Written vs. Spoken French: A Clustering Approach to Grammatical Category Analysis

Ioana-Madalina Silai Sylvain Kahane
MoDyCo, University Paris Nanterre, CNRS
imsilai@parisnanterre.fr, skahane@parisnanterre.fr

MOTS-CLÉS : Universal Dependencies, treebanks, catégories grammaticales, français écrit, français oral, clustering, méthodes non supervisées, annotation syntaxique, variabilité selon la modalité, chevauchement catégoriel.

KEYWORDS: Universal Dependencies, treebanks, part-of-speech categories, written French, spoken French, clustering, unsupervised methods, syntactic annotation, modality variation, category overlap.

Abstract

— This study compares how lexemes cluster in spoken and written French, using Universal Dependencies (UD) treebanks as a starting point. Using clustering algorithms, we group lexemes according to their syntactic and morphosyntactic distributions in dependency structures.

— Modality-specific patterns emerge, suggesting that modality shapes grammatical categorisation and supports the use of this methodology for exploring annotation consistency and linguistic variation.

— More generally, our findings suggest that UD part of speech categories only partially capture distributional behaviour and that unsupervised clustering can serve as a tool for re-examining and refining grammatical categorisation. By exploiting syntactic structure, this approach also demonstrates that meaningful patterns can be detected with relatively small annotated corpora, unlike large-scale distributional models.

1 Introduction

UD treebanks aim to provide consistent syntactic annotation across languages and modalities (de Marneffe *et al.*, 2021). However, written and spoken language differ substantially in lexicon, syntax, and discourse structure (Biber, 2012; Chafe & Tannen, 1987; Halliday, 1985). In this study, we use the term *modality* to refer specifically to the medium of language production : written text versus transcribed speech. This distinction may affect how POS categories are distributed and internally structured. Understanding these differences is relevant both for linguistic theory and for improving the reliability of parsing systems.

This study uses unsupervised learning methods to compare how lexemes cluster in two French UD treebanks, one of them representing written texts, and the other speech transcripts.

The UD POS tags serve as a reference point, but the clustering is driven by the syntactic distributions of lexemes within dependency structures. The goal is not simply to test the coherence of existing categories, but to explore whether alternative groupings emerge from usage patterns. In doing so, the study contributes to discussions on annotation consistency and on how modality shapes syntactic categorisation.

Prior work has addressed similar issues from different perspectives : Poiret & Liu (2020) examine distribution and word order differences across POS in spoken versus written French, while Dobrovoljc (2025) compares syntactic structures across modalities in English and Slovenian. Our approach differs by focusing on lemma-level behaviour, providing both a global view of category organisation and a fine-grained view of which lexical items deviate from canonical patterns.

2 Data

The written data come from the UD Sequoia treebank (70,545 tokens) (Candito & Seddah, 2012). The spoken data come from the UD ParisStories treebank, a corpus of transcribed conversational French containing 42,789 tokens (Kahane *et al.*, 2021). Sequoia and ParisStories were selected because they are closer in size and more homogeneous in genre than larger corpora such as French GSD (Guillaume *et al.*, 2019) or Rhapsodie (Lacheret *et al.*, 2014).

3 Methodology

The basic unit of analysis is a *lexical unit*, defined as a `(lemma, POS)` pair. This allows us to capture a single representation per lemma while distinguishing between homographs. All POS categories are included, with items occurring fewer than ten times excluded to reduce noise. Each lexical unit is represented by the morphosyntactic and contextual features of all its occurrences in the treebank.

Context is defined following Herrera *et al.* (2024) as the node itself, its parent and children, and the immediately preceding and following nodes. From these positions we extract all morphosyntactic features (POS, dependency relations, number, gender, tense, etc.), excluding lemma and surface form to reduce noise. The result is a feature matrix where rows correspond to lexical units and columns to contextual features.

Feature selection reduces sparsity, and clustering is performed using hierarchical agglomerative clustering and k-means. The resulting structures are visualised with t-SNE and PCA. Clusters are evaluated using a combination of internal metrics and qualitative linguistic inspection.

4 Results

While a variety of clusterings can be produced depending on the choice of algorithm and hyperparameters, we focus here on a representative case. Lexical units are represented using a coverage score (inspired by Herrera *et al.* (2024)), which reflects the probability that a word w co-occurs with a feature f. Clustering is performed with hierarchical agglomerative clustering, and the resulting structures are visualised using t-SNE. Because t-SNE is designed to preserve neighbourhoods rather than global distances, the relative proximity of nearby points is meaningful, but the apparent separation of larger groups should be interpreted with caution.

The t-SNE visualisations for both modalities are shown in Figures 1 and 2. Each point represents a lemma, with colours indicating cluster membership assigned by the unsupervised algorithm. Lemmas mentioned in the analysis below are annotated for reference, and certain ones are annotated with their POS category for clarity. The clusters are labelled according to their most frequent POS category.

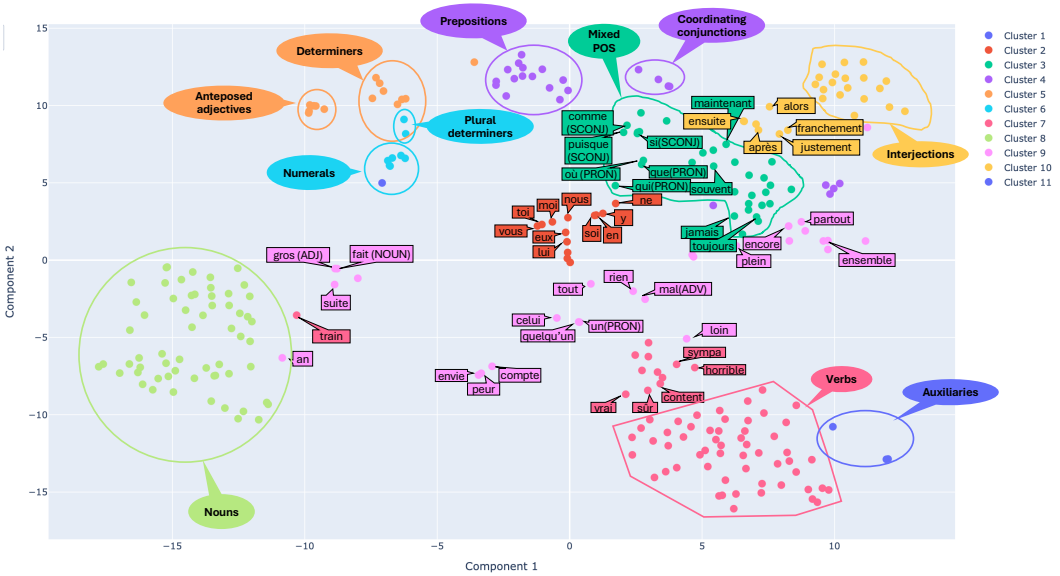

FIGURE 1 – t-SNE visualisation of lexical units from the UD_French-ParisStories treebank (spoken French).

The spoken French treebank produces more clusters, spread further apart in the visualisation space, often containing intermixed lemmas from different UD categories. In the ParisStories treebank, discourse adverbs such as *alors* 'so', *ensuite* 'then', and *franchement* 'honestly' cluster with interjections, showing their pragmatic role in structuring conversation. Adjectives like *vrai* 'true', *content* 'happy' or *horrible* 'horrible' cluster with verbs, reflecting their strong distributional preference for post-copular position with *être* 'to be' in our corpus.

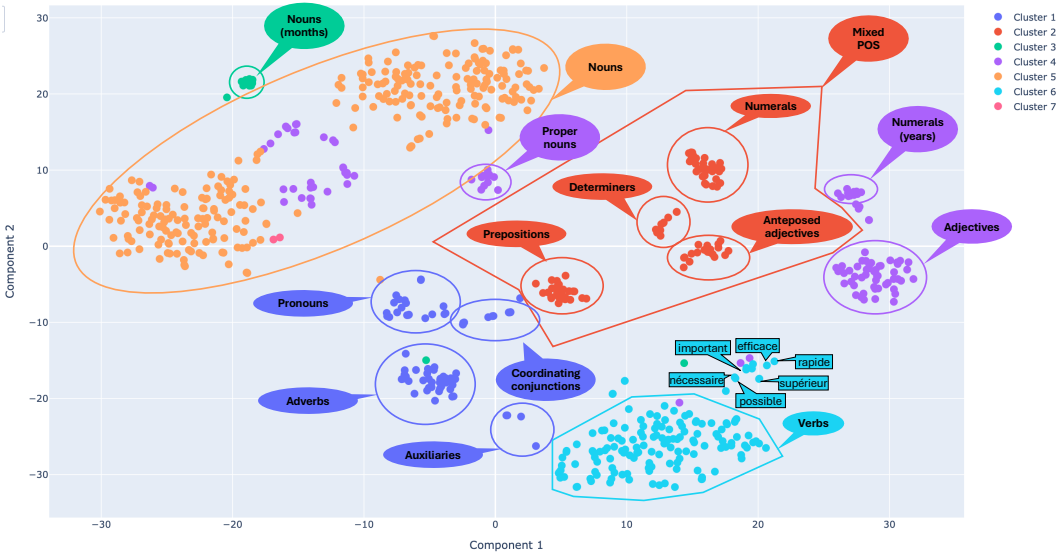

FIGURE 2 – t-SNE visualisation of lexical units from the UD_French-Sequoia treebank (written French).

Written French also shows adjectives clustering with verbs, but the items differ, e.g., *important* 'important', *nécessaire* 'necessary', *rapide* 'fast'. Numerals tend to merge with determiners and anteposed adjectives in both corpora, though annotation inconsistencies in the spoken data (e.g., plural *les* lemmatised separately) occasionally alter this pattern.

Pronouns in spoken French split into two groups, demonstrating the variety of their contexts. One cluster contains personal pronouns (lemmatised as *moi* 'me', *toi* 'you(sg.)', *vous* 'you(pl.)', *nous* 'we') plus clitics (*en*, *y*) and the adverb *ne* (first part of a negative item). The other contains demonstratives/indefinites (*celui* 'the one', *quelqu'un* 'someone', *tout* 'everything', *rien* 'nothing'), which cluster with certain nouns (*peur* "fear", *compte* "account") and adjectives (*gros* 'big') reflecting frequent idiomatic expressions such as *avoir peur/envie de* 'to be afraid of / in the mood to', *se rendre compte de* 'to realise' or *en gros* 'basically'. The same cluster also contains adverbs like *encore* 'still' or *partout* 'everywhere' which appear across a wide range of contexts in the corpus.

Mixed clusters of prepositions, conjunctions, adverbs, and pronouns appear in both corpora but differ in composition. In spoken French, for instance, subordinating conjunctions (*comme* 'as', *puisque* 'since' or *si* 'if') cluster with relative pronouns (*où* 'where', *que* 'that', *qui* 'who') and temporal adverbs (*maintenant* 'now', *jamais* 'never', *toujours* 'always'). In written French, by contrast, a comparable cluster combines most prepositions with determiners, anteposed adjectives, and numerals.

Written French also displays modality-specific clusters : months of the year cluster separately

with the noun *rubrique* 'section', linked by their shared use in numeral-modified contexts (e.g., *septembre 2001*, *voir rubrique 4.3*). Numerals denoting years form their own cluster, closer to adjectives. In spoken French, the noun *train* clusters with verbs, reflecting its role in the fixed expression *en train (de Vinf)* 'in the process of'.

5 Discussion

The results indicate that modality shapes the organisation of lexical units. Spoken French displays greater overlap : items belonging to different UD POS categories cluster together more frequently, and clusters are more dispersed. This likely reflects the prevalence of discourse markers, ellipsis, and multifunctional word uses in conversation. These findings suggest that category boundaries in spoken French are flexible, as distributional behaviour often places lemmas from different UD categories in close proximity, such as adverbs that align with interjections. Written French, by contrast, produces more compact clusters, as items from the same UD POS category tend to group together more closely, even if a single cluster may still contain a mix of UD categories.

This has implications for annotation, raising the question of whether UD guidelines should explicitly account for modality-driven variation. For NLP, the results suggest that parsing spoken data is more challenging, at least partly because models must account for cross-category distributional patterns, making it harder to assign consistent categorical labels.

6 Limits and Future Work

Several limitations remain. First, clustering reveals groupings but not the precise features that drive them, which currently requires manual inspection. Second, the corpora are relatively small, raising the possibility that some patterns are corpus-specific.

A promising extension is to represent tokens as triplets of (lemma, POS, treebank) and compare their clustering across modalities. This would allow direct analysis of whether lexical items behave consistently or diverge between spoken and written French. Another promising direction is to make the clustering process iterative. In the current study, UD POS categories serve as the starting point for grouping lexical units. A second stage could then take the emergent categorisation as input for a new round of clustering. Such an iterative procedure might reveal increasingly refined categories, moving beyond the constraints of UD annotation and providing a clearer view of how distributional behaviour shapes grammatical organisation.

7 Conclusion

This exploratory study has compared lexeme clustering in spoken and written French, identifying both shared and modality-specific patterns. The results demonstrate that the internal structure of grammatical categories can differ markedly across modalities, and

that these differences can be detected through unsupervised methods. The approach is reproducible and could be extended to other languages or to specific categories within a language or treebank.

Références

BIBER D. (2012). *Variation across Speech and Writing*. Cambridge, England : Cambridge University Press.

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus : Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In G. ANTONIADIS, H. BLANCHON & G. SÉRASSET, Éds., *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 321–334, Grenoble, France : ATALA/AFCP.

CHAFE W. & TANNEN D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, **16**, 383–407.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. DOI : 10.1162/coli_a_00402.

DOBROVOLJC K. (2025). Counting trees : A treebank-driven exploration of syntactic variation in speech and writing across languages.

GUILLAUME B., DE MARNEFFE M.-C. & PERRIER G. (2019). Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Revue TAL : traitement automatique des langues*, **60**(2), 71–95. HAL : hal-02267418.

HALLIDAY M. A. K. (1985). *Spoken and Written Language*. Geelong, Victoria : Deakin University Press. (Language and Learning Series). Also : Oxford University Press, London, 1989.

HERRERA S., CORRO C. & KAHANE S. (2024). Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 15114–15125, Torino, Italia : ELRA and ICCL.

KAHANE S., CARON B., STRICKLAND E. & GERDES K. (2021). Annotation guidelines of UD and SUD treebanks for spoken corpora : A proposal. In D. DAKOTA, K. EVANG & S. KÜBLER, Éds., *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, p. 35–47, Sofia, Bulgaria : Association for Computational Linguistics.

LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : a prosodic-syntactic treebank for spoken French. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 295–301, Reykjavik, Iceland : European Language Resources Association (ELRA).

POIRET R. & LIU H. (2020). Some quantitative aspects of written and spoken French based on syntactically annotated corpora. *J. Fr. Lang. Stud.*, **30**(3), 355–380.