Polysemy and Ambiguity: the case of the French modal verb Devoir

Anna Colli, Delphine Battistelli
Laboratoire MoDyCo (Université Paris Nanterre, CNRS)
{acolli, dbattist}@parisnanterre.fr

Abstract

This article focus on a methodology for representing the semantics of polysemous markers whose meanings cannot (or do not have to) be disambiguated, even in context. We name this task (multi-)sense representation and present here the French modal verb *devoir* as a case study. Specifically, we reframe this task — traditionally treated as a multi-class problem — as a multi-label classification problem to account for instances that remain ambiguous due to contextual and intentional factors. In order to fine-tune our model (CamemBERT), we implement an active learning loop to enhance the annotation process and we demonstrate that combining global and local features yields the best results (F1-micro = 0.83; F1-macro = 0.79). The model is then applied on two distinct corpora, showing that the automatic analysis of *devoir*'s modal senses provides deeper insights into modal verb usage and facilitates comparisons across corpora differing in *medium* (spoken vs. written) or genre (e.g. legal discourse). Furthermore, our multi-label approach enables the detection and analysis of double-labeled instances, offering valuable applications, as for example legal discourse interpretation and second language acquisition.

Keywords: modality, polysemy, amibiguity, multi-label classification

1. Introduction

Modal verbs, like other modal cues, give information about the speaker's attitude toward a propositional content (Lyons, 1977; Quirk et al., 1985; Gosselin, 2010). For example, in the English translation of the French sentence (1), the modal verb *must* can signify that Paul is probably at home (in this case, it is an epistemic reading); it can also signify that Paul has the obligation to be at home (in this case, it is a deontic reading). The ambiguity between the two modal readings cannot be resolved without other information. The same ambiguity is already present in the French version, with the modal verb *doit*, and persists in its translation.

(1) Paul doit être à la maison
(en) Paul must be at home

While additional context can sometimes clarify an intended modal sense, corpora often leave such ambiguities unresolved—either because the speaker does not complete the sentence or (more or less) consciously maintains the ambiguity. Let's take the examples (2) to (4), extracted from two spoken corpora [1], which use the French modal verb *devoir*.

(2) je pense qu'elle elle devait se dire que ça craignait rien quoi. (13_11)
(en) I think she must have been telling herself it wasn't risky or anything. (13_11)

(3) pourtant il y avait, il devait y avoir des médias, avant le match, mais, mais ça a été beaucoup moins relayé que le reste, quoi (13_11)
(en) Still, there must have been some media attention before the match, but it didn't get nearly as much coverage as everything else. (13_11)

(4) La recherche de la puissance euh financière doit être un un moyen d'assurer le contrôle des autres et on y est prédestiné. (CFPP)
(en) The pursuit of financial power um must be a way of ensuring control over others and we're destined for it. (CFPP)

In (2), the speaker adds the modal verb *devait* to mark that he is not certain, but there is a high probability that his friend (elle) thought the situation was not scary (epistemic meaning). In (3), the role of *devait* is ambiguous : it could either express the speaker's assumption that there was likely media coverage for the match (epistemic meaning) or that media presence is required by norms in these situations (deontic meaning). Similarly, in (4), the use of *doit* is ambiguous : the speaker might be presenting his declaration as highly probable (epistemic meaning) or as an objective truth based on natural necessity (alethic meaning).

Modal verbs appear very frequently in corpora, whatever the type of corpus (either when regarding its *medium* (spoken vs. written) or its genre (journalistic, encyclopedic, etc.)) are. The fact remains that both their overall frequency and the distribution of their modal senses appear to vary depending on the corpus. Disambiguating their occurrences, including identifying instances where multiple modal senses coexist (ambiguous instances), is then essential when pursuing a diverse type corpus-based analysis of modality. The present study presents a step toward this issue. It presents a model for the French modal verb *devoir* that aims at (multi-)sense representation rather than simple disambiguation,

[1] More details will be give about those corpora in Section 3.1)

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 13–21
12 May 2026.

in both written and spoken corpora. This approach enables the model to disambiguate only when possible and to represent persistent ambiguity when disambiguation is not feasible, whether due to limited context or the possible speaker's/writer's deliberate choice. This article is organized as follows. Section 2 presents different kinds of needs for corpus-based analysis of modality and the state of the art for automatic analysis of modality. Section 3 introduces the corpora we use and the linguistic model employed to describe the modal values of the modal verb *devoir*. Section 4 presents the annotation procedure and the integration of an active learning loop to improve the efficiency and quality of the annotation process. Section 5 presents our BERT-based model and its performance. Finally, in Section 6 we test our model on two corpora (a spoken one and a written one) and we discuss the findings in Section 7.

2. State of the art

Research in various domains examines modal verbs frequency and sense distribution through corpus-based approaches by manually disambiguating modal verbs instances. In second-language acquisition, modal verbs are used to compare learner and native corpora to improve textbook design. In fact, modals are essential for communication but challenging to teach due to their polysemy and lack of direct equivalents in learners' native languages as stated by Li (2024). For example, Bouhlal et al. (2018) compared English spoken corpora Nation (2012) with Québécois learner corpora Martini (2012), finding discrepancies in modal verb use, particularly in the underuse of the deontic *might* and epistemic *should* and *must*. Legal discourse studies likewise analyse modal verbs to highlight their role in shaping legal meaning and authority. For instance, Wu et al. (2025) examined epistemic *may*, must, and related markers in war crimes tribunal trials. Other studies on legal language, for example Jaskot and Wiltos (2017), focus on the translation of modal verbs, given the complexity of their senses and the lack of direct equivalence across languages. For example, the English modals *must*, *may*, *might*, and *should* are all translated into French with the verb *devoir*. As these kinds of studies use manual procedures for disambiguating modal verbs instances, it remains a challenge to do this automatically. Our study aims to bridge this gap by proposing an automatic approach that, on one hand, disambiguates the French verb *devoir* when possible, and on the other, represents ambiguity when it persists.

The earliest studies on modal verbs in NLP relied on rule-based systems or feature-based Support Vector Machines (SVM) aiming to disambiguate modal senses via multi-class classification. The aims of these studies in NLP were varied; for example, enhance modal verbs translation (Baker et al., 2012), analyze their epistemic usage as hedges in the biomedical domain (Light et al., 2004), extract rules from legal regulations (Wyner and Peters, 2011). From a methodological point of view, we can distinguish diverse studies which have followed the evolution of NLP, for example: Ruppenhofer and Rehbein (2012) which proposed an annotation scheme for each modal verb in English, an annotated news domain corpus and logistic regression models based on an ensemble of hand-crafted features, Marasović et al. (2016) which extended this original feature set and applied it to a CNN architecture. Finally, more recent studies which have attempted to solve the problem as a classical modal sense classification task by probing BERT architecture Devlin et al. (2019). As examples for this last case, we can mention Wagner and Zarrieß (2023) which showed that BERT, given the same semantic value, encodes it differently for each modal verb. For this reason, individual classifiers for each verb perform better than a classifier for each modal sense. Finally, Dehouck and Denis (2023) performed classification on BERT's last hidden layer representations of the English modal verbs and their context showing that BERT-based models outperform the frequency baseline and previous models. However, as Owan et al. (2022) noted, this remains a non-trivial task, even for expert annotators. They compare annotations based on two different frameworks and highlight proximity in the interpretation of some labels, even when grounded on distinct theoretical frameworks, and the ambiguity that often leads to multiple acceptable interpretations. Regarding French, little research has focused on the disambiguation of modal verbs using a machine learning or deep learning approach. Nissim et al. (2013) proposed an annotation scheme for French and Italian for modality markers at large and use it to annotate on a spoken corpus. (Colli et al. (2024)) propose a fine-tuned CamemBERT for the modal sense disambiguation on the verb pouvoir. Nissim et al. (2013) and (Colli et al. (2024)) highlight the challenges of disambiguating modal verbs in spoken contexts. Limited contextual cues and intentional speaker ambiguity often make interpretation difficult also for human annotators, increasing task complexity and leading to a scarcity of annotated data. Moreover, although not all instances can be linguistically disambiguated, all methods treat the task as a multi-class classification problem.

In this context, Active Learning provides an effective strategy to overcome data scarcity in annotation. Active Learning is a method for reducing annotation costs and effort by selecting the most informative instances to label for the training pro-

cess (Settles, 2009). Recent studies demonstrate its effectiveness in boosting BERT's performance, especially in low-resource and class-imbalanced settings (Ein-Dor et al., 2020). Active Learning has proven particularly effective, even outperforming larger language models, in domain-specific tasks (Lu et al., 2023), or multi-label classification task, for example, user's intent classification (Zhang and Zhang, 2019) where annotation is expensive and time-consuming.
This study develops a French-specific model that focuses on the (multi-)sense representation of the modal verb *devoir* across written and spoken corpora. By representing modal senses, our approach both disambiguates *devoir* when the context allows only one interpretation and represents ambiguity when it persists. The model's task is framed as a multi-label classification problem to capture cases of contextual or intentional ambiguity, using a BERT-based architecture combined with active learning to enhance annotation efficiency and quality.

3. Corpus and linguistic framework

In this section we present our corpus (3.1) and the linguistic model (3.2) on which the annotation scheme is based.

3.1. Corpus

We train our model on two French corpora, one spoken - named here ES_CP - and one written - named here T2K.

- ES_CP (approximately 250.000 tokens) is composed of 112 semi-structured interviews and short monologues extracted from two different corpora. In the first corpus, named Eslo (Université d'Orléans and CNRS - Laboratoire Ligérien de Linguistique (LLL), 2015), we selected 20 interviews featuring questions to the citizens of Orléans about their habits and feelings regarding their city as well as some conversations. In the second one, named CFPP (CLESTHIA, 2024), we selected 6 interviews containing similar questions but focusing on the city of Paris.

- T2K (approximately 100.000 tokens) is composed of 100 written texts of two different genres : newspaper (60 texts) and encyclopedic (40 texts). The corpus is part of Battistelli et al. (2022).

Lately, we applied our model on two other corpora, 13_11 and legal_CODE in order to show real life applications of our model.

- 13_11 is a corpus of 1750 transcribed interviews (approximately 20 million tokens) about the November 13, 2015 terrorist attacks in Paris and in the Île-de-France region. These interviews are collected as part of the interdisciplinary study Étude 1000 within the Programme 13-Novembre.[2]

- legal_CODE is a corpus of 70 French legal codes (approximately 3 million tokens) extracted from a site that offers an open-source updated, enhanced version of the main French legal codes.[3]

3.2. Linguistic framework

In French, several studies have focused on elucidating the various meanings of the modal verb *devoir*, e.g. (Kronning, 1996; Gosselin, 2010; Barbet, 2012; Vetters and Barbet, 2006). We choose to rely on the model of Gosselin (2010) that retains three possible modal categories for *devoir*. Table 1 presents those three semantic values of our annotation scheme with some examples.

Our task is defined as multi-label annotation. Each instance of the verb *devoir* is assigned one or more labels. Instances annotated with multiple labels reflect either insufficient contextual information for disambiguation or inherent semantic ambiguity, which persists even when the surrounding context is available. For example, in (5) ambiguity persists between deontic ("At that moment, I had to follow my daughters") and epistemic sense ("At that moment, I probably followed my daughters") for *dû*.

> (5) À ce moment là, j'ai dû suivre une des mes filles (CFPP).

4. Corpus annotation

In order to annotate our corpus, we conducted manual annotation in two steps. In the first step (4.1), we performed multi-label annotation on a portion of the corpus. In the second step (4.2), after assessing annotator agreement, we enhanced annotation quality by incorporating an active learning framework into the process.

4.1. Annotation procedure

Firstly, we annotated 207 instances from our training corpus described in Section 3.1. The annotation was carried out by three expert annotators using Glozz (Widlöcher and Mathet, 2012). We then calculated Inter-Annotator Agreement (IAA) using Krippendorff alpha obtaining a score of 0.8.

[2] https://www.memoire13novembre.fr/
[3] https://codes.droit.org

Label	Definition and Examples
deontic	**Definition:** *devoir* with a sense of obligation whose source is a human being, norms or conventions, or external circumstances. **Example:** "Je crois qu'après la troisième euh les enfants doivent passer un un examen obtenir un diplôme" — (en) *I think that after middle school uh children must take an exam to obtain a diploma.* (ESLO_ENT_089)
alethic	**Definition:** *devoir* with a sense of necessity grounded in a law of nature. **Example:** "Pour moi la voiture est un un outil c'est un véhicule rien de plus ça doit être pratique" — (en) *For me a car is a tool it's a vehicle nothing more it has to be practical.* (CFPP)
epistemic	**Definition:** *devoir* with a value of strong probability. **Example:** "Il devait être 22 heures, entre, ouais à peu près, 22h ou 10 heures et quart." — (en) *It must have been around 10 p.m., yeah, roughly 10 or a quarter past ten.* (13_11)

Table 1: Annotation scheme of *devoir* based on Gosselin (2010)'s approach.

4.2. Active Learning loop

After the first manual annotation step, we observed that many instances of *devoir*, particularly in the ES_CP corpus, required extensive reflection by the annotators. In fact, as noted by Owan et al. (2022), the modal verb disambiguation annotation task is challenging even for expert annotators, as it requires more attention than many other linguistic features. To address this, we integrated an active learning loop into the annotation procedure to prioritize the most informative examples, to improve both annotation speed and quality. We chose to implement an uncertainty-based active learning loop as this strategy shows better performances over random sampling for a classification task for BERT models (Jacobs et al., 2021) and LLMs (Lu et al., 2023). Uncertainty-based Active Learning aims to identify the most uncertain instances for subsequent training iterations. Within each iteration, the model operates on a randomly sampled subset of the training data locating instances with the highest entropy in model predictions. We provide a pool of 150 unannotated texts from our corpus. At each iteration, the model (described in Section 5) —initially trained on the dataset annotated in Step 1 (4.1)—is used to infer probabilities over the pool. The *n* instances for which the model is least confident are then selected for annotation. Annotators label these examples and can stop the loop at any point. After each iteration, the model is retrained on the newly annotated batch and is ready for the next iteration. To further validate the effectiveness of this approach over a random sampling baseline, we plan an additional experiment in which annotators will label a comparable set of instances selected at random, allowing us to directly compare model performance improvements across the two sampling strategies. During this active learning step we annotate 214 devoir instances reaching a total of 418 instances with the following distribution (see Figure 1).

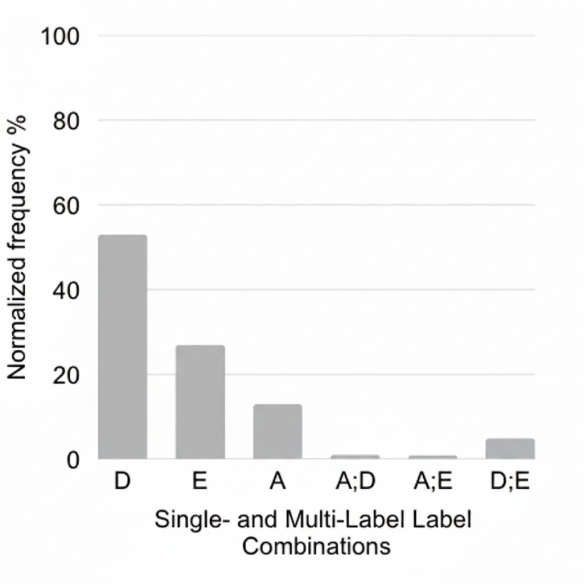

Figure 1: Distribution of annotated instances.

5. Experiments and results

In our experiments, the primary objective was to provide a multi-label classification of *devoir* in order to take into account instances that cannot be disambiguated due to lack of context or intentional ambiguity by the speaker.

5.1. Model architecture

Before fine-tuning the model, the dataset was segmented into individual sentences, and instances of *devoir* were automatically identified using Stanza syntactic analyzer (Qi et al., 2020). The positions of these instances within each sentence were recorded and later used to extract their corresponding embeddings for classification. In sentences containing multiple *devoir* instances, the positions of all instances were considered separately, as each occurrence may convey a distinct modal meaning.

This phenomenon is particularly common in the spoken corpus, where sentences tend to be longer. In example (6), the same sentence contains an epistemic devoir (*dû*) and a double-labeled epistemic-deontic one (*devais*).

> (6)Avec le mode de garde de cette époque-là je devais l'avoir sûrement en milieu de semaine à ces moments-là donc j'ai dû passer du temps avec ma fille et aussi passer à autre chose (ESLO_ENT)
> (en) *With the custody arrangement at that time, I must have had her probably in the middle of the week back then so I must have spent time with my daughter and also moved on to other things.* (ESLO_ENT)

We fine-tuned the CamemBERT model [4] (Martin et al., 2019) for multi-label classification performing 5-fold cross-validation using 80% of the data for training, 10% for validation and 10% for testing. The model was fine-tuned using a learning rate of 2e-5 and a batch size of 8. Training was conducted for a maximum of 10 epochs, with early stopping. To address class imbalance and well-calibrate the model, we incorporated class-specific weights in the BCEWithLogitLoss function. We experimented two embeddings architectures :

- *local*: in this setup, each sentence was encoded using CamemBERT, and the last hidden layer representation (12th layer) corresponding to the target *devoir* token was extracted. This 768-dimensional token embedding was then fed directly into a classification head.
- *global_local*: in this setup, after encoding each sentence, we extracted the embedding of the target *devoir* token and the sentence embedding [CLS]. These embeddings are then concatenated and passed through the classifier.

local embeddings follow the standard contextual embedding architecture. They are designed to capture local information about the target token (*devoir*), while inherently encoding some contextual information. In contrast, *global_local* embeddings enrich the local embeddings by incorporating global sentence-level information via the [CLS] token. As demonstrated by Miaschi and Dell'Orletta (2020), although many sentence-level properties are implicitly encoded within individual word embeddings, sentence representations are generally more effective at capturing syntactic and structural features, whereas token-level embeddings better encode raw text and morphosyntactic properties. By combining both, the *global_local* leverages the complementary strengths of local and sentence-level representations, improving the model's ability to capture both token-specific meaning and sentence context. Our approach was inspired by Zhang et al. (2022), who demonstrated that a BERT-based model that combines the [CLS] embedding (representing global features) with selected token embeddings (representing local features) outperforms a model that uses only the [CLS] token for multi-label text classification.

[4] https://huggingface.co/almanach/camembert-base

Model	Micro F1	Macro F1	D	E	A
majority baseline	0.56	0.24			
camemBERT	0.66	0.60	0.76	0.65	0.39
local before AL	0.74	0.73	0.75	0.80	0.64
global_local before AL	0.78	0.76	0.80	0.81	0.67
***global_local* after AL**	**0.83**	**0.79**	**0.83**	**0.86**	**0.67**

Table 2: Results for automatic multi-label classification

5.2. Model performances

Results are presented in Table 2, comparing *local* and *global_local* models against two baselines: a majority baseline based on the most frequent labels combination and CamemBERT in its original setup. Overall, the *global_local* configuration yields the best results across all evaluation metrics. For this reason, we selected the *global_local* configuration and used it to implement the active learning loop described in Section 4.2. Table 2 reports the model's performance before (*global_local* BEFORE AL) and after (*global_local* AFTER AL) the active learning step.

6. Application on 13_11 and legal_CODE

We applied our model to the 13_11 and legal_CODE corpora (described in Section 3.1) to automatically classify instances of *devoir*. We expected different distributions due to differences in the *medium* (spoken versus written) of these corpora and their genre (testimonies related to terrorist attacks in 13_11 versus legal rules in legal_CODE). Specifically, we expected a higher prevalence of epistemic *devoir*, which marks uncertainty, in 13_11, and a higher prevalence of deontic *devoir*, which marks obligation, in legal_CODE. In 13_11, we observe that half of instances of *devoir* express an epistemic meaning. It shows a higher frequency compared to our annotated corpus (see Section 4.2) where epistemic instances were 27% and to the legal_CODE corpus where the number of epistemic values is minimal (0.02%). However, frequency of the double label epistemic-deontic (5%) remains stable between the annotated corpus and 13_11. On the other hand, in the legal_CODE corpus, we observe that almost all instances of *devoir* (85%) convey a deontic meaning, followed by the alethic–deontic double label, which indicates

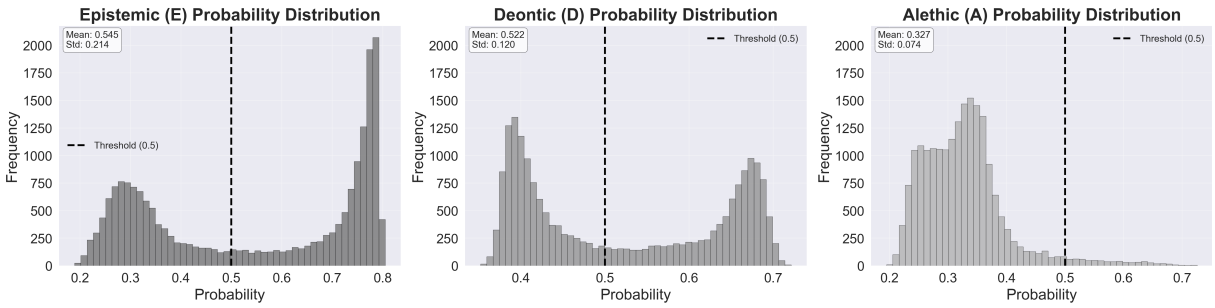

Figure 2: Probabilities distributions for epistemic (E), deontic (D), and alethic (A) labels in 13_11

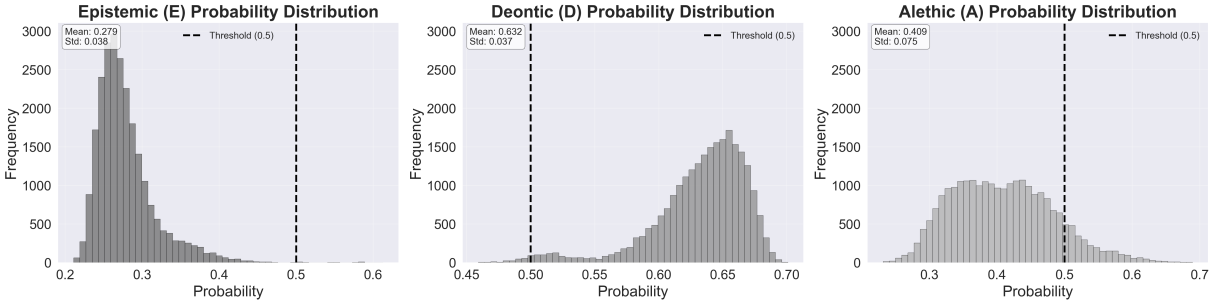

Figure 3: Probabilities distributions for epistemic (E), deontic (D), and alethic (A) labels in legal_CODE

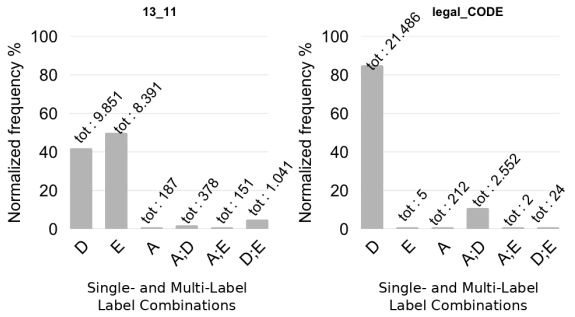

Figure 4: Distribution of (automatically) annotated instances in 13_11 (left) and legal_CODE (right).

ambiguity between the two senses. Comparing this distribution with that of our manually annotated corpus, we find that - as expected due to the legal nature of these texts - deontic *devoir* is much more frequent in legal_CODE, dominating the other categories. Ambiguity between deontic and alethic persists, with a higher frequency in the legal_CODE corpus (11%) compared to the manually annotated corpus (1,1%). This highlights the proximity between the two labels, less evident in the manual annotated corpus, and the difficulty of distinguishing between them.

We also examined, for each corpus, the extent to which the model's label assignments were decisive. Figures 2 and 3 show the distribution of predicted probabilities for the E, D, and A labels in 13_11 (figure 2) and legal_CODE (figure 3). In 13_11, the model appears highly decisive for epistemic classifications, often assigning either very high or very low probabilities, with relatively few instances receiving intermediate values. In contrast, for deontic classifications, the model is less decisive, as a larger number of instances fall within the middle probability range. In the legal_CODE corpus, the model is decisive for both deontic and epistemic classifications— assigning very high probabilities to deontic instances and very low probabilities to epistemic ones—but less decisive for the alethic label, where a greater proportion of instances receive intermediate probabilities.

7. Discussion

In this article, we fine-tuned a BERT-based model to represent the modal (multi-)senses of *devoir* instances. This methodology corresponds to a modal sense disambiguation or, when disambiguation is not possible, to an ambiguity representation. In first place, we performed manual annotation on a dataset of written and spoken French. After obtaining a stable IAA (0.8) we integrated an active learning loop based on uncertainty sampling in our annotation procedure to improve annotation speed and quality. Concerning the BERT-based model, we tested two embedding architectures for classification: a single token embedding of *devoir* (*local*) and a concatenation of the sentence-level [CLS] embedding with the *devoir* token embedding (*global_local*). The *global_local* architecture

yielded better performance reaching an F1-micro of 0.83 and F1-macro of 0.79. We did not attempt classification relying solely on the [CLS] embedding, because some sentences contain multiple *devoir* instances with different modal values (e.g., Example 6) that needed to be treated individually. Finally, we applied our model to 13_11 corpus and legal_CODE corpus. The results of the automatic (multi)-sense representation of *devoir* demonstrate how modal sense classification can capture corpus-specific characteristics by contributing to the definition of each corpus's modal profile. As introduced in (*anonymous*), a modal profile characterizes a corpus according to how modality is expressed, including which modal markers are used and the frequency of different modality categories. On the one hand, 13_11 is characterized by a high frequency of epistemic *devoir*, a marker of uncertainty; on the other hand, legal_CODE is dominated by deontic *devoir* and marked by a persistent ambiguity between deontic and alethic uses. Automatic (multi-)sense representation of all instances also allows us to perform some linguistic analysis in order to find, for example, correlations between modal senses and moods. In 13_11, 99% of epistemic *devoir* instances are indicative and only 0,2% are in a conditional mood. On the contrary, conditional mood covers 18% of deontic meaning. This result shows that, although the conditional mood is usually considered a marker of uncertainty, in this context it appears to be linked to deontic *devoir*, modulating obligation to express a suggestion (7) or what is considered appropriate by the speaker (8).

(7) Mais vous ne devriez pas y aller (13_11)
(en) *But you shouldn't go there.* (13_11)

(8) On devrait pas, on devrait pas mourir parce que quelqu'un a décidé qu'on devrait mourir. (13_11)
(en) *We shouldn't — we shouldn't die because someone decided that we must die.* (13_11)

On the contrary, in the legal_CODE corpus, deontic *devoir* appears predominantly in the indicative mood (94%), with only 1% of instances in the conditional mood. This is expected, as normative texts rarely express modulated obligations. These results suggest that the relationship between a specific modal value (deontic) and a particular mood (conditional) is corpus-specific.

8. Conclusion

This study represents a first step in the automatic classification of the French verb *devoir* in both spoken and written corporain order to compare how *medium* or genre are related or not to specific semantic values of this modal verb. Traditionally, the task has been framed as a multi-class classification problem, but we decided to reframe it as a multi-label classification problem. This allows us to account for instances in which *devoir* cannot be disambiguated due to insufficient context or intentional ambiguity. Furthermore, we show that a combination of global and local features—achieved by concatenating the *devoir* token embedding with the [CLS] sentence embedding—yields better performance than classification relying solely on the *devoir* embedding. In future work, we will explore additional representation strategies, particularly classification based on the [CLS] embedding by introducing special tokens around the targeted *devoir* instance. In addition, to better identify the appropriate context for *devoir* instances and to reduce the number of multiple occurrences within the same sentence—common in spoken corpora—we will experiment with segmenting the data into discourse units rather than relying on punctuation-based sentence boundaries, as proposed in (Prevot and Muller, 2025). However, we argue that (multi-)sense representations in spoken corpora may not benefit from a shift from sentence-level to discourse-level analysis, due to the rapid nature of conversation. The application of our model to two corpora, 13_11 and legal_CODE, which differ in *medium* (spoken vs. written) and genre (testimonies about a terrorist attack vs. legal code), demonstrates how correctly representing the modal senses of *devoir* instances can help enrich the modal profile of a corpus, thereby enabling comparisons with other corpora that differ in *medium* or genre. This model will be useful for studies in second language acquisition and legal discourse that focus on the use of modal verbs and and currently depends on manual annotation for their analysis. In addition, as in the NLP domain, research in these two fields generally frames the task as a multi-class classification problem, where each *devoir* instance is assigned to only one modal category. Our multi-label classification approach provides richer insights into how modal verbs are used in corpora. Moreover, in the context of learner corpora, our model can help identify and focus on double-labeled instances—that is, cases where ambiguity persists—in order to reformulate their expression for pedagogical purposes. Reframing the disambiguation task as a (multi-)sense representation problem may be more appropriate for cases traditionally treated as disambiguation, but where ambiguity can still persist. This perspective, which is closer to linguistic theory, could be relevant for general sense-labeling tasks.

9. Limitations

There are certain limitations to our work. First, for active learning, we only tested one sampling

method. In future work we aim to explore others methods and compare their performances with the current approach and random sampling to further validate the effectiveness of the active learning approach. Second, we aim to enhance data diversity and include additional written texts in order to balance the distribution of *devoir* instances between written and spoken corpora. Finally, we would like to experiment other methods to concatenate embeddings to improve our model's performances.

10. Bibliographical References

K. Baker, M. Bloodgood, B. J. Dorr, C. Callison-Burch, N. W. Filardo, C. Piatko, L. Levin, and S. Miller. 2012. Modality and negation in SIMT: Use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics*, 38(2):411–438.

C. Barbet. 2012. Devoir et pouvoir, des marqueurs modaux ou évidentiels? *Langue française*, 173(1):49–63.

D. Battistelli, A. Etienne, R. Rahman, C. Teissèdre, and G. Lecorvé. 2022. Une chaîne de traitement pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique (a processing chain to explain the complexity of texts for children from a linguistic and psycho-linguistic point of view). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 236–246, Avignon, France. ATALA.

F. Bouhlal, M. Horst, and J. Martini. 2018. Modality in ESL textbooks: Insights from a contrastive corpus-based analysis. *The Canadian Modern Language Review*, 74(2):227–252.

A. Colli, D. Rossini, and D. Battistelli. 2024. A modal sense classifier for the French modal verb pouvoir. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 233–243, Pisa, Italy. CEUR Workshop Proceedings.

M. Dehouck and P. Denis. 2023. Revisiting modal sense classification with contextual word embeddings. In *Models of Modals: From Pragmatics and Corpus Linguistics to Machine Learning*, chapter 8, pages 225–253. De Gruyter Mouton, Berlin, Boston.

J. Devlin, M. W. Chang, K. Lee, and K. 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.

L. Ein-Dor, A. Halfon, A.n Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Laurent Gosselin. 2010. *Les modalités en français: La validation des représentations*. Rodopi.

P. F. Jacobs, G. Maillette de Buy Wenniger, M. Wiering, and L. Schomaker. 2021. Active learning for reducing labeling effort in text classification tasks. In *Benelux Conference on Artificial Intelligence*, pages 3–29. Springer International Publishing, Cham.

M. P. Jaskot and A. Wiltos. 2017. An approach to the translation of deontic modality in legal texts: The case of the polish and english versions of the "charter of fundamental rights of the european union". *Cognitive Studies| Études cognitives*, (17).

Hans Kronning. 1996. *Modalité, cognition et polysémie : sémantique du verbe modal 'devoir'*.

L. X. Li. 2024. Developing strategies to improve textbook design using synergy of native and learner corpora. *Journal of Psycholinguistic Research*, 53(6):76.

M. Light, X. Y. Qiu, and P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pages 17–24.

Y. Lu, B. Yao, S. Zhang, Y. Wang, Pe. Zhang, T. Lu, T. J.-J. Li, and D. Wang. 2023. Human still wins over LLM: An empirical study of active learning on domain-specific annotation tasks.

John Lyons. 1977. *Semantics: Volume 1*. Cambridge University Press.

A. Marasović, M. Zhou, A. Palmer, and A. Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *Linguistic Issues in Language Technology*, 14.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, E. V. de La Clergerie, B. Sagot, et al.

2019. Camembert: A tasty french language model.

Juliane Oliveira Martini. 2012. High frequency vocabulary in a secondary quebec esl textbook corpus. Unpublished thesis.

A. Miaschi and F. Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

M. Nissim, P. Pietrandrea, A. Sansò, and C. Mauri. 2013. Cross-linguistic annotation of modality: A data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14. Association for Computational Linguistics.

R. Owan, M. Gini, and D. Kang. 2022. Quirk or palmer: A comparative study of modal verb frameworks with annotated datasets. *arXiv preprint arXiv:2212.10152*.

L. Prevot and P. Muller. 2025. A few shades of supervision for discourse segmentation: Experiments on a french conversational corpus. *Dialogue & Discourse*, 16(2):35–73.

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*. Association for Computational Linguistics.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

J. Ruppenhofer and I. Rehbein. 2012. Yes we can!? annotating the senses of english modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1538–1545.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

C. Vetters and C. Barbet. 2006. Les emplois temporels des verbes modaux en français: Le cas de devoir. *Cahiers de Praxématique*, 47:187–210.

J. Wagner and S. Zarrieß. 2023. Probing BERT's ability to encode sentence modality and modal verb sense across varieties of English. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 28–38, Nancy, France. Association for Computational Linguistics.

A. Widlöcher and Y. Mathet. 2012. The glozz platform: a corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, page 171–180, New York, NY, USA. Association for Computing Machinery.

S. Wu, M. Amini, and O. H. A. Mahfoodh. 2025. Unveiling certainty and doubt: A systemic functional exploration of epistemic modality in courtroom discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*.

A. Wyner and W. Peters. 2011. On rule extraction from regulations. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press.

L. Zhang and L. Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 107–111. Association for Computing Machinery (ACM).

R. Zhang, Y.-S. Wang, Y. Yang, T. Vu, and L. Lei. 2022. Exploiting local and global features in transformer-based extreme multi-label text classification.

11. Language Resource References

CLESTHIA. 2024. Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Nation, Paul. 2012. *The BNC/COCA word family lists v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, ISLRN 532-206-426-067-2. PID https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/paul-nations-publications/publications/documents/Information-on-the-BNC_COCA-word-family-lists.pdf.

Université d'Orléans and CNRS - Laboratoire Ligérien de Linguistique (LLL). 2015. *Corpus ESLO: Enquête Sociolinguistique à Orléans*. LLL / CNRS / Université d'Orléans. ORTOLANG (Open Resources and TOols for LANGuage), ESLO resources, 2.0.