Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines

Pablo Ruiz Fabo, Delphine Bernhard, Carole Werner

Création d'un corpus FAIR de théâtre en alsacien et normalisation de variétés non-contemporaines

Pablo Ruiz Fabo Delphine Bernhard Carole Werner
Université de Strasbourg, LiLPa UR 1339, 67000 Strasbourg, France
{ruizfabo,dbernhard,wernerc@unistra.fr}

RÉSUMÉ

Nous présentons des travaux en cours vers la création d'un corpus diachronique de pièces de théâtre en alsacien pour la période 1870-1940, publiquement disponible, encodé selon les recommandations de la Text Encoding Initiative (TEI) et suivant les principes FAIR pour la création de données de la recherche. Le corpus sera utile aux recherches en sociolinguistique historique et analyse dramatique. Nous décrivons le travail effectué en vue des pratiques FAIR et introduisons des questions de recherche en modélisation TEI de variables pertinentes pour l'analyse linguistique et dramatique. De façon générale, la création du corpus est un exemple des difficultés du travail avec les langues peu dotées. Particulièrement, le corpus présente de l'alternance codique et d'énormes défis pour l'identification automatique des variantes orthographiques, sur lesquels nous aimerions échanger avec la communauté.

ABSTRACT

Creating a FAIR corpus of Alsatian theater and orthographic normalization of non-contemporary varieties

We present work in progress towards creating a diachronic corpus of theater plays in Alsatian. The corpus is publicly available under an open license, encoded according to the Text Encoding Initiative (TEI) guidelines and strives to follow FAIR principles for scholarly data development. We describe our work towards FAIR practices and introduce research questions on the TEI modeling of variables relevant for sociolinguistic and drama analysis. This corpus creation effort exemplifies difficulties related to working with low-resource languages. The corpus shows code-switching and huge challenges for the automatic identification of orthographical variants, which we would like to discuss with the community.

MOTS-CLÉS : corpus, variation, alternance codique, langues peu dotées, TEI, théâtre alsacien.

KEYWORDS: corpus, variation, code-switching, under-resourced languages, TEI, Alsatian theater.

1 Introduction

Le projet MeThAL, « Vers une macroanalyse du théâtre en alsacien [1] », est en train de créer un corpus encodé en TEI (TEI Consortium, 2020) de pièces de théâtre en alsacien pour la période 1870-1940 [2] ;

1. Site du projet : https://methal.pages.unistra.fr/

2. Entre 1871 et 1918, l'Alsace est politiquement rattachée à l'Empire allemand. Le besoin d'auto-détermination des Alsaciens « par rapport au reste du monde allemand » (Huck *et al.*, 2007, 12) passera notamment par le théâtre alsacien et la mise en scène et création de l'Alsace. La date-borne supérieure correspond à l'annexion de l'Alsace au III[e] Reich.

la pièce fondatrice du théâtre dialectal en alsacien, le *Pfingstmontag* de J. G. Arnold (1816), fait également partie du corpus du fait de son importance et son influence dans les pièces plus récentes. Un volume de 50 pièces ou 400 000 tokens est visé. Le corpus est public [3] et suit des principes FAIR ou *Findable*, *Accessible*, *Interoperable*, *Reusable* (Wilkinson *et al.*, 2016). Dans la mesure où le corpus permet de documenter les pratiques langagières de son époque, il aidera à examiner des questions de sociolinguistique historique de l'Alsace (cf. Huck *et al.*, 2007; Huck, 2015). L'encodage permettra une analyse des types de personnages et de la variation linguistique telle que représentée dans leurs paroles selon leur âge, sexe, statut social ou origine, et facilitera aussi l'étude d'aspects formels de la technique dramatique.

Nous présentons des travaux en cours sur la modélisation des données et sur l'identification de variantes orthographiques, nécessaire à cause de l'énorme variabilité dans la représentation écrite de l'alsacien. Des questionnements se posent concernant la création de données linguistiques ouvertes, l'encodage de ressources multilingues qui présentent de l'alternance codique et les méthodes de traitement des langues peu dotées, notamment sur l'identification de variantes orthographiques dans un contexte de ressources linguistiques limitées.

L'article est structuré comme suit : La section 2 présente notre procédure d'encodage TEI et démarche FAIR et nos questionnements autour de la modélisation de variables sociales décrivant les personnages. La section 3 décrit le degré de variation orthographique présent dans le corpus (3.1) ainsi que des cas d'alternance codique (3.2). La section 4 aborde la question de l'identification automatique des variantes dans ce type de corpus.

2 Modélisation et FAIRisation des données

Cette section décrit nos sources, notre procédure d'encodage TEI et nos efforts d'adoption des principes FAIR. La modélisation des descripteurs socio-économiques des personnages est ensuite abordée, ainsi que des possibilités d'encodage TEI de la variation orthographique et de l'alternance codique.

2.1 Sources du corpus

La source principale du corpus est une collection représentative d'environ 150 pièces en alsacien numérisées en 2019 par la Bibliothèque nationale et universitaire (Bnu) à Strasbourg [4]. C'est une ressource électronique fondamentale mais qui demande des améliorations afin de faciliter la recherche linguistique et littéraire : les pièces sont disponibles comme des fichiers d'image, sans balisage, et sans OCR pour la plupart. Nous avons sélectionné un sous-ensemble des pièces visant la variété d'époques et de sous-genres dramatiques [5] et nous avons commencé son océrisation et encodage TEI.

3. Le corpus est mis à jour graduellement sur `https://git.unistra.fr/methal/methal-sources`
4. Voir `https://numistral.fr/fr/theatre-alsacien` (lien *[Découvrir]* pour explorer la collection)
5. Le rendu sur *Drama Corpora* de nos pièces encodées en donne un aperçu : `https://dracor.org/als`

2.2 Procédure d'encodage TEI

Le standard TEI permet la modélisation d'éléments d'analyse dramatique ainsi que de phénomènes linguistiques comme la variation et l'alternance codique. Après océrisation et validation manuelle du texte reconnu, notre encodage TEI s'effectue par une transformation automatique d'une sortie hOCR [6] de Tesseract [7]. Des indices typographiques et de mise en page dans cette sortie reflètent les divisions en acte et scène, répliques et didascalies. Le format est plus variable pour les listes de personnages et les pages de titre, qui fournissent des renseignements essentiels pour les analyses sociolinguistiques et thématiques, ainsi que pour les métadonnées bibliographiques. Afin de gérer ces contenus, nous les avons transcrits manuellement dans une base de données. Nos scripts d'encodage fusionnent ces informations avec la sortie hOCR pour créer les versions TEI. La figure 1 présente la chaîne de traitement.

Notre automatisation de l'encodage TEI repose sur des règles de transformation créées manuellement. Nous voudrions à l'avenir évaluer l'applicabilité de méthodes d'apprentissage automatique, en nous inspirant des travaux de Khemakhem *et al.* (2017, 2018) pour l'encodage TEI de dictionnaires avec des CRF (champs aléatoires conditionnels), qui exploitent la typographie et la mise en page pour la prédiction de la structure TEI. Il serait pertinent de comparer la productivité permise par une telle approche et par notre chaîne de traitement actuelle.

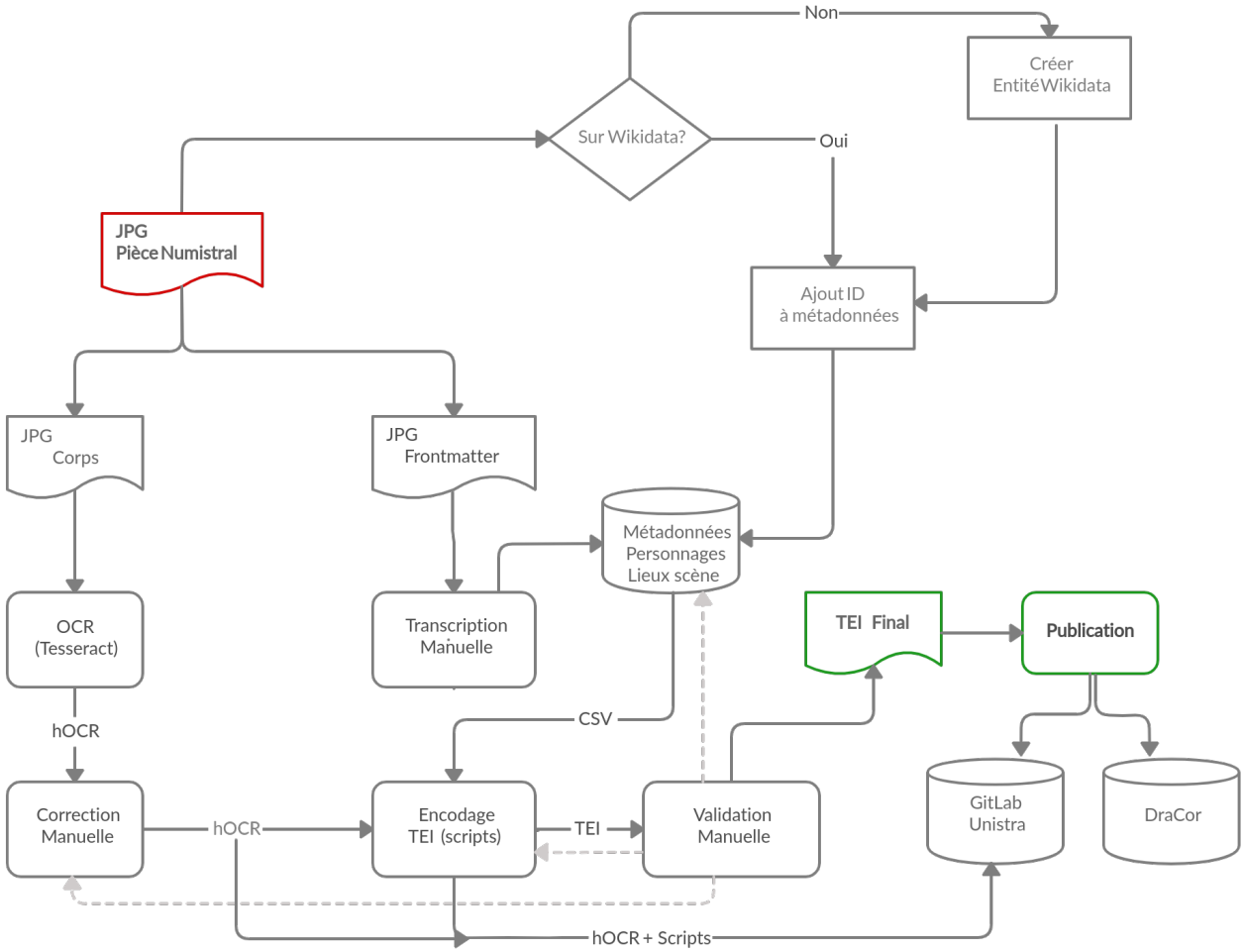

FIGURE 1 – Chaîne de traitement

6. Pour le standard hOCR, voir `http://kba.cloud/hocr-spec/1.2/`
7. `https://github.com/tesseract-ocr/tesseract`

2.3 FAIRisation

Nous visons la création d'un corpus FAIR. Nous avons travaillé sur son interopérabilité et réutilisabilité, et entrepris de premiers pas vers la trouvabilité et accessibilité. Ont contribué à l'interopérabilité l'adoption du standard TEI et l'utilisation d'identifieurs Wikidata pour les pièces et les auteurs, incluant notre création des nouvelles entités Wikidata nécessaires [8]. Concernant la réutilisabilité, chaque pièce est publiée sous une licence ouverte. Pour promouvoir la transparence du processus de prétraitement et d'encodage, les scripts et ressources créés pour traiter chaque pièce, ainsi qu'un wiki pour documenter nos pratiques, sont publiés sur nos dépôts git [9].

Le corpus a des métadonnées riches, en accord avec les requis FAIR pour la trouvabilité des ressources (Wilkinson *et al.*, 2016, 4). Or, il manque à ce jour des identifieurs persistants (DOI ou semblables). Deux options seront considérées dans ce sens : le service d'exposition de données Nakala (Huma-Num, 2020) et le service TEI2Zenodo (Wagner, 2020). C'est aussi par le biais d'une plate-forme d'exposition de données que le corpus sera rendu conforme au critère FAIR d'accessibilité, qui met l'emphase sur l'accès aux données et métadonnées par des programmes informatiques, avec des protocoles de communication standard. Une accessibilité dans un sens moins technique est déjà garantie car le corpus est disponible sur GitLab et sur la plate-forme DraCor[5] (Fischer et Börner, 2019). Celle-ci permet, profitant du balisage TEI, l'accès programmatique à des éléments structurels des pièces (p. ex. toutes les répliques par des femmes ou toutes les didascalies) via une API HTTP.

2.4 Encodage de la variation orthographique et de l'alternance codique

Le corpus doit permettre la comparaison du contenu des pièces, de sorte à faciliter l'analyse de tendances dans les sujets abordés selon diverses variables. À cette fin, la variation orthographique des pièces (voir section 3 pour des exemples) doit être neutralisée ; l'identification automatique des variantes d'un même lexème [10] est un vrai défi, discuté dans la section 4.

Une fois le lexème identifié, la TEI propose des façons naturelles d'encoder la relation entre la variante et son lexème. Une option serait de créer des identifieurs uniques pour les lexèmes du corpus et les donner dans un attribut `@xml:id`. Une autre option serait d'effectuer une normalisation des variantes vers un norme concrète et d'utiliser un élément `<choice>` dont les fils `<orig>` et `<reg>` contiendront la variante originale et normalisée respectivement.

Concernant l'alternance codique, un encodage de base consiste à créer des éléments `<seg>` avec un attribut `@xml:lang` pour le code de la langue de la séquence ; nous avons déjà utilisé cette option dans l'encodage de *D'r Poetisch Oscar* par Marie Hart :

```
<sp who="#oscar">
  <speaker>OSCAR:</speaker>
  <p>Un Sie han m'r wieder d'rzue verholfe, Madame Lewermann,
  Sie ellein verstehn min poetisch Empfinde.
  <seg xml:lang="fre">Vous êtes ma muse</seg>.</p>
</sp>
```

8. Nous avons travaillé sur un sous-ensemble des entités montrées par cette [requête SPARQL] sur Wikidata.
9. Pour scripts/ressources, voir le dossier [work] du dépôt.
10. Suivant Bernhard (2014), nous utilisons *lexème* dans le sens de *lexeme* chez Bauer (2003) : Un mot du dictionnaire ; une unité abstraite du vocabulaire, réalisée par des mot-formes représentant le lexème et sa morphologie flexionnelle. Une des formes est choisie par convention afin de nommer le lexème dans une entrée de dictionnaire ou ouvrage similaire.

2.5 Modélisation des descripteurs sociaux des personnages

Une question de modélisation qui se pose avec le corpus concerne la formalisation des variables sociales qui décrivent les personnages et les relations entre eux ; il s'agit d'attributs des personnages pertinents pour l'analyse linguistique et dramatique. Des typologies pour modéliser les personnages, formalisables en TEI, existent déjà (Galleron, 2017). Or, elles peuvent être complétées concernant la description des professions des personnages. Nous avons commencé à développer une typologie multilingue de professions avec des termes en alsacien, français et allemand (langues des professions dans les listes de personnages du corpus) ainsi qu'en anglais, car notre recherche cible un public intéressé aux langues régionales mais qui ne maîtrise pas forcément l'alsacien, et souvent anglophone. Une question de recherche est de savoir comment représenter la typologie de façon à faciliter l'interopérabilité et son intégration dans l'encodage TEI. Tant les « feature structures » (hiérarchies de caractéristiques) proposées par Galleron que le formalisme sur la base d'attributs RDFa (un format web sémantique) intégrés dans la TEI (Ruiz Fabo *et al.*, 2020) peuvent être considérés.

3 Variation et alternance codique

Le corpus présente une énorme variabilité orthographique. À ceci s'ajoute l'alternance codique entre alsacien, allemand et français. Cette section montre des exemples de chaque phénomène, qui donnent une indication des défis que pose le corpus pour des tâches de TAL comme l'identification de variantes orthographiques.

3.1 Variation

Les parlers dialectaux d'Alsace sont caractérisés par une grande variation à l'oral, qui se traduit par autant de variation à l'écrit. Dans les pièces de théâtre, la variation dans la scripturalisation dépend de la variété dans laquelle s'exprime le dramaturge, mais aussi des variations 'internes' à la pièce, c'est-à-dire relatives aux personnages, en fonction de leur origine géographique et partant, sociale. On peut citer comme cas d'école le discours prêté aux personnages présents dans le *Pfingstmontag* (1816) de J.G. Arnold, première pièce de théâtre en alsacien, dont le but est de « dresser un petit monument linguistique alsacien [11] ». Dans cette comédie, ce sont les dialectes et autres variétés linguistiques présentes en Alsace à l'époque (allemand 'standard' et français) qui sont véritablement mis en scène par leurs personnages. Dans cette pièce on retrouve des représentants de la bourgeoisie strasbourgeoise, s'exprimant dans la variété dialectale de la ville, mais aussi des représentants stéréotypés de la paysannerie du Kochersberg (une région rurale proche de Strasbourg).

La variation graphique peut donc varier d'un personnage à l'autre, comme c'est le cas chez Claus, le paysan du Kochersberg s'exprimant dans sa variante dialectale et chez Wolfgang, magister ès philosophie, s'exprimant quant à lui en allemand 'standard'. Les deux personnages emploient le verbe *(an)fragen* (questionner), ce qui donne les variations de scripturalisation dans (1) et (2). [12]

11. « [D]ie Bestimmung eines kleinen alsatischen Sprachdenkmals », comme l'exprime Arnold dans sa préface au *Pfingstmontag*.

12. La graphie qui représente la racine du verbe, sans préfixes ou suffixes, est identifiée en caractères gras. Les traductions vers le français sont données avec les exemples. Des versions encodées en TEI pour les pièces citées sont disponibles sur notre dépôt public, sauf dans le cas du *Herr Maire* (disponible sans encodage sur Numistral).

(1) I **fröau** ob err no' brüche d' Pfärd
Je demande si vous avez encore besoin des chevaux

(2) Wir sollten doch zuerst bei ihr zu Haus an**frag**en
Nous devrions d'abord aller poser la question chez elle

Dans le *Chrischtowe* de Clemens (3), ainsi que dans *Sainte Cécile* de Julius Greber (4), la racine du même verbe présente tant la graphie *frö* que *fröu* ; cette dernière est aussi trouvée dans *In's Ropfer's Apothek* par Gustave Stoskopf (5).

(3) Äi sie ruefe mr alli e Üwername. Dr Schuelmäischter hett mi g**frö**jt wie „der Ofen" häisst — no hawi gsäit „Furneau"
Ils me donnent tous un surnom. Le maître d'école m'a demandé comment on dit « der Ofen » [le four] — j'ai dit « Furneau »

(4) Do kannsch lang **fröu**je — — er saat nix, ken Wort schnüüft er
Tu peux redemander sans cesse — — Il dit rien, il ne pipe mot

(5) Ich hab e schoene Schrecke bekumme, wie 'r mich waje d'r Susanne g'**fröu**it hett
J'ai vraiment eu peur quand il m'a demandé par rapport à Susanne

Dans *D'r Herr Maire* (1898) de Stoskopf, différentes variétés sont également mises en scène et un même lexème peut à nouveau prendre des graphies divergentes à l'extrême. Dans (6), *Daö* représente l'adaptation phonographique au dialecte du Kochersberg de *Tag* (*jour*).

(6) Un dass dich guet schicksch un Savuar-Wiewr an de **Daö** leisch !
Tu as intérêt à bien te comporter et à faire preuve de savoir-vivre

Les occurrences ***Daa*** et ***Tag*** apparaissent dans la même pièce ; la première est prononcée par le fils du riche épicier strasbourgeois Pfeffer, qui s'exprime dans sa variante strasbourgeoise et la seconde apparaît dans une lettre, écrite en allemand standard, faisant également état de la diglossie médiale alors en vigueur.

3.2 Alternance codique

Le corpus présente de l'alternance codique entre variétés alsaciennes, français et allemand ; dans certains cas d'autres variétés régionales sont également présentées, comme c'est le cas de l'allemand de Saxe chez *D'r Hoflieferant* par Stoskopf, à travers le personnage Hans Grinsinger.

À part le mélange d'autres langues avec le français, une caractéristique additionnelle dans certaines pièces est l'écriture du français 'à l'alsacienne'. Dans le *Pfingstmontag*, le personnage du licencié, Alsacien âgé essayant de montrer l'étendue de ses connaissances en français, est particulier, dans la mesure où son discours est truffé de termes français, dont la prononciation est largement adaptée au dialecte alsacien, comme le révèlent les graphies dans (7a-e) :

(7a) Nong
Non

(7b) Pardong
Pardon

(7c) Wui wui
Oui, oui

(7d) Sannebawrä
Ça n'est pas vrai

(7e) Ong nangtang riäng ... Pong, Pong ... Mongtong dong sangfassong
On n'entend rien ... Bon, bon ... Montons donc sans façons

Le *Herr Maire* de Stoskopf (1898) reprend l'idée de transcrire le 'français-alsacien' déjà utilisé par Arnold en 1816 : On le voyait dans l'expression *Savuar-Wiewr* pour *savoir-vivre* de l'exemple (6) ci-dessus.

D'r Hoflieferant de Stoskopf (1905) est un autre exemple des subtilités qui peuvent être représentées dans le corpus concernant l'alternance de variantes. Dans cette pièce, les personnages utilisent parfois la prononciation française ou allemande des noms de famille pour exprimer leur identité et leur proximité à leur interlocuteur ou leur rejet de celui-ci ; l'utilisation de la prononciation française est alors indiquée en italiques, comme dans l'exemple suivant par le personnage Fritz Grinsinger :

(8) *Pardon*, dass ich Sie unterbrech, erschtens bin ich noch lang nit Ihr Liewer und zweitens heiss ich nit Grinsinger [avec prononciation allemande], ich heiss *Grinsinger* [avec prononciation française, en italiques dans l'original].
Pardonnez mon interruption, mais premièrement je ne suis pas votre cher [monsieur Grinsinger] et deuxièmement je ne m'appelle pas Grinsinger, mais *Grinsinger*.

Comme le montrent les exemples dans cette section, le corpus va au-delà de cas 'simples' d'alternance codique. Nous prévoyons une représentation TEI basique du phénomène avec des éléments `<seg>` et des attributs `@xml:lang`, comme vu en (2.4). La possibilité d'encoder plus de détails (ce qui serait évidemment permis par le standard TEI) est une question ouverte. La détection automatique des cas d'alternance codique est un autre sujet de recherche possible sur le corpus.

4 Identification automatique de variantes orthographiques

La neutralisation des variantes est incontournable pour comparer le contenu des pièces et faire des analyses thématiques, p. ex avec le *topic modeling* (Blei, 2012) ou des méthodes de textométrie (Lebart *et al.*, 2019) ; ces méthodes demandent une représentation orthographique homogène du vocabulaire. En outre, cette neutralisation pourra bénéficier à la recherche en texte intégral, une fonction de base de l'interface d'exploration du corpus qui sera développée dans la phase finale du projet.

Ces questionnements ne sont pas nouveaux et deux approches différentes peuvent ici être envisagées : soit les variantes sont normalisées vers une forme correspondant à une norme choisie, soit elles sont tout simplement reconnues comme étant des variantes, sans qu'il y ait pour autant une normalisation explicite.

La normalisation orthographique automatique, en tant que tâche de Traitement Automatique des Langues (TAL) a notamment été appliquée pour l'analyse de textes du web social (Han et Baldwin, 2011; Alegria *et al.*, 2015; Doval *et al.*, 2020). Dans ce cas précis, la normalisation des mots hors vocabulaire (fautes d'orthographe, orthographe non conventionnelle, abréviations) se fait généralement vers la forme standard. La normalisation est aussi utilisée pour les variétés historiques (Etxeberria *et al.*, 2016; Bollmann *et al.*, 2017; Bollmann, 2019). La norme est alors souvent la forme standard contemporaine, même si cela pose la question des formes disparues, qui n'ont pas d'équivalent dans la variété contemporaine.

La deuxième approche consiste à identifier les variantes sans chercher pour autant à les normaliser : en effet, pour de nombreuses applications, comme la recherche dans un corpus, la normalisation n'est pas nécessaire. Il s'agira ainsi de repérer les variantes, par exemple à l'aide de méthodes non

supervisées de *clustering* (Dasigi et Diab, 2011; Rafae *et al.*, 2015) ou des méthodes supervisées qui déterminent si deux formes sont des variantes ou non (Barteld *et al.*, 2019). Nous nous orientons également vers ce type d'approche, pour faire suite à de premières expériences visant à identifier les variantes dans des lexiques bilingues alsacien-français (Bernhard, 2014) [13]. Il n'y a en effet pas de "norme" orthographique stable à laquelle nous pourrions nous référer pour les dialectes alsaciens. Même si l'allemand est souvent considéré comme la forme écrite à privilégier pour l'alsacien, cela ne reflète pas la réalité de nos corpus, comme nous avons pu le montrer dans la section précédente.

Nous testons actuellement des méthodes de classification supervisée (cf. Barteld *et al.*, 2019) et les résultats sont en cours d'analyse. Nous aimerions échanger avec la communauté sur des approches permettant de profiter au mieux d'un nombre limité de données d'entraînement, avant de nous engager dans la création de nouvelles données (annotées) pour la tâche.

5 Perspectives

Après avoir encodé les premières pièces du corpus, plusieurs intérêts de recherche, en partie évoqués *supra*, sont les suivants : d'un côté, implémenter la modélisation TEI des variables sociales décrivant les personnages. D'un autre côté, l'application possible de méthodes d'apprentissage automatique à la détection des éléments structurels des pièces (répliques, didascalies) pour leur encodage TEI automatique. Finalement, nous sommes en train d'évaluer l'application de méthodes de TAL à l'identification automatique de variantes, ce qui constituerait un bon apport à l'exploitabilité du corpus pour des analyses linguistiques et de contenu. En outre, la FAIRisation du corpus sera complétée par sa mise à disposition sur des plateformes ouvertes d'exposition de données.

Remerciements

Ce travail a bénéficié d'un financement dans le cadre de l'IdEx Université de Strasbourg. Nous remercions également les stagiaires ayant participé à l'encodage des pièces : Audrey Deck et Soihira El-Kabir. Merci aux relecteur·trice·s pour leurs commentaires détaillés qui ont aidé à améliorer l'article.

Références

ALEGRIA, I., ARANBERRI, N., COMAS, P. R., FRESNO, V., GAMALLO, P., PADRÓ, L., SAN VICENTE, I., TURMO, J. et ZUBIAGA, A. (2015). TweetNorm : a benchmark for lexical normalization of Spanish tweets. *Language Resources and Evaluation*.

BARTELD, F., BIEMANN, C. et ZINSMEISTER, H. (2019). Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation*, pages 1–30.

BAUER, L. (2003). *Introducing linguistic morphology*. Edinburgh University Press Edinburgh.

13. Les habitudes de scripturalisation du corpus (utilisation du eszett par exemple, ou utilisation du graphème simple <u> pour rendre le <ou> français) sont obsolètes par rapport aux pratiques actuelles, ce qui demande l'adaptation des méthodes ; des ressources pour le TAL en alsacien ont été développées par le projet ANR RESTAURE (Bernhard *et al.*, 2019) mais un corpus diachronique de théâtre demande d'élargir les ressources.

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)*, pages 23–29, Reykjavík, Iceland.

BERNHARD, D., BRAS, M., ERHART, P., LIGOZAT, A.-L. et VERGEZ-COURET, M. (2019). Language Technologies for Regional Languages of France : The RESTAURE Project. *In International Conference Language Technologies for All (LT4All) : Enabling Linguistic Diversity and Multilingualism Worldwide*, Paris, France.

BLEI, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

BOLLMANN, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. *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)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

BOLLMANN, M., BINGEL, J. et SØGAARD, A. (2017). Learning attention for historical text normalization by learning to pronounce. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 332–344, Vancouver, Canada. Association for Computational Linguistics.

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

DOVAL, Y., VILARES, J. et GÓMEZ-RODRÍGUEZ, C. (2020). Towards robust word embeddings for noisy texts. *arXiv :1911.10876 [cs]*. arXiv : 1911.10876.

ETXEBERRIA, I., ALEGRIA, I., URIA, L. et HULDEN, M. (2016). Evaluating the Noisy Channel Model for the Normalization of Historical Texts : Basque, Spanish and Slovene. *In LREC*.

FISCHER, F. et BÖRNER, I. (2019). Programmable Corpora : Introducing DraCor, an Infrastructure for the Research on European Drama. *In Digital Humanities 2019*, page 5, Utrecht.

GALLERON, I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm : Needs, Problems and Foreseen Solutions. *Human and Social Studies*, 6(1):88–108.

HAN, B. et BALDWIN, T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a #twitter. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

HUCK, D. (2015). *Une histoire des langues de l'Alsace*. La Nuée Bleue.

HUCK, D., BOTHOREL-WITZ, A. et GEIGER-JALLET, A. (2007). L'Alsace et ses langues. Eléments de description d'une situation sociolinguistique en zone frontalière. *In Aspects of Multilingualism in European Border Regions : Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, pages 13–101. EURAC Research (Europäische Akademie / Accademia Europea / European Academy), Bozen/Bolzano.

KHEMAKHEM, M., FOPPIANO, L. et ROMARY, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *In electronic lexicography, eLex 2017*, Leiden, Netherlands.

KHEMAKHEM, M., ROMARY, L., GABAY, S., BOHBOT, H., FRONTINI, F. et LUXARDO, G. (2018). Automatically Encoding Encyclopedic-like Resources in TEI.

LEBART, L., PINCEMIN, B. et POUDAT, C. (2019). *Analyse des données textuelles*. Presses de l'Université du Québec, 1 édition.

RAFAE, A., QAYYUM, A., MOEENUDDIN, M., KARIM, A., SAJJAD, H. et KAMIRAN, F. (2015). An Unsupervised Method for Discovering Lexical Variations in Roman Urdu Informal Text. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 823–828.

RUIZ FABO, P., BERMÚDEZ SABEL, H., MARTÍNEZ CANTÓN, CLARA et GONZÁLEZ-BLANCO, ELENA (2020). The Diachronic Spanish Sonnet Corpus (DISCO) : TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. *Digital Scholarship in the Humanities*.

TEI CONSORTIUM (2020). TEI P5 : Guidelines for Electronic Text Encoding and Interchange. Publisher : Zenodo.

WAGNER, A. (2020). TEI XML to Zenodo service published : Automatic depositing the project's TEI files at a long-term archive – Die Schule von Salamanca.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3(1).