

The MeThAL Alsatian theater corpus and related resources: Work done and perspectives

Pablo Ruiz Fabo

Université de Strasbourg – Laboratoire LiLPa UR 1339, 67000 Strasbourg, France
ruizfabo@unistra.fr

ABSTRACT

As the MeThAL project nears completion, we present resources created and future perspectives. The first large TEI corpus of theater plays in Alsatian varieties was created (>500,000 words) and distributed under an open licence; a program for semi-automatic TEI-encoding based on OCR output was also released. A character prosopography in TEI was constructed; characters are annotated with metadata of potential sociolinguistic relevance, like their gender or social class. An Alsatian emotion lexicon was developed, and a program to analyze the distribution of emotion terms in the plays, aggregating them by metadata like the characters' social variables or the plays's genre. Work related to Alsatian's large scriptolinguistic variation was performed, especially regarding emotion terms. To open the resources to a non-specialist public, a corpus exploration interface was created.

RÉSUMÉ

Le corpus MeThAL de théâtre en alsacien et ressources reliées : Réalisations et perspectives

Nous présentons les ressources développées et les perspectives, près de la clôture du projet MeThAL, qui a créé le premier grand corpus de théâtre alsacien en TEI (>500 000 mots), distribué sous licence ouverte. Un programme d'encodage TEI semi-automatique sur la base de sorties OCR a été distribué. Une prosopographie TEI de personnages a été construite ; elle décrit les personnages avec des variables sociales potentiellement pertinentes pour une analyse sociolinguistique (p. ex. genre des personnages, classe sociale). Un lexique d'émotions a été développé, ainsi qu'un programme permettant de l'appliquer pour analyser la distribution de ses termes dans le corpus. Le lexique est adapté à l'énorme variation scriptolinguistique des variétés alsaciennes. Pour ouvrir les ressources à un public non spécialisé, une interface d'exploration de corpus a été mise en ligne.

MOTS-CLÉS : théâtre alsacien, corpus TEI, analyse d'émotion, variation, sociolinguistique.

KEYWORDS: Alsatian theater, TEI corpus, emotion analysis, variation, sociolinguistics.

1 Introduction

The MeThAL project¹ has developed the first large public TEI-encoded corpus for Alsatian theater and related linguistic resources; the project’s goals were presented at an early stage at LIFT’s 2020 workshop. As the project nears completion, we would like to share with the same community the work done and perspectives.

2 Resources developed

In this section, we present the TEI-encoded corpus (2.1), the character prosopography (2.2), emotion analysis resources (2.3) and work on orthographic variation (2.4). Finally, we present the TEI-encoding workflow and corpus navigation interface (2.5).

2.1 TEI-encoded corpus

The first resource is the **TEI corpus**, with 77 plays (623,977 tokens), published in different locations and representing different varieties, by 27 authors. We included minor authors and well-known ones, both male and female, and several dramatic genres. Regarding genres, the corpus reflects comedy’s predominance in the Alsatian tradition, but we also included dramas, popular dramas (*Volksstücke*) and tales (*Weihnachtsmärchen*). Out of the 77 plays, 51 are plays for which no previous electronic text existed; we carried out OCR based on image-mode digitizations at Strasbourg’s national library (Bnu)² and a manual correction of the OCR prior to semi-automatic TEI encoding (section 2.5 describes the workflow). The remaining 26 plays had been published in wikitext format on Wikisource and, in those cases, our work consisted in converting the wikimarkup to TEI, using rules.³

As per FAIR practices, rich metadata were collected in the TEI header; an ODD with embedded schematron rules ensures metadata consistency. We assign a DOI to each resource via the Nakala data repository. The resources are exposed in a Nakala collection and via GitLab.⁴

¹<https://methal.pages.unistra.fr/en>

²See <https://www.numistral.fr/fr/theatre-alsacien>

³The Wikisource transcriptions were carried out by wikipedian MireilleLibmann https://als.wikipedia.org/wiki/Text:August_Lustig/A._Lustig_S%C3%A4mtliche_Werke:_Band_2

⁴The Nakala collection is at <https://nakala.fr/collection/10.34847/nkl.feb4r8j9>. The GitLab repository is at <https://git.unistra.fr/methal/methal-sources>. Note that, as of this writing, the Nakala collection contains only 37 of the corpus plays. The rest are published on GitLab and are usable in their current version, having already been used for emotion analyses (Bernhard & Ruiz Fabo, 2022; Liu *et al.*, 2023; Ruiz Fabo *et al.*, 2024), but await more detailed verification before we assign a DOI to them via Nakala.

2.2 TEI feature structures for character social annotations

A **character prosopography in TEI** was also created (Ruiz Fabo *et al.*, 2024),⁵ using the feature structures formalism (Romary, 2015), improving upon proposals for character description by Galleron (2017). Formalizing annotations with feature structures presents some degree of complexity, and consistency was ensured with the FS-VALIDATOR tool (Bermúdez Sabel, 2022), which generates ODD statements based on the feature structure declaration.

In the prosopography, characters are annotated with social variables such as their gender, social class, profession or professional category; we built a taxonomy of socioprofessional groups appropriate for the corpus context (Ruiz Fabo & Werner, 2021). These social variables have potential relevance for a sociolinguistic description of character speech. In section 3, we will discuss challenges related to analyzing such data.

2.3 Emotion analysis

A resource related to the corpus is Bernhard’s ELAL (*Emotion Lexicon for Alsatian*);⁶ it is based on French-Alsatian bilingual lexica and existing emotion lexica for French and German, aggregated into a large network. Emotion term variants are identified with graphical similarity methods. Thanks to this, the lexicon handles Alsatian’s huge orthographic variability. The MeThAL corpus was used to extract emotion terms’ variants (Bernhard & Ruiz Fabo, 2022), and the lexicon was then applied for first analysis of emotion vocabulary in Alsatian plays.

We applied the ELAL lexicon in EDYTHA (*Emotion Dynamics in Theater in Alsatian*), a program by Liu *et al.* (2023), which detects emotion vocabulary trends in a TEI corpus. It is inspired by Vishnubhotla & Mohammad’s (2022) *TED* tool,⁷ but improves its scoring via corpus-driven term-weighting schemes that prevent potential skews due to frequent lexicon terms. Taking advantage of the MeThAL character prosopography, EDYTHA aggregates emotion scores based on metadata like the character’s gender, social class or professional group. This is interesting for purposes like assessing whether characters in a socially disfavoured position use more negative emotional terms than other groups. The tool also aggregates scores based on the plays’ genre (e.g. comedies vs. dramas). The program was released under GPL.⁸

⁵<https://git.unistra.fr/methal/methal-sources/-/tree/master/personography>

⁶See <https://nakala.fr/10.34847/nkl.40cex998> for the lexicon items and <https://nakala.fr/10.34847/nkl.39b7617v> for their emotion scores

⁷*TED* (Tweet Emotion Dynamics): <https://github.com/Priya22/EmotionDynamics>

⁸<https://git.unistra.fr/methal/edytha>

2.4 Approaches to orthographic variation

We also worked on Alsatian’s large **orthographic variation**. Bernhard performed successful variant detection experiments, reported in Ruiz Fabo *et al.* (2024), using the ELAL lexicon as a training corpus with classical machine learning models and neural methods. Yang (2022) induced scriptural variation rules via methods in Millour (2020), applicable to multiple sequence alignment.

2.5 TEI-encoding automation and corpus navigation interface

A workflow for **semi-automatic TEI encoding** based on OCR outputs, using rules, a conditional random fields (CRF) model, and manual revision was described in Ruiz Fabo *et al.* (2024). We released the software and the CRF model under an open license (Briand & Ruiz Fabo, 2023). For OCR, we used Tesseract v4 (Smith, 2018), combining Fraktur and Latin script models, both language-specific and independent; we found that the tool and said models work well for the print sources from 1850 onwards used in the project. The CRF was implemented with `sklearn-crfsuite` (Korobov, 2019).

Finally, to engage the non-specialist public with this digitally underrepresented tradition, we developed a **navigation interface** that allows exploring the texts and metadata.⁹

3 Outlook: Towards sociolinguistic analyses?

The previous section shows that the corpus was useful for *computational literary studies* approaches, previously unattempted for Alsatian given lack of a corpus, but important because they add empirical variety to the field. Now the question arises whether sociolinguistic description could also benefit from the resources created.

The first caveat is of course that character speech is fictional and can only be sociolinguistically relevant insofar as it reflects actual linguistic practice. Beyond that, in terms of geographic variation, it is unclear how to analyze the data: Do scriptural practices reflect the author’s variety? The publisher’s location? Fictional characters’ intended varieties? Other types of variation may be more promising: Thanks to the prosopography metadata, we have over 28,000 speech turns annotated for character gender, 14,000 with the character’s professional category and 11,000 with the social class.¹⁰ These data may allow us to assess some variation trends in the future. The approach in ŠeĽa *et al.* (2023), which has successfully identified distinctive features in character speech, sometimes tied to social factors, could be applied.

⁹The interface is at <https://methal.eu/ui/>.

¹⁰For these data and the way they were prepared, see the repository at <https://git.unistra.fr/methal/alsatian-character-speech>

Acknowledgements

This research was supported by Université de Strasbourg’s IdEx program (Attractivité 2020 call).

I thank further project members, Delphine Bernhard, Dominique Huck, Pascale Erhart and Carol Werner for all collaborations, besides Alice Millour, who co-supervised one of the project internships.

We also thank our interns: Nathanaël Beiner, Andrew Briand, Lena Camillone, Hoda Chouaib, Audrey Deck, Barbara Hoff, Valentine Jung, Salomé Klein, Audrey Li-Thiao-Té, Qinyue Liu, Kévin Michoud, Alexia Schneider, Vedisha Toory, Heng Yang.

We also acknowledge the High Performance Computing Center of the University of Strasbourg for scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d’Avenir) and the CPER Alsacalcul/Big Data.

References

- BERMÚDEZ SABEL H. (2022). FS-Validator. <https://github.com/HelenaSabel/FS-Validator>.
- BERNHARD D. & RUIZ FABO P. (2022). ELAL: An Emotion Lexicon for the Analysis of Alsatian Theatre Plays. In *Proceedings of LREC 2022*, p. 5001–5010, Marseille: ELRA.
- BRIAND A. & RUIZ FABO P. (2023). FETE: Fast Encoding of theater in TEI. <https://git.unistra.fr/methal/FETE>.
- GALLERON I. (2017). Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions. *Human and Social Studies*, **6**(1), 88–108.
- KOROBOW M. (2019). sklearn-crfsuite. <https://github.com/TeamHG-Memex/sklearn-crfsuite>.
- LIU Q., RUIZ FABO P. & BERNHARD D. (2023). Towards emotion analysis for Alsatian theater. In *Computational Humanities Research (Posters)*. DOI : [10.5281/zenodo.8404253](https://doi.org/10.5281/zenodo.8404253).
- MILLOUR A. (2020). *Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées*. PhD Thesis, Sorbonne Université. HAL : [tel-03083213](https://hal.archives-ouvertes.fr/tel-03083213).
- ROMARY L. (2015). Standards for language resources in ISO – Looking back at 13 fruitful years. *edition - Die Fachzeitschrift für Terminologie*, **11**(2), 13–19.

RUIZ FABO P., BERNHARD D., BRIAND A. & WERNER C. (2024). Computational drama analysis from almost zero electronic text. In M. ANDRESEN & N. REITER, Édts., *Computational Drama Analysis: Reflecting Methods and Implementations*. De Gruyter. <https://univoak.eu/islandora/object/islandora%3A157880>.

RUIZ FABO P. & WERNER C. (2021). Exploration du théâtre alsacien à travers ses listes de personnages pendant la période 1870-1940. In *Humanistica 2021*. DOI : [10.5281/zenodo.4762733](https://doi.org/10.5281/zenodo.4762733).

SMITH R. M. D. (2018). Tesseract (v4.0). <https://github.com/tesseract-ocr/tesseract>.

VISHNUBHOTLA K. & MOHAMMAD S. M. (2022). Tweet Emotion Dynamics: Emotion Word Usage in Tweets from US and Canada. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4162–4176, Marseille, France: European Language Resources Association.

YANG H. (2022). Détection de la variation graphique dans une langue non standardisée : le cas des dialectes alsaciens. Mémoire de master. <https://dumas.ccsd.cnrs.fr/dumas-03794680>.

ŠEĽA A., NAGY B., BYSZUK J., HERNÁNDEZ-LORENZO L., SZEMES B. & EDER M. (2023). From stage to page: language independent bootstrap measures of distinctiveness in fictional speech. DOI : [10.48550/arXiv.2301.05659](https://doi.org/10.48550/arXiv.2301.05659).