A modal sense classifier for the French modal verb pouvoir

Anna Colli, Diego Rossini, Delphine Battistelli

A modal sense classifier for the French modal verb pouvoir

Anna Colli[1], Diego Rossini[2] and Delphine Battistelli[1]

[1]Modyco laboratory, Paris Nanterre University, 200 Av. de la République, 92000 Nanterre, France
[2]Paris Nanterre University, 200 Av. de la République, 92000 Nanterre, France

Abstract

In this paper we address the problem of modal sense classification for the French modal verb *pouvoir* in a transcribed spoken corpus. To the best of our knowledge, no studies have focused on this task in French. We fine-tuned various BERT-based models for French in order to determine which one performed best. It was found that the Flaubert-base-cased model was the most effective (F1-score of 0.94) and that the most frequent categories in our corpus were material possibility and ability, which are both part of the more global alethic category.

Keywords

pouvoir, modal verbs, Modal Sense Classification, BERT, modality, French

1. Introduction

In this paper, we present our research into the automatic disambiguation of the French modal verb *pouvoir* (in English, this verb can be translated by can, could, may or might) in a corpus of semi-structured interviews[1]. This problem statement is part of a broader quantitative and qualitative analysis currently underway on modal markers in order to better understand which kinds of modal categories are prevalent in this kind of corpus. As an NLP task, the problem of the automatic disambiguation of modal markers relies on what is generally called "modal sense classification" (MSC). As far as we know, no studies have focused on disambiguating modal verbs using a machine learning approach in French. Our aim is to fill this gap by finding the best fine-tuned BERT model to classify the semantic values of the French modal verb *pouvoir* in a transcribed spoken corpus. The article is organized as follows. In section 2 we review related work on the task of modal sense classification. Section 3 describes our corpus and our linguistic model. Section 4 presents the annotation of the corpus with an annotation scheme. Section 5 presents our experiments in fine-tuning different BERT models in order to choose the most effective one. Finally, in section 6 we discuss our results and in section 7 we close our contribution with conclusions and suggestions for future research.

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

anna.colli@parisnanterre.fr (A. Colli); 42013189@gmail.com (D. Rossini); dbattist@parisnanterre.fr (D. Battistelli)

[1]The code and the annotated corpus is available on GitHub https://github.com/DiegoRossini/Modal-verbs-modality-detector. The model is available at https://huggingface.co/DiegoRossini/flaubert-pouvoir-modality-detector

2. Related work

The first study to focus exclusively on modal sense classification was [1], who proposed logistic regression models for each modal verb in English, based on an ensemble of hand-crafted syntactic and lexical features. It was also the first study to present an annotation scheme and an annotated news domain corpus. Further studies pointed out the problem of the biased distribution and sparsity of data used in [1]. For example, two of these studies, [2] and [3], suggested creating a larger and balanced dataset using a paraphrase projection approach from German data (English-German parallel corpus of film subtitles and proceedings from the EU Parliament). More specifically, [2] updated the original feature set with semantic features. [3] also updated the original features of [1] with lexical and discourse features to improve the performances of the classifiers; in addition, they explored the influence of genre on the classification of modal verbs. Lastly, [4] proposed the most accurate and flexible alternative to classifiers based on manually engineered features. Their model is based on a CNN architecture and is able to automatically extract features that are relevant for classification (word embeddings). By adapting the model to German, they demonstrated the model's ability to generalize across different languages. [5] introduced another model architecture in which a simple classifier is fed with a combination of three sets of hand-crafted features and a concatenation of pre-trained embeddings of context words. This representation of the modal context was obtained by testing various weighting schemes. More recent studies have attempted to solve the problem as a classical modal sense classification task by probing BERT architecture [6]. BERT-based models do not need a hand-crafted feature set and they are claimed to be better at capturing contextual information than previous models. [7] showed that BERT does not have a unique representation for each modal sense, but, given the same semantic

value, BERT 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, [8] used BERT's last hidden layer representations of the English modal verbs and their context to feed a k-nn and logistic regression model. In addition, they tried to train a single common model for all the modal verbs but they showed that for some of them, including can and could, this does not improve the results. [8] used the [1] and [2] datasets and also introduced a new and richer dataset from COCA[2], characterized by 5 genres including the spoken genre.In general, BERT-based models outperform the frequency baseline and previous models for almost all modal verbs. Regarding French, as far as we know, no research has yet focused on the disambiguation of modal verbs using a machine learning approach. The only NLP approach is [9] which studied the notion of "possible" and adopted a symbolic approach with a set of rules to semantically annotate epistemic possibility. The present paper aims to fill this void by using a BERT architecture to solve the MSC task in a transcribed spoken French corpus. We present here the work carried out for the disambiguation of the modal verb *pouvoir*.

3. Corpus and linguistic model

This section presents our corpus (3.1) and the linguistic model (3.2) on which the annotation scheme is based.

3.1. The ES_CF corpus

Our corpus – named here corpus ES_CF – is composed of 221 semi-structured interviews extracted from two different corpora[3]. In the first corpus, named Eslo[4], we selected 207 interviews featuring questions to the citizens of Orléans about their habits and feelings regarding their city. In the second one, named CFPP[5], we selected 14 interviews containing similar questions but focusing on the city of Paris. An automatic tool, named ModalE, described in ([10]; [11]), was employed to count the different modal categories that are present in these two corpora. The tool is built on the typology proposed by [12]. Each French modal marker is associated with one or more modal categories depending on its more or less polysemous nature. The results indicate that the verb *pouvoir* is among the four most frequent modal markers[6] in the ES_CF corpus which contains globally 150.000 modal markers. The marker *pouvoir* is a "highly polysemous" marker as it can potentially be part of three categories: alethic, epistemic and deontic (see section 3.2 for their examination in detail). In order to determine the semantic value of each instance of polysemic modal markers, we propose a NLP approach for disambiguating the modal verb *pouvoir* in its context. Our approach is based on the linguistic model of [12].

3.2. Linguistic model for analysing semantic values of *pouvoir*

In French, several studies have focused on elucidating the various contextual meanings of the modal verb *pouvoir*, e.g. ([13]; [14]; [12]). In order to build our annotation scheme (see section 4.1), we rely on the analysis presented in [12]. This is the model that was used in the ModalE tool used for extracting modal markers [10]. As mentioned in section 3.1, this tool assigns 3 possible global modal categories to *pouvoir*: alethic, epistemic and deontic. A deeper analysis of pouvoir, based on [12], led us to consider that this modal verb can have 6 possible refined modal categories (see table 6): 4 belong to the alethic category (descriptive judgements on a reality independent of the subject), 1 is part of the epistemic category (descriptive judgements referring to a subjective evaluation of the reality by the subject) and 1 belongs to the deontic one (prescriptive judgements based on institutions or systems of conventions). In [12], the values of "possibilité matérielle" (material possibility) and "capacité" (ability) are first [12, p. 442] presented as two distinct values, and later [12, p. 448] as part of a single one. Since this ambiguity is not resolved in Gosselin's typology, we decided to treat them as two distinct values.

4. Corpus annotation

In order to follow a supervised learning procedure, it is necessary to have a manually annotated corpus. We describe here the process of manual annotation (4.1) and the constitution of 4 different versions of our annotated corpus (4.2) that we used for the experiments detailed in section 5.

4.1. Annotation procedure

Table 2 presents the elements of our annotation scheme based on [12]'s typology summarized in table 6 (for a fuller version with examples and definitions, see A). Table 2 shows the 7 possible modal categories of *pouvoir*

[2] https://www.english-corpora.org/coca/

[3] Among the different types of interviews and recordings which are present in these two corpora, we have extracted only the semi-structured interviews between an interviewer and an interviewee

[4] https://www.ortolang.fr/market/corpora/eslo (700 recordings in total).

[5] https://www.ortolang.fr/market/corpora/cfpp2000 (60 recordings in total).

[6] the others: "bien" (well) (7.3% of the total modal markers), "dire" (to say) (6.9%), "savoir" (to know) (5.6%), "pouvoir" (4.94%).

Table 1
Gosselin [12] categories for *pouvoir*

global modal categories	modal categories	examples
aléthique (*alethic*)	sporadicité (*sporadicity*)	Les alsaciens **peuvent** être obèses. (*Alsaciens may be obese.*) [12, p. 442]
	possibilité matérielle (*material possibility*)	D'ici on **peut** voir la mer. (*From here, one can see the sea.*) [12, p. 442]
	capacité (*ability*)	Maintenant qu'il est déplâtré, il **peut** marcher. (*Now that his cast has been removed, he can walk.*) [12, p. 442]
	possibilité logique (*logical possibility*)	Un triangle isocèle **peut** avoir un angle droit. (*An isosceles triangle can have a right angle.*) [12, p. 448]
épistémique (*epistemic*)	éventualité (*eventuality*)	Il **peut** faire beau cet après midi. (*The weather could be nice this afternoon.*) [12, p. 442]
déontique (*deontic*)	permission (*permission*)	Vous **pouvez** sortir. (*You can go out.*) [12, p. 442]

(the logical possibility category is included in the annotation scheme even though we did not find any examples in our corpus). We have also added an "undetermined" category, which includes the occurrences of *pouvoir* for which an annotator hesitates between two or more values and the ones that we were unable to annotate due to a lack of context. We annotated 24 interviews from the ES_CF (17 from the Eslo corpus and 7 from the CFPP corpus) with an average length of 15,000 tokens. The annotation was carried out by three annotators (first author and two linguistic masters students) using Glozz [15]. We then calculated two inter-annotator agreements using Fleiss' Kappa. The first one is called "strict" and includes the 6 values (excluding logical possibility). For the second one, denominated "broad", we decided to merge "ability" and "physical possibility" into a single category called "physical possibility and ability" because of the ambiguity that persists in Gosselin [12]'s typology (see section 3.2), confirmed also by the frequent disagreement between annotators on these two categories. We obtained a result of 0.6 for the strict inter-annotator agreement and 0.66 for the broad inter-annotator agreement. Since the result of the broad inter-annotator agreement was better, we decided to adopt this version of the annotated corpus for training. The model was trained on all the categories except for logical possibility and the "undetermined" category. The total number of occurrences of *pouvoir* manually annotated in the corpus is 879[7].[8]

4.2. Corpus preparation

In order to effectively train and evaluate our classifier for detecting the semantic value of the French verb *pouvoir*, we prepared 4 distinct datasets, each crafted to address specific challenges and enhance performance (see examples in C).

Table 2
The 7 categories of *pouvoir* in the annotation scheme

global modal categories	modal categories
alethic	sporadicity
	material possibility
	ability
	logic possibility
epistemic	eventuality
deontic	permission
undetermined	undetermined

- **Corpus Base**: this dataset contains 776 sentences with at least one occurrence of *pouvoir*. Serving as our foundational dataset, it suffers from an imbalance in the distribution of modality categories. This imbalance could bias the classifier toward more common categories, making it essential to address this issue in subsequent datasets.
- **Corpus Base Augmented**: to rectify the imbalance observed in the "corpus base", we created this augmented dataset containing 1716 sentences. We employed data augmentation using the cc.fr.300.bin model and the gensim library for lexical substitution. This process balanced the distribution of modality categories, resulting in a more evenly distributed training set for our classifier.
- **Corpus Context**: considering the significant influence of surrounding context on the meaning of the modal verb *pouvoir* we constructed a third dataset (776 sentences with context). This dataset includes sentences with *pouvoir* along

[7] sporadicity (71 occurrences), material possibility or ability (448), eventuality (131), permission (229)
[8] The annotated corpus is available on GitHub: https://github.com/DiegoRossini/Modal-verbs-modality-detector

with one speaker's phrase before and after, offering a broader contextual framework to help the classifier better understand the modal sense of *pouvoir* and make more accurate predictions (see .

- **Corpus Context Augmented**: this fourth and final dataset combines the benefits of both data augmentation and expanded contextual framing (1716 sentences with context).

5. Experiments and results

In our experiments, the primary objective was to identify the most effective configurations regarding training data and model selection for the token classification of the French modal verb *pouvoir*. We chose to perform token classification to isolate occurrences of *pouvoir*, enabling us to label them with the specific categories we developed. The primary evaluation metric used across these tests was the F1-score, which harmonically combines precision and recall. This metric is particularly crucial in scenarios such as ours where class imbalance is significant; over 97% of the dataset constituted the non-*pouvoir* class labeled "O". This label was used to mark all tokens that did not correspond to instances of *pouvoir*, allowing the model to focus specifically on identifying and classifying the modality of *pouvoir*'s occurrences.

5.1. Training Data selection

Initially, the corpus listed in 4.2 was experimented upon using the camembert-base model with a stratified train-validation-test split of 80-10-10 over seven epochs to determine the most effective training data. This split allowed us to monitor model performance on a small validation set during training, and the augmented context corpus (corpus_context_augmented) proved to be superior, achieving an F1-score of 0.90 in evaluation and 0.88 when the "O" class was excluded. These results indicated that data balancing coupled with contextual enhancements significantly benefits model performance. After identifying the corpus_context_augmented dataset as the optimal choice, we applied a 5-fold cross-validation strategy to evaluate the model's robustness. This cross-validation process was conducted on the 80% training portion of the dataset, while the 20% test set remained untouched. Cross-validation yielded further improvements in model performance, solidifying the combination of the corpus_context_augmented dataset and the camembert-base model as our most reliable setup.

5.2. Model performance comparison

After determining the optimal training data setup, we tested various pre-trained models to assess their effectiveness in the modal classification of the French verb *pouvoir*. Throughout this phase, we maintained the stratified 80-20 split for training and testing, ensuring that the 20% test set remained unseen for final evaluations. For all models tested, the training set was subjected to 5-fold cross-validation during training to leverage its demonstrated benefits. As shown in table 3, the best performing model was the flaubert-base-cased which achieved an F1-score of 0.94 and 0.92 when the "O" class was excluded[9]. One possible reason for its superior performance could be attributed to the extensive and diverse pretraining corpus it was trained on, which is specifically designed to capture various nuances of the French language. Given that our dataset is based on oral corpora, the flaubert-base-cased model may be particularly well-suited for this type of data, as the other models have been trained on less diversified data forms. In the final evaluations, the flaubert-base-cased model demonstrated strong performance in identifying non-modal occurrences and distinguishing specific modalities such as "eventuality" and "permission" (see confusion matrix and results per category in appendix B). However, it encountered some challenges with the "material possibility or ability" category, indicating slight semantic overlaps. The confusion matrix corroborates these findings, showing minimal misclassifications, particularly between categories such as "material possibility or ability". This final analysis highlights that holistic advancements in both model selection and detailed category definition refinement are crucial. By leveraging models optimized for the French language such as FlauBERT, alongside meticulously curated and balanced training data, the task of modality classification for *pouvoir* is approached with an increasingly nuanced understanding and precision, promising further enhancements and consistency in future NLP applications of the same kind.

Table 3
Best model selection experiment result

model[10]	F1-score	F1-Score without "O" category
roberta-base	0,89	0,86
distilbert-base	0,89	0,87
distilbert-multilingual-base	0,89	0,86
bert-multilingual-base	0,92	0,9
camembert-large	0,89	0,86
camemberta-base	0,90	0,88
flaubert-base-uncased	0,92	0,90
flaubert-base-cased	**0,94**	**0,92**
flaubert-large-cased	0,92	0,90

[9]The model is available at https://huggingface.co/DiegoRossini/flaubert-pouvoir-modality-detector

[10]for RoBERTa see https://huggingface.co/FacebookAI; for DistilBERTseehttps://huggingface.co/distilbert; for CamemBERT see https://huggingface.co/almanach; for FlauBERT see

6. Discussion

The semantic substitution process was particularly challenging due to the resource-intensive nature of available models such as FastText[11] and the complexity of handling text derived from spoken language. Our approach involved using Spacy to capture verbs, determining the most semantically similar verbs with FastText, and then conjugating them to match the form of the original verbs. This sequence of operations proved extremely resource-demanding and difficult to implement. Additionally, Spacy and FastText both demonstrated significant difficulties with the French language, leading to several inconsistencies during lexical substitution. These findings underscore the need for more robust, language-specific tools to improve the accuracy and efficiency of semantic substitution in NLP tasks involving French, particularly with spoken text.

If we take a closer look at the model's results, we notice that "permission" is the second best classified category with an f-score of 0.95. However, a qualitative analysis of the classified sentences revealed some incongruences. Among the various uses of *pouvoir* with the value of permission, there are two that are very frequent (40% of permission annotations) and have a typical structure. These are the "*pouvoir* of politeness" (see Ex. 1.), a question that allows the subject to express a request politely, and the expression "je/nous/on" (*I/we/impersonal pronoun "on"*) + "pouvoir" + "dire" (*to say*) , called "pouvoir_dire" (see Ex. 2.).

> (1) Euh attends j'ai un train de retard tu **peux** répéter ? (*Uh, wait, I'm a bit behind, can you repeat that?*) (ESLO2_ENTJEUN_1235)

> (2) Enfin j'ai fait essentiellement des mesures on **peut** dire (*Well, I mostly took measurements, one could say [...]*) (ESLO2_ENT_1014)

Our model is biased by the fact that most of the permission *pouvoir* follow one of these two patterns that are characterized by a fixed structure: the model is not able to identify as *pouvoir* of permission any use that is different from 1. or 2.

> (3) Je suis nommé par le siège qui **peut** du jour au lendemain si je ne fais pas le travail me me basculer. (*I am appointed by headquarters, which can, from one day to the next, if I don't do the job, toss me out.*) (ESLO1_INTPERS_438)

For example, the model classifies Example 3. as "possibilité matérielle et capacité" even though the institution (i.e., "headquarters") granting permission to the subject is clearly mentioned. The solution will be to enrich the data of deontic *pouvoir* with some examples of different structures. To address this problem, it would be necessary to enrich and to vary, in terms of structures, the examples in the deontic category. Finally, we tested our model on all the 221 interviews in the ES_CF corpus. The results show that most instances of *pouvoir* belong to the category of physical possibility or ability (51% of pouvoir instances), followed by permission (35%), eventuality (9%) and sporadicity (5%). In general, the most representative modal category is the alethic one (value of material possibility and ability and sporadicity: 56%). These results are consistent with those we obtained in the manually annotated portion of the ES_CF corpus presented in section 4.1.

7. Conclusion

This study demonstrates significant first progress in the automatic classification of the French verb *pouvoir* by finding the best fine-tuned BERT model. Moderate to substantial inter-annotator agreement led to merging some subcategories for more streamlined annotations. The flaubert-base-cased model, with contextual data augmentation, achieved an impressive F1-score of 0.94 with cross-validation, highlighting the importance of context (see section 4.2 "Corpus Context"). However, challenges persist, such as limited training data and the need for better annotation tools and more powerful computational resources. The model struggles with certain deontic usages that humans easily identify. Intentional ambiguity by the speaker also poses a challenge for both annotators and the model. Future work should expand and enrich the dataset and consider training on full texts instead of isolated sentences to capture context better. [8] propose a similar approach, emphasizing the importance of taking a large context around the target token and advocating for the use of full texts as context. In the future, we will also experiment with an augmented context window of 10 lines before and after the target token. These enhancements will improve model robustness and set the stage for further advancements in natural language processing, particularly for classifying semantic values of French modal verbs. This is the first step in a larger project that will soon include the verb *devoir* (*must*). More globally, the ultimate goal of our approach is to be able to identify which modal categories are prevalent in any given corpus [16]. Indeed, given that the verb *pouvoir* is present in all types of texts, the ability to identify its modality becomes a necessary tool for refining the overall analysis of modality in different tasks such as sentiment analysis

https://huggingface.co/flaubert; for BERT-base-multilingual: https://huggingface.co/google-bert

[11] https://fasttext.cc/

([17] or hedge detection ([18]).

References

[1] J. Ruppenhofer, I. Rehbein, Yes we can!? annotating english modal verbs, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 1538–1545.

[2] M. Zhou, A. Frank, A. Friedrich, A. Palmer, Semantically enriched models for modal sense classification, in: M. Roth, A. Louis, B. Webber, T. Baldwin (Eds.), Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 44–53. doi:10.18653/v1/w15-2705.

[3] A. Marasović, M. Zhou, A. Palmer, A. Frank, Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations, Linguistic Issues in Language Technology 14 (2016) 191–214.

[4] A. Marasović, A. Frank, Multilingual modal sense classification using a convolutional neural network, in: Proceedings of the 1st Workshop on Representation Learning for NLP, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 111–120. doi:10.18653/v1/W16-1613.

[5] B. Li, M. Dehouck, P. Denis., Modal sense classification with task-specific context embeddings, in: ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Association for Computational Linguistics, Bruges, Belgium, 2019, pp. 1–6.

[6] J. Devlin, C. Ming-Wei, L. Kenton, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, volume 1, Association for Computational Linguistics, Minneapolis, MN, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[7] J. Wagner, S. Zarrieß, Probing bert's ability to encode sentence modality and modal verb sense across varieties of english, in: M. Amblard, E. Breitholtz (Eds.), Proceedings of the 15th International Conference on Computational Semantics, Association for Computational Linguistics, Nancy, France, 2023, pp. 28–38.

[8] M. Dehouck, P. Denis, Revisiting modal sense classification with contextual word embeddings, in: Models of Modals: From Pragmatics and Corpus Linguistics to Machine Learning, De Gruyter Mouton, Berlin, Boston, 2023, pp. 225–253. doi:doi:10.1515/9783110734157-009.

[9] A. Vinzerich, La sémantique du possible: approche linguistique, logique et traitement informatique dans les textes, Ph.D. thesis, Paris 4, Paris, France, 2007.

[10] D. Battistelli, A. Étienne, La modalité au prises des émotions et vice versa, Presented at Marqueurs modaux, énonciation et argumentation, Cerlico, Nantes, France, 24-25 mai, 2024.

[11] D. Battistelli, A. Etienne, R. Rahman, C. Teissèdre, G. Lecorvé, 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: Y. Estève, T. Jiménez, T. Parcollet, M. Z. Boito (Eds.), Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 of *JEP/TALN/RECITAL*, ATALA, Avignon, France, 2022, pp. 236–246.

[12] L. Gosselin, Les modalités en français: la validation des représentations, volume 1, Brill, Leiden, The Netherlands, 2010. doi:10.1163/9789042027572.

[13] C. Barbet, Devoir et pouvoir, des marqueurs modaux ou évidentiels ?, Langue française 173 (2012) 49–63. doi:10.3917/lf.173.0049.

[14] C. Vetters, Modalité et évidentialité dans pouvoir et devoir : typologie et discussions, Langue française 173 (2012) 31–47. doi:10.3917/lf.173.0031.

[15] A. Widlöcher, Y. Mathet, The glozz platform: a corpus annotation and mining tool, in: Proceedings of the ACM Symposium on Document Engineering (DocEng'12), Paris, France, 2012, pp. 171–180. doi:10.1145/2361354.2361394.

[16] A. Colli, D. Battistelli, M. Chagnoux, Quel usage des marqueurs modaux dans les discours post-traumatiques ?, Presented at Marqueurs modaux, énonciation et argumentation, Cerlico, Nantes, France, 24-25 mai, 2024.

[17] Y. Liu, X. Yu, B. Liu, Z. Chen, Sentence-Level Sentiment Analysis in the Presence of Modalities, in: D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Gelbukh (Eds.), Computational Linguistics and Intelligent Text Processing, volume 8404, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 1–16. URL: http://link.springer.com/10.1007/978-3-642-54903-8_1. doi:10.1007/978-3-642-54903-8_1, series Title: Lecture Notes in Computer Science.

[18] S. Agarwal, H. Yu, Detecting hedge cues and their scope in biomedical text with conditional ran-

dom fields, Journal of Biomedical Informatics 43 (2010) 953–961. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046410001140. doi:10.1016/j.jbi.2010.08.003.

A. Annexe A: Extended version of annotation examples of the 7 semantic values of *pouvoir*

Table 4
Extended version of annotation examples of the 7 semantic values of *pouvoir*

Global modal categories	**Modal categories**	**Definitions**	**Examples**
aléthique (*alethic*)	**sporadicité** *sporadicity*)	Occurrences of pouvoir used to indicate the contingency of a state or process	Parfois dramatique comme les les romans qui **peuvent** rappeler des situations plus ou moins pénibles. (*Sometimes dramatic, like novels that can evoke more or less painful situations*) (ESLO1_ENT_003_C)
	possibilité matérielle (*material possibility*)	Occurrences of pouvoir where the source of the possibility they express is material conditions external to the subject.	C'est un un personnage donc il y a des choses que vous ne **pouvez** pas faire uniquement avec du verre et du plomb par exemple ces cheveux-là le nez la bouche oui. (*It is a character, so there are things you cannot do with just glass and lead, for example, the hair, the nose, the mouth, yes.*) (ESLO1_ENT_002_C)
	capacité (*ability*)	Occurrences of *pouvoir* where the source of the possibility they express is inherent characteristics of the subject.	À l'intérieur on a une galette on a un gâteau on le partage en X morceaux on peut pas le le faire grandir par le le un coup de baguette magique. (*Inside, we have a cake, we share it into X pieces, we cannot make it grow with a wave of a magic wand.*) (ESLO1_INTPERS_421_C)
	possibilité logique (*logical possibility*)	Occurrences of *pouvoir* used to indicate statements that are true by convention.	ø
épistémique (*epistemic*)	**éventualité** (*eventuality*)	Occurrences of *pouvoir* that indicate assumptions or personal judgments on the part of the speaker.	Les payer pour qu'ils euh fassent leur boulot et euh qu'on donne un un prix euh au meilleur grapheur money price et on prend cinq mille euros ça pourrait être pas mal. (*Pay them so they, uh, do their job and, uh, give a, uh, prize, uh, to the best graffiti artist, money prize, and we take five thousand euros, that could be nice*) (ESLO2_ENTJEUN_1228_C)
déontique (*deontic*)	**permission** (*permission*	Occurrences of *pouvoir* that indicate permission granted to the subject by an animate being, an institution, or by social or ethical laws.	Euh les gens sont libres de venir consulter quelque médecin que ce soit et ils **peuvent** en changer à tout moment et que donc euh après être venus me consulter euh si je ne leur plais pas. (*Uh, people are free to consult any doctor they choose and they can change at any time, and so, uh, after coming to see me, uh, if they don't like me.*) (ESLO1_ENT_003_C)
indeterminé (*undetermined*)	**indeterminé** (*undetermined*)	Occurrences of *pouvoir* for which the annotator hesitates between two or more values.	C'est ça ? justement je me dis comment est-ce que je vais **pouvoir** utiliser mes capacités informatiques ? (*That's it? Exactly, I'm wondering how I will be able to use my computer skills?*) (ESLO2_ENTJEUN_1235_C)
		Occurrences of *pouvoir* that are impossible to annotate due to lack of context (incomplete statements).	Parce que sinon on aurait **pu** ... (*Otherwise, we could have...*) (CFPP, Catherine_Lecuyer)

B. Annexe B: confusion matrix of the best model's results

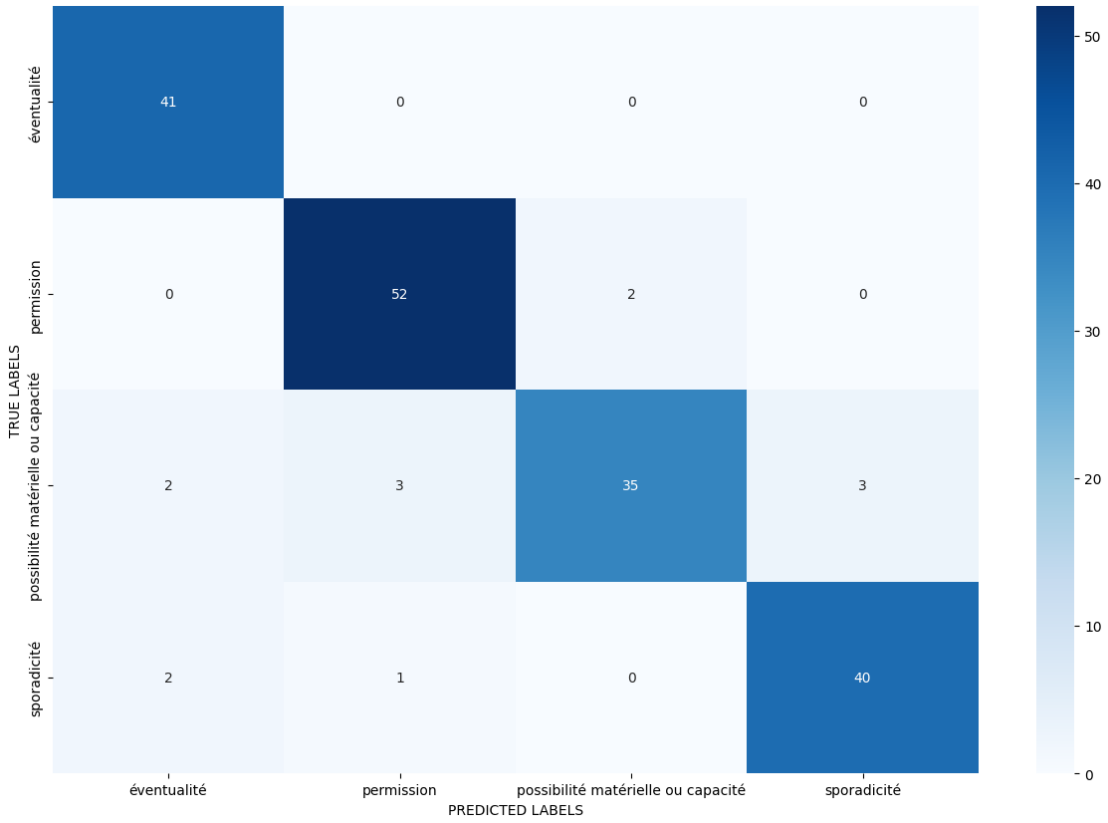

Figure 1: confusion matrix of the best model's results

C. Annexe C:

Table 6
Examples from each corpora

Datasets	Examples
Corpus_base (1 example = 1 oral speech turn)	Benjamin Franklin mais c'était le bonheur quand même est-ce qu'il y a beaucoup d'enfants qui peuvent dire ou même moi je prenais mon vélo (*Benjamin Franklin, but it was happiness all the same. Are there many children who can say, or even me, I would take my bike...*) [ESLO1]
Corpus_Base_Augmented (from a Corpus Base example another is created performing lexical substitution)	Benjamin Franklin, mais c'était le bonheur tout de même, est-ce qu'il y a beaucoup d'enfants qui peuvent s'exprimer ou même moi j'utilisais mon vélo (*Benjamin Franklin, but it was happiness all the same. Are there many children who can express themselves, or even me, I used to ride my bike*)
Corpus_Context (1 exemple = 1 oral speech turn + the oral speech turn before and the oral speech turn after)	Quand même hein la collègue un peu plus loin bon le lycée il l'a fait sur Orléans à hm + Benjamin Franklin mais c'était le bonheur quand même est-ce qu'il y a beaucoup d'enfants qui dire ou même moi je prenais mon vélo hm hm hm aller au travail en vélo + non mais c'était euh enfin bon puis nous sommes partis mon mari il a été à la retraite donc ça nous a fait une occasion aussi pour partir mais je veux dire que la vie à Olivet ne me plait pas du tout donc on doit pas se maquiller donc on est plus ou moins mal dans notre peau vu qu'on est sans cesse complexé on peut pas porter une jupe ouais c'est vrai hm qu'il y a beaucoup d'enfants qui dire ou même moi je prenais mon vélo hm hm hm aller au travail en vélo non mais c'était euh enfin bon puis nous sommes partis mon mari il a été à la retraite donc ça nous a fait une occasion aussi pour partir mais je veux dire que la vie à Olivet ne me plait pas du tout. (*Still, you know, the colleague a little further away, well, he went to high school in Orléans, um, + Benjamin Franklin, but it was happiness all the same. Are there many children who can say that, or even me, I used to ride my bike, um, um, um, to go to work by bike + No, but it was, well, then we left when my husband retired, so that gave us an opportunity to move, but I mean, life in Olivet doesn't appeal to me at all. So, we don't wear makeup, so we feel more or less uncomfortable in our own skin, constantly self-conscious. You can't wear a skirt, yeah, it's true. Um, are there many children who can say that, or even me, I used to ride my bike, um, um, um, to go to work by bike? No, but it was, well, then we left when my husband retired, so that gave us an opportunity to move, but I mean, life in Olivet doesn't appeal to me at all.*)
Corpus_Context_Augmented (from a Corpus Context exemple another is created performing lexical substitution)	Quand même hein la collègue un peu plus loin bon le lycée il l'a réalisé sur Orléans à hm + Benjamin Franklin mais représentait le bonheur quand même est-ce qu'il y a beaucoup d'enfants qui affirmer ou même moi je prenais mon vélo hm hm hm se rendre au travail en vélo + non mais représentait euh enfin bon puis nous avons départis mon mari il a été à la retraite donc ça nous a fait une occasion aussi pour partir mais je veux affirmer que la vie à Olivet ne me agrée pas du tout donc on devrait pas se maquiller donc on est plus ou moins mal dans notre peau vu qu'on est sans cesse complexé on ne peut pas mettre une jupe ouais c'est vrai hm qu'il y a beaucoup d'enfants qui affirmer ou même moi je prenais mon vélo hm hm hm se rendre au travail en vélo non mais représentait euh enfin bon puis nous avons départis mon mari il a été à la retraite donc ça nous a fait une occasion aussi pour partir mais je veux affirmer que la vie à Olivet ne me agrée pas du tout. (*Still, you know, the colleague a little further away, well, he finished high school in Orléans, um, + Benjamin Franklin, but it represented happiness all the same. Are there many children who can say that, or even me, I used to ride my bike, um, um, um, to go to work by bike + No, but it represented, well, then we left when my husband retired, so that gave us an opportunity to move, but I want to say that life in Olivet doesn't suit me at all. So we shouldn't wear makeup, so we feel more or less uncomfortable in our own skin, constantly self-conscious. You can't wear a skirt, yeah, it's true. Um, are there many children who can say that, or even me, I used to ride my bike, um, um, um, to go to work by bike? No, but it represented, well, then we left when my husband retired, so that gave us an opportunity to move, but I want to say that life in Olivet doesn't suit me at all.*)