Stage Direction Classification in French Theater: Transfer Learning Experiments

Alexia Schneider and **Pablo Ruiz Fabo**
Université de Strasbourg, LiLPa UR 1339
67000 Strasbourg, France
alexia.schneider4@etu.unistra.fr,ruizfabo@unistra.fr

Abstract

The automatic classification of stage directions is a little explored topic in computational drama analysis (CDA), in spite of their relevance for plays' structural and stylistic analysis. We developed a 13-class stage direction typology, based on annotations in the FreDraCor corpus (French-language plays), but abstracting away from their huge variability while still providing classes useful for literary research. We fine-tuned transformers-based models to classify against the typology, gradually decreasing training-corpus size to compare model efficiency with reduced training data. A result comparison speaks in favour of distilled monolingual models for this task, and, unlike earlier research on German, shows no negative effects of model case-sensitivity. The results have practical relevance for computational literary studies, as comparing classification results with complementary stage direction typologies, limiting the amount of manual annotation needed to apply them, would be helpful towards a systematic study of this important textual element.

1 Introduction

Machine learning methods have brought important contributions to Computational Literary Studies (CLS). To name just one monograph-length work, Underwood (2019) used such methods to provide insights on complex issues like the long-term evolution of genres and of literary prestige criteria, focusing mainly on fiction and poetry. Drama has also benefited from such approaches. The recent *Computational Drama Analysis* workshop[1] featured work on the automatic classification of dramatic situations, character types, and emotions in drama. In this paper, we approach a little explored dramatic analysis topic: the automatic classification of stage directions, using a French theater corpus. Stage directions introduce indications about performance, decoration or other information to complement character speech, but can sometimes be largely independent from it (Pfister, 1988, 15). Several typologies have been proposed for them in literary studies (see 3.1). However, their automatic classification has scarcely been studied and poses challenges, given types hard to distinguish from each other. Stage directions on characters' entrance and exit indicate changes to character copresence on stage and are thus tied to play structure and dramatic technique. The frequency and length of stage directions and their types can be stylistic parameters related to author groups or subgenres. Automatically classifying stage directions facilitates large-scale quantitative analyses of this element's structural and stylistic role.

For French theater, the FreDraCor corpus (Milling et al., 2021), based on Fièvre (2007) and covering mostly the 16th to 20th centuries, offers over 38,000 annotated stage directions. Given the large number of categories (over 5,000), exploiting these annotations for supervised learning is a challenge, that we address in the paper.

Pre-trained language models were a game changer in NLP, allowing for transfer learning (e.g. Devlin et al. 2019) that yields viable classifiers even with a reduced number of examples, or, in the case of larger language models (Brown et al., 2020), in-context learning from (almost) zero examples. In our study, we work with specialized categories for which we could develop annotations, and we opted for transfer learning. We examine the extent to which we can reduce manually annotated training data for the supervised classification of stage directions, in French. As producing manual annotations can be costly, and exploring literary questions may require comparing the results of classifying against several, complementary typologies, the question addressed has practical relevance for CLS. The paper's contributions are:

- A new stage direction typology (3.1) based on the related literary theory and on the FreDra-

[1] https://page.hn/anuvah

Proceedings of LaTeCH-CLfL 2024, pages 278–286
March 22, 2024

Cor types, but abstracting away from some of their large variability to obtain a category set amenable to supervised learning and still useful for addressing literary questions.

- Experiments to clarify which language models (LMs) learn most efficiently on such data, focusing on model characteristics that may generalize beyond our corpus language (French).

An overall goal is to start a reflection on good practices to develop methods to classify this textual element, across languages.

The paper is structured as follows: Section 2 reviews related work. Section 3 describes our typology and classification workflow. Section 4 presents results, and section 5 outlines future perspectives.

2 Related Work

Heterogeneous criteria have been used to design stage direction typologies (Dahms, 1978; Gallèpe, 1997; Issacharoff, 1981; Martinez Thomas, 2007; Pfister, 1988). Among other aspects, taxonomies pay attention to whether stage directions refer to verbal/speech-related or visual information, to characters or setting, whether they describe movements (including entrance and exit) or character interactions. Another feature used is their narrative vs. descriptive nature, their relatedness with or independence from spoken text, or their impact on the play's plot. The Text Encoding Initiative (TEI) guidelines (TEI Consortium, 2023) reflect this heterogeneity in their description of possible `@type` attribute values for the TEI `<stage>` element, used for stage directions. Galleron (2021) attempts to reconcile inconsistencies in earlier typologies, systematizing types via a set of values for the `@ana` attribute of `<stage>`.

Automatic stage direction classification was performed in German by Dennerlein (2016), with four classes (*exit*, *entrance*, *dead* and *aside*); per-class results ranged between 0.75 and 0.88 F1 using random forests. The typology is less complex than our 13-class typology and the classes are more distinct. Maximova and Fischer (2019), working with Russian, developed a model to classify against the TEI guidelines' 9 categories, trained on 6,569 manually annotated examples and reaching ca. 0.75 F1.

Pagel et al. (2021) performed a related but not equivalent task. In their study on predicting German plays' structural elements in TEI, one of the five classes was stage directions, besides act and scene divisions, speaker names and speeches. They thus classified stage directions vs. other structural elements, reporting that a binary classification between stage directions and character speech was not trivial (0.81 F1). Stage directions were also the worst performing class in the 5-way classification (0.84 F1, while other classes were above 0.9). To assess the role for language-specific knowledge, they fine-tuned both English and German BERT cased and uncased models, with best results for the German uncased model. Their experimental setup informed our own (3.2).

3 Methods

3.1 Stage Direction Typology

We wanted to start assessing to what extent it is possible to automatically classify stage directions against different typologies useful for literary analysis. We do not intend the typology here to be the only choice, but in a way a testbed for fine-tuning and a means to assess the potential of the models to classify this type of material, using these classes or similar ones according to researchers' needs.

To develop the typology (table 1),[2] we started off from FreDraCor, which has 38,306 `<stage>` elements with 5,109 unique types.

We grouped semantically the 87 most frequent values, covering 25,823 stage directions, into our 13-class typology, creating a mapping between FreDraCor original labels and our own (Appendix B). E.g. FreDraCor labels *location*, *decor* and *décor* yield class *Setting* in our typology, and *kill*, *fight*, *hit*, *suicide*, *threat* yield *Aggression* in our typology. We only considered single-type FreDraCor labels, for simplicity; the 87 categories selected correspond to classes with at least 50 examples.

The typology contains classes that can be very ambiguous, the vocabulary of which is likely to represent different semantic fields, like *Action* or *Narration*, which are intended to be difficult for classifiers. Other classes can often be detected with surface lexical cues, like the presence of certain prepositions in *Toward*. Our choice of classes is meant to reflect different interests that a scholar studying stage directions may have. E.g. *Aggression* stage directions may be more present in a serious subgenre like the tragedy, *Music* stage directions in the *vaudeville*, or long *Narration* types in plays from the 19th century onwards or experimental work. Thus, the detection of such types is

[2] See corpus examples and English glosses in appendix A.

Class	Scope
Action	General character action category
Aggression	Violent action
Aparté	Aside (character addresses audience or is alone)
Delivery	Delivery manner (e.g. *laughs*, *sobs*)
Entrance	Character enters stage
Exit	Character exits
Interaction	Non-verbal character interaction
Movement	Character movement (but not exit/entrance)
Music	Tune names (in plays with songs)
Narration	Long, "narrative quality", for readers
Object	Describes object or interaction with it
Setting	E.g. *the stage represents a bar*
Toward	Indicates the addressee of a speech

Table 1: Stage direction typology

relevant for subgenre characterization. Some types are relevant for dramatic structure, e.g. *Exit* and *Entrance*, which give information about configuration (character co-presence on stage). *Aparte* is related to knowledge distribution in the play, an active CDA research area (Andresen et al., 2024).

After removing duplicates, we obtained a set of 14,613 examples and used it, gradually decreasing training set size, for our fine-tuning experiments (3.2). Label distribution is imbalanced (table 2).

Class	N. examples	Class	N. examples
Music	2863	Delivery	962
Action	2467	Entrance	646
Toward	2144	Movement	583
Exit	1295	Interaction	565
Object	1130	Narration	554
Setting	982	Aggression	350
		Aparté	72

Table 2: Number of examples per class (training corpus)

3.2 Classification Workflow

We first implemented classical machine learning (ML) models, with version 1.0.2 of `scikit-learn` (Pedregosa et al., 2011): Logistic Regression, Ridge Classifier, Random Forest and SGD. For the last two, results reported (macro-F1) are averaged over 5 runs. The implementation used for the first two is deterministic. Hyperparameters were default. The features were character length and the number of sentences, tf-idf-weighted token unigrams, bigrams, and part-of-speech unigrams (obtained with the `fr_core_news_md` module of spaCy by Honnibal et al. 2020) and 2- to 4-character ngrams, with 1% minimum document frequency.

Models for fine tuning were **(1)** **`camembert-base`** (Martin et al., 2020), a French monolingual model trained on the OSCAR corpus (Ortiz Suárez et al., 2019); **(2)** **`distilcamembert-base`** (Delestre and Amar, 2022), a version of (1) that is distilled, i.e. that attempts to preserve result quality while reducing model complexity (thus size and fine-tuning time); **(3)** **`bert-base-multilingual-cased`** and **(4)** **`bert-base-multilingual-uncased`**, case-insensitive version of (3), both by Devlin et al. (2019), which include French among the 102 languages covered. Model **(5)** was **`distilbert-base-multilingual-cased`** (Sanh et al., 2019), a distilled cased version of (3). The final model **(6)** was based on the **SetFit** architecture (Tunstall et al., 2022). This first fine-tunes a Sentence Transformer (S-BERT) model (Reimers and Gurevych, 2019) using contrastive training on positive and negative training pairs, which helps it learn effectively based on a small number of examples. Then, based on the fine-tuned S-BERT model (`distiluse-base-multilingual-cased-v1` in our case), it trains a classifier for the task, with logistic regression in our setup. Fine-tuning was carried out with version 4.34.1 of the `transformers` library (Wolf et al., 2020) on a V100 GPU. Learning rate was 2e-5 and batch size was 16. The trainer was configured with a maximum of 40 epochs, with early stopping monitoring validation loss and a patience of 3, but only `camembert-base` went over 10 epochs on average. For SetFit, as its API has no early stopping callbacks, we chose 6 epochs with 20 contrastive learning iterations each. We report macro-F1 (mean over 5 runs) in table 4.

Our model choices are justified thus: CamemBERT **(1)** is a leading monolingual LM for French. A distilled version **(2)** was also tested because, given that distilled models are smaller and faster, should there be no important difference in result quality, the distilled model is to be preferred. The multilingual BERT models, cased **(3)** and uncased **(4)** were chosen to compare with the monolingual ones, to get an indication of the extent to which language-specific knowledge helps classification, especially as training data is reduced. We chose the distilled version **(5)** for the same reason as (2). Finally, we tested SetFit **(6)** because its contrastive learning approach allows it to learn from limited data, which fits our study's goal. For our experiments, the annotated corpus was increasingly reduced (table 3), and split into training and validation sets. The test-set was always the same (2,923 examples). This was meant to help assess the extent to which models generalize to the test set, even when fine-tuned on a training set smaller than it.

	100%	**50%**	**25%**	**10%**	**5%**
train	9352	4676	2337	935	467
validation	2338	1169	585	234	117
test			2923		

Table 3: Number of examples at each corpus size

4 Results and Discussion

As table 4 and figure 1 show, LMs performed better than classical ML; improvement increases as training set size decreases. With the full corpus, several LMs reach 0.81 F1, and at 50% (5,845 manually annotated examples), macro-F1 ranges between 0.73 and 0.79. Monolingual LMs show an advantage over multilingual ones from 50% on, more so at 25% (2,922 training examples), 10% (1,169 examples) and 5%. Distilled models performed very closely to full ones, either in mono- or multilingual cases. As their fine tuning was taking between 40% and 75% of the time needed by the full model, they are a better choice. SetFit is the only case where a multilingual and distilled model was competitive with monolingual models at 10% (only 0.01 points below distilled CamemBERT), which suggests the effectiveness of its contrastive learning approach. However, fine-tuning time with our setup was much higher than with all other models.

We see no consistent difference between cased vs. uncased models, contrary to Pagel et al. (2021), who observed that a cased model may generalize less well on their German data. The relevance of language-specific knowledge seen in their study is also seen in ours with the higher performance of French language-specific models. The 10% results are interesting, as they were obtained with less than 1,200 manual annotations. The monolingual models provide ca. 0.7 F1 (vs. approx. 0.8 F1 with ca. 6,000 examples). A 0.7 F1 may not be

Model	**100%**	**50%**	**25%**	**10%**	**5%**
Classical ML					
LogisticReg	0.70	0.64	0.57	0.51	0.41
RidgeClassif	**0.73**	**0.71**	**0.62**	**0.55**	**0.49**
RandomForest	0.65	0.61	0.57	0.49	0.45
SGD	**0.73**	0.69	**0.62**	0.54	**0.49**
Transfer learning					
(1) cam-base	0.77	0.77	**0.75**	**0.71**	**0.68**
(2) d-cam-base	**0.81**	**0.79**	0.74	0.7	0.67
(3) mbert-cas	**0.81**	0.74	0.69	0.61	0.51
(4) mbert-unc	**0.81**	0.74	0.7	0.61	0.57
(5) d-mbert-cas	0.8	0.73	0.67	0.61	0.53
(6) setfit-dmc	0.78	0.74	0.69	0.69	0.62

Table 4: Macro-F1 with different training-set sizes, testing on the 2,923 example test-set

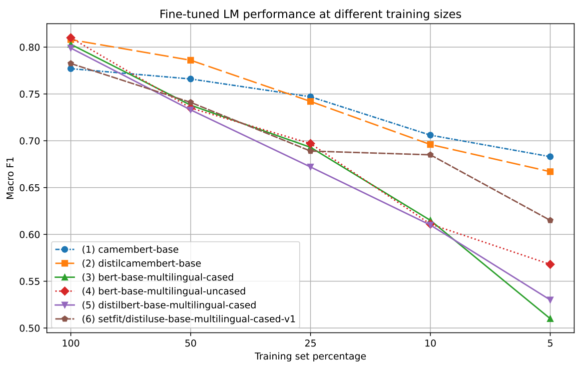
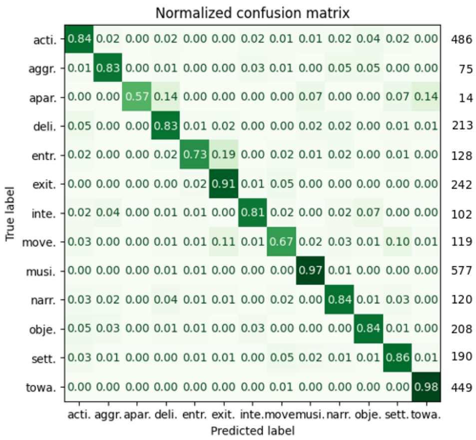


Figure 1: LM performance as training set is reduced

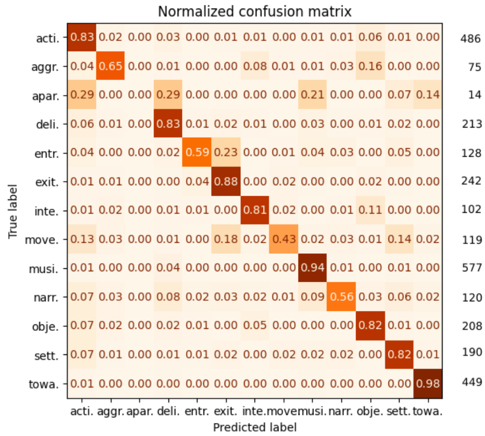
enough for a number of literary studies questions. However, an iterative workflow may be attempted with manual correction of the model outputs and fine-tuning of a better model while keeping the number of manual annotations low. Manual corrections could focus on the worst performing categories. A possible use of the best models would be to automatically annotate FreDraCor stage directions beyond the most frequent 87 single-type labels, which we did not handle here (section 3.1) and account for ca. 30% of the corpus, to assess whether the model's predictions could be a viable way to simplify the corpus' wide variety of labels. The same could be done to corpus examples that bear mixed-type labels.

Confusion matrices for distilled CamemBERT fine-tuned on 100% and 10% of the data are in figure 2,[3] and per-category results in table 5. Cate-

[3]Best F1 of 5 runs; Std Dev 0.012 at 100%, 0.009 at 10%.



(a) `distilcamembert-base`, 100% of data



(b) `distilcamembert-base`, 10% of data

Figure 2: Confusion matrices for distilled CamemBERT. Values are normalized. The right column shows number of examples per class

	FT on 100% of data			FT on 10% of data				
Class	**P**	**R**	**F1**	**P**	**R**	**F1**	**N**	**Diff**
Action	0.907	0.844	0.874	0.821	0.829	0.825	486	-0.049
Aggression	0.713	0.827	0.765	0.662	0.653	0.658	75	-0.107
Aparte	1	0.571	0.727	0.000	0.000	0.000	14	-0.727
Delivery	0.859	0.831	0.845	0.724	0.826	0.772	213	-0.073
Entrance	0.809	0.727	0.765	0.807	0.586	0.679	128	-0.086
Exit	0.833	0.909	0.87	0.762	0.884	0.818	242	-0.052
Interaction	0.791	0.814	0.802	0.741	0.814	0.776	102	-0.026
Movement	0.702	0.672	0.687	0.761	0.429	0.548	119	-0.139
Music	0.969	0.971	0.97	0.931	0.938	0.934	577	-0.036
Narration	0.765	0.842	0.802	0.744	0.558	0.638	120	-0.164
Object	0.826	0.841	0.833	0.725	0.822	0.770	208	-0.063
Setting	0.823	0.858	0.84	0.757	0.821	0.788	190	-0.052
Toward	0.978	0.984	0.981	0.978	0.978	0.978	449	-0.003

Table 5: Per-category precision, recall, macro-F1 with `distilcamembert-base`, fine-tuned on 100% and 10% of the data. Column *N* is the number of test items per class, and *Diff* is the 10% macro-F1 minus the 100% one.

gories *Movement* and *Entrance* are regularly misclassified as *Exit*; a challenge here is that French verb *rentrer* is a contronym, meaning both to go onstage or offstage. The models have trouble to tease apart *Interaction*, *Aggression*, *Object* and *Movement* from each other; *Agression* and *Movement* are among those most affected by reduced fine-tuning data. Misclassification of various categories towards *Action* also happens, more so as data for fine-tuning decreases or with classical ML.

5 Conclusion and Outlook

The results are encouraging towards automatic large-scale stage-direction classification: 0.7 F1 with a 13-class typology using less than 1,200 manually annotated examples; a more costly set of ca. 5,900 annotations allowed for 0.81 F1. A distilled monolingual model was the best choice, offering satisfactory results and faster fine-tuning than the full model. Besides comparing with large language models and in-context learning, a relevant future task is the detection and classification of *internal* stage directions, implicit from character speech (Galleron, 2018). Reliable multilingual stage direction classification would open the door to large-scale quantitative comparative and diachronic work on an important element of dramatic texts.

Ethics Statement

The study involved the use of GPUs. Given potential carbon footprint, we assessed the necessity of their use. We consider their use justified given that we observed substantially better results at the task with transfer learning than with classical machine learning models. We compared distilled and full models, ascertaining that the distilled ones perform largely equivalently at the task. We thus propose the use of distilled models for the task described, which will mean less data transfer and GPU usage time in fine-tuning.

The study involves a large corpus of theater in French. A bias in this corpus is the underrepresentation of women authors. Relevant future work to counter this bias would be to encode in TEI and publicly release plays by women authors. Resources such as the database by Bourdic (2022) will facilitate the related bibliographic research.

Data and Code Availability

The dataframe derived from the FreDraCor TEI documents is at `https://doi.org/10.34847/nkl.fde37ug3`. Code to run the experiments is at `https://doi.org/10.5281/zenodo.10594104`, both under open licenses.

Acknowledgements

We thank the anonymous reviewers for their insightful comments, which helped improve the paper.

This work was supported by IdEx University of Strasbourg, call Attractivité 2020 (MeThAL project) and Maison Interuniversitaire des Sciences de l'Homme Alsace (MISHA) 2023-24 program (Thealtres project).

References

Melanie Andresen, Benjamin Krautter, Janis Pagel, and Nils Reiter. 2024. Knowledge Distribution in German Drama. *Journal of Open Humanities Data*, 10(1).

Maïwenn Bourdic. 2022. Pièces de théâtre écrites par des femmes, et représentées à Paris entre 1809 et 1906.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gisela Dahms. 1978. *Funktionen der Ascriptionen zum Sprechtext im russischen Drama von 1747 bis 1903. Eine Typologie*. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.

Cyrile Delestre and Abibatou Amar. 2022. DistilCamemBERT : une distillation du modèle français CamemBERT. In *CAp (Conférence sur l'Apprentissage automatique)*, Vannes, France.

Katrin Dennerlein. 2016. Automatic recognition of configurations in plays – training a machine learning algorithm.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Fièvre. 2007. Théâtre classique.

Ioana Galleron. 2018. Quel encodage pour les didascalies internes? In *NACLA2 Corpus et textes de la représentation / Corpora and performance texts*, Avignon, France.

Ioana Galleron. 2021. Pour un balisage sémantique des textes de théâtre : le cas des didascalies. *Sens public*. Publisher: Département des littératures de langue française.

Thierry Gallèpe. 1997. *Didascalies : les mots de la mise en scène*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Michael Issacharoff. 1981. Texte théâtral et didascalecture. *MLN*, 96(4):809–823. Publisher: Johns Hopkins University Press.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Monique Martinez Thomas. 2007. Typologie fonctionnelle du didascale. In Florence Fix et Frédérique Toudoire-Surlapierre, editor, *La didascale au dans le théâtre du XXème siècle : Regarder l'impossible*, Ecritures, pages 35–46. EUD.

Daria Maximova and Frank Fischer. 2019. Using Machine Learning for the Automated Classification of Stage Directions in TEI-Encoded Drama Corpora. In *TEI Conference*, page 123, University of Graz, Austria.

Carsten Milling, Frank Fischer, and Mathias Göbel. 2021. French Drama Corpus (FreDraCor): A TEI P5 Version of Paul Fièvre's "Théâtre Classique" Corpus.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Janis Pagel, Nidhi Sihag, and Nils Reiter. 2021. Predicting Structural Elements in German Drama. In *Proceedings of the Second Conference on Computational Humanities Research (CHR2021)*, online.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Manfred Pfister. 1988. *The theory and analysis of drama*. European studies in English literature. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

TEI Consortium. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.6.0.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. In *Efficient Natural Language and Speech Processing Workshop at NeurIPS*.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

A Stage direction examples and glosses

For each class, we provide examples from FreDraCor (separated with pipes), followed by their respective English glosses.

Class	Scope	FreDracor examples and English glosses
Action	General character action category	Il désigne le garçon de café ǀ Il lit ǀ Elle s'assied He points to the waiter ǀ He reads ǀ She sits down
Aggression	Violent action	Il tire son épée ǀ Il se donne un coup He draws his sword ǀ He strikes himself
Delivery	Delivery manner, e.g. regarding voice or vocal expression of emotion	En riant ǀ À demi-voix Laughing ǀ In a low voice
Entrance	Character enters stage	Ils entrent en scène ǀ Il rentre chez lui They enter the stage ǀ He enters his home
Exit	Character exits	Il sort ǀ Il rentre She exits ǀ He re-enters
Interaction	Non-verbal character interaction	Elle va aussi pour l'embrasser She moves to kiss him ǀ He takes her hand
Movement	Character movement (but not exit/entrance)	Il continue sa marche ǀ Il recule d'un autre côté ǀ Il veut sortir He continues his walk ǀ He retreats to the other side ǀ He wants to exit
Music	Tune names (plays with songs); music description	Air en duo ǀ Musique céleste Duet melody ǀ Celestial music
Narration	Long, "narrative quality", for readers	Cependant VENDE, qui avait été mandée, survient après les acclamations du peuple, elle commande à son Chancelier de déclarer ses intentions à l'Assemblée However, VENDE, who had been summoned, appears after the cheers of the people; she commands her Chancellor to declare her intentions to the Assembly
Object	Describes object or interaction with it	Il lui donne un écu ǀ Elle froisse la lettre He gives her a coin ǀ She crumples the letter
Setting	Stage description or play location	Le théâtre représente un salon ǀ À Sicilie The theater represents a living room ǀ In Sicily
Toward	Indicates the addressee of a speech	À Julie ǀ Au commandeur et au comte Toward Julie ǀ To the commander and the count

B Mapping between FreDraCor classes and our typology

We list the FreDraCor types that we assigned to each class in our typology. We only worked with single-type labels in FreDraCor, leaving aside stage directions annotated with more than one type (see section 3.1).
For the small number of cases where there was an obvious typo in a FreDraCor label (e.g. *title* spelled as *ttitle*, or *decor* also spelled as *décor*), we accepted both as variants of the same label.

Our types	Corresponding FreDraCor types
Entrance	entrance, entrée
Exit	exit, escape
Setting	location, decor, décor [sic]
Narration	narration, meteo, noise
Toward	toward
Aparte	aparte, alone
Delivery	together, call, interrupt, loud, low, laugh, silence, quiet, cry, shout, nervous, ironic, anger, serious, happy, hesitate, enthousiasm, emotion, emphasis, friendly, grimace, feeling, furious, continue, sing, repeat
Interaction	kiss, touch, help, pull, push
Aggression	kill, fight, hit, suicide, threat
Action	action, watch, show, paint, pray, jump, read, kneel, fall, knock, write, drink, search, open, eat, sleep, stand, sit, move, listen, ring
Movement	closer, away, walk, follow, back
Object	costume, throw, tear, get, give, dress, drop, close
Music	music, title, ttitle [sic], bis