de *n*-gram de lettres que nous avons développé ne souffre pas de cette limite.

Résultats expérimentaux Les performances de notre modèle d'analyse morpho-syntaxique sont comparées à trois modèles de référence : deux analyseurs syntaxiques appris uniquement sur les modèles monolingues (c.-à-d. un analyseur appris uniquement sur le corpus anglais et un autre appris uniquement sur le corpus espagnol) ainsi qu'un analyseur appris lorsque les phrases du corpus anglais sont mélangées aux phrases du corpus espagnol. Nous considérons également, comme point de comparaison, un résultat oracle correspondant à une situation dans laquelle la langue est systématiquement correctement identifiée.

La Table 3 rapporte le taux d'erreur obtenu par chacun de ces modèles sur les deux corpus considérés. Ces taux d'erreurs sont moyennés sur 10 apprentissages.

Les résultats très faibles obtenus par les modèles monolingues montrent qu'il est clairement nécessaire de prendre en compte la présence d'alternance codique. Une simple concaténation des corpus monolingues semblent par contre déjà permettre une réduction forte du nombre d'erreurs, la différence avec les performances obtenues sur les corpus de test de l'UD pouvant s'expliquer par la nature des données considérées : l'UD contient essentiellement des textes journalistiques et issus de wikipédia alors que les corpus `Conversation` et `Twitter` contiennent de la parole spontanée.

De manière très surprenante, la méthode par spécialisation du modèle que nous proposons obtient des résultats légèrement plus mauvais que la simple concaténation des corpus, même lorsque la langue est connue de manière certaine. Ce résultat montre que la connaissance de la langue d'un mot n'apporte pas une information pertinente à la prédiction de son étiquette morpho-syntaxique.

5. Au final, les corpus présentant de l'alternance codique ne sont utilisés que pour l'évaluation de notre modèle : nous apprenons deux modèles d'analyse morpho-syntaxique indépendant, l'un sur le corpus UD espagnol et l'autre sur le corpus UD anglais. Lors de la phase de test, ces deux modèles sont utilisés alternativement en fonction du résultat de l'identification de la langue de chaque mot.

méthode	Conversation	Twitter	UD espagnol	UD anglais
Analyseur anglais	45,5%	23,8%	67,5%	6,5%
Analyseur espagnol	39,0%	60,2%	5,0%	69,7%
Analyseur anglais+espagnol	13,2%	18.2%	5,1%	7,1%
Sélection (`langid`)	37,4%	30,5%	—	—
Sélection	17,3%	25,8%	—	—
Sélection oracle	13,1%	19,7%	—	—

TABLE 3: Taux d'erreurs obtenus par les différents modèles sur les deux corpus considérés.

Plusieurs raisons peuvent expliquer ce résultat. En particulier, le nombre de mots identiques en anglais et en espagnol et dont l'étiquette morpho-syntaxique diffère n'est peut-être pas suffisant pour avoir un impact sur le taux d'erreur global. En pratique, sur les ensembles d'apprentissage des corpus `UD_English` et `UD_Spanish` il n'y en a que 2 424 types communs (pour 16 568 mots anglais et 44 739 mots espagnols) et seulement 583 d'entre eux ont des catégories morpho-syntaxique différente dans les deux langues. De plus, l'annotation des langues semble ne pas toujours être de très bonne qualité et présente de nombreuses décisions arbitraires (par exemple, au niveau des interjections et des noms propres). Il faut également noter que le modèle anglais+espagnol est appris sur un corpus deux fois plus grand que les modèles monolingues.

Comme on pouvait s'y attendre, les performances chutent de manière significative lorsque la langue d'un mot est déterminée de manière automatique : lorsque la langue est prédite par `langid.py`, les taux d'erreur obtenus sur les corpus `Twitter` et `Conversation` sont, respectivement, de 30,5% et de 37,4%. En utilisant notre classifieur à réseau de neurones pour la détection de la langue, le taux d'erreur est de 17,3% pour le corpus `Conversation` et 25,8% sur le corpus `Twitter`.

4 Conclusion

Nous avons présenté dans ce travail deux corpus contenant des énoncés avec de l'alternance codique et annotés en partie du discours. C'est, à notre connaissance, l'une des première fois qu'un aussi gros volume de données présentant ce phénomène est annoté avec des informations morpho-syntaxiques ce qui ouvre la voie à beaucoup de perspectives pour analyser ce phénomène.

Nous avons également présenté des modèles d'analyse morpho-syntaxiques simples, mais conçus pour prendre en compte les phénomènes d'alternance codique et analyser leurs performances. Ces résultats montrent la difficulté de la tâche.

Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (projet PARSITI, ANR-16-CE33-0021). Nous remercions les relecteurs pour leurs commentaires et suggestions.

Références

P. AUER, Ed. (1998). *Code-Switching in Conversation : Language, Interaction and Identity*. Routledge.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

COLLINS M. (2002). Discriminative training methods for hidden markov models : Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, p. 1–8 : Association for Computational Linguistics.

ISURIN L., WINFORD D. & DE BOT K. (2009). *Multidisciplinary Approaches to Code Switching*. John Benjamins Publishing.

LUI M. & BALDWIN T. (2012). langid.py : An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, p. 25–30, Jeju Island, Korea : Association for Computational Linguistics.

MOLINA G., ALGHAMDI F., GHONEIM M., HAWWARI A., REY-VILLAMIZAR N., DIAB M. & SOLORIO T. (2016). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, p. 40–49, Austin, Texas : Association for Computational Linguistics.

MYERS-SCOTTON C. (1997). *Duelling Languages : Grammatical Structure in Codeswitching*. Clarendon Press.

NIVRE J., AGIĆ Ž., AHRENBERG L. & OTHER (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

RAMANARAYANAN V. & SUENDERMANN-OEFT D. (2017). Jee haan, i'd like both, por favor : Elicitation of a code-switched corpus of hindi-english and spanish-english human-machine dialog. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, p. 47–51.

SOTO V. & HIRSCHBERG J. (2017). Crowdsourcing universal part-of-speech tags for code-switching. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, p. 77–81.

STRAKA M., HAJIČ J. & STRAKOVÁ J. (2016). UDPipe : trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia : European Language Resources Association.

TSURUOKA Y., MIYAO Y. & KAZAMA J. (2011). Learning with lookahead : Can history-based models rival globally optimized models ? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL'11, p. 238–246, Portland, Oregon, USA : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014a). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014b). Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 173–183, Marseille, France : Association pour le Traitement Automatique des Langues.

ÖZLEM ÇETINOĞLU (2016). A turkish-german code-switching corpus. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France : European Language Resources Association (ELRA).