Classifying TEI Encoding for DutchDraCor with Transformer Models

Florian Debaene, Véronique Hoste
LT[3], Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

Computational Drama Analysis relies on well-structured textual data, yet many dramatic works remain in need of encoding. The Dutch dramatic tradition is one such an example, with currently 180 plays available in the DraCor database, while many more plays await integration still. To facilitate this process, we propose a semi-automated TEI encoding annotation methodology using transformer encoder language models to classify structural elements in Dutch drama. We fine-tune 4 Dutch models on the DutchDraCor dataset to predict the 9 most relevant labels used in the DraCor TEI encoding, experimenting with 2 model input settings. Our results show that incorporating additional context through beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens greatly improves performance, increasing the average macro F1 score across models from 0.717 to 0.923 (+0.206). Using the best-performing model, we generate silver-standard DraCor labels for EmDComF, an unstructured corpus of early modern Dutch comedies and farces, paving the way for its integration into DutchDraCor after validation.

1 Introduction & Related Work

The Drama Corpora Project (DraCor) is a rapidly growing open database that employs TEI XML encoding to standardize language-independent, digitally readable formatting of dramatic texts (Fischer et al., 2019). This encoding facilitates computational and comparative research on drama across historical periods, languages, and cultures. However, manually encoding texts according to the Text Encoding Initiative (TEI Consortium, 2025) is a labor-intensive and time-consuming process, which presents a major bottleneck in the expansion and scalability of DraCor. This challenge is evident for the Dutch dramatic tradition among others, which has only recently been incorporated into DraCor. Currently, DutchDraCor contains 180 encoded plays, while hundreds of historical Dutch plays remain unencoded (Debaene et al., 2024), which complicates further structural and comparative analysis. Accelerating the structural encoding of these plays would not only advance research in Dutch literary studies but also support the emerging field of Computational Drama Analysis (Andresen and Reiter, 2024), enabling large-scale, cross-linguistic, and diachronic comparisons of dramatic traditions.

To address this bottleneck, recent research has explored the use of Machine Learning (ML) to support or automate aspects of TEI annotation in digital literary corpora. Pagel et al. (2021) investigate the automatic enrichment of German dramatic text with structural TEI elements. Using fine-tuned BERT-based models, they predict 5 elements (*"act"*, *"scene"*, *"stage"*, *"speaker"*, *"speech"*) and achieve promising results in identifying these structural features from plain text after sentence tokenization. Similarly, Schneider and Fabo (2024) focus on the fine-grained classification of stage directions in French theater. They propose a detailed 13-class typology of stage directions and fine-tune transformers to classify these, demonstrating that even with limited training data, transfer learning techniques can support the structural annotation tasks relevant for computational literary studies.

Building on these approaches, this work aims to automatically annotate historical Dutch drama with structural DraCor labels by leveraging the existing DutchDraCor as a dataset. Assigning a label from the most fundamental set of TEI elements to each line of text from DutchDraCor, we model this task as a multiclass classification problem. Innovatively, we experiment with incorporating additional contextual information as adjacent lines in the model input, introducing beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens, to operationalize the structurally repetitive nature of dramatic texts. To our knowledge, this feature of drama has not been put to use in similar classification

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 136–140
July 5–10, 2020.

contexts, as related work focuses on classifying individual textual instances, often sentences. We hypothesize, however, that expanding the context will improve models' performance for this task, as it might help models to classify speakers, spoken text, act divisions and stage directions when the immediately preceding and subsequent context is given. The ultimate aim of this research is to support the semi-automated annotation of unstructured dramatic texts for DutchDraCor, reducing the manual workload for human annotators. After validation, the automatically annotated labels following from this work in other unstructured plays can serve as gold-standard TEI markup and facilitate DraCor integration. This work presents a methodology that offers scalable solutions to support the incorporation of dramatic literary traditions into DraCor, even if no specifically historically adapted language models exist, as we expect it to be transferable to encoding drama in other languages and contexts. Our contributions to automatically encode drama therefore include:

1. **Operationalizing DutchDraCor for ML:** We create and release the DutchDraCor4ML dataset, enabling supervised learning for TEI encoding classification in historical Dutch.

2. **Fine-tuning Dutch transformer models for TEI encoding classification:** We apply 4 Dutch transformer-based encoder models, both historical and contemporary, to classify TEI elements in historical Dutch drama. We release the best performing fine-tuned model, GysBERT4DutchDraCor.

3. **Improving classification by increasing context:** We enhance classification performance by increasing the model input context and by introducing BOS and EOS tokens, improving the average macro F1 score from 0.717 to 0.923 (+0.206) across models.

4. **Application on EmDComF corpus:** We apply GysBERT4DutchDracor to EmDComF (Debaene et al., 2024), an unstructured corpus of early modern Dutch comedies and farces, generating silver-standard TEI labels, and release EmDComF4DutchDraCor.

2 Operationalizing DutchDraCor

Given that DutchDraCor contains 180 manually annotated plays with TEI encoding, we can operationalize these annotations to create a fine-tuning dataset for TEI encoding classification. In total, TEI files in DutchDraCor contain 52 unique labels. However, predicting all 52 labels is unnecessary, as rule-based approaches can help create some of the umbrella TEI elements, such as speaker turns containing a speaker and their spoken text, or the list of characters containing all roles of the play. We therefore focus on extracting the most relevant labels from the DutchDraCor plays on the condition that a label contains text. After manual inspection, the following 9 labels seemed to encode all textual instances of a play: *"line"*, for spoken lines by each *"speaker"*; *"stage"* for stage directions; *"head"* for structural indications such as act and scene divisions; *"persName"* for author names and the list of characters, which is in some plays annotated with *"role"*; *"paragraph"* elements indicating legal clauses regarding ownership, dedications, or other prefaces; and *"title"* and *"titlePart"* elements, which marks statements from the title page regarding place of publishing and the editor. Creating random 70-20-10% splits based on the 180 DutchDraCor plays, all text contained in the aforementioned labels was extracted per split for training, testing and development respectively (Section 3), resulting in the label distribution showed in Table 1.

	Train	Test	Dev
line	175,807	64,175	24,857
speaker	40,395	12,986	6,357
stage	3,819	1,304	601
head	2,044	904	316
persName	1,453	444	219
role	1,323	436	203
paragraph	1,211	385	167
titlePart	327	147	63
title	310	97	42

Table 1: Label distribution of the DutchDraCor dataset.

3 Model Fine-Tuning

We leverage the operationalized DutchDraCor dataset to fine-tune existing language models for classification. For this, we choose language models trained on Dutch. These include GysBERT (Manjavacas Arevalo and Fonteyn, 2022), fine-tuned on historical Dutch, RobBERT (Delobelle et al., 2020) and BERTje (de Vries et al., 2019), both fine-tuned on contemporary Dutch, and finally GysDRAMA, a GysBERT model fine-tuned by continuing full-

model pre-training on Dutch dramatic texts (Debaene et al., Forthcoming). Each of these models are given the dataset for fine-tuning in 2 model input settings. In setting T, extracted text is given and the model is tasked to predict the correct label. In setting T+C, extracted text is contextualized with adjacent lines, namely the preceding and subsequent line, and delimited with beginning-of-sequence (BOS) and end-of-sequence (EOS) tokens. The model is then tasked to predict the correct label. An example from the opening scene of Vondel's Gysbreght van Aemstel (1637), with both model input settings:

model input	**label**
1T. Gysbreght van Aemstel.	*head*
2T. Het eerste bedryf.	*head*
3T. Gysbreght van Aemstel	*speaker*
4T. Het hemelsche gerecht heeft zich...	*line*
1T+C. [BOS] Gysbreght van Aemstel. [EOS] Het eerste bedryf.	*head*
2T+C. Gysbreght van Aemstel. [BOS] Het eerste bedryf. [EOS] Gysbreght van Aemstel	*head*
3T+C. Het eerste bedryf. [BOS] Gysbreght van Aemstel [EOS] Het hemelsche gerecht heeft zich...	*speaker*

Using both input settings, the models were fine-tuned using the transformers library (Wolf et al., 2020) on 4x NVIDIA A100-SXM4 (40 GB GPU memory) GPUs for 5 epochs with batchsize 8. To prevent overfitting, we implemented early stopping if the eval_F1 did not increase after 3 evaluations on the dev set. We evaluated every 2000 steps, which coincided with a quarter epoch roughly. After training, model performance was evaluated on the test set.

4 Results

Table 2 presents the F1 scores of the 4 fine-tuned transformer encoder models (BERTje, GysBERT, GysDRAMA, and RobBERT) for predicting the 9 labels in the DutchDraCor dataset. Each model was evaluated with the 2 input settings: (1) using only the extracted text (T), and (2) incorporating additional context from adjacent lines with BOS and EOS tokens (T+C).

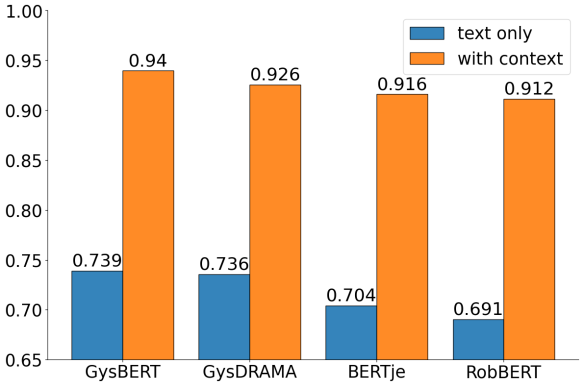

Figure 1: Macro averaged F1 scores on test set.

4.1 Performance Improvement with Context

Across all models, providing contextual information (T+C) greatly improves classification performance for almost all labels. The average macro F1 score increases from 0.717 to 0.923 (+0.206), demonstrating the importance of contextualization in TEI encoding classification. This increase is particularly pronounced for labels that are often ambiguous without additional textual cues, such as *"persName"* and *"role"*, and *"title"* and *"titlePart"*, where classifiers in the text-only setting struggle due to limited information. By explicitly marking the sequence boundaries and incorporating surrounding lines, models gain a better understanding of which textual cues lead to the correct TEI label, resulting in more accurate predictions. Figure 1 visualizes these improvements, showing a consistent trend where contextualization benefits all models, regardless of whether they were initially pre-trained on historical or contemporary Dutch. This suggests that the improvement is not merely due to domain adaptation but rather an inherent advantage of the structurally repetitive nature of dramatic texts.

4.2 Model Comparisons

GysBERT consistently performs best when using contextualized input (T+C), achieving the highest F1 scores for 7 of the 9 labels, including *"head"* (0.951), *"line"* (0.997), *"paragraph"* (0.813), *"persName"* (0.966), *"speaker"* (0.986), *"stage"* (0.918), and *"titlePart"* (0.906). GysDRAMA, which was specifically pre-trained on Dutch dramatic texts, follows closely behind, especially for *"role"* (0.950), *"title"* (0.984) on par with GysBERT, and *"speaker"* (0.979). BERTje and RobBERT also show strong improvement with context but slightly trail behind GysBERT and GysDRAMA in

	BERTje		GysBERT		GysDRAMA		RobBERT	
	T	T+C	T	T+C	T	T+C	T	T+C
line	0.992	0.996	0.991	**0.997**	0.993	0.995	0.992	0.996
speaker	0.940	0.983	0.852	**0.986**	0.882	0.979	0.909	0.985
stage	0.757	0.898	0.838	**0.918**	0.831	0.894	0.821	0.900
head	0.932	0.904	0.936	**0.951**	0.936	0.921	0.913	0.925
persName	0.362	0.939	0.176	**0.966**	0.172	0.956	0.237	0.940
role	0.661	0.913	0.680	0.936	0.697	**0.950**	0.668	0.904
paragraph	0.608	0.774	0.644	**0.813**	0.716	0.756	0.687	0.779
titlePart	0.647	0.848	0.451	**0.906**	0.702	0.896	0.488	0.801
title	0.723	**0.990**	0.646	0.985	0.723	0.984	0.623	0.974

Table 2: Detailed F1 scores on test set after fine-tuning on text only (T) and text with context (T+C).

several categories, as the latter are domain-adapted to historical Dutch. However, BERTje achieves the highest score for *"title"* (0.990), and RobBERT maintains competitive performance across labels but does not outperform GysBERT or GysDRAMA in any class. These results emphasize the benefit of domain-specific model fine-tuning for TEI encoding classification, as models like GysBERT and GysDRAMA demonstrate a stronger ability to capture the textual patterns inherent in historical Dutch dramatic texts leading to the correct TEI label. Nevertheless, the fact that even the contemporary Dutch language models BERTje and RobBERT benefit from the added context suggests the generalizability of our approach.

4.3 Label-Specific Insights

"Line" is classified with near-perfect accuracy by all models, with scores reaching up to 0.997. By far the largest class, spoken text follows easily discernible patterns in Dutch drama. Structural elements (*"head"*, *"stage"*, *"speaker"*) show strong classification improvements when context is provided, particularly *"speaker"*, where model performance improves from 0.852 (GysBERT, T) to 0.986 (T+C). Less frequent labels (*"persName"*, *"role"*, *"paragraph"*, *"titlePart"*) benefit the most from context. For example, the classification performance for *"persName"* improves dramatically in GysBERT (from 0.176 to 0.966), suggesting that surrounding textual cues help identify named entities. Finally, while performance improves notably with context to predict *"paragraph"* (GysBERT, 0.813), it remains one of the weaker classes. This suggests that legal clauses, dedications, and prefaces in historical Dutch drama may vary significantly in structure, making them harder to classify.

5 Conclusion & Future Work

This work suggests that incorporating contextual information substantially enhances TEI encoding classification in historical Dutch drama, improving performance across both historical Dutch models (GysBERT, GysDRAMA) and general-purpose Dutch models (BERTje, RobBERT). By expanding the input beyond isolated text segments, transformer-based encoder models achieve a deeper understanding of dramatic structures, leading to more accurate predictions. Notably, even models not pre-trained on historical language successfully classify TEI labels when given additional context, highlighting fine-tuning and contextualization as effective strategies for adapting modern NLP techniques to this specific annotation task for historical and literary corpora. Beyond Dutch drama, these findings suggest broader applications for Machine Learning and deep learning techniques in TEI encoding, particularly in other dramatic traditions facing similar challenges in encoding standardization and accessibility. Transformer encoder models, with contextualized input, offer a scalable approach to facilitating Computational Drama Analysis across languages and periods, even when domain-specific language models are not readily available. Future work should explore cross-linguistic adaptations and deeper integration with TEI workflows, advancing the intersection of NLP and digital humanities for more comprehensive literary and theater studies.

Limitations

In this work, we researched whether context improves TEI encoding classification, but did not investigate the impact of context quantity on model

performance. Although we found that adding contextual input improves classification performance, transformer models have a fixed context window, which may limit their ability to capture distant dependencies beyond the three-sample input. We base our findings on fine-tuning with a single random seed. This means that the observed performance differences between models, such as GysDRAMA performing slightly worse than GysBERT, may be due to randomness rather than inherent model differences. Given that these differences are small, it is possible that they are not statistically significant. Future work should investigate this more systematically. However, model comparison was not the main focus of this study; rather, our goal was to explore how to effectively structure an automatic annotation task for TEI encoding historical drama with existing resources, making detailed benchmarking somewhat beyond our current scope. Furthermore, since our experiments focus exclusively on Dutch drama, the generalizability of this approach to other dramatic traditions or languages with perhaps different structural conventions seems feasible, but remains untested. Inconsistencies in TEI annotations across historical texts, including variations in editorial practices and incomplete markup, pose additional challenges that may introduce noise and affect model reliability. Future research should address these limitations by exploring multilingual validation, improving long-text processing, and refining TEI standardization to support broader applications in Computational Drama Analysis.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. This work was supported by the Research Foundation Flanders (FWO) under grant G032123N.

References

Melanie Andresen and Nils Reiter. 2024. *Computational Drama Analysis: Reflecting on Methods and Interpretations*. De Gruyter.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582.

Florian Debaene, Aaron Maladry, Pranaydeep Singh, Els Lefever, and Véronique Hoste. Forthcoming. Unlocking domain knowledge: Model adaptation for non-normative dutch. *Computational Linguistics in the Netherlands Journal*, 14.

Florian Debaene, Kornee van der Haven, and Veronique Hoste. 2024. Early Modern Dutch comedies and farces in the spotlight: Introducing EmDComF and its emotion framework. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 144–155, Torino, Italia. ELRA and ICCL.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. Zenodo.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Janis Pagel, Nidhi Sihag, and Nils Reiter. 2021. Predicting Structural Elements in German Drama. In *Proceedings of the Second Conference on Computational Humanities Research*, volume 1613, page 0073.

Alexia Schneider and Pablo Ruiz Fabo. 2024. Stage direction classification in french theater: Transfer learning experiments. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 278–286.

TEI Consortium. 2025. Tei p5: Guidelines for electronic text encoding and interchange. 4.9.0.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.