Computational drama analysis from almost zero electronic text: The case of Alsatian theater

Pablo Ruiz Fabo, Delphine Bernhard, Andrew Briand, Carole Werner

Pablo Ruiz Fabo, Delphine Bernhard, Andrew Briand, and Carole Werner

Computational Drama Analysis from Almost Zero Electronic Text: the Case of Alsatian Theater

Abstract: At the MeThAl project, we are creating the first large TEI corpus of Alsatian theater; Alsatian refers to Germanic varieties spoken in Alsace (Eastern France). The corpus, covering mainly the 1870-1940 period, will have above 500 000 tokens (51 plays) for which no previous electronic text existed. We present our automatic TEI encoding workflow assisted by a Conditional Random Fields model based on OCR sources, followed by manual correction. As the corpus shows large orthographic variation (there is no standard spelling) and NLP resources for Alsatian are scarce, several text analyses are challenging; we discuss our approach to tackle this. We developed detailed character metadata using TEI feature structures, encoding characters' social variables like their socio-professional group and social class. This provided an overview of the evolution of social groups in the plays across time, complementing earlier, smaller-sample studies of the tradition. Besides metadata-based analyses, with a view to performing emotion detection, we present an emotion lexicon which handles orthographic variation in the corpus, and how it can be used to train automatic variant detection. We also outline how we are publicly sharing the resources, on open data repositories, on DraCor, and via a corpus navigation interface. An overall goal is starting off computational literary research (CLS) on Alsatian theater and helping compare Alsatian theater with other traditions for which rich CLS results already exist, including the two dominant ones which influenced Alsatian theater most: German and French.

1 Introduction

Quantitative drama analysis, assisted in recent years by automatic computational means, has delivered new insights on many issues. Some of the analyses performed pose certain preconditions: first, the availability of electronic text in adequate quantity, ideally with the relevant structural markup (e.g., TEI). Second,

Pablo Ruiz Fabo, Delphine Bernhard, Carole Werner, Université de Strasbourg
Andrew Briand, University of Washington

https://doi.org/10.1515/9783111071824-004

(NLP-based) tools to perform automatic annotations which can help operationalize a drama analysis research question.

Alsatian theater is a tradition for which these conditions were not fulfilled until very recently, even if dramatic works in Alsatian (i.e., Germanic varieties spoken in Alsace, Eastern France) have been steadily produced for over two centuries. A large-scale quantitative study of this tradition was so far impossible, given the absence of an electronic corpus. Within the MeThAl project,[1] we are taking the first steps towards enabling such analyses, focusing on the 1870–1940 period. We present the challenges encountered for corpus development and analysis, and the solutions adopted, paying attention to how this work may generalize to other low-resource settings. One of the challenges we discuss is the huge orthographic variation in the corpus, given the lack of a standard variety. This hinders statistical analyses typical of Computational Literary Studies (CLS); we will present the first resources developed to tackle this problem. An overall goal is to help start a dialogue between CLS research on Alsatian theater and existing CLS research on the dominant traditions that have influenced Alsatian theater (German and French), for which major results have already been achieved. Given the limited range of electronic text and NLP resources for Alsatian varieties, there is a gap between the CLS questions that can be addressed in this tradition and in larger ones, and our project seeks to start bridging this gap.

The paper is structured as follows: section 2 introduces the specificities of Alsatian theater and our corpus selection criteria. Section 3 outlines our encoding workflow. Since part of the corpus is still being encoded, and given orthographic variation in the text, we focused on detailed metadata as a way to reach generalizations about the material. Within this effort, section 4 describes our encoding of character social variables into a TEI personography, which gives an overview of the evolution of social groups present in the corpus. Deeper, text-based analyses require dealing with the corpus's large scriptural variation. In this respect, section 5 presents a lexicon for emotion analysis which is able to handle this task, besides methods for spelling variant detection that can be developed based on the lexicon. In Section 6, we outline how we are sharing the project resources, targeting both scholars and the general public. Section 7 concludes.

1 https://methal.pages.unistra.fr/en.

2 Alsatian Theater Particularities and Corpus Selection

A continuous dramatic production in Alsatian has been existing since the early nineteenth century (Gall 1974, pp. 5–6). Comedic subgenres predominate (from refined satires to farces), without excluding other genres like the *Volksstück* (popular drama) and *Weihnachtsmärchen* (Christmas tale). Alsatian theater is also interesting due to its ties to the two dominant traditions surrounding it, German and French, such that a polysystem analysis (Even-Zohar 1990) might be attempted.

Alsatian theater's foundational play is *Der Pfingstmontag* by Johan Georg Daniel Arnold, published in 1816. The tradition, however, experienced its golden age around 1900; Alsace was part of the German Empire between 1871 and 1918, and writing dialect theater was one means whereby Alsatian authors sought to affirm their identity as apart from the rest of the German-speaking world (Huck et al. 2007, p. 12). Our corpus, stretching mainly from 1870 to 1940, covers this period.

A salient characteristic of the corpus is its vast scripto-linguistic variation. No standard orthography existed during the corpus period, and practices vary across authors and also for characterization purposes within the same author; representing sociophonetic variation also leads to different scripturalizations for the same lexeme. Besides this variation, code-switching between Alsatian, French, and standard German is also characteristic in the corpus, reflecting a situation of language contact (and diglossia, see Huck 2015, pp. 151–161). Besides, the paratext (e.g., stage directions) can be in either Alsatian or German.

2.1 Corpus Selection

We are creating the first electronic-text corpus for Alsatian theater. The overall goal is to provide digital text amenable to large-sample quantitative analyses that complement existing knowledge of the tradition (see section 4.3 for some findings). In this scenario, we are aiming for breadth, including minor authors and a variety of author origins and publisher locations, rather than attempting an exhaustive coverage of the best-known authors. Besides, earlier literature has covered major authors based on samples specific to them, using non-quantitative approaches (Cerf 1972, 1975; Huck 1998; Hülsen 2003; Huck 2005). Regarding genres, we focus on comedies since this is the tradition's main genre, but are also encoding important *Volksspiele* or dramas (including socially engaged dra-

mas from the 1890s by Julius Greber).[2] As a starting point for our encoding, we prefer digital sources if available, such as the collection of image-mode digitizations at the Numistral heritage portal by Strasbourg's National and University Library.[3] Other platforms offering public-domain content like the Internet Archive or Google Books also complement our sources, as does a Wikisource collection with the complete works of Alsatian dramatist August Lustig (transcribed by Mireille Libmann in 2013).[4] Besides, by consulting secondary literature on the history of Alsatian theater (Cerf 1972; Gall 1974; Hülsen 2003; Huck 2005), we identified a small number of plays unanimously considered important by those sources but not yet digitized, and obtained digitizations.

When selecting plays for encoding, we prefer those mentioned by the secondary literature just cited, and, more generally, we strive for representing all decades in the corpus period, different geographical origins for authors and publishers and different subgenres, with a majority of comedic works of various types. Even if we seek variety, a limitation of the corpus is that male authors predominate; increasing the share of women authors is a challenge, since they are underrepresented in digital sources and are minoritary in preserved sources. We intend to counter this bias with deeper searches in non-digitized collections (e.g., at city archives).

We have so far performed OCR on, and TEI-encoded, 51 plays (30 plays with 329 969 text tokens already published and 21 more with 188 275 text tokens awaiting validation). We also converted to TEI the 26 plays by August Lustig on Wikisource (136 675 text tokens); this involved no OCR and only wiki-markup conversion to TEI with rule-based methods, thus being a much simpler process than the workflow for plays for which no electronic text existed.

3 TEI Encoding Workflow

We adopted the Text Encoding Initiative's (TEI) recommendations (TEI Consortium 2022). The TEI very naturally matches our project goals: it allows for encoding structural divisions, which are crucial for classical quantitative drama analyses like configuration matrices (Pfister [1977] 2001, p. 236), and it has rich possi-

2 The distribution can be seen in a metadata table in our repository: https://git.unistra.fr/methal/methal-sources#plays-available-in-tei.

3 https://www.numistral.fr/fr/theatre-alsacien. Some of the plays have OCR, but it has not been corrected, and we produce our own OCR and its correction (see section 3).

4 https://als.wikipedia.org/wiki/Text:August_Lustig/A._Lustig_Sämtliche_Werke:_Band_2.

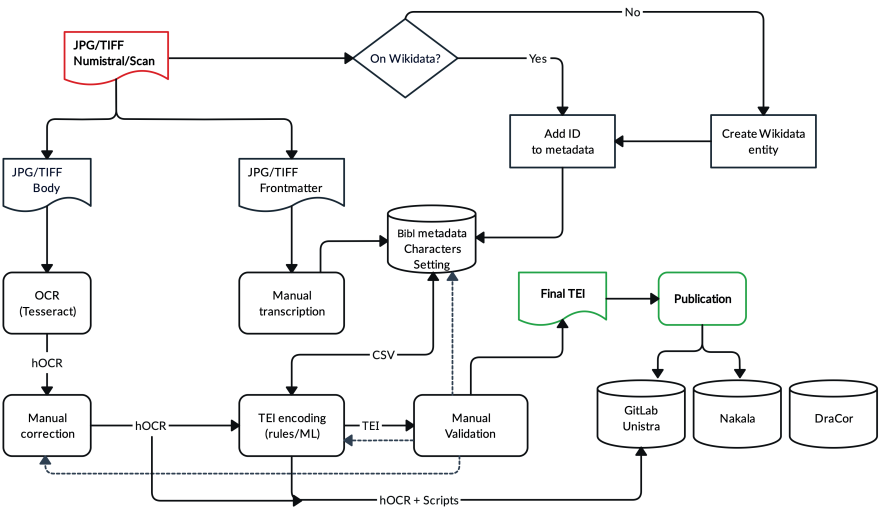


Fig. 1: TEI encoding workflow.

bilities for encoding author and character metadata. Besides, there is an ecosystem of corpora and tools for working with dramatic texts in TEI, e.g., the DraCor platform (Fischer and Börner 2019) or the DramaAnalysis R package (Reiter and Pagel 2020). It is thus an advantageous format to help start a dialogue between research on other traditions and on Alsatian theater.

Our workflow combines manual transcription of bibliographic and character metadata (3.1), automatic TEI encoding of the plays' body based on OCR output (3.2), and manual revision for both the OCR and the automatic TEI-encoding. We believe that our encoding approach could be used in similar projects where little or zero electronic text exists. The workflow is summarized in figure 1.

3.1 Transcription of Bibliographic and Character Metadata

Based on initial tests, we decided to treat differently the plays' front matter from their body. The front matter, i.e., the cover page with bibliographic metadata, and other material before a play's beginning, such as the cast list, takes many different formats depending on the publisher or series. Given this challenge, we manually transcribed bibliographic and character information into a delimited format and later generated the relevant TEI elements (e.g., `teiHeader`, `castList`, `listPerson`) based on this; we also annotated characters with social attributes describing them (section 4). As the volume of manual transcription amounts to one or two pages

per play, the approach scaled well to over three hundred plays and over 2000 characters.[5] The manual work also includes creating Wikidata entities for plays and authors to refer to them in the TEI versions.

3.2 Automatic OCR to TEI Conversion With Conditional Random Fields

The bodies of the plays (from the first act or scene to the final curtain) are quite regular and lend themselves to automated encoding, followed by manual corrections. As we start off from image-mode digitizations, the first step is optical character recognition (OCR), for which we use Tesseract (v4, Smith 2018). We found that for Alsatian text from the corpus period, best results are obtained by combining German and French models plus the "script" model matching each play's typeface, i.e., a language-independent model trained on either blackletter (Fraktur model) or roman typeface ("Latin" model). For plays typeset in blackletter, German Fraktur models were used in addition to the language-independent ones.[6] A possible reason why combining these models works best is that Alsatian scripturalization in the corpus includes characters and diacritics typical of German (eszett and umlaut), but also typical of French (acute and grave accent). Besides, there is code switching involving both languages, and the paratext is often in German. We chose hOCR (Baierer 2020) as our output format; this format encodes word-position information on the page (bounding boxes), which can be used to compute features for TEI-element prediction (details below). OCR output was manually corrected, generally by a single person.[7]

A first challenge is converting hOCR output into TEI. Several projects have performed automatic TEI generation from OCR sources, often implemented as a sequence labelling task: Khemakhem et al. (2017) present GROBID dictionaries, which encodes dictionary entries in TEI based on layout and typographic cues present in OCR outputs, using a cascade of Conditional Random Fields (CRF)

5 We have bibliographic references for over 350 plays, but digitizations (and thence character lists) for ca. 230.

6 In other words, we used a combination of models from https://github.com/tesseract-ocr/tessdata, language codes `deu` and `fra` in general, plus `deu_frak` and `frk` for blackletter plays, and, among models in the `script` directory, `Fraktur` or `Latin` depending on the play's typeface.

7 Double keying has been shown to handle most OCR and transcription errors in electronic corpus creation (Geyken et al. 2012, §3), reaching >99.99 % accuracy (Deutsche Forschungsgemeinschaft 2016, p. 32), however for this first Alsatian theater corpus, we preferred to obtain first versions of a larger number of plays rather than a smaller number of double-keyed versions.

models (Lafferty et al. 2001), each of which handles a different TEI element. Khemakhem et al. (2018) also use GROBID to encode encyclopedia entries and auction catalogues in French, as do Gabay et al. (2021) for exhibition catalogues. Still using CRFs (a single model rather than a cascade), Erjavec et al. (2021) TEI-encoded academic theses' structure (e.g., abstract vs. acknowledgements) using simple surface features (most frequent words per page, page and word length).

Neural methods have also been used for TEI generation. Pagel et al. (2021) predict the TEI structure for plays' body elements in German, assuming one label per sentence (5-way classification task predicting act and scene divisions, speaker names and their speech, and stage directions). The input is not OCR, but rather plain text. The annotated corpus contains ca. 1.4 million sentences, the smallest class is *Act* with 1458 sentences, *Scene* has over 11 000 and the rest over 175 000.[8] They obtain best results (>0.97 F1 for all classes but stage directions, with 0.84 F1) by fine-tuning the German `bert-base-uncased` model. The results are higher than their baseline's, a CRF with surface features (token string, case, presence of digits) and lexical ones (presence of triggers indicating act/scene divisions), which got >0.92 F1 for all classes but stage directions, with only 0.44 F1.

In our project, we considered that the structure of a play's body is less complex than the text-types for which GROBID has been used (dictionaries, encyclopedias or catalogues). Accordingly, we implemented a simpler architecture: a single CRF model was created to predict, based on OCR input in hOCR or ALTO formats,[9] six types of content: act and scene divisions, speaker names and their speech, stage directions, and verse lines. The *verse lines* class refers to verse within character speech (e.g., a song or poem), typographically distinct from "prose-based" character speech. The models were created with `sklearn-crfsuite` (Korobov [Nov. 26, 2015] 2019). The system is publicly available at https://git.unistra.fr/methal/FETE.

At project outset, we had no TEI-encoded plays at all, so we produced a first set of seven plays using dictionary- and rule-based methods; a manual correction of the TEI-encoded documents was also carried out. These seven plays and their OCR were used as the initial training set for a CRF model. The initial model was used to tag some more plays, which were then manually corrected, up to the 16 plays that make up the training data described in table 1, comprising ca. 150 000 tokens.

8 For the *Act* or *Scene* classes, a "sentence" would be the expression that indicates an act or scene onset, e.g., *Erster Akt* for *First Act*.

9 ALTO: Analyzed Layout and Text Objects, (Alto Editorial Board 2022); hOCR: Baierer (2020).

Tab. 1: CRF training-data distribution (number of examples per class).

Content type	TEI element predicted	Train	Test
Act	`<div type="act">`	25	9
Scene	`<div type="scene">`	408	68
Speaker	`<speaker>`	9,675	978
Speech	`<p>`	104,504	12,239
Verse line	`<l>`	1,390	236
Stage direction	`<stage>`	21,371	3,192

The model predicts a label for each corpus token. Three sets of features were implemented in the CRF and computed for the current, previous and following token:

- **Set 1 (Token-level features):** These include the token string (also lowercased), initial case, the presence of digits, the presence of punctuation often used to structure plays (period, colon, parentheses). Based on hOCR or ALTO output, font size and token horizontal position (normalized per play), and an indication of font-size difference between previous and current token.
- **Set 2 (Set 1 + Heading features):** Includes the features in Set 1, plus new features to detect act/scene headings. In view of the small number of examples to learn from, we favored lexical features, i.e., words that mean *act* or *scene* in either Alsatian, German or French (paratext is often not in Alsatian).
- **Set 3 (Sets 1 + 2 + Verse line features):** Includes the features in sets 1 and 2 plus features to detect verse lines. As layout or token-based cues are inconclusive, the following two features were implemented. (1) Based on a comparison of lines' final characters, a boolean to indicate whether the current line rhymes with the previous or next. (2) An estimation of the difference in syllable count between consecutive lines, obtained by counting sequences of consecutive vowels, as an approximation to the number of syllable nuclei.

After prediction, the TEI hierarchy is recreated from the predicted labels. The TEI header and cast list are generated automatically based on our manual transcription of the relevant text (3.1) and added to the document. Errors made by the automatic encoding are corrected manually.

CRF prediction results are in table 2. Even if for some categories (table 1) the number of test-items is small, overall the results suggest that a simple CRF model producing token level annotations is a viable means to speed up TEI encoding of dramatic texts when little training data are available, also helping with segmenting verse-lines among prose.

We see the following limitation: the model uses punctuation as a feature and overfits to the main punctuation cues used in the training data to indicate speak-

Tab. 2: Precision (P), Recall (R) and F1 for the three feature-sets. Best F1 per class bold, second-best italicized.

	Set 1			Set 2			Set 3		
Content	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Act	0.889	0.889	0.889	1	1	**1**	1	1	**1**
Scene	1	0.941	0.97	1	1	**1**	1	0.956	*0.977*
Speaker	0.988	0.995	**0.991**	0.988	0.994	**0.991**	0.988	0.994	**0.991**
Speech	0.98	0.993	0.986	0.985	0.991	*0.988*	0.989	0.991	**0.99**
Verse-line	1	0.475	0.644	0.921	0.695	*0.792*	0.898	0.898	**0.898**
Stage direction	0.971	0.875	0.908	0.972	0.964	**0.968**	0.971	0.962	*0.967*
Weighted mean	0.979	0.979	0.978	0.982	0.982	*0.982*	0.984	0.984	**0.984**

er/speech divisions (a colon in an overwhelming majority of examples) or stage directions (almost invariably in parentheses). While encoding new plays, we saw that performance is poor when those cues are absent (e.g., with older plays who use a period after the speaker instead of a colon, or with a play delimiting stage directions via square brackets). A simple workaround is to preprocess plays so that those delimiters are the majority ones prior to prediction and postprocess them back into the original delimiters after it. However, this does not solve the more difficult (although rare) case where no delimiter is used to indicate stage directions; for this, Pagel et al.'s contextual embeddings approach, which classifies entire sentences based on content, should help, even if stage directions were the hardest category for their model. In our model, the hardest class is verse-lines, which suffer from a similar problem: their lexical content is difficult to distinguish from non-verse lines.

A model which uses plain-text as its input like the one developed by Pagel et al. (2021) has the advantage that it requires no prior OCR output with layout and word-position information. Such a model can thus be used to TEI-encode already available electronic transcriptions of plays (e.g., manually performed transcriptions or manually corrected plain-text OCR for which no hOCR or ALTO output was produced). We would like to attempt this approach in the future; our corpus has ca. 10 times fewer tokens than theirs and it would be interesting to examine the impact of training data volume.

4 Character Social Annotations: A Corpus Overview From Its *Dramatis Personæ*

The TEI-encoding of the complete corpus, which will help address CLS questions related to the plays' structural properties and character configurations per genre, is still ongoing. Besides, the corpus' huge scripto-linguistic variation poses challenges for many text-based analyses. In this situation, a feasible effort is to obtain an overview of the social groups portrayed in the corpus, thanks to creating detailed character metadata. Since earlier literature on Alsatian theater has also focused on social groups represented in the plays, based on smaller samples, we can compare our findings to what is known from earlier studies.

Available studies on Alsatian theater have examined how social status relates to plot, interacting with setting (rural or small town vs. urban) and character gender. Cerf (1972, pp. 339–348) speaks of rural plots in which parents (rich farmers) wish for their daughter to marry into the same social class, whereas the daughter prefers a less wealthy farmer whom she likes or an urban suitor; this is one of the plays' sources of conflict. In urban plots, parents wish for their daughters to marry someone with a better socio-economic status; this is punished in late nineteenth century plays (the young women are abandoned and become outcasts with a child to raise) but the daughters' ascent by marriage to a social class immediately above their own is possible in early twentieth-century plays. Sons' social ascent takes place via education. Hülsen (2003, p. 104) also documents marriage plots showing a clash between the parents' will and the younger generation's choices, besides tension between French and German cultures (used for comic purposes), as sources of dramatic conflict.

Cerf (loc. cit.) provides an overview of socio-professional groups in the plays. In small towns or villages, most characters are farmers, rich or poor, and domestic employees at the farms. In city settings, mostly the middle classes (small and middle bourgeoisie) are represented: craftspeople, office workers, liberal professions, rentiers. Domestic workers are also frequent characters. Less frequently, we find very poor characters (e.g., unemployed). Hülsen (2003, p. 140) reaches similar conclusions, noting also that both the lowest classes and French speaking nobility are absent from the plays.

4.1 Character Social Variable Annotations

The generalizations reported in the literature were arrived at by an examination of relevant text passages in samples focusing on the *Théâtre alsacien de Strasbourg*'s

Tab. 3: Character social variables annotated (besides *age*). Gender *both* is for group characters (not individuals).

Category	Values
Socio-professional groups	professionals, scientific, technical; intermediate professions; service and sales; crafts; industry and transportation; agriculture; elementary professions; *rentiers*; clergy; military; government officials; *volunteer positions*
Social class	upper class; upper middle class; lower middle class; lower class
Gender	both; female; male; unknown

repertoire. Cerf (1972) analyzed 39 plays between 1898 and 1939, and Hülsen (2003) covered 13 plays between 1898 and 1914. As a first step to complement such knowledge, we annotated socioprofessional groups, social class, age and gender for 2386 characters in 231 plays, creating a TEI personography. As a rule, only information that is explicit or can be deduced from the *dramatis personæ* of the plays was annotated, without recourse to the plays' text.

Regarding socio-professional groups, although classification schemes for historical professions exist, such as HISCO (Van Leeuwen et al. 2004) or an ontology by Moeller and Nasarek (2018) for German-speaking settings, these are intended for historical research rather than for describing literary characters, and their applicability to the corpus is not necessarily immediate. Fièvre (2017) presented a taxonomy for French classical drama, but its categories are obsolete for our period.

Given that existing taxonomies were not a good fit, we developed the one shown in table 3, which is inspired by the above but adapted to our corpus situation. In the table, groups in italics represent social rather than professional groups important for the corpus period. *Rentiers* are rich enough to not need to work for a living; in the corpus, this often happens after many years as a successful tradesperson. The *volunteer positions* group refers to occupations that are not professional, but that are important to define a character's social status, such as being an association leader (e.g., of a patriotic association or of a sports club) or in some cases being a member of such associations or societies.

Character professional groups were annotated by a single person (one of the authors or an intern) and were then revised by another author. Disagreement was rare (this was not measured more precisely). Social class annotations were carried out by one of the authors or by an intern and revised by another author. Based on

the sample discussed in 4.3, there was a disagreement in 8.2% of cases.[10] As future work, it could be considered whether adding an intermediate category *middle class* is a better option than a binary division of middle class characters into *upper* and *lower*.

There are some limitations to our social variable annotations: First, many female characters in the corpus are not described with a profession, but rather by reference to a male character (e.g., *wife of*); more detailed data about this are shown in in 4.3. Accordingly, quantitative generalizations about female characters using our socioprofessional taxonomy are limited, although when described as associated to another character, they could "inherit" class from it. A second limitation is that our annotations are static and refer to character state in the *dramatis personae*. Character end-state and social class changes (e.g., by education or marriage) are not directly captured. Third, given Alsace's status as alternating between French and German political rule (including during the corpus period), characters' origin is important as an indication of power relations, as is the language variety they use (Alsatian, other dialects, standard German, or French). At this point, we do not have these annotations.

4.2 TEI Personography

In order to share the social annotations in an interoperable format, we provide them as a TEI personography, using the *feature structures* (FS) formalism (Romary 2015).[11] FS represent feature hierarchies, using key-value pairs. Values can be either atomic, or a new feature structure; recursivity allows us to choose our description granularity. The data model is described in a feature system declaration `<fsdDecl>`, composed of feature declarations `<fDecl>` and feature structure declarations `<fsDecl>`.

Galleron (2017) created a feature library composed of *characterisemes* or characterization units, for describing the characters of French theater (1630 to 1810), including mimetic traits, which provide characters with human-like features, and synthetic traits related to characters' role in the play (e.g., an antagonist) (Phelan 1989).

10 The annotations for 2386 characters are now released in TEI format (see section 4.2). However, only a smaller sample with 1103 characters was available when we obtained the results discussed in 4.3 below. We have shared these initial annotations and analyses in spreadsheet format at https://page.hn/8ynxsp.

11 https://tei-c.org/release/doc/tei-p5-doc/fr/html/FS.html.

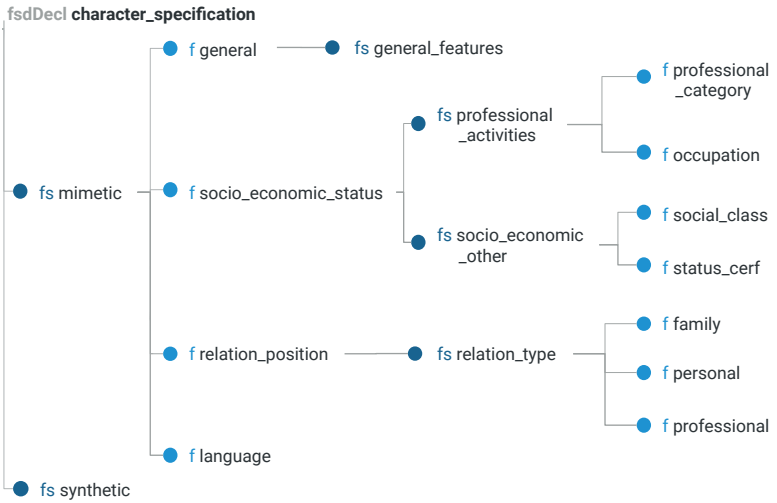


Fig. 2: Informal representation of the mimetic features in our feature library for character description. Node `general_features` refers to age, gender and origin. Dark nodes are feature structures, light nodes are features. The project repository gives the formal declaration. [13]

We believe that this characteriseme approach facilitates comparative studies across dramatic corpora. One of our goals is helping to compare Alsatian theater to the two hegemonic traditions surrounding it (German and French), so we adopted the approach.

In order to make the feature library applicable to new corpora, including Alsatian theater, we added new feature values (e.g., family relations such as *ex-husband* or *divorced wife*). When annotating, we abstracted away from variation in job titles by normalizing them into a list of ca. 350 professions; we added these as feature values as well. We also added intermediate levels to the FS for a more modular analysis of mimetic feature groups: (1) basic characteristics like gender or age, (2) features related to socio-economic status (including socioprofessional group and profession) and (3) characters' position in a relation (whether they are presented as e.g., a spouse, a relative or subordinate of another character).[12] Personography compliance was ensured with a schema automatically derived from the feature system declaration using the approach in (Bermúdez Sabel 2019).

Figure 2 shows an informal diagram of our FS declaration and figure 3 gives an example character annotated with its features. The complete personography

12 The relation itself (who is related to whom) is not part of the personography, although available in the plays' TEI via `<listRelation>`.

```
<person xml:id="mtl-per-0890">
 <bibl corresp="#mtl-090"/>
 <persName>Alice Sandel</persName>
 <note type="roleDesc">Dactylo</note>
 <occupation>Dactylo</occupation>
 <fs type="character_specification">
  <f name="specification_type">
   <fs type="mimetic_features">
    <f name="general">
     <fs type="general_features">
      <f name="sex">
       <symbol value="F"/>
      </f>
     </fs>
    </f>
    <f name="socio_economic_status">
     <vColl>
      <fs type="professional_activities">
       <f name="occupation">
        <symbol value="typist"/>
       </f>
       <f name="professional_category">
        <symbol value="intermediate_professionals"/>
       </f>
      </fs>
      <fs type="socio_economic_other">
       <f name="social_class">
        <symbol value="lower_class"/>
       </f>
      </fs>
     </vColl>
    </f>
    <f name="language">
     <default/>
    </f>
   </fs>
  </f>
 </fs>
</person>
```

Fig. 3: Personography entry for the character *Alice Sandel*.

is at the project repository.[13] Findings enabled by these annotations are presented in the next section.

4.3 Findings

Our character social variable annotations allowed us to both reproduce trends already described in the available literature and arrive at new generalizations. We

13 https://git.unistra.fr/methal/methal-sources/-/tree/master/personography.

Tab. 4: Character social class distribution.

Social class	Count	%
Upper	61	10.85
Upper middle	74	13.17
Lower middle	209	37.19
Lower	218	38.79
Total	562	100

present our results on character gender distribution, character social class distribution, and character professional group distribution across time. Note that the complete personography (231 plays, 2386 characters) presented in 4.2 was not yet available when we carried out the analyses in the present section and the sample we report about here includes 108 plays with 1103 characters.

Regarding character gender, 24.59 % of characters were female, 71.32 % were male, while for the rest, the gender cannot be determined from the *dramatis personæ* or we are dealing with mixed-gender group characters. One salient difference between female and male characters in the corpus is that most female characters are not described with a profession, while the opposite holds for male characters, as shown in figure 4. As we discuss at the end of the section, there is, however, a characterization difference in this respect depending on whether the author is female or male.

Regarding social class distribution, it was possible to annotate 562 characters with the classification introduced in table 4; we considered that no sufficient cues were present in the *dramatis personæ* to assign a class to the remaining characters. We found a majority of middle class characters, followed by lower class ones. "Lower class" in this corpus mostly refers to characters in roles such as domestic worker; marginalized characters are absent or very rare, as was pointed out in earlier literature (Cerf 1972, p. 344; Hülsen 2003, p. 140). Some high class characters with socially prestigious occupations appear. However, our corpus also confirms observations in earlier literature that the French-speaking high-bourgeoisie and nobility are rare or absent (Hülsen 2003, p. 140).

We examined the distribution of socioprofessional groups according to our taxonomy (see 4.1) and their temporal evolution (Ruiz Fabo and Werner 2021). We focused on the 1890 to 1939 period. Earlier decades show only 10 characters or less; given the small data volume, we do not report results about them. Several of our results are compatible with trends described in (Cerf 1972; Hülsen 2003). Figure 5 synthesizes our results.

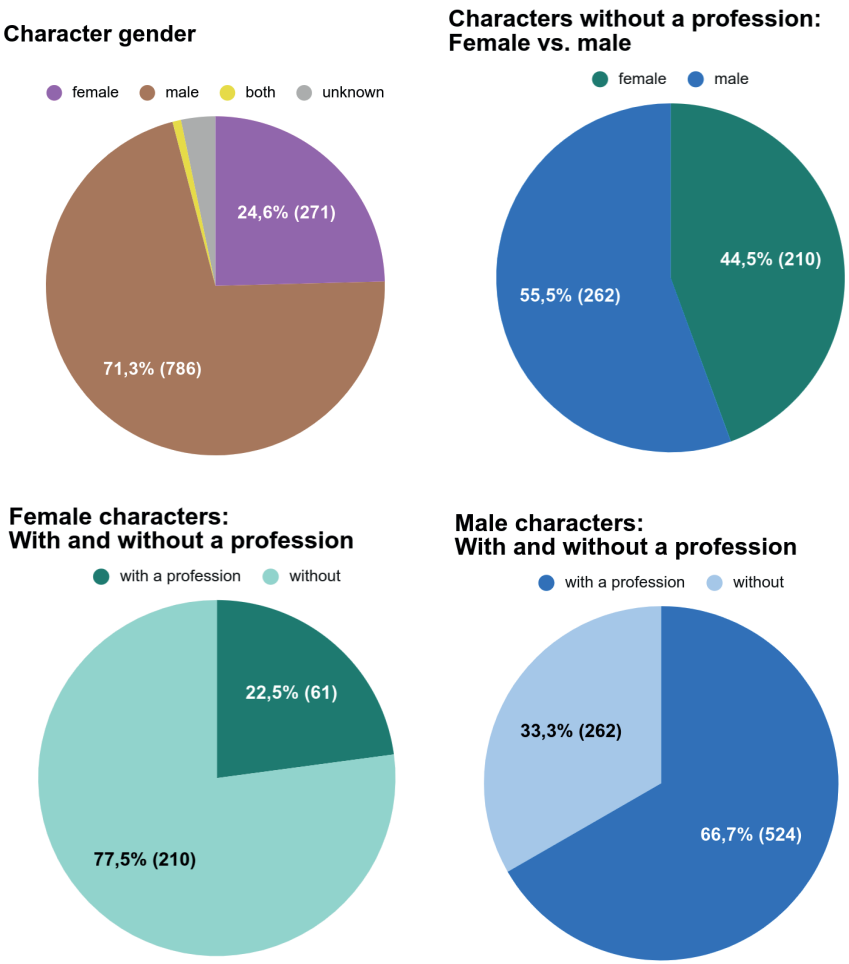


Fig. 4: Character gender and characters described with or without a profession.

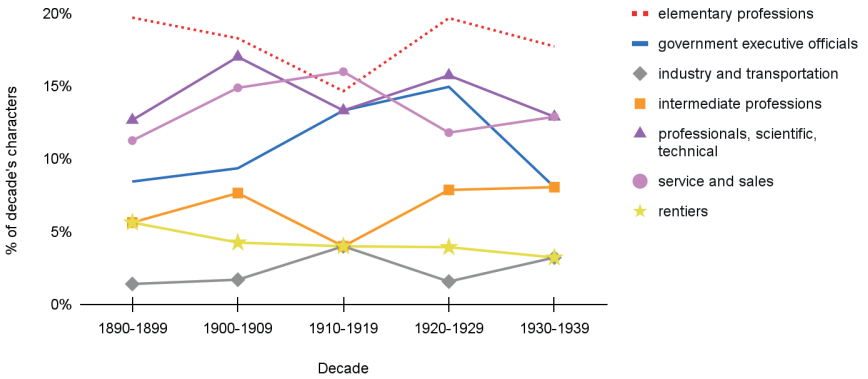
Subfigure 5a shows several groups whose representation is steady throughout the corpus period; the difference between the period minimum and maximum is no more than 5 percentage points, aggregating per decade. Most of these groups belong to the small and middle urban bourgeoisie. There are also two groups (*elementary professions* and *government*) whose interactions with the bourgeoisie groups can be a source of dramatic conflict in the plays. Characters in the elementary professions group are often domestic employees whose disagreements with their employers contribute to the plays' plots. Besides that, (comical) conflict between political authority figures and members of bourgeoisie groups also appears in the plays.

A further important bourgeoisie group, craftspeople, is shown in subfigure 5b. Unlike the groups displayed in 5a, this group's representation is not just steady across time, but rising; our sample shows a clear increase of professionals from the crafts group between the beginning and the end of the corpus period. This contrasts with the clear decline in character occupations related to agriculture between the beginning and the end of the period, a second trend also shown in 5b. As Cerf (1972, p. 344) notes, speaking of the Théâtre Alsacien de Strasbourg's (TAS) repertoire, it describes not only rural Alsace, but also the urban social groups that form its public. Our sample's increase in craftspeople during the corpus period points in the same direction.

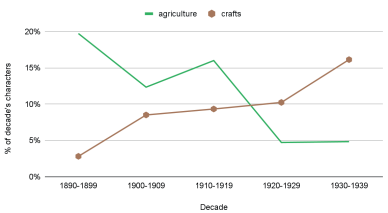
Cerf (1972, pp. 352–354) reports that Alsatian theater is careful not to promote controversy about religion and politics. This may be reflected in the fact that the clergy and the military only have a discontinuous presence in the plays; these and other character groups that have a minor presence in the plays, not appearing in all decades, are shown in subfigure 5c.

We also looked at the distribution of female characters described with a profession in the *dramatis personæ* (subfigure 5d); this character subgroup had not been addressed in earlier literature. It should be noted that data are scarce. As mentioned, only 24.6 % of characters are female and only 22.5 % of these are described via a profession in the plays' character lists. In other words, only about one character out of 20 in the corpus is a female character with a profession, which yields a total number of only 56 characters represented in subfigure 5d.[14] Even if the data are not abundant, a generalization that emerges is that most female characters with a profession are part of the elementary professions group (i.e., domestic workers in most cases). Several examples of female characters with professions

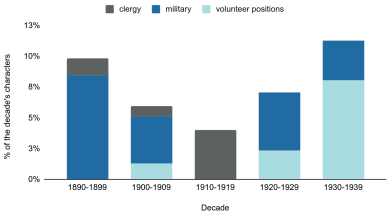
14 The count for female characters with a profession here (56) is slightly smaller than the one in figure 4 (61), because figure 4 includes all corpus decades, and we only consider plays from 1890 onwards when analyzing professional group distribution, given the small number of plays overall before 1890.



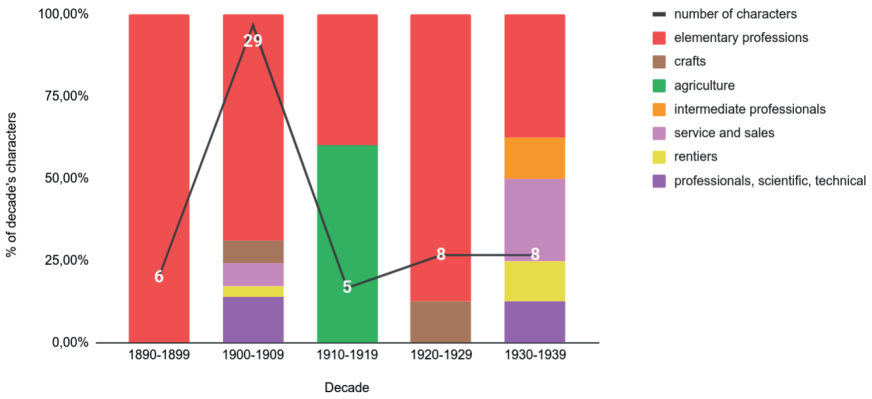
(a) Bourgeoisie groups (other than crafts), plus elementary professions and government.



(b) Crafts vs. agriculture professions.



(c) Minor groups.



(d) Female characters with a profession in the *dramatis personæ*: professional group distribution.

Fig. 5: Evolution of professional groups in the corpus, based on a sample of 108 plays and 1103 characters.

unrelated to domestic work exist in the 1900 and 1920 decades, e.g., a philosophy student, a textile merchant, a typist, an actress and a midwife. However, given the small amount of data, it is not possible to establish trends.

When crossing character-level metadata with play-level metadata we arrive at new generalizations. We annotated each play's scene for its urban or rural/small-town setting. Cerf (1972, p. 340), based on 39 TAS plays, states that craftspeople in small-town settings are absent from the repertoire. When looking at our larger sample this no longer holds, which attests to the usefulness of obtaining larger-scale annotations. We also annotated author gender and found a difference in the characterization of female characters in female vs. male authors. Among characters presented via a profession in the *dramatis personæ*, 46 % are female in plays written by female authors, vs. only 11 % in plays written by men. Such observations were not possible so far for this tradition, as the annotations that underlie them were not available.

5 Spelling Variants Matching

In Alsatian, all analysis methods that require a homogeneous representation of the corpus's vocabulary (such as corpus search tools or topic modelling) need to manage scriptural variation first, so as to detect and merge variants. For instance, we would like to perform automatic emotion analysis in the MeThAl corpus in the future. As a first step, we have built an emotion lexicon for the Alsatian dialects, named ELAL (Bernhard and Ruiz Fabo 2022).[15] In order to cater to the large amount of graphical variants observed in the MeThAl corpus, we have included graphical variants of Alsatian lexical items in this lexicon. Alsatian spelling variants (e.g., *Arbeewa* and *Ardbewa*) and closely related Alsatian and German word forms (e.g., Alsatian *Erdbewe* and German *Erdbeben*) have been identified automatically and corrected manually.[16]

As a by-product of the manual correction of the ELAL emotion lexicon, we obtained pairs of true, correct variants and false, incorrect variants. We have used this dataset to test and compare two different approaches to spelling variants matching. In the rest of this section, we first present these two approaches, then detail our training and test datasets and finally discuss the results obtained.

15 https://nakala.fr/10.34847/nkl.40cex998.

16 We refer the reader to (Bernhard and Ruiz Fabo 2022) for details about the process.

5.1 Methods for Spelling Variants Matching

The goal of spelling variants matching is to decide whether a given pair of forms are spelling variants or not. This can be considered as a binary classification task with two outcomes: true or false. The features traditionally used rely on string similarity measures or character n-grams, which are applicable out of context. In addition, features extracted from large corpora, e.g., word embeddings, make it possible to take the context of occurrence into account for validating variants. We have compared two methods using only word-internal features, since we operate on the vocabulary and not on the corpus level. The first method, proposed by Barteld et al. (2019), aims at filtering spelling variant candidates and was proposed for dealing with variation in texts from Middle Low German. The filter uses n-grams extracted from the pairs of word-forms given as input and word embeddings.[17] For example, for the pair (*beschtadiga*, *bschtadiga*), the following n-grams are extracted, based on the alignment of both words and the detection of the mismatches in the alignment; here, there is a single mismatch (insertion of 'e' between 'b' and 's'):

- 2-grams: [bb, -e], [-e, ss],
- 3-grams: [$$, bb, -e], [bb, -e, ss], [-e, ss, cc]
- 4-grams: [$$, bb, -e, ss], [bb, -e, ss, cc], [-e, ss, cc, hh]

These features are then used to train a Support Vector Machine (SVM).

The second method, DeezyMatch (Coll Ardanuy et al. 2020; Hosseini et al. 2020), uses a deep learning approach relying on a Siamese classifier consisting of two parallel recurrent layers and supporting the following architectures: Elman Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). DeezyMatch was originally developed for toponym matching in the context of entity linking. The system converts input pairs to dense vectors after preprocessing steps which include tokenising the forms into smaller units (in our case, into characters and n-grams of characters).

5.2 Dataset

We have used the dataset which was manually annotated for building the ELAL emotion lexicon to generate training data. The development set contains 52 239 pairs, out of which 42 209 are positive instances (true variant pairs) and 10 030 are negative instances (non variant pairs). The held-out test set contains 22 389

17 We only used n-gram features, since we lack large corpora for building dense representations of word forms.

Tab. 5: Excerpts from the development dataset.

Form 1	Form 2	Matching
Gebortsdoe	Geburtsdäuj	TRUE
Pischtol	Pistol	TRUE
Gueter	Güeta	TRUE
unruehich	unruhig	TRUE

Form 1	Form 2	Matching
Schlöüi	schlag	FALSE
blute	blutt	FALSE
Grond	Grénd	FALSE
féjle	féle	FALSE

pairs, out of which 18 090 are positive instances and 4299 are negative instances. The dataset is imbalanced and contains about 4 times as many positive as negative instances. Figure 5 shows some examples from the development dataset,[18] which illustrate the difficulty of the task: very similar word forms can be negative instances (e.g., "féjle" – " féle") while more distant word forms can be positive instances (e.g., "Gebortsdoe" and "Geburtsdäuj").

5.3 Experimental Settings

In addition to the SVM evaluated in (Barteld et al. 2019), we tested two other classifiers: Ridge Classifier[19] and Logistic Regression.[20] We used the implementation provided by Barteld et al. (2019)[21]. for extracting the n-gram based features `ng` and the Python Scikit-learn library for the supervised classification experiments (Pedregosa et al. 2011). The best hyperparameters were chosen using 5-fold cross-validation on the development dataset.[22] The final models were trained on the entire development set. We also tested `other` features in addition to the n-grams: the frequency in the MeThAl corpus of each of the words in the pair, the number of French translations in the bilingual French-Alsatian lexicons for each of the words in the pair, the number of French translations shared by the two words in the pair.

We used the default parameters of DeezyMatch, except for the maximum n-gram length, which we changed from 3 to 4, and we did not lowercase the forms, so as

18 We considered inflectional variants, such as *Gueter* and *Güeta* to be valid pairs.

19 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html.

20 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

21 Available at https://github.com/fab-bar/SpellVarDetection.

22 The best hyperparameters used in the experiments are as follows: Ridge Classifier alpha=1.0, SVM C=1.0, gamma=1.0, kernel='linear', Logistic Regression C=10000.

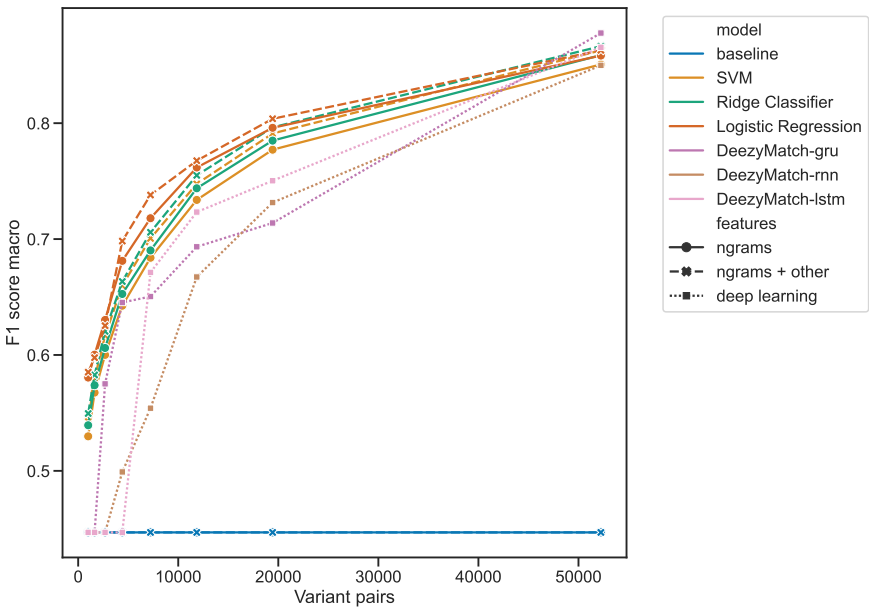


Fig. 6: F1-scores (macro average) on the test set for different amounts of training pairs.

to match the parameters used for Barteld et al.'s feature extractor. We used 85 % of the development dataset as training data and the rest for validation: the final model is the one with the lowest validation loss.

5.4 Results and Discussion

We compared the performance of the different models by varying the size of the training corpus (see Figure 6). The F1-scores increase logarithmically. The machine learning models SVM, Ridge Classifier and Logistic Regression overall perform better than DeezyMatch for the smallest training datasets, but obtain lower or equivalent results for the largest training dataset. The additional features bring some small improvements, which are more notable when less training data are available.

Table 6 shows the final results obtained on the held-out test set. The best model overall is DeezyMatch-gru. We compared the performances of the best performing methods on the macro F1 score using the paired bootstrap test (Berg-Kirkpatrick et al. 2012). The differences between DeezyMatch-gru and SVM, Ridge Classifier and Logistic Regression using the `ng+other` features are statistically sig-

Tab. 6: Evaluation results on the held-out test set. Scores are reported as macro-averages, since the dataset is imbalanced. We also include the F1 score for the minority class (non-variant NV). The baseline corresponds to always returning the majority class.

Model	Features	Precision	Recall	F1	F1 (NV)
Baseline		0.4	0.5	0.45	0
SVM	ng	0.88	0.83	0.85	0.75
Ridge Classifier	ng	**0.9**	0.83	0.86	0.77
Logistic Regression	ng	0.88	0.84	0.86	0.77
SVM	ng+other	0.89	0.84	0.86	0.77
Ridge Classifier	ng+other	**0.9**	0.84	0.87	0.78
Logistic Regression	ng+other	0.87	0.86	0.86	0.78
DeezyMatch-gru		0.88	**0.88**	**0.88**	**0.8**
DeezyMatch-rnn		0.85	0.85	0.85	0.76
DeezyMatch-lstm		0.87	0.86	0.87	0.78

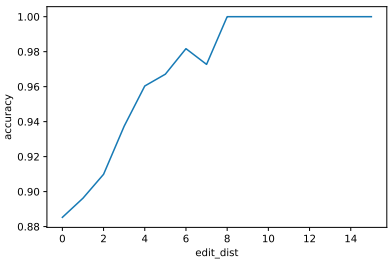


Fig. 7: Accuracy of DeezyMatch-gru as a function of the edit distance between the words in the pair.

nificant ($p<0.01$). The differences between DeezyMatch-gru on the one hand and DeezyMatch-rnn and -lstm on the other are also statistically significant ($p<0.01$). We can therefore conclude that DeezyMatch-gru is the best performing classifier for our dataset.

Finally, we analyze the accuracy of DeezyMatch-gru as a function of the edit distance between words (see Figure 7). As could be expected, word pairs with a larger edit distance (and thus more spelling differences) are easier to classify than word pairs with a lower edit distance (short words, or word pairs with few but distinctive differences).

The results obtained with these experiments are encouraging and we plan to utilize these methods to allow better access to the MeThAl corpus and to perform analyses based on the vocabulary and textual contents of the plays. Moreover, due

to limited workforce, we were only able to validate about 2/3 of the initial data used for building the ELAL emotion lexicon. We will try and use a semi-automatic method based on DeezyMatch to correct the data that have not been manually annotated yet.

6 Sharing the Project Resources

The resources are encoded in standard formats to facilitate reuse. They are also public, shared through several means: (1) our institutional GitLab repository;[23] (2) Nakala data repository, hosted at the French national infrastructure for digital humanites;[24] (3) the DraCor platform, which has accepted our first 25 plays, and where we plan on sharing the rest of the collection.[25] Thanks to a comprehensive TEI header and the DOI assigned by Nakala, our resources follow the FAIR principles.

We hope to promote interest in Alsatian in a wider public beyond the research community. For this reason, we developed a corpus navigation interface.[26] The usefulness of interfaces in literary research is disputed: whereas Schuwey (2019, pp. 12–15) has stressed their importance, Reiter et al. (2017, p. 1184) warn that their impact is not always positive, in the sense that they provide ready results rather than promoting a focus on the exact methodological choices used to arrive at them. Since our project targets not only specialists but also the general public, we believe that an interface can help make the tradition better known.

The interface allows for a structured navigation of the plays based on our character social annotations, besides on the bibliographic metadata (see figure 8). It is possible to filter the corpus based on the presence and co-occurrence of professional groups in the *dramatis personæ*, and filtering with other social variables like class will also be implemented. This can help find plays which show a conflict between certain groups and we hope it is an attractive way for the public to engage with the collection. A full-text search that handles variation in the corpus is under development.

23 https://git.unistra.fr/methal/methal-sources.
24 https://nakala.fr/collection/10.34847/nkl.feb4r8j9.
25 https://dracor.org/als.
26 https://methal.eu/ui.

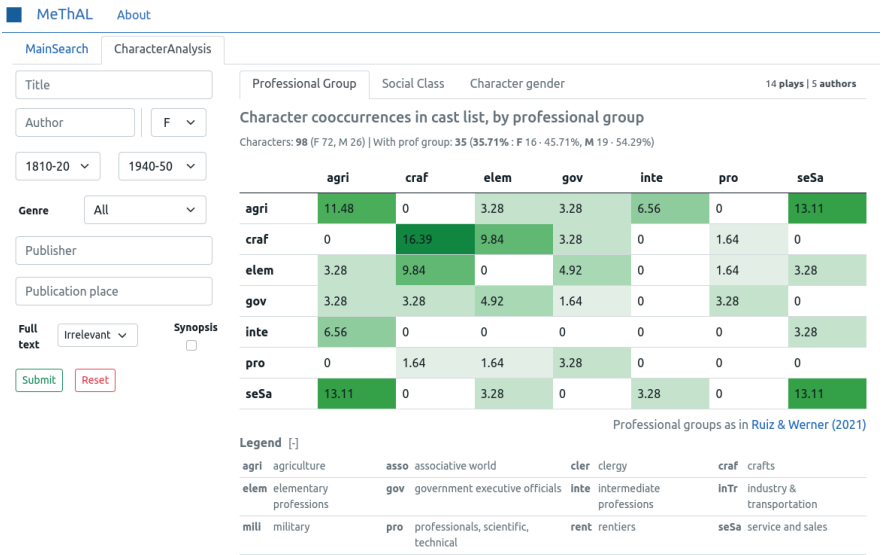


Fig. 8: Character view of the interface. The corpus is filtered to plays by female authors. It can be filtered further according to the character-group co-occurrences in the central pane.

7 Conclusion and Perspectives

We presented our work towards creating the first large-scale TEI corpus of Alsatian theater covering the 1870 to 1940 period. This is a dramatic tradition for which close to zero electronic text existed. NLP resources for Alsatian are also scarce, which limits the CLS questions that can be addressed. TEI encoding was aided by creating a simple sequence labelling model (Conditional Random Fields) which predicts elements based on OCR outputs. It was possible to train the model with our first seven plays only, which we had encoded using rules. This method can be applied to other dramatic corpora for which no electronic text existed and performing OCR was a requirement. Earlier literature on this tradition based on small samples had examined social groups in the plays. We performed an annotation of character social variables for all plays' *dramatis personæ*, later shared as a TEI personography implemented with feature structures which reflect characterisemes or characterization units. This gave an informative overview of the evolution of social groups in the plays which complements what was known about the tradition. It also attests to the usefulness of annotating character metadata systematically in a collection where the text is not yet encoded and where content analyses are hindered by large scripto-linguistic variation. In the future, characters from the

major traditions surrounding Alsatian theater (French and German) could be annotated with the feature library in order to compare the plays' character makeup. As a first step towards emotion analysis, we created an emotion lexicon that handles variation in the corpus. We presented variation detection methods that could be applied to similar scenarios and that can facilitate full-text search. As more linguistic resources get developed for Alsatian varieties (cf. DIVITAL project, Bernhard and Vergez-Couret (2022)), the range of possible NLP-based literary analyses of Alsatian theater will broaden.

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

Acknowledgment: We thank further project members, Dominique Huck and Pascale Erhart, for advice on corpus selection and other project aspects.

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

The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing 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

Alto Editorial Board (2022). "ALTO: Technical Metadata for Layout and Text Objects." URL: https://www.loc.gov/standards/alto/.

Baierer, Konstantin, ed. (2020). *hOCR - OCR Workflow and Output Embedded in HTML*. URL: http://kba.cloud/hocr-spec/1.2/.

Barteld, Fabian, Chris Biemann, and Heike Zinsmeister (2019). "Token-based spelling variant detection in Middle Low German texts." In: *Language Resources and Evaluation*, pp. 1–30. DOI: 10.1007/s10579-018-09441-5.

Berg-Kirkpatrick, Taylor, David Burkett, and Dan Klein (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*, pp. 995–1005.

Bermúdez Sabel, Helena (2019). "Encoding of Variant Taxonomies in TEI." In: *Journal of the Text Encoding Initiative* Issue 11. DOI: 10.4000/jtei.2676.

Bernhard, Delphine and Pablo Ruiz Fabo (2022). "ELAL: An Emotion Lexicon for the Analysis of Alsatian Theatre Plays." In: *Conference on Language Resources and Evaluation (LREC 2022)*, pp. 5001–5010. URL: https://hal.archives-ouvertes.fr/hal-03655148.

Bernhard, Delphine and Marianne Vergez-Couret (2022). *DiViTal: Increase the DIgital VITALity and Visibility of Languages of France: Linguistic Descriptions and Annotated Corpora*. URL: https://divital.gitpages.huma-num.fr/en/.

Cerf, Eve (1972). "Le théâtre alsacien de Strasbourg, miroir d'une société (1898-1939)." In: *Saisons d'Alsace* 43.

Cerf, Eve (1975). "Les contes merveilleux du théâtre alsacien de Strasbourg." In: *Revue des sciences sociales de la France de l'Est* 4, pp. 3–30.

Coll Ardanuy, Mariona, Kasra Hosseini, Katherine McDonough, Amrey Krause, Daniel van Strien, and Federico Nanni (2020). "A Deep Learning Approach to Geographical Candidate Selection through Toponym Matching." Tech. rep. arXiv:2009.08114. arXiv. DOI: 10.48550/arXiv.2009.08114.

Deutsche Forschungsgemeinschaft (2016). "DFG Practical Guidelines on Digitisation." Deutsche Forschungsgemeinschaft. URL: https://www.dfg.de/formulare/12_151/12_151_en.pdf.

Erjavec, Tomaž, Darja Fišer, and Nikola Ljubešić (2021). "The KAS Corpus of Slovenian Academic Writing." In: *Language Resources and Evaluation* 55.2, pp. 551–583. DOI: 10.1007/s10579-020-09506-4.

Even-Zohar, Itamar (1990). "Polysystem Theory." In: *Polysystem Studies [=Poetics Today]* 11.1, pp. 9–26. URL: https://www.tau.ac.il/~itamarez/works/books/Even-Zohar_1990--Polysystem%5C%20studies.pdf.

Fièvre, Paul (2017). *Rubrique Concernant Les Personnages*. Théâtre Classique. URL: http://www.theatre-classique.fr/pages/PagePersonnages.html.

Fischer, Frank and Ingo Börner (2019). "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." In: *Digital Humanities 2019*, p. 5. URL: https://dev.clariah.nl/files/dh2019/boa/0268.html.

Gabay, Simon, Barbara Topalov, Caroline Corbières, Lucie Rondeau Du Noyer, Béatrice Joyeux-Prunel, and Laurent Romary (2021). "Automating Artl@s – Extracting Data from Exhibition Catalogues." In: *EADH 2021 - Second International Conference of the European Association for Digital Humanities*.

Gall, Jean-Marie (1974). *Le théâtre populaire alsacien au XIXe siècle*. Istra.

Galleron, Ioana (2017). "Conceptualisation of Theatrical Characters in the Digital Paradigm: Needs, Problems and Foreseen Solutions." In: *Human and Social Studies* 6.1, pp. 88–108. DOI: 10.1515/hssr-2017-0007.

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, and Frank Wiegand (2012). "TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv." In: *Jahrbuch für Computerphilologie*. URL: http://computerphilologie.digital-humanities.de/jg09/geykenetal.html.

Hosseini, Kasra, Federico Nanni, and Mariona Coll Ardanuy (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*, pp. 62–69. DOI: 10.18653/v1/2020.emnlp-demos.9.

Huck, Dominique (1998). "D'r Herr Maire (1898) de Gustave Stoskopf. Entre ethnologie et littérature : les Alsaciens en auto-représentation." In: *Recherches Germaniques* 28, pp. 163–190.

Huck, Dominique (2005). "Le 'Théâtre Alsacien de Strasbourg' et la production dramaturgique de ses fondateurs (1898-1914)." In: *Culture et histoire des spectacles en Alsace et en Lorraine : De l'annexion à la décentralisation (1871-1946)*. Ed. by Jean-Marc Leveratto, Jeanne Benay, Olivier Thomas, and Séverine Wuttke. Peter Lang, pp. 198–222.

Huck, Dominique (2015). *Une Histoire Des Langues de l'Alsace*. La Nuée Bleue.

Huck, Dominique, Arlette Bothorel-Witz, and Anemone Geiger-Jallet (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*. Ed. by Andrea Abel, Mathias Stuflesser, and Leonhard Voltmer. EURAC Research (Europäische Akademie / Accademia Europea / European Academy), pp. 13–101.

Hülsen, Bernhard von (2003). *Szenenwechsel im Elsass: Theater und Gesellschaft in Strassburg zwischen Deutschland und Frankreich 1890-1944*. Leipziger Universitätsverlag.

Khemakhem, Mohamed, Luca Foppiano, and Laurent Romary (2017). "Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields." In: *electronic lexicography, eLex 2017*. URL: https://hal.archives-ouvertes.fr/hal-01508868.

Khemakhem, Mohamed, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, and Giancarlo Luxardo (2018). "Automatically Encoding Encyclopedic-like Resources in TEI." In: *Proceedings of the annual TEI Conference and Members Meeting*. URL: https://hal.inria.fr/hal-01819505.

Korobov, Mikhail [Nov. 26, 2015] (2019). *sklearn-crfsuite*. TeamHG-Memex. URL: https://github.com/TeamHG-Memex/sklearn-crfsuite.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In: *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289.

Moeller, Katrin and Robert Nasarek (2018). "Die Ontologie historischer deutschprachiger Berufs- und Amtsbezeichnungen. Interoperationalität und Berufsklassifizierung durch semantisches Topic Modeling." In: *DHd 2018. Kritik der Digitalen Vernunft. Konferenzabstracts*. DHd 2018, pp. 173–177.

Pagel, Janis, Nidhi Sihag, and Nils Reiter (2021). "Predicting Structural Elements in German Drama." In: *Second Conference on Computational Humanities Research*, pp. 217–227.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pfister, Manfred [1977] (2001). *Das Drama: Theorie und Analyse*. 11th ed. W. Fink.

Phelan, James (1989). *Reading People, Reading Plots: Character, Progression, and the Interpretation of Narrative*. University of Chicago Press.

Reiter, Nils, Jonas Kuhn, and Marcus Willand (2017). "To GUI or Not to GUI?" In: *INFORMATIK 2017*. URL: https://dl.gi.de/bitstream/handle/20.500.12116/3880/B14-8.pdf?sequence=1&isAllowed=y.

Reiter, Nils and Janis Pagel (2020). *quadrama/DramaAnalysis: 3.0.2*. Version v3.0.2. DOI: 10.5281/zenodo.4051004.

Romary, Laurent (2015). "Standards for Language Resources in ISO – Looking Back at 13 Fruitful Years." In: *edition - Die Fachzeitschrift für Terminologie* 11.2, pp. 13–19.

Ruiz Fabo, Pablo and Carole Werner (2021). "Exploration du théâtre alsacien à travers ses listes de personnages pendant la période 1870-1940." In: *Humanistica 2021*. DOI: 10.5281/zenodo.4762733.

Schuwey, Christophe (2019). *Interfaces. L'apport des humanités numériques à la littérature*. Alphil éditions.

Smith, Ray (main developer) (2018). *Tesseract (v4.0)*. URL: https://github.com/tesseract-ocr/tesseract.

TEI Consortium (2022). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version v4.4.0. DOI: 10.5281/zenodo.6482461.

Van Leeuwen, Marco H. D., Ineke Maas, and Andrew Miles (2004). "Creating a Historical International Standard Classification of Occupations An Exercise in Multinational Interdisciplinary Cooperation." In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 37.4, pp. 186–197. DOI: 10.3200/HMTS.37.4.186-197.