Transfert zero-shot pour l'étiquetage morphosyntaxique : analyse de l'impact de la transformation des données à étiqueter pour les dialectes alsaciens

Delphine Bernhard

Actes des journées LIFT 2023
Nancy, France, 20-21 novembre 2023

Transfert *zero-shot* pour l'étiquetage morphosyntaxique : analyse de l'impact de la transformation des données à étiqueter pour les dialectes alsaciens

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

RÉSUMÉ

Nous présentons et évaluons une méthode de transformation des données pour améliorer les performances du transfert *zero-shot* pour l'étiquetage morphosyntaxique des dialectes alsaciens. Le corpus à annoter est transformé à l'aide de trois procédures simples, reposant notamment sur des lexiques bilingues alsacien-allemand. Les résultats obtenus avec des modèles entraînés pour l'étiquetage morphosyntaxique en diverses langues (de Vries *et al.*, 2022), plus ou moins proches de l'alsacien, montrent des gains importants sans ré-entraînement.

ABSTRACT

Zero-shot Transfer for POS Tagging : an Analysis of the Impact of Target Data Transformations for the Alsatian Dialects

We present and evaluate a data transformation method to improve the performance of *zero-shot* transfer for POS tagging of Alsatian dialects. The corpus to be annotated is transformed using three simple procedures, based in particular on bilingual Alsatian-German lexicons. The results obtained with models trained for POS tagging in various languages (de Vries *et al.*, 2022), more or less similar to Alsatian, show substantial gains without retraining.

MOTS-CLÉS : dialectes alsaciens ; transfert *zero-shot* ; étiquetage morphosyntaxique.

KEYWORDS: Alsatian dialects ; zero-shot transfer ; POS tagging.

1 Introduction

Les grands modèles de langue multilingues se prêtent aux approches par transfert ne nécessitant pas de ressources annotées pour la langue cible (*zero-shot*). Ces approches sont supposées être particulièrement utiles pour les langues disposant de peu de ressources ; nous nous intéresserons ici plus particulièrement à l'étiquetage morphosyntaxique pour les dialectes alsaciens.

Les travaux de de Vries *et al.* (2022) ont montré que différents éléments sont à prendre

en compte pour l'étiquetage morphosyntaxique par transfert inter-langues et notamment la présence de textes de la langue cible dans le corpus utilisé pour le pré-entraînement du modèle de langue. Or, les modèles multilingues les plus utilisés que sont mBERT (Devlin *et al.*, 2019; Devlin, 2019) et XLM-R (Conneau *et al.*, 2020) ont été pré-entraînés pour 104 et 100 langues, respectivement. Ainsi, la grande majorité des langues sont absentes des données utilisées pour produire ces modèles. Ceci peut expliquer en partie pourquoi la performance mesurée par de Vries *et al.* (2022) pour 65 langues sources et 105 langues cibles se limite à une exactitude moyenne de 57,4%, ce qui est faible pour une tâche réputée "simple" comme l'étiquetage morphosyntaxique.

Afin de contrebalancer l'absence de pré-entraînement pour la langue cible, plusieurs approches ont été proposées, de manière à tirer parti des ressemblances entre langues. Elles consistent à transformer des données d'une langue connue du modèle pour les rapprocher au mieux de la langue cible ou, au contraire, transformer les données de la langue cible pour les approcher d'une langue du modèle. Dans cet article, nous évaluons ce dernier type d'approche, en utilisant notamment des lexiques bilingues alsacien-allemand.

2 Transformation des données pour les méthodes par transfert

En transfert inter-langues sans ressources, la transformation des données peut se faire à différents niveaux : corpus de pré-entraînement du modèle de langue, données d'affinage pour la tâche cible, données cibles.

Ainsi, Hana *et al.* (2011) décrivent une méthode d'étiquetage morphosyntaxique pour le vieux tchèque dont la stratégie consiste à transformer un corpus de tchèque moderne pour qu'il ressemble au vieux tchèque et, inversement, à transformer un corpus de vieux tchèque pour qu'il ressemble au tchèque moderne. Ces transformations font notamment appel à des règles de modification phonémique et graphémique. Bernhard & Ligozat (2013) reprennent certaines de ces idées pour l'étiquetage morphosyntaxique de l'alsacien, en remplaçant les règles de changement par un lexique bilingue allemand-alsacien limité aux mots grammaticaux, qui sont les plus fréquents. Ce lexique est utilisé pour transposer les textes cibles en alsacien vers l'allemand, avant des les étiqueter avec des outils entraînés pour l'allemand. Cette méthode simple augmente les performances de l'étiquetage.

Wang *et al.* (2022) proposent également l'utilisation de lexiques bilingues pour synthétiser des données, en justifiant cette approche par la plus grande disponibilité de lexiques bilingues par rapport aux corpus monolingues de grande taille pour une large majorité des langues de la planète. Les lexiques sont utilisés pour traduire des mots d'une langue bien dotée vers la langue cible, afin de générer des données permettant de poursuivre le pré-entraînement du modèle de langue multilingue ou encore des données annotées pour l'affinage pour la tâche cible. Les résultats montrent une augmentation significative de la performance pour les tâches d'étiquetage morphosyntaxique, analyse syntaxique et reconnaissance d'entités

nommées.

La méthode consistant à générer des données monolingues synthétiques à l'aide d'un lexique bilingue est également utilisée par Lothritz *et al.* (2022). Leur objectif est d'entraîner un modèle BERT pour le luxembourgeois en augmentant les données par des textes partiellement traduits de l'allemand vers le luxembourgeois. Cette traduction se limite aux mots outils. Toujours pour le luxembourgeois, Song *et al.* (2023) produisent un "pseudo" corpus parallèle luxembourgeois-anglais à partir d'un corpus allemand-anglais en utilisant lexique bilingue Le pseudo-corpus luxembourgeois-anglais est ensuite utilisé pour entraîner un modèle de traduction. dont les résultats restent toutefois largement inférieurs au modèle multilingue NLLB-large (Costa-jussà *et al.*, 2022).

D'autres approches font totalement l'impasse sur les lexiques bilingues et génèrent des données synthétiques par l'injection aléatoire de bruit dans les données disponibles pour la langue source mieux dotée. Aepli & Sennrich (2022) augmentent les données de pré-entraînement du modèle de langue en injectant du bruit au niveau des caractères (suppression, insertion, remplacement) afin de générer de la variation orthographique. Dans la mesure où les mots sont découpés en sous-mots par les tokéniseurs des modèles de langue, ces perturbations conduisent à des modifications dans la segmentation des mots. Cette méthode permet d'obtenir une augmentation de l'exactitude de 22 points de pourcentage pour l'étiquetage morpho-syntaxique des dialectes suisses allemands, par rapport à une méthode sans injection de bruit utilisant uniquement des données en allemand. Blaschke *et al.* (2023) reprennent cette méthode et en font une analyse détaillée pour l'étiquetage morphosyntaxique de 7 langues appartenant à 3 familles linguistiques, incluant l'alsacien. Leur étude montre que la différence entre les proportions de mots qui ont été découpés en sous-mots dans les données source et cible a une corrélation négative avec la performance : plus cette différence est faible, plus l'exactitude est élevée. Les expériences décrites dans notre article visent à comparer, pour l'alsacien, l'approche de Blaschke *et al.* (2023), qui suppose d'entraîner un nouveau modèle à partir des données transformées, à une approche *zero-shot* par transformation des données cibles.

3 Méthode

Nous utilisons un corpus de textes alsaciens étiqueté en parties du discours selon les catégories Universal Dependencies (De Marneffe *et al.*, 2021) et comportant 12 582 tokens de surface et 12 907 mots syntaxiques (Bernhard *et al.*, 2018, 2023). Ce corpus est transformé de manière à s'approcher de l'allemand à l'aide de trois procédures simples :

1. Accentuation (`A`) : suppression des diacritiques de voyelles spécifiques aux dialectes alsaciens et conversion vers la forme non accentuée. Seuls les umlauts ⟨ä, ö, ü⟩ sont conservés, car ils sont utilisés en allemand. Les apostrophes sont également normalisées vers la forme ⟨'⟩.

2. Classes fermées (C) : utilisation d'un lexique de conversion de l'alsacien vers l'allemand de formes appartenant aux classes fermées. Nous réutilisons directement le lexique constitué par Bernhard & Ligozat (2013), sans modification. Ce lexique contient 133 entrées et a été constitué par étude d'un petit corpus de 5 textes. Un seul de ces textes se trouve également dans le corpus annoté utilisé pour l'évaluation (soit 396 tokens)
3. Classes ouvertes (O) : utilisation d'un lexique de conversion de l'alsacien vers l'allemand de formes appartenant aux classes ouvertes. Nous réutilisons le lexique produit par Bernhard (2014, 2021). Si un mot en alsacien a plusieurs traductions possibles, seule la plus fréquente est conservée (fréquence dans le corpus `deu_news_2022_1M`[1] (Goldhahn *et al.*, 2012)). Le lexique final comporte 6 699 paires de mots alsacien-allemand.

Ces procédures peuvent également être combinées entre elles, de manière à augmenter le nombre de transformations sur le corpus d'entrée. La Table 1 récapitule le nombre et le pourcentage de mots transformés par chaque procédure tandis que la Table 2 donne des exemples. Nous indiquons également le nombre moyen de sous-mots par mot après tokénisation avec XLM-R-base (Conneau *et al.*, 2020), qui est le modèle de langue utilisé pour l'entraînement des modèles d'étiquetage morphosyntaxique de de Vries *et al.* (2022) que nous utilisons dans nos expériences. Ces modèles, qui ont été entraînés pour 65 langues sources, sont ensuite appliqués pour étiqueter les différentes versions du corpus alsacien.

Toutes les expériences et analyses sont réalisées à l'aide des principaux outils et bibliothèques Python suivants : *Hugging Face*[2] pour les modèles[3], les bibliothèques *Transformers v. 4.30.2* et *Datasets v. 2.13.0* (Tunstall *et al.*, 2022), *PyTorch v. 2.0.1*[4], *pandas v. 2.0.3* (pandas development team, 2023), *scikit-learn v. 1.3.0* (Pedregosa *et al.*, 2011), *matplotlib v. 3.7.2* (Hunter, 2007) et *seaborn v. 0.12.2* (Waskom, 2021).

4 Résultats

La Table 3 détaille les résultats, en terme d'exactitude, pour les 10 langues sources qui obtiennent les meilleurs résultats en moyenne pour l'alsacien et pour les différentes transformations. A titre de comparaison, nous faisons également figurer les résultats obtenus pour les dialectes suisses allemands (Aepli & Clematide, 2018) avec les mêmes langues sources (sans transformation des données) : ces dialectes sont très proches des dialectes alsaciens, en particulier l'aire haut alémanique au sud de l'Alsace, en zone limitrophe de la Suisse.

Ces résultats montrent un impact important des transformations simples sur les résultats

1. https://wortschatz.uni-leipzig.de/en/download/German#deu_news_2022
2. https://huggingface.co
3. https://huggingface.co/wietsedv/xlm-roberta-base-ft-udpos28-[codedelangu
4. https://pytorch.org/

Traitement	# transformations	% mots	# sous-mots / mot
Aucun	0	0%	1,92
A	2 586	20%	1,70
C	2 475	19%	1,82
O	749	6%	1,88
AC	4 344	34%	1,66
AO	3 108	24%	1,68
CO	3 127	24%	1,78
ACO	4 804	37%	1,64

TABLE 1 – Nombre et pourcentage de mots syntaxiques transformés selon chaque procédure. La dernière colonne indique le nombre moyen de sous-mots par mot après tokénisation avec XLM-R-base.

Texte original	Mìt	dr	Jugend	ìsch	nit	loos	!
sous-mots	M␣ì␣t	dr	Jugend	ì␣sch	nit	loo␣s	!
A	Mit	dr	Jugend	isch	nit	loo␣s	!
C	Mit	der	Jugend	ist	nicht	loo␣s	!
O	M␣ì␣t	dr	Jugend	ì␣sch	nada	loo␣s	!
AC	Mit	der	Jugend	ist	nicht	loo␣s	!
AO	Mit	dr	Jugend	isch	nada	loo␣s	!
CO	Mit	der	Jugend	ist	nada	loo␣s	!
ACO	Mit	der	Jugend	ist	nada	loo␣s	!

TABLE 2 – Exemples de sous-mots en fonction des pré-traitements.

obtenus. La suppression des accents à elle seule permet de gagner 7,2 points d'exactitude en moyenne par rapport aux données brutes pour l'ensemble des 65 langues sources, soit plus que l'utilisation d'un lexique de mots de classes ouvertes (gain de 2,1 points en moyenne). C'est d'ailleurs cette dernière ressource qui a l'impact le plus faible sur les résultats. Le lexique de mots grammaticaux permet d'augmenter le score d'exactitude de 14,1 points en moyenne sur l'ensemble des langues sources, confirmant ainsi les observations de Bernhard & Ligozat (2013). Enfin, les meilleurs résultats sont obtenus par combinaison de l'ensemble des ressources (ACO) : exactitude moyenne de 61,0 (+18,7) pour les 65 langues sources, et de 72,1 (+21,7) pour les 10 langues présentées dans la Table 3.

Si l'on met en regard les données des Tables 1 et 3, on constate que, globalement, plus le pourcentage de mots transformés augmente, plus le nombre moyen de sous-mots par mot diminue et plus le score d'exactitude augmente également. La diminution du nombre moyen de sous-mots par mot indique, indirectement, que les données se rapprochent davantage de celles utilisées pour pré-entraîner le tokéniseur du modèle de langue.

Ces résultats sont inférieurs mais très proches du score de 78 % d'exactitude obtenu par

Langue source	suisse	alsacien	A	C	O	AC	AO	CO	ACO
afrikaans	55,2	53,0	61,0	68,6	55,1	71,6	62,2	69,3	72,0
allemand	50,2	49,9	58,7	69,5	52,5	73,0	60,4	70,6	73,6
arménien	46,2	47,6	58,9	65,8	51,2	71,1	61,2	67,8	72,2
arménien occidental	58,2	55,6	65,6	71,0	59,0	74,8	67,7	72,4	75,5
bulgare	50,3	48,5	57,8	66,0	51,5	69,8	59,9	67,4	70,8
féroïen	54,1	50,5	59,5	67,5	52,9	71,3	61,2	68,6	72,0
gallois	49,9	50,4	57,6	66,9	52,4	69,5	58,8	67,8	70,0
lituanien	49,7	48,8	57,9	66,4	51,4	69,5	59,4	67,8	70,5
roumain	53,0	51,8	60,6	69,2	54,9	73,2	62,8	70,4	74,0
tchèque	50,8	49,6	58,1	67,9	52,1	71,3	59,4	69,4	72,2

TABLE 3 – Exactitude (en %) pour différents prétraitements. Les 10 langues sources représentées sont celles qui obtiennent les meilleurs résultats en moyenne pour l'alsacien. Les langues présentes dans les données de pré-entraînement pour XLM-R sont soulignées.

Blaschke *et al.* (2023) pour le même corpus alsacien, en manipulant le corpus allemand avant l'entraînement du modèle d'étiquetage morphosyntaxique. Dans notre cas, nous n'avons pas ré-entraîné les modèles et les avons utilisés tels quels.

5 Discussion

Conclusion Nous avons analysé une méthode peu coûteuse, simple à mettre en œuvre et ne nécessitant pas de ré-entraînement de modèles pour le transfert inter-langues de modèles d'étiquetage morphosyntaxique. Les ressources requises sont limitées et peuvent être facilement constituées par une étude de corpus ou à partir d'un lexique bilingue, même de taille limitée. Les résultats obtenus pour l'étiquetage de l'alsacien montrent que tous les modèles bénéficient des transformations, pas uniquement le modèle affiné pour l'allemand.

Perspectives Le lexique des mots de classes fermées gagnerait à être étendu, ce qui pourrait contribuer à augmenter encore l'exactitude. Par ailleurs, il serait utile de comprendre et expliquer pourquoi de si bons résultats sont obtenus avec l'arménien occidental et le roumain. Les bonnes performances globales du roumain, pour un large ensemble de langues cibles, avait déjà été remarqué par de Vries *et al.* (2022).

Limites Les travaux présentés dans cet article ont été réalisés pour une seule langue cible. Par ailleurs, le corpus alsacien utilisé ne représente qu'une partie de la variation observée dans l'espace dialectal germanique en Alsace et ne saurait être représentatif de l'ensemble des locutrices et locuteurs. Il faudrait donc étendre les expériences à d'autres langues pour vérifier si les conclusions restent valides. Enfin, une seule tâche a été évaluée (classification de tokens, et, plus particulièrement, étiquetage morphosyntaxique) et il faudrait donc vérifier si les transformations proposées ont un impact similaire pour d'autres types de tâches.

Remerciements

Nous remercions le Centre de Calcul Haute Performance de l'Université de Strasbourg pour avoir soutenu ce travail en fournissant un support scientifique et l'accès aux ressources informatiques. Une partie des ressources informatiques a été financée par le projet Equipex Equip@Meso (Programme Investissements d'Avenir) et le CPER Alsacalcul/Big Data.

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

Références

AEPLI N. & CLEMATIDE S. (2018). Parsing Approaches for Swiss German. In *Proceedings of SwissText 2018*.

AEPLI N. & SENNRICH R. (2022). Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In *Findings of the Association for Computational Linguistics : ACL 2022*, p. 4074–4083, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.321.

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.

BERNHARD D. (2021). Lexique multilingue alsacien – français – allemand relié aux synsets de BabelNet. DOI : 10.34847/nkl.3f9b2i11.

BERNHARD D., ERHART P., HUCK D. & STEIBLÉ L. (2023). Annotated Corpus for the Alsatian Dialects. version 3.0, DOI : 10.5281/zenodo.1170128.

BERNHARD D. & LIGOZAT A.-L. (2013). 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, Les Sables d'Olonne, France.

BERNHARD D., LIGOZAT A.-L., MARTIN F., BRAS M., MAGISTRY P., VERGEZ-COURET M., STEIBLE L., ERHART P., HATHOUT N., HUCK D., REY C., REYNÉS P., ROSSET S., SIBILLE J. & LAVERGNE T. (2018). Corpora with Part-of-Speech Annotations for Three Regional Languages of France : Alsatian, Occitan and Picard. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, p. 3917–3924, Miyazaki, Japan.

BLASCHKE V., SCHÜTZE H. & PLANK B. (2023). Does Manipulating Tokenization Aid Cross-Lingual Transfer ? A Study on POS Tagging for Non-Standardized Languages. In

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), p. 40–54, Dubrovnik, Croatia.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D. & MAILLARD J. (2022). No language left behind : Scaling human-centered machine translation. arXiv : 2207.04672.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal dependencies. *Computational linguistics*, **47**(2), 255–308.

DE VRIES W., WIELING M. & NISSIM M. (2022). Make the Best of Cross-lingual Transfer : Evidence from POS Tagging with over 100 Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7676–7685, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.529.

DEVLIN J. (2019). Multilingual bert readme document. https://github.com/google-research/bert/blob/master/multilingual.md.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GOLDHAHN D., ECKART T. & QUASTHOFF U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection : From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 759–765, Istanbul, Turkey : European Language Resources Association (ELRA).

HANA J., FELDMAN A. & 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*, p. 10–18.

HUNTER J. D. (2007). Matplotlib : A 2d graphics environment. *Computing in Science & Engineering*, **9**(3), 90–95. DOI : 10.1109/MCSE.2007.55.

LOTHRITZ C., LEBICHOT B., ALLIX K., VEIBER L., BISSYANDE TEGAWENDE., KLEIN J., BOYTSOV A., LEFEBVRE C. & GOUJON A. (2022). LuxemBERT : Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Language Resources and Evaluation Conference*, p. 5080–5089, Marseille, France : European Language Resources Association.

PANDAS DEVELOPMENT TEAM (2023). pandas-dev/pandas : Pandas v2.0.3. 10.5281/zenodo.8092754, DOI : 10.5281/zenodo.8092754.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

SONG Y., EZZINI S., KLEIN J., BISSYANDE T., LEFEBVRE C. & GOUJON A. (2023). Letz Translate : Low-Resource Machine Translation for Luxembourgish. In *5th International Conference on Natural Language Processing*, Guangzhou, China.

TUNSTALL L., VON WERRA L. & WOLF T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

WANG X., RUDER S. & NEUBIG G. (2022). Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 863–877, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.61.

WASKOM M. L. (2021). seaborn : statistical data visualization. *Journal of Open Source Software*, **6**(60), 3021. DOI : 10.21105/joss.03021.