Comparaison de méthodes de prétraitement pour l'alignement de mots dans des corpus parallèles alsacien-français

Delphine Bernhard

Comparaison de méthodes de prétraitement pour l'alignement de mots dans des corpus parallèles alsacien-français

Delphine Bernhard[1]
(1) Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg
dbernhard@unistra.fr

RÉSUMÉ

L'analyse des corpus de textes dans les dialectes alsaciens est rendue difficile par la grande variation qui se rencontre à l'écrit, en l'absence d'une norme orthographique stable. Dans cet article, nous décrivons des expériences visant à améliorer la qualité de l'alignement automatique des mots français et alsaciens en prétraitant les textes. Ces prétraitements visent à réduire la variation en utilisant diverses stratégies en fonction de la langue : lemmatisation, désuffixation, clés métaphone, normalisation et réduction des tokens à leur préfixe.

ABSTRACT

Comparison of pre-processing methods for word alignment in parallel Alsatian-French corpora.

The analysis of text corpora in Alsatian dialects is made difficult by the large variation that occurs in writing, in the absence of a stable orthographic standard. In this article, we describe experiments to improve the quality of automatic alignment of French and Alsatian words by preprocessing the texts. These preprocessing operations aim to reduce variation by using various strategies depending on the language: lemmatisation, stemming, metaphone keys, normalisation and reducing tokens to their prefix.

MOTS-CLÉS : dialectes alsaciens, variation, alignement de mots.

KEYWORDS: Alsatian dialects, variation, word alignment.

1 Introduction

Les dialectes alsaciens se caractérisent par une scripturalisation non normée selon des standards orthographiques. Bien que des standards aient été proposés, ils ne sont pas largement diffusés. Cette absence de norme complique la manipulation de corpus de textes pour les dialectes alsaciens pour des applications telles que la recherche dans les corpus, l'analyse thématique, ou l'entraînement et l'application d'outils de traitement automatique des langues. Il est donc nécessaire de pouvoir identifier les variantes orthographiques dans ces corpus pour pouvoir les exploiter au mieux. La gestion de la variation a été largement abordée pour divers types de données (textes historiques, médias sociaux, dialectes). En particulier, les approches supervisées telles que celles proposées par Barteld *et al.* (2019) pour des textes en moyen bas allemand ou Hosseini *et al.* (2020) pour la mise en correspondance de toponymes nécessitent de disposer de données d'entraînement, sous la forme de paires de variantes vraies ou fausses.

Nous proposons d'exploiter des textes parallèles pour collecter des données permettant d'entraîner

de tels systèmes de détection automatique de variantes : les mots graphiquement similaires qui sont alignés avec la même traduction en français peuvent être considérés comme des variantes graphiques. Dans cet article, nous décrivons plus particulièrement des expériences visant à déterminer la meilleure stratégie d'alignement de mots pour un corpus parallèle alsacien-français. Nous comparons deux outils d'alignement et appliquons diverses procédures de prétraitement au corpus pour améliorer les alignements. Nous montrons que les procédures de prétraitement visant à réduire la variation sont bénéfiques, en particulier lorsqu'elles sont appliquées aux textes alsaciens.

2 Données et méthode

2.1 Corpus parallèle et lexique d'évaluation

Notre corpus regroupe des textes parallèles de divers genres : contes, pièce de théâtre, textes administratifs, sites web, chansons, recettes. Il comporte 149 119 tokens alsaciens et 155 168 tokens français, pour un total de 12 945 segments alignés. On observe un accroissement du vocabulaire bien plus important dans le corpus alsacien que dans le corpus français (voir Figure 1), ce qui s'explique notamment par l'absence de standard orthographique en alsacien.

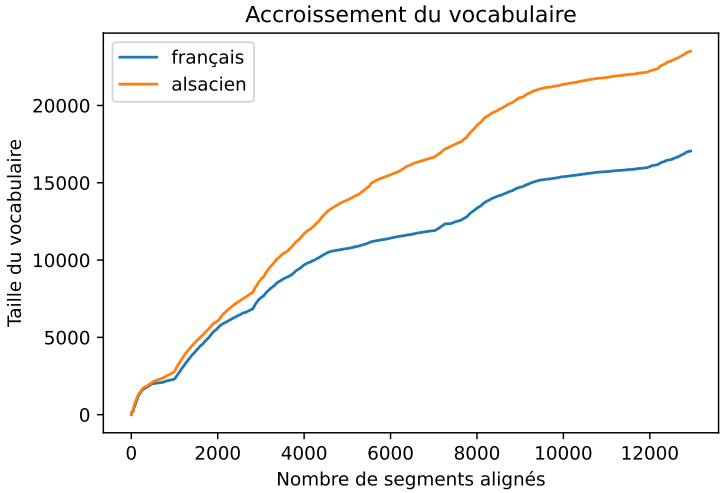


FIGURE 1 – Accroissement du vocabulaire dans les deux corpus.

La Table 1 montre quelques exemples de variantes observées dans le corpus.

français	**alsacien (variantes)**
Strasbourg	Strosburi, Strosbùri, Strossburi, Strossburig, Strossburri, Strossbùri, Stroßbùrri
bientôt	ball, boll, bàll, böl
simple	ainfàch, eifach, einfach, einfàch, ëmfàch, ëënfàch

TABLE 1 – Exemples de variantes en alsacien avec leur traduction en français.

Pour les besoins de l'évaluation, nous avons utilisé divers lexiques bilingues, et n'avons conservé que les paires de formes se trouvant dans un segment aligné de notre corpus parallèle, aboutissant ainsi à un lexique de 3 102 entrées. Nous utilisons la procédure d'évaluation proposée par Lardilleux *et al.* (2010), qui tient compte des probabilités de traduction pour le calcul de la précision du rappel et du score F1.

2.2 Outils d'alignement

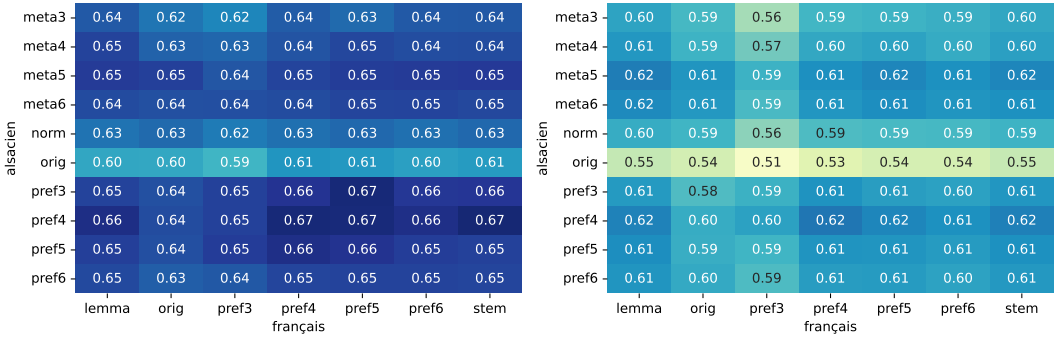
Nous avons comparé deux outils d'alignement de mots : `eflomal` (Östling & Tiedemann, 2016) [1] et `fast_align` (Dyer *et al.*, 2013) [2], avec les paramètres par défaut. Les alignements produits sont asymétriques et peuvent être rendus symétriques par diverses heuristiques. Nous utilisons 1/3 du lexique de référence pour choisir la meilleure heuristique de symmétrisation, à savoir l'intersection. Ceci est cohérent avec les travaux de Steingrímsson *et al.* (2021) qui ont également montré que l'intersection était la meilleure stratégie à la fois pour `eflomal` et `fast_align`.

2.3 Prétraitements

Nous appliquons diverses méthodes de prétraitement au corpus pour réduire la variation et ainsi améliorer la qualité de l'alignement, par rapport aux formes originales dans le corpus (`orig`) :

— français : lemmatisation (`lemma`), désuffixation (`stem`) [3] et tokens réduits à leurs `n` premiers caractères (`pref`).

— alsacien : normalisation par mise en minuscules, suppression des apostrophes, des diacritiques et remplacement des lettres doublées par un seul exemplaire (`norm`), tokens normalisés réduits à leur `n` premiers caractères (`pref`), tokens normalisés remplacés par leur clé métaphone (Philips, 2000; Bernhard, 2014) s'ils ont au moins `m` caractères (`meta`).



(a) `eflomal` (b) `fast_align`

FIGURE 2 – Scores F1 obtenus par `eflomal` et `fast_align`.

1. Version 1.0.0 publiée le 7 avril 2020 sur https://github.com/robertostling/eflomal/releases/tag/v1.0.0

2. https://github.com/clab/fast_align, Git hash `cab1e9aac8d3bb02ff5ae58218d8d225a039fa11`

3. La lemmatisation est effectuée à l'aide de spaCy 2.3.5, modèle `fr_core_news_sm` version 2.3.0. La désuffixation est obtenue à l'aide de `nltk.stem.snowball.FrenchStemmer` (`nltk` version 3.7).

3 Résultats

Les scores F1 obtenus sur les ⅔ restant du lexique d'évaluation sont présentés dans la Figure 2 ; la Figure 3 détaille la précision et le rappel pour `eflomal`. D'une manière générale, `eflomal` obtient de meilleurs résultats que `fast_align`, ce qui confirme les résultats de Steingrímsson *et al.* (2021) qui ont montré qu'`eflomal` était plus performant que `fast_align` pour les corpus de petite taille.

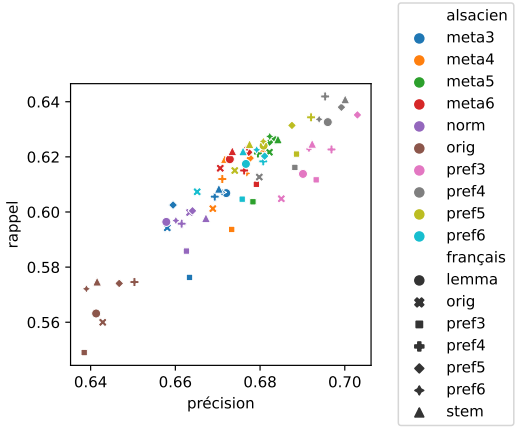


FIGURE 3 – Précision et rappel obtenus par `eflomal` pour les différents prétraitements.

Nous mesurons la significativité statistique des résultats obtenus par `eflomal` à l'aide d'une méthode de ré-échantillonnage bootstrap apparié (« paired bootstrap resampling », Berg-Kirkpatrick *et al.* (2012)) avec 1 000 réplications pour divers types de prétraitements (voir Tables 2 et 3).

Prétraitement alsacien		**Prétraitement français**						
1	**2**	`lemma`	`orig`	`pref3`	`pref4`	`pref5`	`pref6`	`stem`
`meta5`	`orig`	0.000	0.000	0.000	0.000	0.000	0.000	0.000
`pref3`	`orig`	0.000	0.000	0.000	0.000	0.000	0.000	0.000
`pref4`	`orig`	0.000	0.000	0.000	0.000	0.000	0.000	0.000
`pref3`	`meta5`	0.716	0.900	0.041	0.105	0.010	0.390	0.324
`pref4`	`meta5`	0.041	0.154	0.021	0.001	0.005	0.056	0.003

TABLE 2 – Valeurs de p pour les différences entre prétraitements sélectionnés pour le corpus alsacien. Les valeurs correspondant aux seuils $p < 0.05$ et $p < 0.01$ sont mises en évidence.

Prétraitement fr		**Prétraitement alsacien**									
1	**2**	`meta3`	`meta4`	`meta5`	`meta6`	`norm`	`orig`	`pref3`	`pref4`	`pref5`	`pref6`
`lemma`	`orig`	0.008	0.002	0.268	0.280	0.234	0.597	0.065	0.000	0.053	0.016
`stem`	`orig`	0.010	0.021	0.251	0.193	0.427	0.917	0.007	0.000	0.101	0.008
`pref4`	`orig`	0.020	0.108	0.406	0.347	0.302	0.021	0.005	0.000	0.000	0.007
`pref4`	`stem`	0.339	0.184	0.153	0.702	0.267	0.266	0.415	0.379	0.007	0.517
`pref5`	`stem`	0.055	0.289	0.372	0.375	0.496	0.353	0.024	0.316	0.042	0.391
`pref4`	`lemma`	0.366	0.030	0.195	0.576	0.413	0.052	0.064	0.825	0.010	0.315
`pref5`	`lemma`	0.064	0.321	0.413	0.249	0.151	0.057	0.002	0.184	0.086	0.262

TABLE 3 – Valeurs de p pour les différences entre prétraitements sélectionnés pour le corpus français. Les valeurs correspondant aux seuils $p < 0.05$ et $p < 0.01$ sont mises en évidence.

L'application de prétraitements au corpus alsacien (métaphone, préfixe) conduit systématiquement à de meilleurs résultats par rapport au corpus original. L'étude de significativité (Table 2) montre que la différence observée entre les prétraitements `meta5`, `pref3` et `pref4` et le corpus d'origine `orig`, non prétraité, est statistiquement significative ($p < 0.05$). Il est toutefois plus difficile de conclure sur la supériorité du prétaitement par préfixe (`pref3` et `pref4`) par rapport au prétraitement par métaphone (`meta5`) : pour ces prétraitements, la différence n'est statistiquement significative que dans quelques cas, avec certains types de prétraitements pour le français. Globalement, nos résultats pour l'alsacien vont dans le sens des observations faites par Östling & Tiedemann (2016) qui montrent qu'une méthode de désuffixation approximative consistant à ne conserver que les 4 premières lettres d'un mot permet généralement d'améliorer les résultats en traduction automatique statistique pour les langues à suffixation.

Par ailleurs, si l'on observe des résultats légèrement meilleurs avec un prétraitement pour le français, les différences ne sont généralement pas statistiquement significatives (Table 3), en particulier lorsque l'on compare les trois méthodes de prétraitement proposées : préfixes, lemmatisation et désuffixation.

Enfin, la Table 4 donne quelques exemples de mots alsaciens avec leur traduction la plus probable en l'absence de prétraitement des corpus et avec réduction à un préfixe de 4 caractères. La traduction attendue est également indiquée.

alsacien	**traduction attendue**	**sans prétraitement**	**préfixation à 4**
Biehn	grenier	∅	grenier
Dràche	dragon	dragon	dragon
fàrwig	coloré	original	coloré
schloeje	battre	jusqu'	puis
schwàsiere	choisir	Nouvel, désirés	∅
versteckle	cacher	cacher	cacher, Cacher
wascha	laver	∅	laver

TABLE 4 – Exemples de mots alsaciens avec leur traduction la plus probable.

4 Conclusion et perspectives

Nous avons présenté des expériences visant à comparer deux outils d'alignement de mots en appliquant diverses procédures de prétraitement à des corpus parallèles alsacien-français. Les résultats montrent une supériorité d'`eflomal` sur `fast_align`, ainsi que la pertinence d'un prétraitement simple consistant à normaliser les mots alsaciens et à les réduire à leurs préfixes, les meilleurs résultats étant obtenus pour un préfixe de 4 lettres. L'utilisation des clés métaphone permet aussi d'obtenir de bons résultats. L'avantage du prétraitement est moins évident pour le français, même si les divers types de prétraitements comparés (préfixe, lemmatisation, désuffixation) peuvent conduire à des améliorations en combinaison avec certains prétraitements pour l'alsacien.

Dans la suite de ce travail, nous souhaitons améliorer la prise en compte de la variation graphique en nous inspirant de la méthode proposée par Burlot & Yvon (2017) pour l'alignement entre langues morphologiquement complexes et langues plus simples. Les lexiques bilingues nous servirons ensuite de données d'entraînement pour la détection automatique de variantes en alsacien.

Remerciements

Ces travaux ont été réalisés dans le cadre du projet ANR-21-CE27-0004 DIVITAL soutenu par l'Agence Nationale de la Recherche.

Nous remercions les étudiantes ayant participé à la collecte et à l'alignement des phrases du corpus parallèle : Natália Leščišínová, Camille Meyer et Johana Libman.

Références

BARTELD F., BIEMANN C. & ZINSMEISTER H. (2019). Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation*, p. 1–30. DOI : 10.1007/s10579-018-09441-5.

BERG-KIRKPATRICK T., BURKETT D. & KLEIN D. (2012). An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 995–1005, Jeju Island, Korea.

BERNHARD D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources : the Example of Alsatian. In *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, p. 23–29, Reykjavík, Iceland. HAL : hal-00966820.

BURLOT F. & YVON F. (2017). Learning Morphological Normalization for Translation from and into Morphologically Rich Languages. *The Prague Bulletin of Mathematical Linguistics*, **108**(1), 49–60. DOI : 10.1515/pralin-2017-0008.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 644–648, Atlanta, Georgia : Association for Computational Linguistics.

HOSSEINI K., NANNI F. & COLL ARDANUY M. (2020). DeezyMatch : A Flexible Deep Learning Approach to Fuzzy String Matching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 62–69, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-demos.9.

LARDILLEUX A., GOSME J. & LEPAGE Y. (2010). Bilingual lexicon induction : Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, p. 252–256. HAL : hal-00488768.

PHILIPS L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*, **18**(6), 38–43.

STEINGRÍMSSON S., LOFTSSON H. & WAY A. (2021). CombAlign : a Tool for Obtaining High-Quality Word Alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, p. 64–73, Reykjavik, Iceland (Online) : Linköping University Electronic Press, Sweden.

ÖSTLING R. & TIEDEMANN J. (2016). Efficient Word Alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics*, **106**(1), 125–146. DOI : 10.1515/pralin-2016-0013.