Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand

Delphine Bernhard, Anne-Laure Ligozat

▶ **To cite this version:**

Delphine Bernhard, Anne-Laure Ligozat. Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. TALARE 2013, Jun 2013, Les Sables d'Olonne, France. pp.209-220. hal-00838355

HAL Id: hal-00838355

https://hal.archives-ouvertes.fr/hal-00838355

Submitted on 4 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Es esch fàscht wie Ditsch, oder net? *
Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand

Delphine Bernhard[1] Anne-Laure Ligozat[2, 3]
(1) LiLPa, Université de Strasbourg
(2) LIMSI-CNRS, Orsay
(3) ENSIIE, Évry
dbernhard@unistra.fr, annlor@limsi.fr

RÉSUMÉ

L'analyse morphosyntaxique est une pierre angulaire de nombreuses applications du traitement automatique des langues. Elle nécessite toutefois des efforts de développement conséquents, qu'il s'agisse d'annoter des corpus ou de produire des lexiques et des outils. Pour les langues moins dotées, il peut être utile d'exploiter leur proximité avec des langues disposant d'outils et de ressources. Dans cet article, nous nous intéressons plus particulièrement aux dialectes alsaciens, qui présentent de nombreuses similitudes avec l'allemand standard. Nous montrons qu'il est possible d'utiliser des outils développés pour l'allemand afin de réaliser l'analyse morphosyntaxique de textes en alsacien. La méthode consiste à transposer les mots outils des textes alsaciens vers leurs équivalents en allemand standard. Cette transposition nécessite pour seule ressource un lexique bilingue des mots outils.

ABSTRACT

Es esch fàscht wie Ditsch, oder net ? POS-Tagging the Alsatian Dialects through German

Morphosyntactic analysis is a cornerstone of many natural language processing applications. However, it requires substantial development efforts, in order to annotate corpora or produce lexicons and tools. For under-resourced languages, it may be useful to exploit their proximity to languages for which tools and resources have already been developed. In this article, we focus on the Alsatian dialects, which have many similarities with standard German. We show that it is possible to use tools developed for the German language in order to perform the morphosyntactic analysis of Alsatian texts. The method consists in transposing function words in the Alsatian texts to their translations in standard German. This transposition requires as only resource a bilingual lexicon of function words.

MOTS-CLÉS : étiquetage morphosyntaxique, dialecte, alsacien, allemand.

KEYWORDS: POS tagging, dialect, Alsatian, German.

∗. C'est presque comme de l'allemand, non ?

1 Introduction

Les dialectes alsaciens, parlés dans le nord est de la France, appartiennent aux familles des langues alémaniques et franciques (Huck *et al.*, 2007). Selon une étude récente, 43% de la population alsacienne se déclare dialectophone (OLCA / EDinstitut, 2012). Toutefois, la proportion des locuteurs de l'alsacien décroit régulièrement depuis les années 1960.

Le traitement automatique des dialectes alsaciens présente plusieurs difficultés :

– Il n'existe pas de standard graphique utilisé et adopté par tous (voir section 2) ;
– L'alsacien correspond à un continuum de dialectes, se caractérisant par des variantes lexicales et phonétiques ;
– Les ressources et corpus écrits sont rares.

Ces contraintes, qui se retrouvent pour nombre de dialectes et langues minoritaires, rendent difficile le développement d'outils de Traitement Automatique des Langues (TAL) selon les méthodes classiques. En effet, l'absence de corpus annotés rend impossible l'utilisation de méthodes par apprentissage supervisé. Par ailleurs, il est également difficile de trouver une main d'œuvre suffisante et qualifiée pour développer des ressources ou annoter des corpus.

Dans cet article, nous nous intéressons à l'une des premières étapes dans toute chaîne de traitement de données textuelles et présentons une méthode simple mais néanmoins efficace pour réaliser l'analyse morphosyntaxique de textes alsaciens. Nous proposons de transposer les mots grammaticaux (déterminants, pronoms, prépositions, conjonctions) et les auxiliaires dans leur équivalent en allemand standard puis d'utiliser des étiqueteurs morphosyntaxiques existant pour l'allemand. Cette approche permet d'exploiter la proximité de l'alsacien et de l'allemand standard.

L'article est structuré comme suit : la section 2 détaille la problématique de l'alsacien écrit. La section suivante décrit les travaux sur l'étiquetage morphosyntaxique des textes "hors norme". Nous décrivons notre approche dans la section 4 et présentons les résultats de l'évaluation dans la section 5.

2 L'alsacien écrit

L'alsacien appartient aux groupes alémaniques et franciques et se rapproche de fait des dialectes parlés dans les régions limitrophes d'Allemagne et de Suisse. On retrouve des emprunts au français, mais la morphosyntaxe est très similaire à celle de l'allemand standard.

La graphie des dialectes alsaciens n'est pas codifiée. L'OLCA (Office pour la Langue et la Culture de l'Alsace) précise d'ailleurs sur sa page Web de définition de la langue régionale : *«La langue normalisée, écrite et codifiée correspondante à nos dialectes est l'allemand standard.»* [1]. Cette page cite également le recteur Pierre Deyon qui en 1985 affirmait qu'*«[i]l n'existe en effet qu'une seule définition scientifiquement correcte de la langue régionale en Alsace, ce sont les dialectes alsaciens dont l'expression écrite est l'allemand.»*

1. `http://www.olcalsace.org/fr/observer-et-veiller/definition-de-la-langue-regionale`.

Cette définition ne prend toutefois pas en compte les situations dans lesquelles il est nécessaire de passer par l'écrit pour documenter (dictionnaires, lexiques), transmettre (cours de langue), informer (encyclopédie, articles de presse, sites Web) et créer (littérature, pièces de théâtre). On trouve en effet de nombreux exemples d'alsacien écrit, notamment :

– Lexiques et dictionnaires : lexiques thématiques de l'OLCA [2], *De elsässisch Dico* (Bitsch et Matzen, 2004), *Mon premier dictionnaire français-alsacien en images* (Matzen *et al.*, 1997) etc. ;
– Manuels d'apprentissage de l'alsacien : *Wie geht's ?* (Matzen *et al.*, 2004), *Parlons alsacien* (Schimpf et Muller, 1998), *L'alsacien pour les nuls* (Keck *et al.*, 2010), etc. ;
– Articles dans la presse régionale : L'Alsace et Dernières Nouvelles d'Alsace ;
– Sites Web d'entreprises locales, comme par exemple Wattwiller [3] ;
– Pièces de théâtre en alsacien : le théâtre dialectal est très vivace en Alsace et de nombreuses troupes perpétuent cette tradition dans les villages, ainsi que dans les agglomérations plus importantes (comme le Théâtre Alsacien de Strasbourg).

Il faut également souligner qu'il y a eu des tentatives pour définir une graphie propre à l'alsacien :

– Le système ORTHAL (Zeidler et Crevenat-Werner, 2008) se réfère à la norme graphique de l'allemand standard tout en permettant la transcription des spécificités des dialectes alsaciens. On retrouve notamment l'emploi des majuscules pour les substantifs et pronoms personnels à la forme de politesse, comme en allemand [4].
– Le système GRAPHAL-GERIPA (Hudlett et Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) spécifie un ensemble de règles pour passer des sons aux graphèmes.

Il est toutefois difficile d'estimer la diffusion et l'utilisation réelle de ces systèmes de scripturalisation. Par ailleurs, ils laissent naturellement place à la variation, afin de correspondre au mieux aux diverses variantes géolinguistiques rencontrées en Alsace. Ils ne peuvent donc résoudre à eux seuls le problème de la déviance par rapport à une certaine norme qu'il est nécessaire de respecter pour permettre un fonctionnement optimal des outils du TAL.

3 Étiquetage morphosyntaxique de textes "hors norme"

L'absence ou le non respect des conventions d'écriture est un problème récurrent en TAL car les outils sont généralement développés pour des données standard, exemptes de fautes d'orthographe ou de phénomènes déviants.

Giesbrecht et Evert (2009) évaluent des étiqueteurs morphosyntaxiques pour l'allemand pour des données issues du Web et montrent que leurs performances sont dégradées pour ce type de textes. Ils montrent également que la performance est dépendante du genre et que certains genres spécifiques du Web, comme les forums, sont plus difficiles à traiter.

Différentes approches ont été proposées pour contourner ces problèmes. La solution la plus fréquente consiste à normaliser les mots "hors norme". Mosquera *et al.* (2012) présentent une méthode permettant de transformer des textes du Web 2.0 vers leur forme canonique en anglais ou en espagnol. La méthode se décompose en deux étapes : dans un premier temps, les mots

2. `http://www.olcalsace.org/fr/lexiques`
3. `http://www.wattwiller.com/?lang=de`
4. Il faut toutefois noter que tous les parlers alsaciens n'intègrent pas ces formes des politesse.

hors vocabulaire sont repérés, puis ces mots sont convertis vers leur forme normalisée. Cette conversion met en œuvre des mesures de similarité phonétique et graphique. Scherrer (2008) propose une approche à base de transducteurs stochastiques pour passer d'un dialecte de Suisse allemand (bernois) vers l'allemand standard, afin d'induire des lexiques bilingues. Hulden *et al.* (2011) visent à transformer des textes en dialecte Labourdin vers le basque standard. Ils utilisent pour cela des mécanismes d'apprentissage afin d'apprendre des transformations à partir d'un corpus parallèle. Deux méthodes sont proposées : la première extrait des règles phonologiques de remplacement en repérant les sous-chaines différentes dans les mots alignés d'un corpus parallèle ; la seconde est similaire à la première mais utilise également les contre-exemples pour chaque règle afin de déterminer les contextes les plus appropriés pour appliquer une règle de transformation. L'approche de Dasigi et Diab (2011) est légèrement différente puisqu'elle consiste à regrouper les variantes orthographiques d'un même lexème pour des dialectes de l'arabe standard. La méthode repose sur la classification non supervisée (*clustering*) et des mesures de similarité graphique (distance d'édition) et contextuelle (cosinus de vecteurs de co-occurrence).

Dans le contexte de l'étiquetage morphosyntaxique, il est possible d'entraîner un nouvel étiqueteur, adapté aux document à traiter. Cette approche a notamment été adoptée par Dipper (2011) qui décrit le processus d'entraînement du TreeTagger pour différentes versions d'un corpus de textes en moyen haut-allemand (mots normalisés ou non, dialecte). Cette approche n'est toutefois réalisable que si l'on dispose de textes préalablement annotés.

Lorsqu'au contraire les données annotées sont insuffisantes ou inexistantes, il est nécessaire d'avoir recours à des outils et ressources développés pour une langue apparentée et mieux dotée. Hana *et al.* (2011) décrivent un étiqueteur pour le vieux tchèque, basé sur deux stratégies. La première consiste à transformer un corpus annoté de tchèque moderne de manière à le rendre similaire à du vieux tchèque. Les transformations se basent sur un nombre limité de règles de changement graphique. Ce corpus transformé est alors utilisé pour entraîner un étiqueteur, qui sera appliqué à un corpus de vieux tchèque modernisé. Ce dernier corpus est à nouveau transformé sous sa forme originale en vieux tchèque, qui est alors annotée et permet donc d'entraîner un étiqueteur. La seconde stratégie consiste à utiliser une ressource morphosyntaxique pour le vieux tchèque, de manière à avoir une approximation des probabilités d'émission de l'étiqueteur. Une des idées sous-jacentes est qu'il suffit de produire cette ressource pour les mots les plus fréquents, qui auront l'impact le plus important sur l'apprentissage.

Nous reprenons certaines des idées proposées par Hana *et al.* : (i) nous transformons partiellement des textes alsaciens en allemand standard et (ii) nous appliquons ces transformations aux mots les plus fréquents, c'est-à-dire les mots grammaticaux. A la différence de Hana *et al.*, ces transformations ne peuvent pas être basées sur des changements de sons et de graphèmes dans le cas de l'alsacien. En effet, ces changements ne peuvent être prédits de manière systématique, comme le montre la Figure 1 qui représente les variantes graphiques que l'on peut trouver dans différents lexiques du Web pour deux lexèmes alsaciens. Ces deux exemples montrent l'extrême variabilité des scripturalisations : choix des voyelles (*u-i-é-ù-ü*, *a-e-i*, etc.), ajout / suppression de caractères (*g* ou *j* en fin de mots), alternances consonantiques (*ch-sch*, *d-t*).

Français	Allemand	Anglais	Variantes alsaciennes
cuisine	Küche	kitchen	Kuch, Kucha, Kische, Khésche, Kùch, Kücha, Kuche, Kiche, Kuchi
lundi	Montag	monday	Mondàà, Mantig, Mandig, Mondàà, Mondoe, Mondàj, Maandi, Mandi

FIGURE 1 – Variantes alsaciennes pour deux lexèmes trouvées dans divers lexiques sur le Web.

4 Approche

Nous avons collecté un corpus de cinq documents afin d'évaluer la performance d'étiqueteurs allemands pour des textes alsaciens :

- *Alsace* : un article tiré du journal local "L'Alsace", intitulé "Zìmlig beschta Frìnd" et datant du 18 février 2012 [5].
- *Dictionnaire multilingue* : phrases d'illustration tirée du *Dictionnaire comparatif multilingue Français - Allemand - Alsacien - Anglais* de Paul Adolf (2006). Ces phrases présentent la particularité d'être pour la plupart disponibles sous forme de diverses variantes graphiques. De plus, la traduction des phrases en allemand standard est également fournie (voir Figure 2).
- *Duttlenheim* : le résumé d'une pièce de théâtre jouée par la compagnie locale du village de Duttlenheim et datant de 2004 [6].
- *Hoflieferant* : une page de la pièce de théâtre "D'r Hoflieferant" de Gustave Stoskopf, datant de 1906.
- *Wikipedia* : un article de la Wikipédia alémanique, au sujet du musée alsacien de Strasbourg, récupérée le 30 octobre 2012 [7].

FR : Je n'ai pas assez d'argent	Ìch hà nìtt g'nüä Gald.
DE : Ich habe nicht genug Geld.	Ìch hàbb nìtt gnüej Gald.
EN : I don't have enough money.	Ìch hàbb nìtt genüej Gald.
	Ìch hàbb nìtt gnüej Gëld.
	Ìch hàbb nìtt genüej Gëld.

FIGURE 2 – Variantes graphiques de la même phrase dans le *Dictionnaire comparatif multilingue Français - Allemand - Alsacien - Anglais* de Paul Adolf (2006)

Ces textes ont été annotés de manière semi-automatique en utilisant un jeu simplifié d'étiquettes morphosyntaxiques afin d'obtenir une annotation de référence (voir Table 1). Nous avons tout d'abord traduit manuellement les mots en allemand approximatif puis avons corrigé les étiquettes fournies par le TreeTagger (Schmid, 1994) pour le *Dictionnaire multilingue* et le Stanford POS Tagger (Toutanova *et al.*, 2003) pour les autres textes (voir Figure 4). Une méthode d'analyse semi-automatique similaire a été utilisée par Giesbrecht et Evert (2009) pour obtenir leurs données de référence.

5. http://www.lalsace.fr/actualite/2012/02/18/zimlig-beschta-frind
6. http://theatreduttlenheim.free.fr/html/annee2004.htm
7. http://als.wikipedia.org/wiki/Els\%C3\%A4ssisches_Museum_\%28Stra\%C3\%9Fburg\%29

Les textes alsaciens ont été étiquetés par le TreeTagger et le Stanford POS Tagger en utilisant deux types de données :
– Textes originaux découpés manuellement en mots, sans autre pré-traitement ;
– Textes dont les mots grammaticaux et les auxiliaires ont été transposés dans leurs équivalents en allemand standard : *àwer* → *aber*, *fer* → *für*, *isch* → *ist*, etc. (voir Figure 3)

ALS	Brüchsch	kenn	Angscht	ze	han	for	mich	,	papa	.
TRANS	Brüchsch	keine	Angscht	zu	haben	für	mich	,	papa	.
DE-GLOSS	Brauchst	keine	Angst	zu	haben	für	mich	,	Vater	.
EN-GLOSS	Need	no	fear	to	have	for	me	,	dad	.
FR-GLOSS	As besoin	pas	peur	à	avoir	pour	moi	,	papa	.

FIGURE 3 – Exemple de transposition (TRANS) à partir de l'alsacien (ALS) et gloses en allemand (DE-GLOSS), en anglais (EN-GLOSS), et en français (FR-GLOSS).

Etiquette	**Description**	**Equivalence avec le jeu d'étiquettes du TreeTagger et du Stanford Tagger**
ADJ	Adjectif	ADJA, ADJD
ADP	Préposition / Postposition	APPO, APPR, APPRART, APZR
ADV	Adverbe	ADV
CARD	Nombre cardinal	CARD
CONJ	Conjonction	KOKOM, KON, KOUI, KOUS
DET	Déterminant	ART
FM	Mots en langue étrangère	FM
ITJ	Interjection	ITJ
N	Nom	NE, NN
PRN	Pronom	PAV, PROAV, PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWAV, PWAS, PWS
PRT	Particule	PTKANT, PTKNEG, PTKVZ, PTKZU, PTKA
V	Verbe	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP
$,	Virgule	$,
$.	Ponctuation de fin de phrase	$.
$(	Autre ponctuation	$(

TABLE 1 – Tableau d'équivalence des étiquettes morphosyntaxiques.

Le lexique utilisé pour la transposition contient 133 entrées. Ce lexique a été constitué manuellement, à partir du corpus. Les items ambigus, c'est-à-dire les mots grammaticaux pouvant avoir plusieurs traductions différentes, ont été éliminés de ce lexique. La Table 2 détaille le nombre de mots du corpus, ainsi que le nombre et le pourcentage de lemmes inconnus du TreeTagger, avant et après transposition.

Document	# tokens	# transpositions	# lemmes inconnus avant		# lemmes inconnus après	
Alsace	320	101	194	60,6%	104	32,5%
Dict-multi	1 180	353	794	67,3%	484	41,0%
Duttlenheim	166	37	60	36,1%	33	19,9%
Hoflieferant	230	39	74	32,2%	45	19,6%
Wikipedia	396	130	248	62,6%	139	35,1%

TABLE 2 – Statistiques du corpus.

5 Résultats

Pour l'évaluation, nous comparons les textes annotés de manière automatique avec nos annotations manuelles. Les étiquettes détaillées fournies par le TreeTagger et le Stanford Tagger sont automatiquement associées à l'étiquette équivalente dans notre jeu simplifié. Les résultats de l'évaluation sont détaillés dans les Tableaux 3 pour le TreeTagger et 4 pour le Stanford Tagger.

Corpus	Texte original	Texte original, mots non transposés uniquement	Texte transposé	Texte transposé, mots non transposés uniquement
Alsace	0,48	0,69	0,79	0,74
Dict. multi.	0,53	0,61	0,79	0,79
Duttlenheim	0,67	0,78	0,86	0,83
Hoflieferant	0,64	0,73	0,78	0,76
Wikipedia	0,50	0,71	0,80	0,75

TABLE 3 – Précision du TreeTagger

La précision du TreeTagger pour les données originales est plutôt basse et se situe sous la barre de 0,70. Après la transposition, la précision augmente, avec un minimum de 0,78. Nous avons également calculé la précision uniquement pour les mots non transposés, afin de voir si la performance de l'étiquetage augmente uniquement pour les mots transposés ou s'il y a également une incidence sur les autres mots. Nous constatons que les mots non transposés sont également mieux analysés : par exemple, la précision augmente de 0,71 à 0,75 pour les mots non transposés du texte issu de Wikipédia.

Corpus	Texte original	Texte original, mots non transposés uniquement	Texte transposé	Texte transposé, mots non transposés uniquement
Alsace	0,56	0,74	0,86	0,83
Dict. multi	0,62	0,69	0,87	0,87
Duttlenheim	0,77	0,85	0,89	0,88
Hoflieferant	0,68	0,76	0,83	0,80
Wikipedia	0,53	0,70	0,85	0,80

TABLE 4 – Précision du Stanford Tagger.

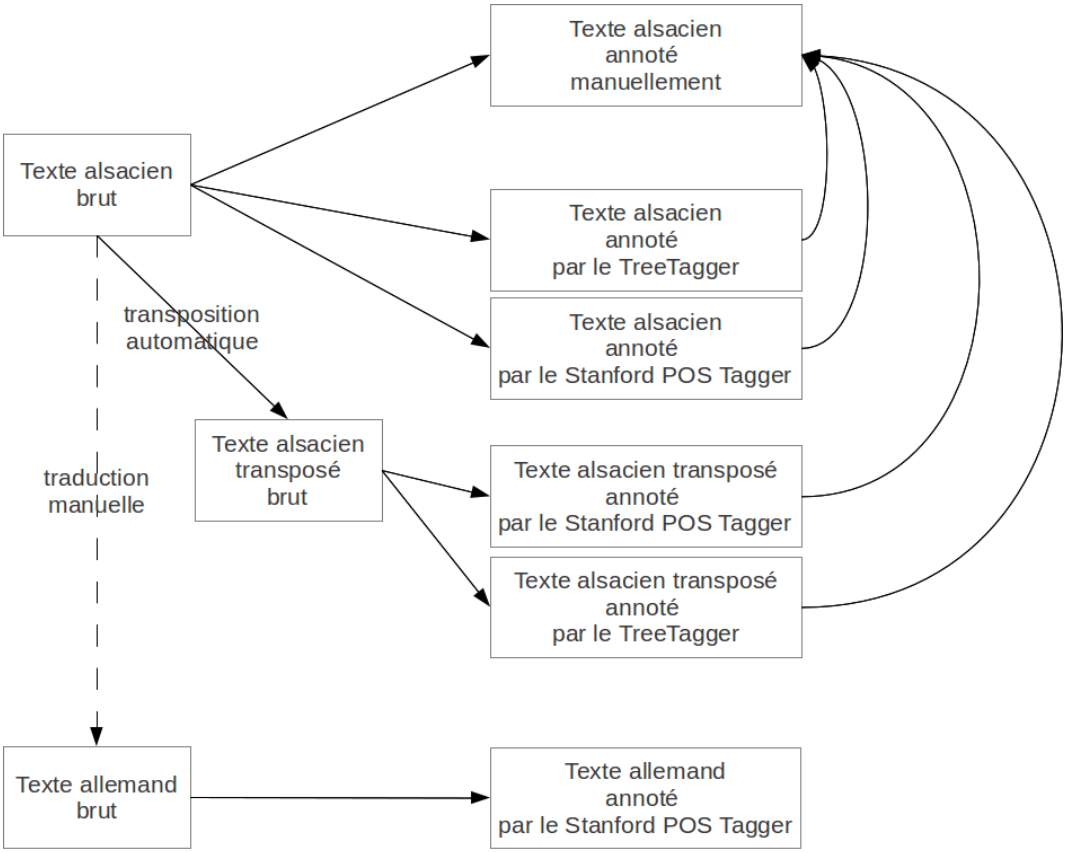

FIGURE 4 – Étiquetage morphosyntaxique de textes alsacien, avec ou sans transposition.

Le Stanford Tagger a des résultats légèrement supérieurs à ceux du TreeTagger, probablement en raison du modèle sous-jacent. Comme pour le TreeTagger, les résultats sont meilleurs pour les textes transposés : par exemple, la précision augmente de 0,53 à 0,85 pour les textes tirés de Wikipédia. La précision pour les mots non transposés est également améliorée.

Nous avons analysé les résultats de manière plus détaillée, en nous focalisant sur certaines différences entre l'alsacien et l'allemand.

Absence de génitif L'allemand standard marque la possession en utilisant soit le génitif saxon, soit les compléments du nom introduits par la préposition "von" (de) suivie du datif. En alsacien, c'est cette dernière construction de datif possessif qui est privilégiée. On trouve ainsi "Wien ìsch d'*Hauptstàdt vun Eschtrich*." [8] en alsacien, alors qu'en allemand standard on peut avoir "Wien ist *Österreichs Hauptstadt*." (génitif saxon) ou "Wien ist die *Hauptstadt von Österreich*." (source :

8. Vienne est la capitale de l'Autriche.

Dictionnaire multilingue). Nous avons repéré six exemples de ce type dans notre corpus. Dans tous les cas, la construction est bien analysée. Cette construction se retrouvant en allemand standard, on peut comprendre qu'elle ne pose aucune difficulté aux analyseurs.

Formes périphrastiques avec l'auxiliaire "düen" Cet auxiliaire est couramment utilisé pour le présent et le conditionnel et fonctionne de manière similaire à l'auxiliaire "do" an anglais. Les formes périphrastiques alternent au présent avec les formes simples du verbe. Ces constructions ont été décrites de manière détaillée par Kleiber et Riegel (1998, 2005). Pour reprendre un des exemples de Kleiber et Riegel (2005, p. 171), on peut trouver la forme périphrastique "Gewiss, e Màmme düet sich fer ihri Kinder ufopfere" [9] ou la forme simple, sans l'auxiliaire "düen" : "Gewiss, e Màmme opfert sich fer ihri Kinder uf." La construction périphrastique ne se retrouve pas en allemand standard. Nous n'avons trouvé que deux exemples de cette construction au présent dans notre corpus, aucune au conditionnel. Si l'auxiliaire "düen" est dans tous les cas analysé comme un verbe, à la fois par le TreeTagger et le Stanford Tagger, l'analyse échoue pour le verbe principal qui est rejeté en fin de proposition. Cette construction est a priori inconnue des analyseurs entraînés pour l'allemand standard, ce qui peut expliquer les problèmes d'étiquetage pour le reste de la phrase. Il faudrait toutefois disposer d'un corpus plus conséquent pour pouvoir tirer des conclusions plus précises.

Utilisation du passé composé à la place de l'imparfait et du passé simple On ne trouve qu'un temps pour exprimer le passé en alsacien : le passé composé ou parfait (Schimpf et Muller, 1998; Keck *et al.*, 2010). Nous avons repéré huit exemples de cette construction dans notre corpus. Dans la majorité des cas, l'auxiliaire et le participe sont correctement analysés : en effet, le parfait est un temps qui existe également en allemand standard et qui peut donc être reconnu par les outils.

Omission du pronom personnel sujet "du" à la deuxième personne du singulier Ce pronom est fréquemment omis dans les dialogues à l'oral (voir Figure 3). Nous avons trouvé quatre exemples d'omission du pronom "du" dans deux documents du corpus. Dans tous les cas, les deux analyseurs ont donné une étiquette erronée au verbe : nom, adjectif ou mot en langue étrangère.

Phénomènes de code-switching On observe régulièrement des incursions du français en alsacien. Par exemple, dans l'extrait de la pièce de théâtre "D'r Hoflieferant" de notre corpus, on trouve la réplique suivante : "Toi, tais-toi !". Cette pratique est toutefois essentiellement liée à l'oral et ce phénomène est donc fréquent dans les médias alsaciens comme la télévision régionale (Erhart, 2012). Le code-switching peut concerner différents types d'unités : mots, syntagmes, phrases complètes. Dans notre corpus, nous avons trouvé quelques exemples de code switching uniquement dans l'extrait de la pièce de théâtre "D'r Hoflieferant" ("Eh bien", "Toi, tais-toi !"). Dans les autres textes, le français n'est utilisé que pour nommer des entités difficiles à traduire (film, association, livre). Le problème se pose toutefois dès l'annotation manuelle : comment étiqueter ce type de phénomènes ? Faut-il donner l'étiquette appropriée en français, ou alors l'étiquette FM (Foreign Material : mots en langue étrangère) ? C'est cette dernière solution que nous avons privilégiée lors de l'annotation.

9. Sûr, une maman se sacrifie pour ses enfants.

L'analyse des principales différences entre l'alsacien et l'allemand standard tend à montrer qu'il faudrait prétraiter certaines constructions, notamment les formes périphrastiques avec l'auxiliaire "düen" et l'absence de pronom personnel sujet à la deuxième personne du singulier, avant analyse par les étiqueteurs. En particulier, l'absence de pronom personnel sujet devrait pouvoir être résolue par l'insertion du pronom avant étiquetage. Les formes périphrastiques avec l'auxiliaire "düen" constituent un autre exemple de difficulté, mais dont la résolution semble moins difficile, en raison du déplacement du verbe principal en fin de proposition.

6 Conclusion et perspectives

L'analyse morphosyntaxique est une étape importante des applications de TAL Nous avons présenté une méthode pour l'analyse morphosyntaxique des dialectes alsaciens qui repose sur une transposition partielle vers l'allemand standard. Les résultats montrent qu'une approche simple qui consiste à transposer uniquement les mots grammaticaux conduit à des performances d'étiquetage améliorées.

Il reste toutefois une grande marge de progression, de manière à atteindre des niveaux de précision au delà des 95%, correspondant aux niveaux de performance attendus pour la tâche.

Une première piste que nous souhaitons explorer est celle de l'identification des cognats en alsacien et en allemand. En effet, les mots pleins alsaciens ont souvent une forte ressemblance graphiques avec leurs équivalents en allemand standard. Ces cognats pourraient ensuite être transposés de manière automatique et améliorer ainsi l'étiquetage des mots pleins.

Une autre possibilité consisterait à intégrer les informations sur la catégorie disponible dans des lexiques alsaciens existants, disponibles sur le Web. Ces informations pourraient permettre de corriger les erreurs de l'analyse automatique. Il faudra toutefois dans ce cas pouvoir résoudre les problèmes de variation graphique car les mots des documents ne se trouvent pas forcément avec la même graphie dans des lexiques rédigés par des auteurs différents.

Enfin, le prétraitement de certaines difficultés syntaxiques identifiées pourrait également conduire à une amélioration des résultats.

Remerciements

Nous remercions Paul Adolf pour nous avoir fourni une version numérique de son dictionnaire multilingue et Pascale Erhart pour sa relecture. Ces travaux ont bénéficié du soutien du conseil scientifique de l'université de Strasbourg dans le cadre du projet COPAL (COrpus Parallèles pour l'ALsacien).

Références

ADOLF, P. (2006). *Dictionnaire comparatif multilingue : français-allemand-alsacien-anglais*. Midgard, Strasbourg, France.

BITSCH, R. et MATZEN, R. (2004). De elsässisch Dico. CD, Editions l'Ami hebdo, Strasbourg et site Web : http://www.ami-hebdo.com/elsadico/index.php.

DASIGI, P. et DIAB, M. (2011). CODACT : Towards Identifying Orthographic Variants in Dialectal Arabic. *In Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 318–326, Chiang Mai, Thailand.

DIPPER, S. (2011). Morphological and Part-of-Speech Tagging of Historical Language Data : A Comparison. *Journal for Language Technology and Computational Linguistics*, 26(2):25–37.

ERHART, P. (2012). *Les dialectes dans les médias : quelle image de l'Alsace véhiculent-ils dans les émissions de la télévision régionale ?* Thèse de doctorat, Université de Strasbourg.

GIESBRECHT, E. et EVERT, S. (2009). Is Part-of-Speech Tagging a Solved Task ? An Evaluation of POS Taggers for the German Web as Corpus. *In Proceedings of the 5th Web as Corpus Workshop (WAC5)*.

HANA, J., FELDMAN, A. et AHARODNIK, K. (2011). A low-budget tagger for Old Czech. *In Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '11)*, pages 10–18.

HUCK, D., BOTHOREL-WITZ, A. et GEIGER-JAILLET, A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions : Insights and Views of Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, page 13–100.

HUDLETT, A. et GROUPE D'ETUDES ET DE RECHERCHES INTERDISCIPLINAIRES SUR LE PLURILINGUISME EN ALSACE ET EN EUROPE (2003). *Charte de la graphie harmonisée des parlers alsaciens : système graphique GRAPHAL - GERIPA*. Centre de Recherche sur l'Europe littéraire (C.R.E.L.), Mulhouse, France.

HULDEN, M., ALEGRIA, I., ETXEBERRIA, I. et MARITXALAR, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. *In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 39–48.

KECK, B., DAUL, L. et KRETZ, P. (2010). *L'alsacien pour les nuls*. Pour les nuls (Éd. de poche), ISSN 1625-0486. Paris, France.

KLEIBER, G. et RIEGEL, M. (1998). Grammaticalisation et auxiliaire modal : L'énigme de düen en alsacien. *In Travaux de linguistique*, volume 36, pages 161–173.

KLEIBER, G. et RIEGEL, M. (2005). Les périphrases düen + verbe à l'infinitif en alsacien : Un auxiliaire modal à tout... faire. *In Les Périphrases Verbales*, volume 25 de *Linguisticae Investigationes Supplementa*, pages 171–184. John Benjamins Publishing Company.

MATZEN, R., DAUL, L. et LAZÉ, C. (1997). *Mon premier dictionnaire français-alsacien en images*. Gisserot J.P, Paris, France.

MATZEN, R., DAUL, L. et WEHRLING, Y. (2004). *Wie geht's ? : le dialecte à la portée de tous*. La Nuée bleue, Strasbourg, France.

MOSQUERA, A., LLORET, E. et MOREDA, P. (2012). Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. *In Proceedings of the Worshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

OLCA / EDINSTITUT (2012). Étude sur le dialecte alsacien. [En ligne, visité le 7 mai 2013 : http://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf].

SCHERRER, Y. (2008). Transducteurs à fenêtre glissante pour l'induction lexicale. *In Actes de RECITAL 2008*, Avignon.

SCHIMPF, J.-P. et MULLER, R. (1998). *Parlons alsacien*. L'Harmattan, Paris, France.

SCHMID, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, page 44–49.

TOUTANOVA, K., KLEIN, D., MANNING, C. D. et 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, page 173–180.

ZEIDLER, E. et CREVENAT-WERNER, D. (2008). *Orthographe alsacienne : bien écrire l'alsacien de Wissembourg à Ferrette*. J. Do Bentzinger, Colmar, France.