Transfer learning and in-context learning for stage direction classification in French

Pablo Ruiz Fabo [(1)] & Alexia Schneider [(2)]
[(1)] Université de Strasbourg, France [(2)] Université de Montréal, Canada

Introduction

Stage directions provide information complementing character speech, e.g. about performance or decoration, and can sometimes be independent from performance, targeting readers (Pfister, 1988). They have been little studied using computational means (Dennerlein, 2016; Maximova & Fischer, 2019; Pagel et al., 2021). However, their automatic annotation on a large scale is useful, as they can give clues about genre stylistics and dramatic structure and technique.

This poster follows up on our previous work on stage direction classification in French via fine-tuning pre-trained language models and prompting large language models (LLM). We expand the range of models, hyperparameters and prompts tested. Both paradigms provide useful results. We undertook a new qualitative analysis of LLM results, showing limits in our reference annotations, and how LLMs can help identify them.

Stage direction typology

The literature proposes several typologies, reviewed by Galleron (2021). Ours (see Schneider, 2024 for details) includes 13 types (Table 1). We mapped 87 types appearing at least 50 times in FreDraCor (Milling et al., 2021, cf. Fièvre, 2007) to our 13 types. Our types are intended to be useful for literary analysis and for testing classification models, as they represent different degrees of challenge for automatic classification; some categories are characterized by a clear vocabulary, some are more ambiguous.

Fine-tuning experiments (transfer learning)

Our earlier work (Schneider & Ruiz, 2024) fine-tuned several models, gradually reducing the number of examples to assess model efficiency. French monolingual models `camembert-base` (Martin et al., 2020) and `distilcamembert-base` (Delestre & Amar, 2022) outperformed multilingual BERT (Devlin et al., 2019). We now test FlauBERT (French monolingual), whose `base`[1] model outperforms CamemBERT's in certain domains (Le et al., 2020). We obtained better results than with CamemBERT, but only when fine-tuning on at least ca. 3,000 examples. We also tested `xlm-roberta-base` (Conneau et al., 2020), a more recent multilingual model, but could not outperform multilingual BERT. Decreasing learning rate or increasing batch size did not improve results. Thus, while it was relevant to fine-tune more recent models for comparison, CamemBERT models remain the best for the task according to our experiments, given their more stable performance when training data is reduced (Table 2 and Figure 1).

Prompting (in-context learning)

In (Ruiz & Schneider, 2024), we tested `gpt4o`, `gpt4o-mini` and `llama-3.1-8B` with zero-shot and few-shot prompting; examples are always in French and we compared results with the rest of the prompt in English or in French. Here we additionally test `mistral-large`

[1] We used both flaubert-base-cased and -uncased

and `mistral-small`.[2] Our original prompts included only one stage direction to classify; the rest of the prompt (category definitions, plus examples in few-shot mode) was repeated for each stage direction, which consumes many tokens. We now tested how many stage directions to classify we can include in a prompt without decreasing result quality. As Table 3 shows, results remained stable when classifying 75 stage directions with a single prompt. We should note however that with many stage directions in the same prompt, it is occasionally impossible to recover the category for a test-item from the model response, as the model may make formatting errors. We made our client and evaluation scripts more robust to handle this, and the huge savings (Table 5) in tokens (>90%) and response time (40-80%) outweigh this issue.

Qualitative analysis

The prompt asked the model to generate an explanation for each classification, and we used this to assist qualitative analysis, noticing several patterns:

- cases counted as LLM errors but that are due to inconsistencies in our reference annotations, e.g. cases of a character interacting with an object sometimes tagged as type *action* but sometimes as type *object* in the reference, whereas the model correctly assigns an *object* category. These cases suggest that LLM output can be used to refine our reference annotations.
- LLM errors that stem from incomplete treatment of some nuances by the model, e.g. stage directions related to non-verbal emotion expression like laughing or crying tagged by the model as a general *action* rather than a *delivery* stage direction. Such cases suggest the limits of prompting.

Our analyses suggest that fine-tuning excels at learning nuanced decision boundaries and will get a better F1 score (Tables 2-4) even with a conceptually flawed typology. Conversely, prompting, with the more limited knowledge about the typology derived from the context given to it, may draw our attention to such flaws and may be used to help improve; bad performance for a category may indicate reference annotation errors and explanations generated by the LLM can help detect error patterns.

Conclusion

The results have practical value regarding stage direction classification via fine-tuning or prompting, and inform next steps regarding an automatic annotation task useful for stylistic and genre analysis.

Once we refine the methods whereby we produce our reference categories, it should be possible to improve classification results and go beyond a classification task. For instance, it could be tested whether LLMs identify characters participating in the event represented by the stage direction (cf. Galleron, 2021). Applying a solid annotation schema would be required before producing reference data to evaluate stage direction classification and related tasks multilingually.

Code and data availability

Publicly available at https://github.com/pruizf/d25-llm-stdir

[2] Versions used were 2407 for `large` and 2409 for `small`.

Table 1. Stage direction typology. Details in Schneider (2024), Schneider & Ruiz (2024).

Category	Definition	Nbr of examples (training corpus)
Action	General character action category. Default class when no other category applies	2,467
Aggression	Violent action	350
Aparté	Aside (character addresses audience or is alone)	72
Delivery	Delivery manner (e.g. *laughs*, *sobs*)	962
Entrance	Character enters stage	646
Exit	Character exits	1,295
Interaction	Non-verbal character interaction	565
Movement	Character movement (but not exit/entrance)	583
Music	Tune names (in plays with songs)	2,863
Narration	Long, "narrative quality", for readers	554
Object	Describes object or interaction with it	1,130
Setting	E.g. *the stage represents a bar*	982
Toward	Indicates the addressee of a speech	2,144

Table 2. Fine-tuning pre-trained language models. Macro-F1 score (average over 5 runs) when fine-tuning with a gradually decreasing number of examples, testing always on 2,923 examples. Classical machine learning methods are shown as baseline. Best model score at each proportion of examples is bolded. Models marked with (+) were added in this study.

	% of available examples used				
model	**100**	**50**	**25**	**10**	**5**
Classical Machine Learning					
Ridge Classifier	0.728	0.705	0.626	0.553	0.496
SGD	0.727	0.694	0.623	0.543	0.491
Transfer Learning					
camembert-base	0.777	0.766	0.747	**0.706**	**0.683**
distilcamembert-base	0.808	0.786	0.742	0.696	0.667
(+) flaubert-base-cased	**0.819**	**0.805**	**0.782**	0.684	0.593
(+) flaubert-base-uncased	0.812	0.767	0.745	0.600	0.461
bert-base-multilingual-cased	0.803	0.738	0.693	0.615	0.510
bert-base-multilingual-uncased	0.81	0.735	0.697	0.611	0.568
(+) xlm-roberta-base	0.754	0.715	0.684	0.611	0.526
number of examples : training	9,352	4,676	2,337	935	467
number of examples : validation	2,338	1,169	585	234	117
testset size			2,923		

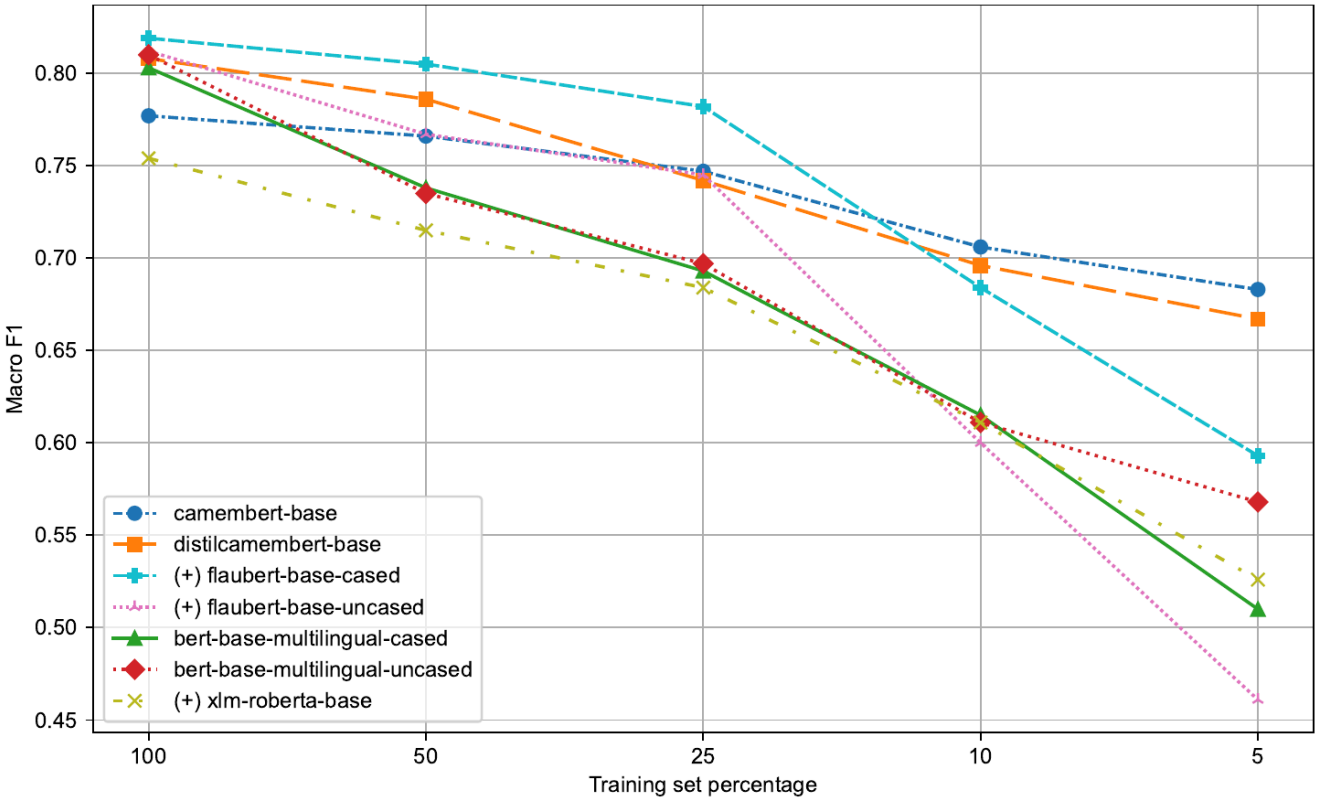

Figure 1. Fine-tuning results as training set is reduced. Models with (+) in the legend were added in this study.

Table 3. Prompting results with a single stage direction to classify in the prompt vs. 75 stage directions, in zero-shot or few-shot mode (20 examples per class). Best results per mode bolded. Examples are always in French; *fr* and *en* are the language in the rest of the prompt.

model	stg dir to classify per prompt	zero-shot		few-shot	
		fr	en	fr	en
gpt-4o-mini	**1**	0.539	0.584	0.619	0.579
	75	0.623	0.623	0.621	0.650
gpt-4o	**1**	0.69	0.709	0.700	**0.734**
	75	0.691	0.696	0.726	0.729
mistral-small	**1**	0.597	0.571	0.604	0.507
	75	0.616	0.61	0.592	0.526
mistral-large	**1**	**0.718**	0.702	0.685	0.684
	75	0.708	0.683	0.707	0.707
examples per prompt (as context)		0		260 (20 per class)	
testset size		2,923		877	

Table 4. Per category results for the best performing models under fine-tuning (transfer learning) and prompting (in-context-learning). Fine-tuning was done with 100% of the available examples (9,352 training + 2,338 validation). N: testset size for each category.

	transfer learning		in-context learning		
category	**distilcamembert-base**	**flaubert-base-cased**	**gpt-4o**	**mistral-large**	**N**
Action	0.87	0.86	0.49	0.68	486
Aggression	0.76	0.76	0.74	0.75	75
Aparte	0.44	0.78	0.9	0.81	14
Delivery	0.85	0.84	0.65	0.62	213
Entrance	0.79	0.82	0.69	0.73	128
Exit	0.86	0.86	0.81	0.81	242
Interaction	0.81	0.82	0.42	0.56	102
Movement	0.69	0.66	0.52	0.56	119
Music	0.97	0.97	0.93	0.92	577
Narrative	0.79	0.71	0.67	0.65	120
Object	0.84	0.83	0.72	0.71	208
Setting	0.85	0.86	0.77	0.69	190
Toward	0.98	0.97	0.9	0.86	449

Table 5. Token count and response times with French prompts, classifying a single stage direction per prompt vs. 75 stage directions per prompt. Token and response time savings in the latter setup provided. Column *lost stg dir* is the number of cases where, given formatting errors, the category for a stage direction cannot be recovered from the model response.

model	**stg dir to classify per prompt**	**zero-shot**					**few-shot**				
		total tokens	**% tokens saved**	**response time (sec)**	**% sec saved**	**lost stg dir**	**total tokens**	**% tokens saved**	**response time (sec)**	**% sec saved**	**lost stg dir**
gpt-4o-mini	1	3,262,985		4,511.95		0	6,983,531		1,340.31		0
gpt-4o-mini	75	217,384	93.34	1,285.97	71.50	5	146,167	97.91	311.51	76.76	0
gpt-4o	1	3,259,042		6,053.86		0	6,985,527		2,413.29		0
gpt-4o	75	217,120	93.34	1,552.17	74.36	0	147,808	97.88	436.84	81.90	0
mistral-small	1	4,278,112		4,502.77		0	8,687,014		1,828.52		0
mistral-small	75	282,212	93.40	2,460.57	45.35	6	190,071	97.81	373.54	79.57	10
mistral-large	1	4,258,542		8,783.03		0	8,667,889		2,078.10		0
mistral-large	75	292,144	93.14	5,203.56	40.75	0	185,284	97.86	736.64	64.55	0
testset size		2,923					877				
examples per prompt (as context)		0					260 (20 per class)				

References

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Delestre, C., & Amar, A. (2022). DistilCamemBERT: Une distillation du modèle français CamemBERT. *CAp (Conférence Sur l'Apprentissage Automatique)*. https://hal.science/hal-03674695

Dennerlein, K. (2016). *Automatic recognition of configurations in plays – Training a Machine Learning Algorithm*. https://comedy.hypotheses.org/category/digitally-assisted-analysis-of-drama

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *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)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fièvre, P. (2007). *Théâtre Classique*. https://www.theatre-classique.fr

Fischer, F., Milling, C., & Göbel, M. (2021). *FreDraCor (French Drama Corpus, TEI P5): A TEI P5 Version of Paul Fièvre's "Théâtre Classique" Corpus.* DraCor. https://github.com/dracor-org/fredracor

Galleron, I. (2021). Pour un balisage sémantique des textes de théâtre: Le cas des didascalies. *Sens public*, 1–23. https://doi.org/10.7202/1089589ar

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). *FlauBERT: Unsupervised Language Model Pre-training for French*. *LREC 2020*. https://doi.org/10.48550/arXiv.1912.05372

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). CamemBERT: A Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Maximova, D., & Fischer, F. (2019). Using Machine Learning for the Automated Classification of Stage Directions in TEI-Encoded Drama Corpora. *TEI Conference*, 123. https://graz-2019.tei-c.org/files/BoATEI2019.pdf#page=123

Pagel, J., Sihag, N., & Reiter, N. (2021). Predicting Structural Elements in German Drama. *Proceedings of the Second Conference on Computational Humanities Research (CHR2021)*.

Pfister, M. (1988). *The Theory and Analysis of Drama*. Cambridge University Press.

Ruiz Fabo, P., & Schneider, A. (2024). Vers une classification automatique de didascalies en français avec des grands modèles de langue. ARIANE Consortium 2024 seminar: L'impact des larges modèles de langue et des agents conversationnels sur les études du texte. Aubervilliers. https://doi.org/10.5281/zenodo.14301375

Schneider, A. (2024). *Analyse computationnelle de textes de théâtre français: Classification automatique du type de didascalies* [MA Thesis, Université de Strasbourg]. https://doi.org/10.34847/nkl.3ecb73zp

Schneider, A., & Ruiz Fabo, P. (2024). Stage Direction Classification in French Theater: Transfer Learning Experiments. In Y. Bizzoni, S. Degaetano-Ortlieb, A. Kazantseva, & S. Szpakowicz (Eds.), *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)* (pp. 278–286). Association for Computational Linguistics. https://aclanthology.org/2024.latechclfl-1.28