Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique

Delphine Bernhard, Lucie Steiblé

Quand l'oral se fait entendre à l'écrit : alignement de lexiques en l'absence de normalisation graphique

Delphine Bernhard Lucie Steiblé
LiLPa, Université de Strasbourg, 14 rue Descartes, F-67084 Strasbourg Cedex
dbernhard@unistra.fr, steiblelucie@gmail.com

Résumé. Les dialectes parlés en Alsace, que l'on regroupe communément sous l'appellation « alsacien », se caractérisent par un manque de ressources numériques, qu'il s'agisse de corpus ou de lexiques. Par ailleurs, les dialectes d'Alsace sont avant tout des langues parlées dans la vie quotidienne, et leur graphie n'est pas encore complètement codifiée : une unité lexicale peut donc avoir plusieurs graphies. Ceci est un défi majeur pour la construction de ressources lexicales, car les variantes orthographiques d'une entrée lexicale doivent être identifiées. Cet article décrit une méthode pour la construction de lexiques bilingues français-alsacien qui vise à résoudre ce problème. Elle consiste à aligner des lexiques bilingues existants, en utilisant l'algorithme phonétique *Double Metaphone* afin de détecter les variantes. En outre, les mots alsaciens sont automatiquement reliés aux entrées de BabelNet, un réseau sémantique multilingue (Navigli & Ponzetto, 2012). La méthode d'alignement des lexiques atteint de bons niveaux de précision, ce qui permet la construction automatique de ressources, avec une intervention humaine limitée à quelques corrections. La principale originalité de ce travail est qu'il ne vise pas la normalisation, qui consisterait à transformer les variantes orthographiques en une norme donnée. Par ailleurs, au lieu d'une simple liste de mots bilingues, les liens vers BabelNet fournissent une couche sémantique supplémentaire reliant les entrées à des sens lexicaux. Enfin, nous utilisons les alignements obtenus pour faire une comparaison entre observations réalisées sur la langue orale et les graphies relevées dans les lexiques.

Abstract.

From Spoken to Written : Lexicon Alignment in the Absence of an Orthographic System

The dialects spoken in Alsace, which are commonly grouped under the name "Alsatian", are characterized by a lack of digital resources, whether corpora or lexicons. Moreover, the Alsatian dialects are primarily spoken in everyday life, and their spelling is not yet completely codified : a given lexical unit can have multiple spellings. This is a major challenge for building lexical resources because alternative spellings of a lexical entry must be identified. This article describes a method for building French-Alsatian bilingual lexicons that aims to solve this problem. It consists in aligning existing bilingual lexicons, using the phonetic algorithm *Double Metaphone* to detect variants. In addition, the Alsatian words are automatically linked to entries in Babelnet, a multilingual semantic network (Navigli & Ponzetto, 2012). The lexicon alignment method achieves good levels of precision, which allows the automatic construction of resources with limited human intervention. The main originality of this work is that it does not target normalization, which would transform the spelling variants to a given standard. Moreover, instead of a simple list of bilingual words, links to Babelnet provide an additional semantic layer which connects the lexical items to senses. Finally, we use the alignments obtained to perform a comparison between phenomena observed in the spoken language and the written forms found in the lexicons.

Mots-clés : alignement de lexiques, variantes orthographiques, alsacien, BabelNet.

Keywords: lexicon alignment, spelling variants, Alsatian, BabelNet.

1 Introduction

La constitution de ressources lexicales est l'une des étapes obligatoires pour le développement d'outils du traitement automatique des langues (TAL). Cette tâche peut sembler triviale, mais dans les faits elle peut être rendue complexe par l'absence de convention orthographique, dans le cas des langues orales ou faiblement normalisées à l'écrit. Dans cet article, nous nous concentrons sur le cas précis des dialectes parlés en Alsace. Les dialectes alsaciens appartiennent aux groupes alémanique et francique (d'où l'utilisation du pluriel pour « les dialectes ») et se rapprochent de fait des dialectes parlés dans les régions limitrophes d'Allemagne et de Suisse (Huck *et al.*, 2007). Selon une étude récente, 43% de la

population alsacienne parle encore le dialecte régional (OLCA / EDinstitut, 2012). Cependant, la proportion de locuteurs alsaciens diminue régulièrement depuis les années 1960, au profit de la langue française. En outre, les dialectes alsaciens sont avant tout des langues parlées dans la vie quotidienne et leur graphie n'est donc pas encore complètement codifiée, ce qui complique toute tentative de développement de ressources et outils pour le traitement automatique des langues. Il y a eu quelques initiatives récentes visant à définir des conventions orthographiques pour l'alsacien. Le système ORTHAL (Zeidler & Crévenat-Werner, 2008) se réfère à l'orthographe allemand standard tout en permettant la transcription des phénomènes qui sont spécifiques aux dialectes alsaciens. Le système GRAPHAL-GERIPA (Hudlett & Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) définit un ensemble de règles pour aller du son au graphème. Cependant, il est difficile d'estimer la diffusion effective et l'utilisation de ces systèmes. Par ailleurs, ils accueillent la variation pour les différentes variantes géolinguistiques rencontrées en Alsace et ne garantissent donc pas une orthographe unique pour la forme d'un mot donné. Enfin, ils ne s'appliquent pas aux écrits plus anciens.

Du point de vue du traitement automatique des langues, il n'existe pas à l'heure actuelle de lexique informatisé pour l'alsacien (Leixa *et al.*, 2014), qui pourrait être utilisé pour diverses applications. Un lexique de ce type devrait idéalement comporter différents types d'informations : liste de formes attestées – pour la reconnaissance optique de caractères –, catégories grammaticales – pour l'étiquetage morphosyntaxique –, et traduction en français – pour l'aide à la lecture, les lecteurs étant souvent plus à l'aise en français qu'en alsacien –.

Pour résumer, les dialectes alsaciens posent plusieurs défis importants pour le TAL :
– Il n'existe pas de convention orthographique utilisée de manière systématique à l'écrit ;
– Le dialecte alsacien est en fait un continuum de dialectes, avec des variantes géolinguistiques tant au niveau lexical qu'au niveau de la prononciation ;
– Il n'y a pas encore de ressources lexicales numériques pour l'alsacien.

Dans cet article, nous présentons une méthode de construction de ressources lexicales numériques pour les dialectes alsaciens qui consiste à aligner plusieurs lexiques français-alsacien bilingues. Par ailleurs, nous souhaitons nous insérer dans le mouvement récent appelé *Linked Open Data* (LOD) qui vise a constituer de grandes ressources linguistiques multilingues inter-reliées. Toutes les langues ne sont pas encore couvertes, en raison du manque d'informations disponibles pour les langues faiblement dotées. Cependant, le LOD constitue une formidable opportunité pour accroître la visibilité des langues minoritaires ou régionales, si tant est qu'elles peuvent y être incorporées. Par ailleurs, le LOD permet d'accéder à de nombreuses ressources lexicales et sémantiques qui pourraient bénéficier au traitement automatique des dialectes alsaciens (définitions, liens sémantiques, etc.). Nous proposons donc d'associer des mots alsaciens à Babelnet, un réseau sémantique multilingue relié au *Linguistic Linked Open Data* (Navigli & Ponzetto, 2012).

Notre méthode s'appuie sur les observations suivantes :
– Les conventions orthographiques adoptées dans les lexiques alsaciens sont très variables, et donc la forme de citation d'une unité lexicale en alsacien peut être représentée par plusieurs graphies [1]. Ces diverses formes peuvent être considérées comme des variantes car elles sont proches phonétiquement et correspondent à la même unité lexicale. Par ailleurs, la plupart des formes de mots alsaciens sont semblables à leur traduction en allemand standard et même parfois en anglais.
– Une unité lexicale en alsacien peut avoir plusieurs traductions en français. Ces traductions peuvent être des synonymes, mais également correspondre à différents sens du mot. Cela complique l'alignement entre les deux langues, ainsi que l'alignement avec une ressource comme Babelnet, dont les unités correspondent à des concepts.

Nous abordons ces questions comme suit :
– Nous proposons d'utiliser une variante d'un algorithme phonétique, *Double Metaphone*, adapté aux dialectes alsaciens, afin d'identifier les variantes orthographiques. L'algorithme prend également en compte l'orthographe de l'allemand standard et de l'anglais afin de trouver des mots apparentés dans ces diverses langues.
– Nous utilisons des ressources externes pour obtenir des informations sur les synonymes dans la langue française et des traductions en allemand et en anglais.

L'article est organisé comme suit : la section suivante récapitule les travaux antérieurs sur l'identification des variantes orthographiques et l'alignement des ressources lexicales. La section 3 détaille les ressources lexicales utilisées dans notre travail. Les méthodes d'alignement sont présentées dans la section 4, qui comprend également une évaluation des alignements obtenus sur la base d'un dictionnaire multilingue publié. Enfin, nous faisons une comparaison entre observations faites sur la langue orale et les graphies relevées à l'écrit, sur la base des alignements obtenus de manière automatique.

1. Dans cet article, la notion d'unité lexicale renvoie à un lexème, dans le sens de Bauer (2003) : « Un lexème est un mot du dictionnaire, une unité abstraite du vocabulaire. Il est réalisé (...) par des mots-formes (*word forms*), de telle sorte que le mot-forme représente le lexème et toutes les flexions (...) qui sont nécessaires. (...) La forme de citation d'un lexème est le mot-forme appartenant au lexème qui est conventionnellement choisi pour nommer le lexème dans les dictionnaires et autres. » (notre traduction)

2 État de l'art

2.1 Identification de variantes orthographiques

Le problème des graphies non standard se rencontre pour différents types de textes, comme par exemple les données issues du Web (en particulier le Web 2.0), les textes correspondant à des états anciens de la langue, et les textes écrits dans des langues qui sont principalement orales et qui n'ont pas de système orthographique. La grande majorité des méthodes consiste à normaliser vers une langue cible, c'est-à-dire, transformer une variante minoritaire en une norme donnée. Par exemple, Scherrer (2008) utilise la distance orthographique de Levenshtein et des transducteurs stochastiques afin de transformer les formes dialectales du Suisse allemand en allemand standard. Hulden *et al.* (2011) présentent deux méthodes qui apprennent automatiquement les transformations d'une forme dialectale vers la forme standard en utilisant un corpus parallèle pour la langue basque et le dialecte basque labourdin. La première méthode s'appuie sur un outil existant, lexdiff (Almeida *et al.*, 2010), qui détecte les différences orthographiques. Ces différences sont transformées en règles de remplacement et compilées sous forme de transducteurs. La deuxième méthode est inspirée par la PLI (programmation logique inductive) et essaie de sélectionner le meilleur ensemble de règles de remplacement, en utilisant des exemples à la fois positifs et négatifs. Dans le contexte de la traduction automatique statistique pour la paire de langues arabe-anglais, Salloum & Habash (2011) décrivent une méthode à base de règles pour générer des paraphrases de l'arabe dialectal en arabe standard. Pour les variantes linguistiques historiques, Porta *et al.* (2013) proposent une méthode pour mettre en correspondance les formes historiques avec leurs homologues modernes. L'approche est basée sur un transducteur de Levenshtein et un transducteur linguistique encodant des règles de réécriture des sons.

Dans l'ensemble, les méthodes de normalisation considèrent que les dialectes ou les formes de mots historiques sont non-standard et doivent être transformées dans des formes contemporaines d'une langue bien dotée en ressources. Même si cette hypothèse est logique dans de nombreux cas, notamment pour faciliter le traitement ultérieur par les outils de TAL, ce n'est pas la seule solution. Par exemple, Dasigi & Diab (2011) présentent un algorithme de clustering qui vise à regrouper les variantes orthographiques dialectales qui correspondent au même mot. Ce type d'approche est particulièrement pertinent dans notre contexte, car il ne normalise pas nécessairement les variantes dialectales. En effet, la normalisation n'est pas souhaitable dans le cas des dialectes alsaciens pour plusieurs raisons. Tout d'abord, il n'y a pas consensus sur la norme de scripturalisation des dialectes alsaciens et il est donc difficile de décider quelle forme doit prévaloir. En outre, même si les dialectes alsaciens sont étroitement liés à l'allemand, qui pourrait être considéré comme le standard, il existe un certain nombre de différences lexicales (notamment des emprunts au français (Matzen, 1985)) et syntaxiques (voir par exemple (Kleiber & Riegel, 1998)) qui doivent être prises en compte. Ajouté à cela, considérer l'allemand comme la norme pour les dialectes alsaciens est une question très sensible du point de vue sociolinguistique, voire politique [2]. Compte tenu de toutes ces raisons, notre méthode ne cherche pas à normaliser les variantes graphiques mais conserve leur diversité en considérant des groupes (ou *clusters*) de variantes comme des entrées du lexique.

2.2 Alignement de ressources lexicales

L'objectif principal de notre travail n'est pas seulement d'identifier les variantes orthographiques, mais aussi d'aligner les entrées issues de différents lexiques bilingues et de mettre en correspondance ces alignements avec les concepts d'un réseau sémantique. Beaucoup de travaux ont été consacrés récemment à l'alignement de ressources collaboratives, comme Wikipedia, et de bases de connaissances lexicales plus classiques, comme WordNet. Niemann & Gurevych (2011) détaillent une méthode pour l'alignement des sens des entrées de WordNet et Wikipedia, qui a ensuite été utilisée pour la ressource lexicale sémantique UBY (Gurevych *et al.*, 2012). La méthode repose sur l'apprentissage automatique afin de classer les alignements comme valides ou non valides. La similitude des sens candidats est calculée sur la base d'une représentation "sac de mots" des sens, puis fournie au classifieur. Pour la ressource UBY, des alignements translingues sont induits de la même manière, en traduisant tout d'abord automatiquement les représentations textuelles des sens. Navigli & Ponzetto (2012) proposent une méthode pour relier les pages Wikipédia aux synsets de WordNet, qui a été utilisée pour la construction de la ressource BabelNet. La méthode applique plusieurs stratégies en séquence. En particulier, elle réutilise une technique proposée dans le cadre de la désambiguïsation lexicale qui consiste à définir un contexte de désambiguïsation pour chaque page Wikipedia et chaque sens dans WordNet. Le contexte utilisé est un ensemble de mots obtenus à partir des informations fournies dans les ressources (par exemple, les noms des pages, les liens, les redirections et les catégories dans Wikipedia ; les synonymes, hyperonymes / hyponymes, gloses dans WordNet). Un score de similarité

2. Il est vrai que cet argument n'a pas beaucoup de poids du point de vue pratique pour les outils de TAL, mais il s'ajoute aux autres.

peut alors être calculé sur la base de ce contexte.

Quand il n'y a aucune ressource lexicale dans une langue donnée, la traduction automatique des ressources d'une autre langue est souvent la meilleure option, en terme de coût de construction. Dans ce cas, une ressource existante est étendue avec les lexicalisations d'une autre langue et la structure est conservée. Le WOLF (Wordnet Libre du Français) a été construit par Sagot & Fišer (2008) en utilisant le Princeton WordNet et plusieurs ressources multilingues. Les principales hypothèses qui sous-tendent leur approche sont que les différents sens d'un mot ambigu dans une langue correspondent souvent à différentes traductions dans une autre langue, et les mots qui sont traduits par le même mot dans une autre langue ont souvent des significations similaires. Ils appliquent ces idées en recueillant un lexique multilingue comportant 5 langues à partir d'un corpus parallèle et en assignant le synset le plus probable à chaque entrée du dictionnaire, en s'appuyant sur les intersections entre les synsets associés à chaque mot non-français du lexique dans le Princeton WordNet ou dans les wordnets du projet BalkaNet. Hanoka & Sagot (2012) ont étendu la ressource WOLF en utilisant une nouvelle approche qui s'appuie sur un grand grand graphe de synonymes et de traductions construit à partir de Wikipedia et de Wiktionary. Le graphe est interrogé avec les littéraux de wordnets multilingues alignés pour obtenir le meilleur candidat de traduction, en utilisant à la fois la traduction et des relations de traduction inverse.

Dans notre travail, nous appliquons également l'idée d'étendre une ressource lexicale sémantique existante (BabelNet) avec des lexicalisations d'une autre langue, à savoir l'alsacien. Nous utilisons le français comme langue pivot pour obtenir une mise en correspondance entre les variantes alsaciennes et Babelnet. En outre, nous exploitons la proximité entre l'alsacien, l'allemand et l'anglais afin d'enrichir les vecteurs de caractéristiques et effectuer la désambiguïsation.

3 Ressources

3.1 Lexiques bilingues français-alsacien

Nous avons récupéré trois lexiques bilingues français-alsacien disponibles sur le Web :
– OLCA : les lexiques produits par l'OLCA (Office pour la Langue et la Culture d'Alsace) [3]. Ces lexiques sont spécifiques à des domaines particuliers (l'artisanat, l'automobile, la bière, les courses, l'équitation, le football, les livres, la médecine, la météo, la nature, la petite enfance, la pêche, la pharmacie, le vélo, la vigne) et fournissent dans certains cas des variantes pour les départements alsaciens du Bas-Rhin et du Haut-Rhin ;
– WKT : un lexique extrait d'une page utilisateur du Wiktionnaire [4] ;
– ACPA : un lexique bilingue disponible sur la page Web d'une association locale [5].
Les lexiques contiennent essentiellement des lemmes, ainsi que quelques expressions. Par ailleurs, ces lexiques, bien que numériques, ne sont pas disponibles dans un format standard. Ils ont été pré-traités avec des analyseurs spécifiques pour extraire les paires de mots français-alsacien. Lorsqu'elles sont disponibles, les informations sur la partie du discours sont conservées [6]. Sinon, nous avons utilisé deux heuristiques pour trouver la partie du discours : (a) utilisation du TreeTagger français (Schmid, 1994) pour obtenir la catégorie des mots simples français [7] ; (b) pour les noms, vérification de la présence d'un déterminant à côté de la forme alsacienne.

La Table 1a répertorie le nombre d'entrées dans la partie française des lexiques après pré-traitement. Le tableau montre que la couverture des différentes parties du discours est inégale, et que les lexiques se concentrent principalement sur les noms, les verbes et les adjectifs.

Les lexiques suivent différentes conventions orthographiques comme le montre la Table 1b [8], qui énumère les traductions trouvées dans les lexiques pour plusieurs mots. Beaucoup de traductions dans la table sont en fait des variantes graphiques du même mot alsacien (par exemple "Kràb" et "Kràpp"). Toutefois, ces variantes graphiques peuvent être très dissemblables, si on ne considère que les caractères utilisés.

3. http://www.olcalsace.org/
4. http://fr.wiktionary.org/wiki/Utilisateur:Laurent_Bouvier/alsacien-fran\%C3\%A7ais
5. Compilé par André Nisslé, http://culture.alsace.pagesperso-orange.fr/dictionnaire_alsacien.htm
6. Nous avons utilisé la liste de catégories suivante : verbe, adjectif, adverbe, préposition, locution, conjonction, pronom, interjection, nom propre, participe passé, déterminant, abréviation, nom (féminin, masculin, neutre, pluriel).
7. Nous utilisons le module TreeTaggerWrapper de Laurent Pointal disponible à http://perso.limsi.fr/pointal/dev:treetaggerwrapper. Nous préférons l'utilisation de l'étiqueteur à un lexique morphosyntaxique du français car il donne l'étiquette la plus probable, alors que dans un lexique morphosyntaxique toutes les étiquettes sont équiprobables.
8. Voir également la Table 4, p. 9.

	OLCA	WKT	ACPA
adjectif	224	122	1 898
adverbe	14	49	295
déterminant	1	20	15
nom	5 106	1 049	15 770
participe passé	63	59	476
pronom	1	38	47
verbe	445	292	3 017
catégorie indéterminée	943	393	2 015
TOTAL	6 797	2 022	23 533

(a) Nombre de mots français dans les lexiques français-alsacien.

Français	**corbeau**	**jambe(s)**	**grenier**
Anglais	crow	leg	attic
Allemand	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	Hejbodde Dàchstüel **Behn**

(b) Exemples de traductions trouvées dans les lexiques. Les variantes qui se retrouvent de manière identique dans au moins deux lexiques sont en gras.

En plus des lexiques bilingues, nous avons également utilisé deux réseaux sémantiques : JeuxDeMots et Babelnet.

3.2 JeuxDeMots

JeuxDeMots (Lafourcade, 2007) est un réseau lexical français disponible gratuitement et construit à l'aide de jeux en ligne [9]. Nous avons utilisé la version datée du 12 Juin 2014 [10], qui contient 178 569 occurrences de la relation de synonymie (le réseau contient également de nombreux autres types de relations, par exemple association, domaine, hyperonymie, hyponymie, etc.). JeuxDeMots est utilisé pour relier des entrées dont les traductions en français sont synonymiques, comme par exemple 'drãckïg' - *sale* et 'trackig' - *malpropre*.

3.3 BabelNet

BabelNet (Navigli & Ponzetto, 2012) est un réseau sémantique multilingue, qui intègre des données issues de WordNet et Wikipedia, entre autres. Babelnet est composé de synsets, qui correspondent à des concepts avec des lexicalisations en plusieurs langues. Les lexicalisations multilingues ont été obtenues soit grâce aux liens inter-langues de Wikipédia ou à la traduction automatique. Nous avons utilisé la version 2.5 de Babelnet [11].

4 Alignement des lexiques

Dans cette section, nous présentons notre méthode pour aligner les lexiques. Elle repose sur une adaptation de l'algorithme phonétique *Double Metaphone* aux dialectes alsaciens.

4.1 *Double Metaphone* pour les dialectes alsaciens

Compte tenu de l'absence de convention orthographique, ainsi que des différences dues à la variation géolinguistique, il n'est pas possible d'aligner les entrées issues de différents lexiques en fonction de la similarité graphique des formes (Considérons par exemple "Grâb" et "Kràbb" de la table 1b, qui n'ont que deux caractères communs : 'r' et 'b'). Afin de résoudre ce problème, nous avons développé un algorithme *Double Metaphone* pour les dialectes alsaciens. *Double Metaphone* (Phillips, 2000) a été proposé à l'origine pour la recherche d'information, afin de trouver des noms orthographiés

9. Voir `http://www.jeuxdemots.org`
10. Disponible sur `http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR`
11. Disponible à `http://www.babelnet.org`

différemment, mais faisant référence à la même entité. *Double Metaphone* appartient à la classe des algorithmes phonétiques, car il transforme la chaîne d'entrée en une clé qui est identique pour les mots qui sont prononcés d'une manière similaire. Par exemple, la clé metaphone est STFN pour les trois noms suivants : "Stephan", "Steven" et "Stefan". Afin de prendre en compte les ambiguïtés, *Double Metaphone* retourne en fait deux clés dans certains cas. *Double Metaphone* a par exemple été utilisé pour la normalisation de textes du Web 2.0 (Mosquera *et al.*, 2012).

Les transformations *Double Metaphone* implémentées pour l'alsacien sont basées sur des transformations proposées à l'origine pour l'anglais [12] et une analyse de nos lexiques. Nous avons constitué un jeu de test comprenant 141 formes alsaciennes et les clés métaphones attendues, avec diverses variantes de la même unité lexicale, afin de vérifier que les clés métaphones sont bien identiques dans ce cas. Nous avons également pris en compte l'allemand standard, afin d'obtenir des clés identiques pour les mots apparentés (cognats) allemands et alsaciens. La table 2 donne quelques exemples des clés metaphone obtenues pour plusieurs mots alsaciens et allemands.

Mot	Traduction en français	Clé metaphone 1	Clé metaphone 2
Schloofwàga	wagon-lit	XLFVK	XLFVY
Schlofwaawe		XLFVV	XLFVY
Rüejdàà	jour de repos	RT	/
Rüaijtààg		RTK	RT
beschtadiga	confirmer	PXTTK	PXTTY
Uffschtànd	insurrection	AFXTNT	/
Iwereinsschtimmung	concordance	AFRNXTMNK	AVRNXTMNK
bestätigen	confirmer	PXTTK	/
Aufstand	insurrection	AFXTNT	/
Übereinstimmung	concordance	APRNXTMNK	AVRNXTMNK

TABLE 2: Exemples de clés metaphone. Les mots alsaciens sont dans la partie supérieure de la table, et les mots allemands sont dans la partie inférieure.

4.2 Méthode d'alignement

Notre premier objectif est d'aligner les entrées dans plusieurs lexiques bilingues français-alsacien. Dans une première étape, toutes les entrées des trois lexiques utilisés sont ajoutées à un graphe. Les nœuds correspondent aux mots alsaciens et à leurs traductions en français. Les mots alsaciens sont connectés à leurs traductions en français dans les lexiques par une arête. En outre, deux mots alsaciens sont reliés par une arête si toutes les conditions suivantes sont remplies :

1. ils ont la même traduction en français ;
2. ils ont une clé métaphone en commun ;
3. ils appartiennent à la même partie du discours [13].

Nous utilisons également des informations obtenus à partir des ressources décrites à la section 3 afin d'assouplir la condition 1. La liste des synonymes français issue de JeuxDeMots est utilisée pour connecter deux mots alsaciens qui ont des traductions françaises synonymes dans cette ressource. Les sens français de BabelNet sont utilisés de la même manière que les synonymes de JeuxDeMots, pour connecter les mots alsaciens qui ont des traductions françaises ayant le même sens.

4.2.1 Alignement des variantes alsaciennes

Après la construction du graphe, les variantes alsaciennes sont regroupées. Les formes qui sont des variantes sont récupérées par la détection de composantes connexes dans le sous-graphe contenant uniquement les mots alsaciens. La figure 1 montre une portion du graphe initial. Les traductions en français, allemand et anglais sont également présentées. Dans le sous-graphe formé par les mots alsaciens, il y a trois composantes connexes : (1) ["Winkäller", "Winkeller", "Winkaller"], (2) ["Wikaller"] et (3) ["Kaller"]. Les formes "Winkäller", "Winkeller" et "Winkaller" sont donc regroupées dans un cluster et considérées comme des variantes graphiques de la même unité lexicale.

12. Notre implémentation de Double Metaphone repose sur un module Python existant : `http://www.atomodo.com/code/`

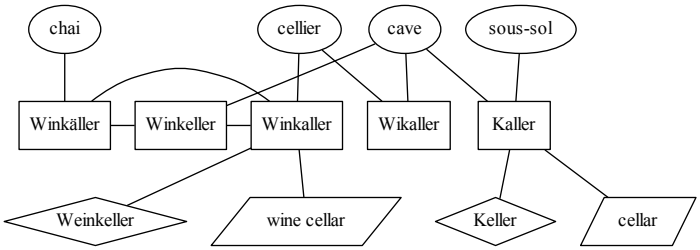

FIGURE 1: Vue simplifiée d'un sous-graphe. Les mots français figurent dans des ellipses, les mots alsaciens dans des rectangles, les mots anglais dans des parallélogrammes et les mots allemands dans des losanges.

4.2.2 Mise en correspondance avec les synsets de BabelNet

Notre deuxième objectif est de mettre en correspondance les variantes alsaciennes alignées avec les synsets de Babelnet. Par exemple, en prenant l'exemple de la figure 1, le cluster formé par ["Winkäller", "Winkeller", "Winkaller"] doit être mis en correspondance avec le synset ayant pour identifiant `bn:00017041n` (voir Figure 2).

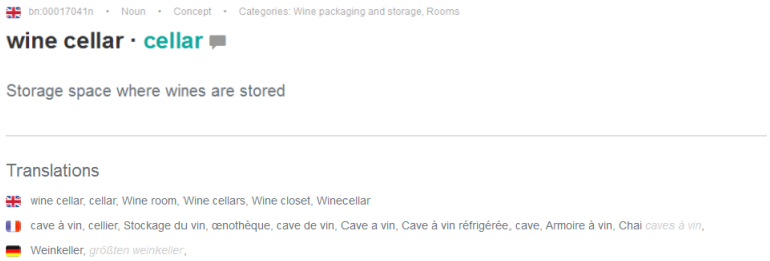

FIGURE 2: Synset `bn:00017041n` dans l'interface de recherche de BabelNet

La mise en correspondance se fait en calculant la similarité cosinus entre des représentations "sac de mots" binaires des synsets de BabelNet et des variantes alsaciennes alignées. Dans le cas le plus simple, la représentation utilisée pour les synsets de BabelNet se compose de leurs lexicalisations français. Les variantes alsaciennes sont représentées par leurs traductions en français : dans l'exemple de la Figure 1, le cluster formé par ["Winkäller", "Winkeller", "Winkaller"] sera représenté par le sac de mots ["chai", "cellier", "cave"]. Les représentations sac de mots peuvent être étendues en utilisant les traductions disponibles dans BabelNet. En effet, il a été montré que l'utilisation des traits multilingues a un effet positif sur la tâche de désambiguïsation (Banea & Mihalcea, 2011). Cependant, il faut éviter l'ambiguïté lorsque l'on sélectionne les traductions en anglais et en allemand pour les mots alsaciens. Cette question a été abordée dans les travaux sur l'acquisition de dictionnaires bilingues pour une paire de langues en utilisant une troisième langue comme un pivot : dans notre cas, le français est la langue pivot, l'alsacien la langue source et l'allemand et l'anglais les langues cibles. Plusieurs méthodes ont été proposées, qui reposent principalement soit sur la structure des lexiques bilingues disponibles soit sur la similarité distributionnelle (Salloum & Habash, 2011; Tanaka & Umemura, 1994). Dans notre cas particulier, nous exploitons la proximité entre l'alsacien et l'allemand, et, à un degré moindre, l'anglais. A partir des traductions en français, les traductions en allemand et / ou en anglais sont ajoutées aux représentations sac de mots des clusters de variantes alsaciennes si les traductions et les mots alsaciens partagent une de leurs clés metaphone. Cette contrainte effectue une sorte de désambiguïsation et assure que seules traductions valides sont sélectionnées. Ainsi, dans l'exemple de la figure 1, le mot allemand "Weinkeller" et le mot anglais "wine cellar" seront ajoutés aux sacs de mots.

`double-metaphone/metaphone.py/view`

13. Les participes passés et les adjectifs sont considérés comme faisant partie de la même catégorie.

4.3 Évaluation et résultats

Afin d'évaluer notre méthode, nous avons produit manuellement 107 alignements de référence entre les lexiques et BabelNet. Dans ce but, nous avons choisi au hasard des entrées d'un dictionnaire multilingue français-allemand-anglais-alsacien (Adolf, 2006). Ce dictionnaire présente plusieurs avantages pour l'évaluation : plusieurs variantes orthographiques sont généralement proposées pour chaque entrée alsacienne ; les traductions en français, allemand et anglais sont fournies, facilitant ainsi la mise en correspondance avec BabelNet ; enfin, le dictionnaire se concentre sur des mots alsaciens qui sont très semblables aux mots allemands et anglais correspondants. S'il est vrai que cela induit un biais dans l'évaluation pour les configurations où les lexiques anglais et allemands sont utilisés, cela permet d'avoir une idée des performances maximales qu'il est possible d'atteindre, et qui seront vraisemblablement inférieures sur l'ensemble du lexique. Pour les données d'évaluation, nous avons vérifié que les entrées se retrouvent dans au moins un des trois lexiques bilingues utilisés (OLCA, WKT et ACPA). En outre, nous avons sélectionné le meilleur synset correspondant de Babelnet. Quand il n'était pas possible de décider, au plus trois synsets de Babelnet ont été choisis.

L'alignement des variantes est évalué en terme de précision, rappel et F-mesure. Pour chaque cluster de mots alsaciens tels qu'une des traductions en français se trouve dans les données d'évaluation, nous mesurons l'intersection entre les alignements automatiques du cluster et les variantes alsaciennes dans les données de référence dans comme des vrais positifs (VP). Les variantes automatiquement alignées qui ne sont pas dans les données de référence sont considérées comme des faux positifs (FP), tandis que celles de la référence qui ne sont pas dans les alignements produits sont considérées comme des faux négatifs (FN). Par exemple, pour le cluster ['Schekbeeschel'/ACPA, 'Scheckbiechel'/Adolf(2006)] [14] (*carnet de chèque* en français), l'alignement de référence est ['Schäckbiachla'/ACPA, 'Schekbeeschel'/ACPA, 'Scheckbiechel'/Adolf(2006)]. Dans ce cas, VP=2, FP=0 et FN=1. Ensuite, la précision (P), le rappel (R) et F-mesure (F) sont calculés comme suit :

$$P = \frac{VP}{VP + FP} \quad ; \quad R = \frac{VP}{VP + FN} \quad ; \quad F = \frac{2 \cdot P \cdot R}{P + R}$$

La mise en correspondance avec BabelNet est également évaluée en termes de précision, rappel et F-mesure. Les synsets de BabelNet étant ordonnés selon la similarité cosinus, nous prenons en compte tous les synsets qui ont le même cosinus au rang 1. Les résultats de l'évaluation pour les différents paramètres sont détaillés dans la table 3. La *baseline* correspond à l'absence de ressources externes. + JDM indique que les synonymes de JeuxDeMots ont été utilisés. + BN indique que BabelNet a été utilisé, avec les lexicalisations en français (FR), allemand (DE) ou en anglais (EN).

	Alignement des lexiques			**Alignement avec BabelNet**		
	P	**R**	**F**	**P**	**R**	**F**
baseline	0,99	0,69	**0,81**	0,18	0,44	0,26
+ BN FR	0,94	0,70	0,80	0,21	0,47	0,29
+ JDM	0,97	0,70	**0,81**	0,18	0,44	0,26
+ BN FR & DE	0,94	0,70	0,80	0,41	0,49	0,45
+ BN FR & EN	0,94	0,70	0,80	0,27	0,50	0,35
+ BN FR, DE & EN	0,94	0,70	0,80	**0,45**	**0,51**	**0,48**
+ JDM + BN FR & DE	0,92	0,70	0,80	0,39	0,48	0,43
+ JDM + BN FR, DE & EN	0,92	0,70	0,80	0,44	0,50	0,47

TABLE 3: Résultats de l'évaluation

Dans l'ensemble, les résultats pour les alignements des variantes issues de différents lexiques sont stables : l'utilisation des ressources externes conduit à une légère baisse de la précision qui est compensée par une très légère hausse du rappel. En outre, le rappel est toujours inférieur à la précision. Pour la mise en correspondance avec les synsets de Babelnet, l'utilisation de traductions en allemand et, à un degré moindre, en anglais, conduit à des améliorations. Dans ce cas, moins de faux positifs sont détectés, parce que les mots allemands et anglais fournissent un contexte de désambiguïsation qui aide à identifier le synset correct. Les résultats peuvent sembler décevants, avec une F-mesure culminant à 0,48. Cependant, le mode d'évaluation est assez strict, car il permet l'alignement avec un seul synset de BabelNet dans la plupart des cas. Les synonymes fournis par JDM ont un effet légèrement négatif sur la performance, très certainement parce que les ensembles de synonymes dans cette ressource sont différents de ceux que l'on trouve dans Babelnet. Le rappel inférieur obtenu

14. Nous indiquons également le lexique d'où est issu la variante après le caractère '/'.

pour les alignements de variantes dans les lexiques est principalement dû à la contrainte qui exige des clés metaphone identiques. Dans certains cas, des variantes ont différentes clés (par exemple “Chilche” - KLX / XLX et “Kirche” - KRX). Cela soulève également une question plus fondamentale : ces formes peuvent-elles encore être considérées comme des variantes, ou correspondent-elles à des unités différentes ? Dans notre construction des alignements de référence, nous avons regroupé les variantes que l’on trouve dans le dictionnaire multilingue, même si elles pouvaient être différentes dans certains cas. Par ailleurs, en plus des clés metaphone, d’autres mesures de similarité orthographique pourraient être utilisées pour aligner les variantes, comme cela se fait pour l’identification de cognats (Inkpen *et al.*, 2005). Ces mesures pourraient aider à l’amélioration du rappel. Certaines erreurs sont également dues à des problèmes dans la récupération des parties du discours pour les entrées de dictionnaire ambiguës. Comme l’une des conditions d’alignement requiert des parties du discours identiques, ces entrées ne sont pas considérées comme des variantes.

Comme le montrent les résultats, l’ajout d’informations multilingues permet d’améliorer la mise en correspondance avec les synsets de BabelNet. Pour le moment, les traductions en allemand et en anglais sont choisies en fonction de leurs clés metaphone, ce qui conduit à des traductions manquantes pour certains “sacs de mots”. À l’avenir, ceci pourrait être amélioré en utilisant des lexiques bilingues supplémentaires, afin d’ajouter des traductions qui ne sont pas nécessairement apparentées aux variantes alsaciennes.

5 Comparaison oral-écrit

L’alignement des variantes alsaciennes trouvées dans différents lexiques permet de réaliser une étude comparative des scripturalisations utilisées, en mettant en évidence les différences les plus fréquentes en termes de remplacements de caractères. Ces différences témoignent des difficultés rencontrées lors de la transcription du dialecte à l’écrit, tant par les lexicologues que par les locuteurs non-spécialistes. La comparaison des formes trouvées dans les lexiques permet de mettre au jour les problématiques majeures de cette mise à l’écrit. Nous avons analysé 581 différences entre les mots des lexiques, plus ou moins fréquentes (la fréquence de remplacement correspond à la colonne “rang” de la Table 4). Ces différences portent essentiellement sur le remplacement d’un graphème par un autre [15].

Rang	Remplacement	Nombre d’occurrences	Exemples
1	a → e	3 307	Wàldbeerla – Wàldbeerle
2	e → i	962	blend – blind
3	e → ä	493	Hardepfel - Hardäpfel
4	a → ä	487	Base – Bäse
5	u → ù	476	Leischtung - Leischtùng
6	a → à	358	hopla – hoplà
7	e → ì	213	Leppel - Lìppel
8	i → ì	211	Kopfkisse – Kopfkìsse
9	d → t	166	dànze – tànze
…	…	…	…
13	g → j	121	Flegel – Flejel
…	…	…	…
21	b → p	61	bolìere – polìere
…	…	…	…
23	g → k	51	stàrig - stàrik
…	…	…	…
124	f → v	8	narfig - narvig
…	…	…	…
458	s → z	2	baidersits - beiderssitz

TABLE 4: Remplacements de graphèmes dans les lexiques

15. On trouve également des remplacements de plusieurs graphèmes mais ils sont plus rares.

5.1 Les voyelles

Les mots *Wàldbeerla* et *Wàldbeerle* (fraise des bois) sont un bon exemple de variabilité : seule la voyelle finale est soumise au changement. De ce fait, ils sont donc comptabilisés dans l'entrée de remplacement « a → e ». Les variantes sur ces deux voyelles sont extrêmement nombreuses : il s'agit de la première entrée du tableau de comparaison, comprenant 3 307 items transformés. Ce chiffre conséquent est corrélé à une réalité sociophonétique : l'une des nuances vocaliques les plus connues concernant l'alsacien est une différence observable entre le Nord et le Sud de l'Alsace. L'aperture des voyelles postérieures augmente en fonction de la provenance géographique des locuteurs : selon notre exemple, *Wàldbeerla* au Sud et *Wàldbeerle* au Nord. Ainsi, les remplacements de graphèmes les plus fréquents sont en accord avec, et même soulignent un fait phonétique. Les voyelles sont en fait très soumises aux variations graphiques : les huit premières entrées des remplacements portent sur ce type de phonèmes, avec au total, 6 507 formes transformées. Ces modifications concernent majoritairement des choix purement graphiques : 1 321 modifications entre des graphèmes tels que « a » et « à », par exemple entre les formes *hoplà* et *hopla*. L'instabilité des voyelles graphiques est parfois correspondante à la variabilité constatée d'un point de vue phonétique. L'utilisation de l'algorithme *Double Metaphone* est donc particulièrement pertinente, puisque les clés utilisées pour l'alignement n'utilisent que les consonnes (voir Section 4.1).

5.2 Les consonnes

Les occlusives de l'alsacien, graphiées `p,t,k` et `b,d,g`, ne sont pas les mêmes que les phonèmes graphiés en français avec les même symboles. En effet, en français, ces consonnes s'opposent selon le trait phonologique de voisement : pour produire les sourdes `/p,t,k/`, les plis vocaux cessent de vibrer pendant l'occlusion, tandis que pour les sonores, la vibration est maintenue. En allemand, cette différence est également présente, mais tend à être remplacée en position initiale de mot et parfois en finale par une opposition d'aspiration, ou de tension (il est souvent fait référence à ce phénomène en tant que *dévoisement* ou *assourdissement*). La vibration des plis vocaux est alors absente pour les deux séries. En alsacien, le voisement n'est pas forcément pertinent pour distinguer ces consonnes, problématique connue en phonétique (Bothorel-Witz & Pétursson, 1972; Erhart, 2012; Pipe, 2014; Woehrling & Boula de Mareüil, 2005). L'analyse acoustique en lecture événementielle des signaux de parole (Steiblé, 2014) a pu apporter des lumières sur ces consonnes, qui s'avèrent être opposées selon leur tension, ou plutôt leur appartenance à une catégorie *fortis* et l'autre, *lenis* (Kohler, 1984). Ainsi, il n'y a pas de correspondance entre les occlusives du français et celles de l'alsacien, ce qui est un vecteur de ce que les francophones perçoivent comme étant un accent alsacien. Notons que cette problématique s'applique également aux fricatives, telles que `/f/` et `/v/` par exemple. Bien entendu, les systèmes graphiques alsaciens utilisent les mêmes symboles qu'en français, mais il n'est pas simple de faire un choix entre les deux graphèmes, aucune des prononciations françaises n'étant correcte. Ces phonèmes posent le problème de consonnes le plus massif : l'analyse des disparités entre les lexiques montre une hésitation fréquente sur l'usage des graphèmes `p,t,k` et `b,d,g`. En effet, ces consonnes sont très souvent utilisées les unes à la place des autres, dans 278 cas au total, par opposition à seulement 10 cas de modifications des fricatives, telles que `/f-v/` ou encore `/s-z/`. La paire apico-alvéodentale, graphiée `t,d`, occupe la neuvième place des modifications les plus fréquentes, après les nombreux changements de voyelles. Elle représente à elle seule 166 modifications, tant en position initiale (ex. *dànze – tànze*, danser) qu'en intervocalique (ex. *Vàder – Vàter*, père) ou qu'en finale de mot (ex. *G'hàlt – Gald*, argent). Un problème certain est soulevé par ce phénomène : il existe de nombreuses paires minimales opposées uniquement par la consonne occlusive, comme *Pump* (pompe) et *Bumb* (bombe), ou *Gàss* (ruelle) et *Kàss* (caisse). Dans ces cas, il serait absolument nécessaire d'opérer une distinction entre les graphèmes utilisés, mais l'étude des lexiques montre que les hésitations sont nombreuses lors de la mise à l'écrit de ces phonèmes spécifiquement. Ainsi, les consonnes elles-mêmes ne sont parfois pas fiables, ce qui est compensé dans l'algorithme *Double Metaphone* par la neutralisation de la différence entre ces consonnes (voir Table 2 : le 'd' dans "Rüejd̲àà" et le 't' dans "Rüaijt̲àag" sont indifféremment transcrits par 'T' dans les clés metaphone). L'utilisation de ressources globales permettant la comparaison, par exemple, avec les graphies choisies en allemand, pourrait contribuer à stabiliser des formes écrites qui respecteraient toujours l'opposition *fortis-lenis*. Cette stabilisation est un enjeu d'importance au vu de l'existence de paires minimales dont l'opposition repose uniquement sur ces consonnes. Il s'agirait donc de tendre à normaliser l'usage de ces graphèmes, à travers une comparaison des formes existantes dans divers dictionnaires, afin de permettre de clarifier les choix à faire dans les ressources futures.

6 Conclusion et perspectives

L'absence de convention orthographique est un problème pour de nombreuses langues peu dotées en ressources linguistiques, ce qui complique encore davantage l'acquisition de ressources lexicales. Nous avons proposé des solutions qui utilisent des ressources disponibles facilement (lexiques bilingues, réseau sémantique BabelNet) et des outils simples à développer et à adapter pour de nouvelles langues (*Double Metaphone*). Les ressources obtenues sont mises en correspondance avec BabelNet, qui ajoute une couche sémantique et donne accès à différents types d'informations supplémentaires : définitions et gloses, traductions dans d'autres langues, des images, etc.

La méthode proposée pour l'alignement des lexiques vise la précision plutôt que le rappel et peut être utilisée pour construire facilement et rapidement des ressources lexicales multilingues avec une intervention humaine limitée pour la correction. Elle pourrait en principe être appliquée à de nombreuses autres langues, car elle nécessite peu de ressources. Son originalité est qu'elle ne cible pas la normalisation, mais vise plutôt à regrouper les variantes graphiques et à les relier à des entrées dans une ressource multilingue. De cette manière, les ressources du *Linguistic Linked Open Data* peuvent être étendues avec des lexicalisations de langues peu dotées et aider à construire et enrichir les ressources pour ces langues. L'alignement des lexiques a mis en évidence les difficultés de scripturalisation des dialectes alsaciens, en lien avec des faits phonétiques avérés. Les consonnes en particulier posent des difficultés résultant en de nombreuses hésitations qui peuvent compliquer l'alignement.

Dans l'avenir, nous prévoyons de fournir le lexique aligné dans un format standard. Nous souhaitons également améliorer les alignements grâce à un meilleur algorithme *Double Metaphone*, une analyse appropriée des formes de mots composés et l'utilisation des ressources supplémentaires avec une meilleure couverture. Enfin, le lexique obtenu pourra être utilisé dans diverses applications, par exemple l'étiquetage morphosyntaxique (Bernhard & Ligozat, 2014).

Remerciements Nous remercions l'OLCA, André Nisslé et Paul Adolf pour nous avoir donné accès à leurs ressources. Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet RESTAURE - convention ANR-14-CE24-0003-01) et du conseil scientifique de l'Université de Strasbourg (projet COPAL).

Références

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

ALMEIDA J. J., SANTOS A. & SIMÕES A. (2010). Bigorna – A Toolkit for Orthography Migration Challenges. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

BANEA C. & MIHALCEA R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics*, p. 25–34.

BAUER L. (2003). *Introducing Linguistic Morphology*. Georgetown University Press. 2nd edition.

BERNHARD D. & LIGOZAT A.-L. (2014). Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, p. 209–220.

BOTHOREL-WITZ A. & PÉTURSSON M. (1972). La nature des traits de tension, de sonorité et d'aspiration dans le système des occlusives de l'allemand et de l'islandais. *Travaux de L'Institut de Phonétique de Strasbourg*, (4).

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

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.

GUREVYCH I., ECKLE-KOHLER J., HARTMANN S., MATUSCHEK M., MEYER C. M. & WIRTH C. (2012). UBY–A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of EACL*, p. 580–590, Avignon, France.

HANOKA V. & SAGOT B. (2012). Wordnet creation and extension made simple : A multilingual lexicon-based approach using wiki resources. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

HUCK D., BOTHOREL-WITZ A. & 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 from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, p. 13–100.

HUDLETT A. & 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*. Mulhouse, France : Centre de Recherche sur l'Europe littéraire (C.R.E.L.).

HULDEN M., ALEGRIA I., ETXEBERRIA I. & 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*, p. 39–48, Edinburgh, Scotland.

INKPEN D., FRUNZA O. & KONDRAK G. (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, p. 251–257.

KLEIBER G. & RIEGEL M. (1998). Grammaticalisation et auxiliaire modal : L'énigme de duen en Alsacien. In *Travaux de linguistique*, volume 36, p. 161–173, Bruxelles, Belgique : Rijksuniversiteit van Gent.

KOHLER K. (1984). Phonetic explanation in phonology : the feature fortis/lenis. *Phonetica*, **41**, 150–174.

LAFOURCADE M. (2007). Making people play for Lexical Acquisition. In *Proceedings of SNLP 2007*, Pattaya, Thailande.

LEIXA J., MAPELLI V. & CHOUKRI K. (2014). *Inventaire des ressources linguistiques des langues de France*. Rapport ELDA/DGLFLF-2013A, ELDA, Paris.

MATZEN R. (1985). Les emprunts du dialecte alsacien au français. In *Le français en Alsace : Actes du colloque de Mulhouse (17-19 novembre 1983)*, Bulletin de la Faculté des lettres de Mulhouse, Mulhouse : Paris : Champion, Genève : Slatkine.

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

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.

NIEMANN E. & GUREVYCH I. (2011). The people's web meets linguistic knowledge : Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, p. 205–214.

OLCA / EDINSTITUT (2012). Etude sur le dialecte alsacien. En ligne : `https://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf`.

PHILLIPS L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*.

PIPE K. (2014). *Accent Levelling in the Regional French of Alsace*. Thèse de doctorat, University of Exeter.

PORTA J., SANCHO J.-L. & GÓMEZ J. (2013). Edit Transducers for Spelling Variation in Old Spanish. In *Proceedings of the Workshop on Computational Historical Linguistics at NoDaLiDa 2013*, volume 87, p. 70–79.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008-Traitement Automatique des Langues Naturelles*.

SALLOUM W. & HABASH N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, p. 10–21.

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

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

STEIBLÉ L. (2014). *Le contrôle temporel des consonnes occlusives de l'alsacien et du français parlé en Alsace*. Thèse de doctorat, Université de Strasbourg.

TANAKA K. & UMEMURA K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, p. 297–303.

WOEHRLING C. & BOULA DE MAREÜIL P. (2005). Identification d'accents régionaux en français : perception et catégorisation. *Bulletin PFC 6*, p. 89–102.

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