Expériences d'étiquetage morphosyntaxique dans le cadre du projet RESTAURE

Pierre Magistry[1] Anne-Laure Ligozat[2] Sophie Rosset[1]
(1) LIMSI, CNRS, Université Paris-Saclay, Bât 508, Campus Universitaire F-91405 Orsay, France
(2) LIMSI, CNRS, ENSIIE, Université Paris-Saclay, Bât 508, Campus Universitaire F-91405 Orsay, France
prenom.nom@limsi.fr.fr

RÉSUMÉ
Le projet RESTAURE vise à outiller en outils TAL trois langues régionales de France : l'alsacien, l'occitan et le picard. Dans cet article, nous abordons la question de l'étiquetage morphosyntaxique et rapportons les performances de différents systèmes proposés dans la littérature. Notre objectif est d'aborder les trois langues de manière homogène afin de pouvoir comparer les méthodes dans la variété de situations que présentent nos données. Ces expériences doivent guider notre réflexion pour le développement d'outils semisupervisés.

ABSTRACT
POS Tagging Experiments in the RESTAURE Project

The RESTAURE project aims at developing resources and NLP tools for three low-resource regional languages of France : Alsacien, Occitan and Picard. In this paper we report preliminary experiments in POS tagging. We benefit from this variety of cases to test methods in realistic situations. These experiments are conducted in order to drive the design of semisupervised methods.

MOTS-CLÉS : étiquetage morphosyntaxique, langues peu dotées.

KEYWORDS: Pos tagging, low-resource languages.

1 Introduction

Nos travaux s'inscrivent dans le projet RESTAURE, qui vise à outiller trois langues régionales de France : l'alsacien, l'occitan et le picard. Avec ces trois langues, nous disposons d'un éventail de cas d'étude pour l'outillage de langues peu dotées. Le projet s'étend de la compilation des corpus et de ressources lexicales à la réalisation d'outils de TAL. Dans cet article nous nous focalisons sur la tâche d'étiquetage morphosyntaxique.

2 Description des corpus

Les ressources mises à notre disposition par les différents partenaires du projet varient grandement d'une langue à l'autre, en fonction de l'état d'avancement de l'outillage de la langue concernée. Nous présentons ici les corpus que nous utilisons dans nos expériences.

2.1 Alsacien

Pour travailler sur l'alsacien, nous disposons d'un corpus annoté par deux expertes qui nous sert de corpus de référence pour l'évaluation de nos systèmes. Ce corpus est composé de quatre textes : deux articles tirés de Wikipedia, des recettes de cuisine et un court extrait d'une pièce de théâtre. Notons que les articles de Wikipedia sont rédigés en haut-rhinois tandis que les deux autres textes sont en bas-rhinois.

Nous disposons aussi de textes non annotés issus de la Wikipedia que nous utilisons comme données brutes.

2.2 Picard

Dans le cadre du projet, les partenaires de l'Université de Picardie travaillent à la constitution d'un corpus de référence pour le picard. Ce travail de construction du corpus est encore en cours ; les chiffres donnés à la table 1 sont donc les comptes de la version du corpus que nous utilisons au moment de la rédaction de cet article, mais ils seront amenés à évoluer rapidement. Ce corpus se compose essentiellement d'œuvres littéraires, et sera prochainement complété par des textes de presse régionale.

Une partie de ce corpus a été annotée en partie du discours. Là aussi, nous utilisons ce sous-corpus comme corpus d'évaluation de nos systèmes, tandis que la partie non annotée constitue l'ensemble de nos données brutes.

2.3 Occitan

Le corpus annoté dont nous disposons pour l'occitan est moins varié que pour les deux autres langues puisqu'il est issu d'une seule œuvre (Escorregudas en Albigés de Sergi Viaule). Nous prévoyons d'enrichir notre corpus avec des textes de presse, mais dans l'immédiat nous limitons nos expériences sur l'occitan à la partie supervisée (Section 4). D'autres expériences sur cette langue sont décrites dans (Vergez-Couret & Urieli, 2015).

2.4 Vue d'ensemble

Des informations quantitatives sur nos corpus sont détaillées dans le tableau 1. Dans toutes les expériences qui suivent, nous utilisons la tokenisation de référence pour les évaluations. Lorsque nous utilisons les corpus bruts, nous utilisons le tokeniseur à base de règles qui a servi à la pré annotation du corpus de référence pour l'alsacien, et une tokenisation « naïve » pour le picard.

Les jeux d'étiquettes varient d'une langue à l'autre. Pour l'alsacien, la liste des étiquettes a été affinée par rapport aux travaux précédents. Ces différences expliquent en partie les variations dans les résultats obtenus. Une conversion vers un jeu d'étiquettes commun aux trois langues est prévue, mais n'est pas encore disponible. La taille des différents jeux est aussi indiquée au tableau 1.

Langue	Texte	Tokens	Types	Jeu d'étiquettes
alsacien	Wikipedia 1	400	210	
	Wikipedia 2	503	252	
	cuisine	364	203	
	théatre	232	141	
alsacien	total annoté	1499	719	16
alsacien	données brutes	70 536	15 362	
occitan	un texte annoté	31 207	5 901	33
picard	lesi ziepe	366	191	
	Philéas Lebesgue	188	124	
	Simons L'Gampe	348	213	
picard	total annoté	11 814	3 843	15
picard	données brutes	1 769 018	140 072	

TABLE 1 – Corpus de référence

3 Étiquetage morphosyntaxique des langues peu dotées

L'étiquetage morphosyntaxique de langues peu dotées est une thématique qui bénéficie déjà d'une littérature abondante. Mais la diversité des situations fait que les méthodes proposées sont rarement applicables directement à un cas de figure particulier.

Ainsi beaucoup ont recours à des corpus alignés (à la suite de (Yarowsky *et al.*, 2001)), et une grande partie de ces expériences utilisent le corpus Europarl (Koehn, 2005). Si les résultats obtenus par ces méthodes sont assez encourageants, elles ne sont applicables qu'à un sous-ensemble assez restreint des langues peu dotées. Cet ensemble exclut les langues auxquelles nous nous intéressons, puisqu'Europarl couvre les langues européennes officielles. Les langues sur lesquelles nous travaillons sont beaucoup moins standardisées et peu ou pas enseignées, ce qui provoque une diversité des pratiques graphiques à laquelle ne sont pas sujettes les langues d'Europarl. De plus, pour utiliser les algorithmes proposés par cette lignée de travaux, il est nécessaire de disposer de corpus alignés. De tels corpus ne sont pas disponibles pour les langues que nous devons traiter.

Une autre pratique que nous laissons de côté dans un premier temps est le recours à des représentations vectorielles des mots par analyse de sémantique distributionnelle (telle que *word2vec*). Pour les travaux qui utilisent de telles méthodes, des corpus de données brutes de plusieurs dizaines de millions de tokens sont généralement requis. Notre corpus le plus grand, celui du picard, n'atteint pas deux millions de tokens.

Dans notre situation, nous sommes poussés à nous intéresser à l'adaptation des ressources ou des outils disponibles pour les langues proches des langues que nous traitons et qui sont mieux dotées. On utilisera ainsi comme langues « sources » : l'allemand pour l'alsacien, le catalan pour l'occitan et le français pour le picard. Nous avons recours à des modèles pré entraînés pour *Treetagger* (Schmid, 1995) ou le *Stanford Tagger* (Toutanova *et al.*, 2003) et nous comparons ceux-ci avec un étiqueteur que nous entraînons en utilisant les corpus Tiger (Brants *et al.*, 2004) pour l'allemand et Sequoia

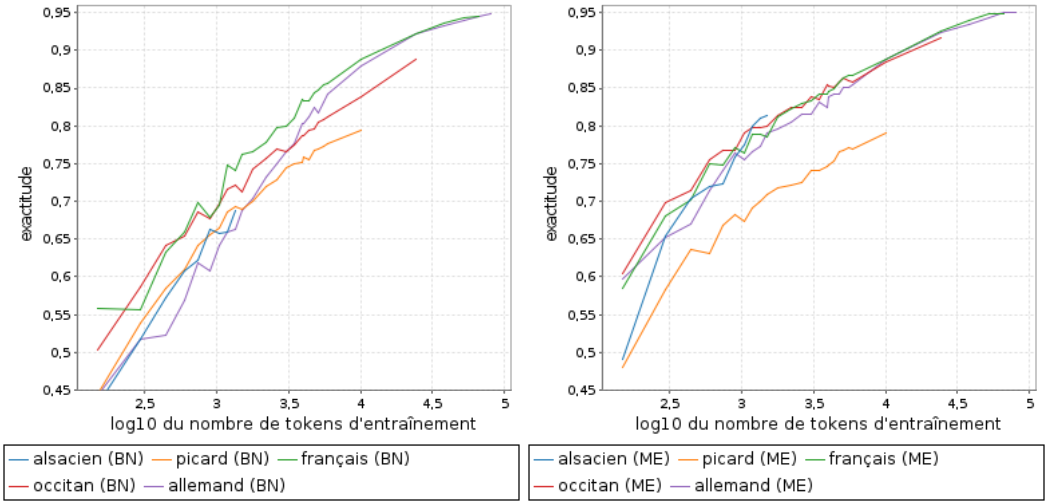

FIGURE 1 – Courbes d'apprentissage sur les différents corpus, en Bayes Naif (BN) ou MaxEnt(ME)

(Candito & Seddah, 2012) pour le français.

Dans un premier temps, nous commençons par observer les résultats que l'on peut obtenir avec le peu de données annotées dont nous disposons et des méthodes supervisées classiques (Section 4). Puis nous testons les méthodes de transfert (Bernhard & Ligozat, 2013) (Section 5), enfin nous essayons l'auto-entraînement (*self-training*) en utilisant une première annotation des corpus par transfert.

4 Apprentissage supervisé

Pour nos expériences, nous utilisons notre propre implémentation d'un étiqueteur très inspiré de MElt (Denis & Sagot, 2009), qui permet d'intégrer facilement de l'information lexicale issue de ressources externes et d'utiliser un classifieur bayésien naïf (BN) ou un Maximum d'Entropie (ME) de façon interchangeable. Dans le cas idéal d'une langue standardisée pour laquelle un large corpus d'entraînement et des lexiques à large couverture sont disponibles (par exemple sur le français ou l'allemand), MElt est au niveau de l'état de l'art. Dans les expériences présentées ici, nous n'intégrons pas de lexiques externes supplémentaires (nous extrayons simplement celui du corpus d'entraînement). Les scores rapportés sont donc légèrement en deçà des performances rapportées pour MElt. La possibilité d'intégrer des lexiques externes nous sera utile dans la suite de nos travaux.

Afin d'observer l'influence de la taille du corpus d'entraînement indépendamment des types de textes considérés, nous procédons par échantillonnages aléatoires sur la totalité du corpus. Pour différentes tailles d'entraînement, nous extrayons cette quantité de phrases pour l'entraînement de notre étiqueteur et nous le testons sur le reste du corpus. Cette opération est réalisée n fois et nous calculons les moyennes des scores obtenus, qui sont reportées à la Figure 1.

Pour l'alsacien et le picard, nous disposons de corpus d'évaluation qui sont composés de textes

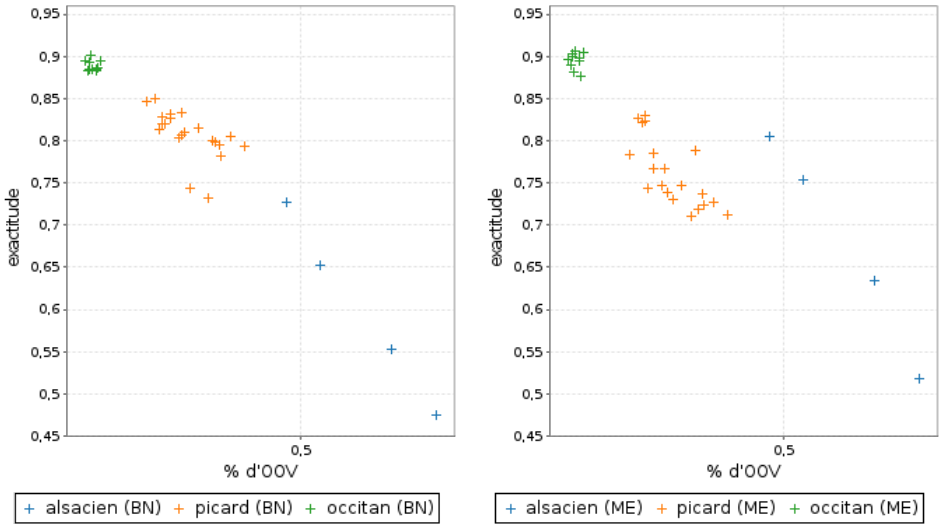

FIGURE 2 – Scores obtenus en fonction de différentes proportions de tokens inconnus. Sur les différents corpus, en Bayes Naif (BN) ou MaxEnt(ME)

d'origines variées. Il nous semble donc intéressant d'observer les différents cas de figure qui se présentent à nous. Nos corpus étant d'une taille très restreinte, nous évaluons indépendamment l'étiquetage de chaque texte avec un étiqueteur entraîné sur l'ensemble des autres textes (de la langue concernée).

Il est ainsi possible d'observer l'influence de différents paramètres, notamment celle de la couverture lexicale du corpus d'entraînement. Les scores obtenus indépendamment sur chaque document sont reportés sur la figure 2 où chaque point représente le score obtenu sur un document en fonction du pourcentage de formes inconnues qu'il contient.

Si ces résultats semblent encourageants au regard de la faible quantité de données dont nous disposons, nos corpus annotés sont coûteux et difficiles à étendre. Soulignons que l'axe du nombre de tokens de la figure 1 est au logarithme. L'effort d'annotation nécessaire pour rejoindre les scores des langues mieux dotées n'est donc pas négligeable. Étant donnée la taille actuelle de nos corpus annotés, nous préférons limiter l'usage de ceux-ci comme corpus de *développement*, pour évaluer et affiner d'autres méthodes.

5 Étiquetage par transfert

Dans cette section nous proposons une évaluation de l'approche par transposition de corpus. On appelle ici « langue source » la langue proche pour laquelle nous disposons de ressources annotées ou de modèles d'étiquetage préentraînés, et « langue cible » la langue que nous cherchons à analyser. Nous procédons en plusieurs étapes :

— utilisation d'un étiqueteur en langue source sans modification ;

— traduction de mots outils fréquents sur la base de listes ne dépassant pas les centaines d'entrées.

Pour ces expériences, nous comparons Treetagger, le Stanford Tagger avec des modèles préentraînés ainsi que notre étiqueteur présenté à la Section 4, que nous entraînons cette fois sur le corpus en langue source. Pour notre étiqueteur, nous testons le classifieur bayésien naïf et la régression logistique. Dans les configurations classiques, nous nous attendons à ce que cette dernière donne les meilleurs résultats, mais ce n'est pas ce que nous observons ci-dessous.

5.1 Étiqueteurs de langue source

Le point de départ de cette approche est d'utiliser un étiqueteur en langue source. Dans cette configuration, les corpus ne sont pas modifiés et nous utilisons les modèles pour la *langue cible* tels qu'ils sont distribués avec l'étiqueteur ou en entraînant notre étiqueteur sur Sequoia (français) ou Tiger (allemand).

langue	Texte	TreeTagger	Stanford	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0.49	0,59	0,48	0,50
	Wikipedia2	0,49	0,51	0,49	0,50
	recette	0,50	0,53	0,49	0,49
	théâtre	0,62	0,65	0,59	0,61
	tous	0,51	0,55	0,50	0,51
picard	lesi ziepe	0,38	0,33	0,43	0,40
	Philéas Lebesgue	0,56	0,58	0,60	65
	Simons L'Gampe	0,83	0,84	0,81	0,83
	tous	0,55	0,53	0,55	0,58

TABLE 2 – Scores d'exactitude de différents étiqueteurs entraînés pour la langue source, testés sur la langue cible sans modification.

En plus d'un score global par langue, nous donnons le résultat pour chaque texte de notre corpus alsacien. Pour le picard, nous ne mentionnons les résultats par texte que pour deux des plus mauvais et le meilleur. Les résultats varient fortement d'un texte à l'autre, particulièrement en picard (plus de 50 points d'écart entre les deux extrêmes). Cela reflète la variété des distances prises par les auteurs des textes avec la langue voisine.

5.2 Transposition des mots outils

Nous reproduisons ici une expérience comparable à (Bernhard & Ligozat, 2013) sur nos nouvelles données. Pour cela nous utilisons les listes de 169 traductions issues de ces travaux pour l'alsacien vers l'allemand. Pour le picard, nous générons cette liste automatiquement à partir d'une partie de nos corpus et d'une annotation avec la traduction mot à mot en français. En partant des 200 formes les plus fréquentes, nous conservons celles qui correspondent à une catégorie grammaticale fermée et qui est différente du français. Nous aboutissons ainsi à une liste de 77 paires de traductions.

Chaque occurrence des formes contenues dans ces mini lexiques est remplacée par sa traduction en langue source dans le texte en langue cible avant qu'il soit soumis à l'étiqueteur. En sortie d'étiquetage,

langue	Texte	TreeTagger	Stanford	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0,78	0,83	0,78	0,78
	Wikipedia2	0,74	0,77	0,73	0,75
	recette	0,65	0,76	0,67	0,67
	théâtre	0,74	0,78	0,71	0,76
	tous	0,73	0,78	0,73	0,74
picard	lesi ziepe	0,63	0,66	0,68	0,66
	Philéas Lebesgue	0,63	0,61	0,63	0,62
	Simons L'Gampe	0,84	0,85	0,82	0,83
	tous	0,70	0,69	0,70	0,71

TABLE 3 – Scores d'exactitude de différents étiqueteurs entraînés pour la langue source, testés sur la langue cible avec traduction des mots outils.

les formes d'origine sont rétablies, afin d'obtenir le corpus étiqueté en langue cible. Par exemple, les formes *ànder ànschtàtt àwer ìwer ùff ùn* de l'alsacien seront remplacées par les formes de l'allemand *ander, anstaat, aber, über, auf, und* que connait l'étiqueteur.

Les résultats des différents étiqueteurs sont donnés dans le tableau 3.

Cette méthode améliore les résultats sur tous les textes, mais elle profite bien plus aux textes les plus éloignés de la langue source, pour aboutir à des scores plus homogènes.

6 Réapprentissage endogène

Pour finir, nous essayons de combiner les différents étiqueteurs afin de tirer partie des données brutes dont nous disposons.

En utilisant l'étiqueteur de Stanford avec transposition comme dans la configuration précédente, nous annotons nos corpus bruts. Ces données annotées automatiquement servent ensuite à entraîner notre étiqueteur que nous évaluons comme précédemment. Les résultats sont donnés dans le tableau 4.

Cette opération n'améliore pas vraiment les scores globaux, mais améliore ou dégrade les scores de différents textes. Il faut noter que les meilleurs scores obtenus sont encore loin d'être satisfaisants et sont comparables à ceux obtenus avec de très petits corpus d'entraînement (cf. Figure 1).

7 Discussion

Nous avons présenté une série d'expériences sur un ensemble de données variées en utilisant différentes approches classiques.

Nous ne parvenons pas à apporter d'amélioration nette à l'état de l'art. Si les approches par corpus parallèles et sémantique distributionnelle sont matériellement difficiles voir impossibles à mettre en œuvre dans les cas concrets de langues peu outillées qui se présentent à nous, nous observons ici que la méthode par transposition des ressources en langue proche amène son lot de difficultés et qu'il n'est pas trivial de l'améliorer. D'un autre côté, l'approche supervisée obtient rapidement de

langue	Texte	Bayes Naïf	MaxEnt
alsacien	Wikipedia1	0,79	0,81
	Wikipedia2	0,80	0,81
	recette	0,75	0,74
	théâtre	0,75	0,71
	tous	0,78	0,78
picard	lesi ziepe	0,64	0,69
	Philéas Lebesgue	0,61	0,68
	Simons L'Gampe	0,81	0,84
	tous	0,70	0,71

TABLE 4 – Scores d'exactitude de différents étiqueteurs entraînés sur des corpus étiquetés automatiquement avec la méthode par transposition.

meilleurs résultats, mais ils sont eux aussi loin d'être satisfaisants et la constitution de corpus annotés de taille suffisante n'est pas envisageable.

En effet, pour les approches supervisées, nous nous attendons à avoir besoin de corpus d'autant plus grands que la variation dialectale, ainsi que les diversités de choix graphiques en l'absence de norme imposée, auxquelles vient s'ajouter la pratique fréquente d'alternance de code entre langues régionales et langues institutionnalisées voisines, accentuent la dispersion des données. La richesse et la variété des phénomènes observables dans nos corpus nous emmènent bien loin des corpus homogènes disponibles pour les « grandes » langues standardisées. (et les moyens engagés pour y travailler sont eux aussi sans comparaison).

En testant la transposition des mots grammaticaux, nous avons observé une grande disparité de résultats d'un texte à l'autre lorsque la différence de distance à la langue source est plus ou moins marquée. Dans certains cas, un algorithme généralement plus performant dans une configuration d'apprentissage supervisé classique (le MaxEnt) peut se révéler moins performant qu'un algorithme plus basique (Bayes Naïf), les features plus lexicalisées de notre étiqueteur « maison » le pénalisant par rapport à celui de Stanford pour cette approche par transposition. À la vue de cette disparité des résultats, il semble difficile, voir impossible de définir une unique bonne stratégie pour l'ensemble d'un corpus. Il sera peut-être nécessaire de se doter de moyens d'identifier des sous-parties du corpus pour entraîner plusieurs modèles et définir une méthode de sélection du modèle le plus pertinent.

Pour la suite de nos travaux, il nous semble nécessaire de combiner les deux approches en nous intéressant aux méthodes semi-supervisées pouvant utiliser la transposition en amorce.

Remerciements

Ces travaux ont bénéficié du soutien de l'ANR (projet RESTAURE - référence ANR-14-CE24-0003).

Références

BERNHARD D. & LIGOZAT A.-L. (2013). Es esch fàscht wie Ditsch, oder net ? Étiquetage

morphosyntaxique de l'alsacien en passant par l'allemand. In *TALARE 2013*, p. 209–220, Les Sables d'Olonne, France.

BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). TIGER : Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, **2**(4), 597–620.

BRAS M. & THOMAS J. (2008). Batelòc : cap a una basa informatisada de tèxtes occitans. In A. R. . D. SUMIEN, Ed., *IXème Congrès International de l'Association Internationale d'Études Occitanes*, p. 661–670, Aachen, Germany : Shaker.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort.

KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, p. 79–86 : Citeseer.

SCHMID H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *In Proceedings of the ACL SIGDAT-Workshop*, p. 47–50.

TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 173–180, Stroudsburg, PA, USA : Association for Computational Linguistics.

VERGEZ-COURET M. & URIELI A. (2015). Analyse morphosyntaxique de l'occitan languedocien : l'amitié entre un petit languedocien et un gros catalan. In *TALARE 2015*, Caen, France.

YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.