Computational Linguistics in the Netherlands Journal 14 (2025) 385-399 Submitted 12/2024; Published 07/2025

Word Sense Discrimination with French Transformer Models

Stef Accou* STEF.ACCOU@STUDENT.KULEUVEN.BE
Tim Van de Cruys* TIM.VANDECRUYS@KULEUVEN.BE

*KU Leuven, Leuven, Belgium

Abstract

This paper investigates unsupervised Word Sense Discrimination using French monolingual transformer models (viz. FlauBERT and CamemBERT), employing clustering and lexical substitution techniques. To investigate approaches that can benefit lower-resource languages, we explore three approaches: (1) clustering contextual embeddings derived through Principal Component Analysis (PCA); (2) a substitute-based method inspired by Amrami and Goldberg (2018), which leverages sparse vectors of model-predicted substitutes; and (3) an enhanced lexical substitution approach adapted from Zhou (2019), designed specifically for BERT-based models and employing embedding dropout to preserve semantic coherence. The evaluation uses two datasets: a manually annotated gold standard comprising 11 homonymous and polysemous target words, and a noisier, augmented corpus sourced from web crawls. Cluster estimation is performed with the Bayesian Information Criterion (BIC), and clustering is conducted using Gaussian Mixture Models (GMM). The gold standard enables comprehensive evaluation across hard-clustering metrics, addressing the lack of consensus on benchmarking Word Sense Discrimination algorithms. Our results show that FlauBERT consistently outperforms CamemBERT on clean datasets, while CamemBERT demonstrates greater robustness to noise. Incorporating Zhou's (2019) lexical substitution technique yields state-of-the-art performance, particularly in substitute-based methods, but at the cost of significantly higher computational demands and variability due to embedding dropout. These findings highlight the trade-offs between precision and scalability in applying advanced lexical substitution methods.

1. Introduction

Word Sense Discrimination (the unsupervised grouping of occurrences of an ambiguous word into distinct senses) is the unsupervised counterpart to Word Sense Disambiguation. While Word Sense Disambiguation relies on annotated data to classify word occurrences into predefined categories, often grounded in lexical resources like WordNet (Fellbaum 1998), this dependence on annotated corpora represents a significant limitation. Such resources are reasonably available for English, but remain far scarcer for many other languages, including French (Scarlini et al. 2020). Consequently, unsupervised approaches like Word Sense Discrimination offer a compelling alternative for resource-poor settings.

The task of Word Sense Disambiguation can be divided into two complementary subtasks: discrimination, which involves identifying and grouping distinct senses of a word, and labelling, which assigns those groups appropriate sense tags (Schütze 1998). Word Sense Discrimination focuses exclusively on the first step, foregoing external sense inventories or annotated datasets. By doing so, it shifts the focus to a model's intrinsic ability to separate meanings, making it a critical benchmark for understanding how models handle lexical ambiguity in a data-independent manner.

Although significant strides have been made in Word Sense Discrimination for English, the task remains far from solved and has seen limited exploration for other languages. French, with its rich morphological and syntactic complexity, presents unique challenges and opportunities for advancing this research. In particular, existing techniques developed for English often assume the availability of clean, annotated datasets and may not generalize effectively to languages with limited resources or differing linguistic structures.

This research aims to bridge this gap by investigating Word Sense Discrimination in French using state-of-the-art transformer models, specifically CamemBERT (Martin et al. 2020) and FlauBERT (Le et al. 2020). We compare multiple approaches, ranging from efficient clustering techniques to computationally intensive, substitution-based methods, to assess their relative strengths and weaknesses. To facilitate this work, we develop two new French datasets: a gold-standard dataset, annotated for sense labels, and a larger, noisier corpus sourced from web crawls[1]. These datasets enable robust evaluations of clustering and discrimination techniques, while also exploring their scalability and adaptability across varying levels of noise and complexity.

Our findings not only highlight the challenges and trade-offs associated with applying Word Sense Discrimination techniques to French but also reveal that certain computationally efficient methods can achieve competitive results. This work underscores the importance of tailoring Word Sense Discrimination methods to the linguistic and resource contexts of different languages, paving the way for broader adoption of unsupervised sense discrimination in multilingual NLP research.

2. Related work

2.1 Early Approaches

Word Sense Disambiguation (WSD) and Word Sense Discrimination (WSDisc) are closely related tasks, but they differ fundamentally in their methodologies and requirements. WSD is a supervised task, relying on annotated corpora to classify word occurrences into predefined sense categories. For instance, the IMS (*It Makes Sense*) classifier (Zhong and Ng 2010) employs machine learning techniques such as Support Vector Machines, trained on manually annotated datasets. These methods have demonstrated high accuracy for resource-rich languages like English, where annotated resources are abundant.

In contrast, Word Sense Discrimination is an unsupervised task that does not depend on predefined sense inventories or annotated data. Instead, WSDisc seeks to group word occurrences into clusters corresponding to distinct senses, typically based on contextual similarity. A seminal work in this area is the context-clustering approach of Schütze (1998), which represents word occurrences as high-dimensional vectors derived from co-occurrence statistics. These vectors are then clustered using statistical techniques, forming groups that correspond to different word senses. This method demonstrated the feasibility of inducing word senses without external annotations and established clustering-based sense discrimination as a foundational approach.

The rise of neural network-based methods brought significant advancements to both WSD and WSDisc. Kågebäck et al. (2015) introduced the use of contextual embeddings derived from neural networks for WSD, enabling more precise sense distinctions by focusing on the specific context of each word occurrence. Building on this, Iacobacci et al. (2016) demonstrated that incorporating skip-gram-based word2vec representations as features significantly improved the IMS model, bridging the gap between traditional statistical methods and modern embedding-based approaches.

Further advancements came with contextually richer models. Context-aware embeddings, such as those provided by *context2vec* (Melamud et al. 2015) and ELMo (Peters et al. 2018), marked a turning point in both WSD and WSDisc. These models generated embeddings that dynamically adjusted based on the surrounding text, enabling more nuanced sense clustering and lexical substitution. Their success underscored the importance of capturing fine-grained contextual information for both tasks and set the stage for the development of transformer-based methods.

2.2 Advances in WSD with Transformer Models

The advent of transformer architectures, notably BERT (Devlin et al. 2019), has marked a significant turning point in Word Sense Disambiguation (WSD) and Word Sense Discrimination (WSDisc).

1. The dataset and experiments were developed as part of the framework outlined in Accou (2024).

These models leverage the self-attention mechanism to capture nuanced contextual relationships between words, providing dynamic embeddings that adjust based on the surrounding text. This capability has allowed transformers to achieve near-human performance in coarse-grained WSD tasks, effectively distinguishing between broad sense categories (Loureiro et al. 2021). However, fine-grained WSD—where closely related senses must be differentiated—remains a challenge. While transformers excel at capturing broader contextual cues, subtle semantic distinctions often require additional fine-tuning or external knowledge resources. For example, distinguishing between the financial and legal senses of a word like "bank" in context can strain even state-of-the-art models.

In the realm of Word Sense Discrimination, transformer-based embeddings have facilitated more accurate clustering by providing rich, context-sensitive representations. Techniques such as substitute-based clustering and dynamic patterns, when paired with transformer embeddings, have shown promise in unsupervised sense discrimination tasks. Nevertheless, these approaches often involve computationally intensive operations, such as generating multiple substitutes or embeddings for each context, which can limit scalability. Additionally, the application of these models to languages such as French has not been extensively explored.

2.3 French Transformer Models

The monolingual French Transformer-based models CamemBERT (Martin et al. 2020) and FlauBERT (Le et al. 2020) have proven to be two competitive models for French NLP tasks. CamemBERT, based on RoBERTa (Liu et al. 2019) employs whole-word masking, it was trained on a relatively small and relatively noisy dataset resulting in good robustness across noisy data. In contrast, BERT-based FlauBERT, trained on curated resources, has shown superiority in certain tasks, particularly those requiring fine-grained semantic distinctions. The performance of these models on Word Sense Discrimination tasks remains under-investigated.

2.4 Challenges in WSD/WSDisc

The scarcity of large, annotated corpora in languages like French poses significant challenges for Word Sense Disambiguation and Discrimination. To mitigate this, unsupervised sense discrimination methods have been developed, such as clustering contextualized embeddings or employing substitute-based approaches. Verbs present greater challenges compared to nouns due to their inherent semantic complexity (Raganato et al. 2017, Segonne et al. 2019). The dynamic symmetric pattern method (Amrami and Goldberg 2018, Başkaya et al. 2013) has significantly improved clustering efficiency, and its adaptation to BERT has achieved state-of-the-art results. Similarly, Zhou et al. (Zhou et al. 2019) introduced a BERT-based lexical substitution model, although its integration into Word Sense Discrimination remains untested.

2.5 SemEval Contributions

The SemEval workshops have played a pivotal role in advancing WSDisc methodologies. For instance, SemEval-2010 Task 14 (Manandhar and Klapaftis 2009), focused on Word Sense Induction and Disambiguation, providing a framework for evaluating unsupervised systems. Subsequent tasks, such as SemEval-2013 Task 13 (Jurgens and Klapaftis 2013), introduced evaluations for graded and non-graded senses, further refining the assessment of WSI systems. These tasks have highlighted innovative clustering approaches that combine contextual embeddings and substitute-based methods, underscoring the difficulty of extending these techniques to morphologically richer languages like French.

2.6 Recent Techniques

Recent studies, such as Yamada et al. (2021), have proposed new methods, including the Bayesian Information Criterion (BIC), to automatically estimate the number of clusters, which is particularly useful for polysemous words. While promising, the cross-linguistic applicability of these methods remains unclear. This review outlines the state-of-the-art and highlights the need for further exploration of Word Sense Discrimination methods tailored to French, bridging gaps between English-centric advances and under-resourced languages.

3. Methodology

This study investigates Word Sense Discrimination (WSDisc) in French using two monolingual transformer-based models: BERT-based FlauBERT and RoBERTa-based CamemBERT. The primary goal is to evaluate how well these models can cluster the senses of homonymous and polysemous words, adapting techniques developed for English to the unique linguistic features of French. To achieve this, we employ both established and novel methods and evaluate performance on two newly developed datasets.

3.1 Target Words

We selected 11 target words comprising both nouns and verbs, chosen to capture a range of polysemy and homonymy in French. The selected words include:

- **Nouns**: *avocat* ("lawyer" / "avocado"), *bien* ("wellbeing" / "good" / "property"), *bureau* ("desk" / "office"), *faculté* ("faculty" / "ability"), *glace* ("ice" / "mirror"), *souris* ("mouse" / "pointer"), *tour* ("tower" / "lap"/ "trick"), and *vol* ("flight" / "theft").
- **Verbs**: *filer* ("to spin" / "to flee"/ "to give"), *supporter* ("to support" / "to endure"), and *tirer* ("to pull" / "to shoot").

To avoid problems related to tokenization, only exact homographs are considered in this study. The words were selected based on their linguistic diversity and their ability to test the methodology across varying levels of semantic complexity. Homonymous words, such as *avocat*, have distinct meanings, while polysemous words, such as *bureau*, involve closely related senses. We predict that models will find it harder to discriminate the less delineated senses of polysemous words. This combination of clear homonyms and polysemous words offers a robust evaluation framework.

3.2 Datasets

We created a gold-standard dataset comprising 100 manually labeled instances for each target word, using the frTenTen web-crawled corpus from ELEXIS (Martelli and Navigli 2020). To ensure semantic diversity, we supplemented this dataset with examples from other sources, including the Gutenberg corpus for French (Sketch Engine 2020). Careful curation was performed to exclude sentences with translation errors, excessive technical terms or jargon and inappropriate content, ensuring the dataset represents the defined meanings of the target words. A second, augmented dataset includes 500 additional sentences per target word, introducing noise such as less coherent sentences, faulty language, and other challenges reflective of real-world data.

3.3 Methodological Approaches

We compare three different methods on each of the words and evaluate their effectiveness on the morphologically more complex French corpus. The methods differ in their way of obtaining homonym representations, but use the same subsequent clustering techniques.

3.3.1 Contextualized Embeddings

Following the approach for the related task of semantic frame induction of Yamada et al. (2021), this method attempts to perform clustering based on their contextualized embeddings. These information-rich embeddings were generated using both FlauBERT and CamemBERT. Principal Component Analysis (PCA) is applied to these embeddings to reduce dimensionality to three dimensions (Greenacre et al. 2022). This choice for three dimensions was made based on preliminary tests on a held-out development set. On the one hand, this approach aims to isolate the most important axis of difference between the different occurrences of the word. If the embeddings contain the necessary information for successful Word Sense Discrimination, this would separate different meanings and collapse the occurrences belonging to the same sense. On the other hand, reducing the embeddings to a three-dimensional representation makes the embeddings more easily interpretable and makes it possible to get insight into the created clusters.

3.3.2 Lexical Substitution with Symmetric Patterns

Inspired byAmrami and Goldberg (2018), this method uses symmetric patterns to extract representative substitutes for target words. Symmetric patterns augment regular masking techniques by considering both *forward* and *backward* masking. For example, in the sentence "J'aime manger des *pommes*" (i.e. "I like eating *apples*"), regular masking of target word *pommes* would result in "J'aime manger des MASK". Forward and backward patterns allow to conserve the target word and predict contextually relevant alternatives by creating coordination patterns. Forward formations would result in "J'aime manger des pommes et MASK" (i.e. "I like eating apples and MASK"), and backward patterns in "J'aime manger des MASK et pommes" (i.e. "I like eating MASK and apples"). These bidirectional patterns, originally conceived for bidirectional recurrent neural networks, have been shown to create more nuanced representations for word senses in English, also when working with BERT-based models (Amrami and Goldberg 2019). For French, however, coordination is not always as straightforward due to, for example, the gendered nature of nouns that is not present in English. The substitutions obtained from the masked predictions across all occurrences of a target word are pooled into a vocabulary, which is used to build up sparse vectors for each target sentence. Dimensionality reduction is again applied to these vectors in a way similar to the embeddings-based experiment, after which the representations are clustered into different word senses. This method provides interpretable clustering vectors based on lexical data: Unexpected results can be investigated by inspecting the proposed substitutes used for clustering the datapoints.

3.3.3 Dropout-Based Lexical Substitution

Based on the state-of-the-art lexical substitution technique from Zhou et al. (2019), this novel approach tries to improve the substitute-based approach of Amrami and Goldberg (2019). Zhou et al. (2019)'s lexical substitution method retrieves high-quality substitutes by applying partial dropout to the embeddings of target words, preserving important semantic information from the sentence. Substitutes are identified based on their contextual coherence and similarity to the original sentence. As the dropout of embeddings applies to randomly chosen dimensions, this method suffers from unpredictable variation which can be more or less damaging to the informational content of the embeddings. By averaging substitution probabilities across multiple iterations, the method mitigates the impact of particularly damaging dropouts. Our application of this method allows to combine the highly informative contextual embeddings with transparent substitute-based clustering, which allows for a nuanced and interpretable representation that reflects the model's understanding of ambiguous words. However, this method comes at the cost of an important increase in computational cost compared to the clustering of contextualized embeddings.

3.4 Evaluation Metrics and Clustering

Following Yamada et al. (2021), clustering was performed using a Gaussian Mixture Model (GMM) informed by a Bayesian Information Criterion (BIC) to estimate the optimal number of clusters. This approach enabled the grouping of word instances based on contextual similarities in a dynamic way, avoiding the need for fine-tuned hyperparameters for each word which further adds to the *unsupervised* character of the study. Clustering efficacy was evaluated using F1, V-measure (Rosenberg and Hirschberg 2007), and Adjusted Rand Index (Hubert and Arabie 1985). For this mapping, the gold standard consisting of sense-labeled data points was used. For the large,"automatic" dataset containing sentences for which no gold standard exists, evaluation happened only on the overlapping data from the curated dataset. The BIC-informed approach was compared to fixed-cluster baselines, consistently outperforming them in identifying meaningful clusters. Difficulties remain in comparing these results with other studies due to differences in task architecture, datasets, and languages. We found no recent studies concerning Word Sense Discrimination in French.

4. Results

4.1 Clustering of contextualized embeddings

4.1.1 Curated Dataset

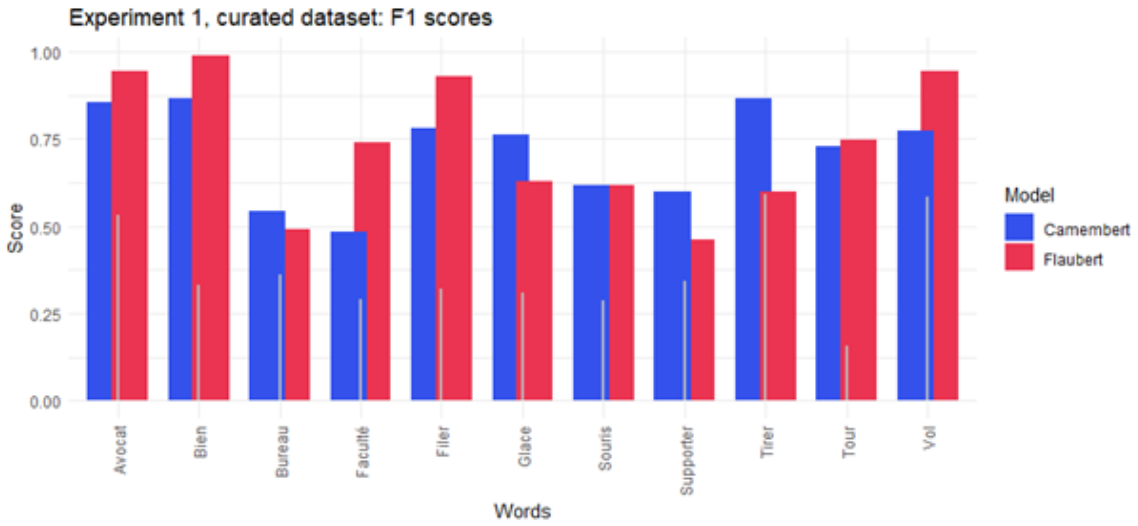

Figure 1: F1 scores for both models, obtained by clustering the contextualized embeddings (Method 1), curated dataset. The grey vertical line indicates the baseline of F1 scores when all points are assigned to the same cluster.

The results of both models on the curated dataset, visualized in Figure 1, demonstrate that unsupervised Word Sense Discrimination tasks in French by clustering contextualized embeddings is feasible. FlauBERT achieved slightly higher results than CamemBERT in most target words, giving a global average F1 score of 0.74. CamemBERT, on the other hand, attained a marginally lower macro-average F1 score of 0.70. Despite this, CamemBERT performed better than FlauBERT on some verb instances and words exhibiting greater polysemy, which can be considered the more challenging cases of WSDisc. The results obtained with both models are comparable to the state-of-the-art outcomes reported previously for the discrimination tasks of the English word sense (Arefyev

et al. 2019). However, it remains difficult to directly compare methods, as no clear consensus exists on clustering evaluation metrics or the most appropriate approaches. As we anticipated,

Figure 2: Two-dimensional representation of clusters based on FlauBERT embeddings for target noun *Bien*. The colour coding corresponds to the gold standard senses labeling, the symbols represent the different identified clusters. The red circle indicates the only point in the dataset that was assigned to an incorrect cluster.

both models perform better on clearly homonymous words but struggle with more polysemous ones. When word senses are distinctly separate, contextual embeddings appear to be sufficient for achieving high-quality discrimination. The two-dimensional representation shown in Figure 2 illustrates the clustering of sentences with target word *Bien* in the curated dataset, which was carefully monitored to include only substantival usages of the word. This visualization reveals that the clustering aligns almost perfectly with the gold standard sense annotations, identifying the senses as *property*, *wellbeing*, and the abstract concept of *good*. The lower scores in cases involving more polysemous words can be attributed to the hard clustering technique employed in this study. While the three-dimensional cluster representations for target words where clustering fails often allow for a clear visual distinction between predefined senses, the clustering method does not always reflect this. For instance, Figure 3 illustrates a perspective where the senses are perfectly visually separable. However, the clustering approach does not consider this configuration optimal. The BIC algorithm identifies only two clusters, and the different senses are randomly distributed across these clusters, with no clear correlation to the senses. This again highlights that contextualized embeddings contain the necessary information for effective sense discrimination, but that issues regarding clustering methodology and granularity hinder the ability to efficiently evaluate this potential.

4.1.2 Automatic Dataset

Using the larger, noisier dataset, a large contrast in model performance can be observed. As shown in Figure 4, CamemBERT demonstrates robust results, with only a minimal decrease in performance compared to the curated dataset (non-significant, $p = 0.51$). FlauBERT, however, shows a more pronounced decline, particularly in its v-measure and Adjusted Rand Index metrics, with statistically significant decreases ($p < 0.01$). While FlauBERT outperformed CamemBERT on the curated dataset, this trend is reversed when using the automatic dataset, with CamemBERT achieving

Figure 3: Three-dimensional representation of clusters based on CamemBERT embeddings for target noun *Faculté*. The colour coding corresponds to the gold standard senses labeling, the symbols represent the different identified clusters.

higher v-measure scores ($p = 0.02$). The better performance of CamemBERT over its competitor continues in all subsequent methods and experiments, which demonstrates CamemBERT's greater robustness when applied to noisy data. Given its consistently inferior performance, FlauBERT's results are omitted for the remainder of this paper, allowing for a direct comparison of methods across the more robust CamemBERT model.

4.2 Symmetric Patterns

Figure 5 contains a comparative overview of the results obtained from embedding-based Experiment 1 and symmetric-pattern-based Experiment 2. The symmetric pattern-based clustering approach, modeled after the methodology of Amrami and Goldberg (2018), yields significantly lower clustering scores compared to contextualized embeddings ($p < 0.01$). This discrepancy can be attributed to the substitute-based method's tendency to group most data points into the same cluster, particularly in the case of the FlauBERT model. Some relatively high F1 scores can be attributed to the preference of this metric for big uniform clusters, even when no good discrimination takes place. The approach employed strict hard-clustering evaluation metrics, which likely contributed to the divergence from earlier state-of-the-art results.

Additionally, the experimental design used in this study may have revealed limitations specific to the application of this method on French transformer models, further explaining the performance gap. Notably, structural and morphological differences between French and English make this method less reliable in French. For example, French gendered nouns will impact the substitution results and decrease performance. Additionally, we have found that the FlauBERT model's masked prediction function performs very poorly on short sentences. FlauBERT was trained on a relatively clean corpus, and seems to struggle more with ungrammatical sentences. As shown in

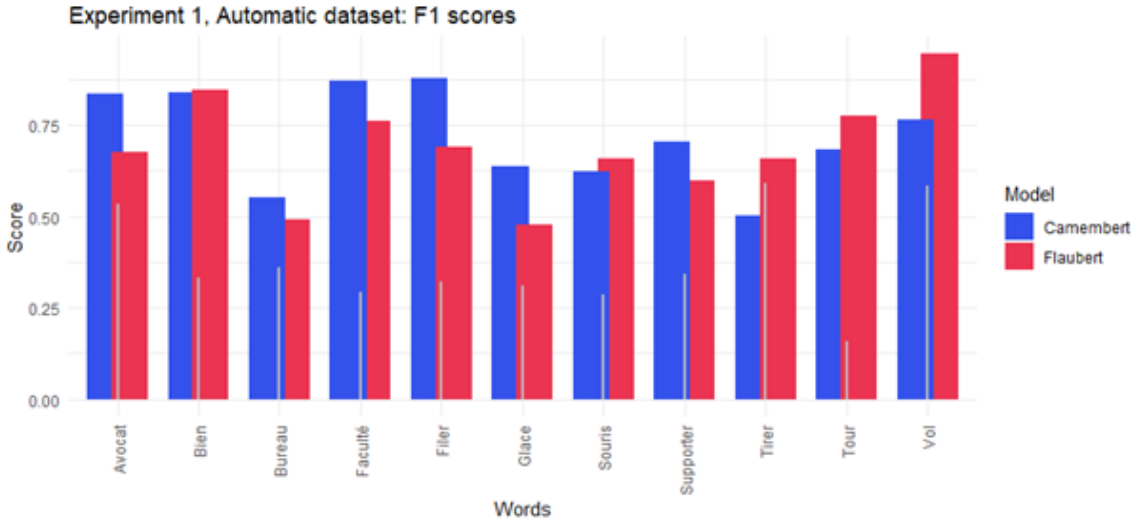

Figure 4: F1 scores for both models, obtained by clustering the contextualized embeddings (Method 1), automatic dataset. The grey vertical line indicates the baseline of F1 scores when all points are assigned to the same cluster.

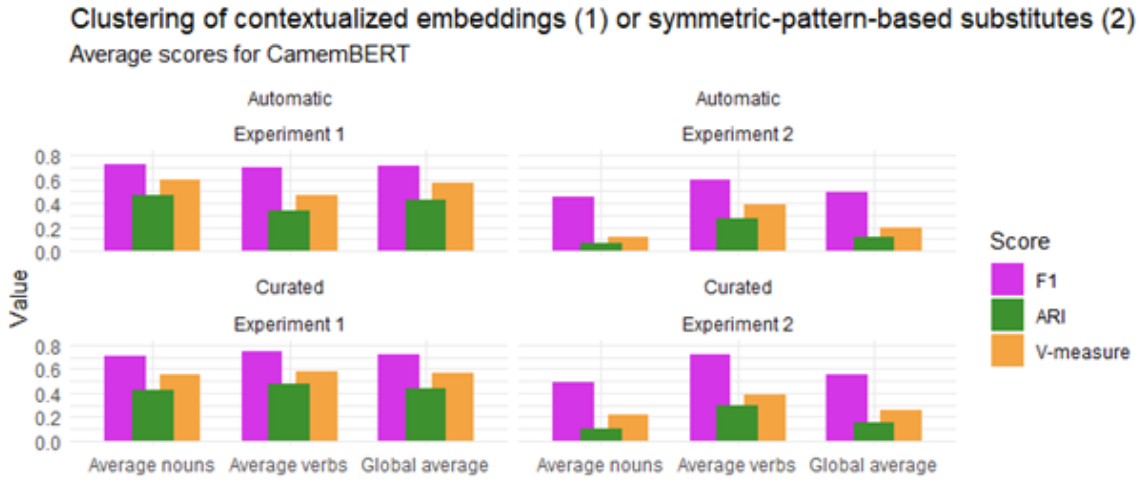

Figure 5: Comparison of scores for experiments 1 and 2

Table 1, even the absence of a full stop can significantly impact the model's performance. In contrast, CamemBERT demonstrates greater robustness to such changes and is able to provide sensible predictions even with little contextual information.

Sentence	**Model**	**Forward Predictions**	**Backward Predictions**
Le chat est sur le tapis *The cat is on the carpet* (no full stop)	CamemBERT	chien / *dog* (0.94), chaton / *kitten* (0.01), lapin / *bunny* (0.01)	souris / *mouse* (0.33), chien / *dog* (0.27), chaton / *kitten* (0.21)
	FlauBERT	leur / *their* (0.04), /, /	/, /, /
Le chat est sur le tapis. *The cat is on the carpet.* (with full stop)	CamemBERT	chien / *dog* (0.92), chaton / *kitten* (0.03), lapin / *bunny* (0.02)	souris / *mouse* (0.34), chaton / *kitten* (0.25), chien / *dog* (0.24)
	FlauBERT	noir / *black* (0.15), jeu / *game* (0.11), jour / *day* (0.10)	moi / *me* (0.23), vous / *you* (0.19), Fred/ *Fred* (0.12)
Le chat est sur le tapis, où il ronronne doucement. *The cat is on the carpet,* *purring softly.*	CamemBERT	chien / *dog* (0.94), chaton / *kitten* (0.03), lapin / *bunny* (0.01)	chien / *dog* (0.27), souris / *mouse* (0.26), chaton / *kitten* (0.25)
	FlauBERT	chien / *dog* (0.23), souris / *mouse* (0.02), chiot / *puppy* (0.02)	souris / *mouse* (0.94), compagnie / *company* (0.02), s'/ *itself* (0.01)

Table 1: Model predictions for sentences with forward and backward context masking.

4.3 Dropout-based substitute representations

4.3.1 Curated Dataset

The implementation of the state-of-the-art dropout-based lexical substitution method (Zhou et al., 2019) resulted in substantial improvements over symmetric pattern-based substitutes. Figure 6 highlights a significant performance boost for CamemBERT, with F1 scores improving by over 20% from Experiment 2.1 to Experiment 3.1 (0.55 to 0.75). This finding underscores the significant impact of high-quality substitutes on clustering outcomes. The lower standard deviation of results for CamemBERT indicates greater consistency in performance compared to FlauBERT. This approach benefits from leveraging linguistically interpretable data, offering a deeper understanding of the underlying reasons for specific clustering errors. However, the method significantly increases computational costs, as contextualized embeddings for all substitutes of a target word must be calculated and averaged over multiple iterations to mitigate the effects of randomness introduced by dropout mechanisms. This results in the compute times for this method to be many times higher than those of the other presented methods. Future research could explore preprocessing techniques to allow for both masculine and feminine forms of nouns in substitutes, potentially further alleviating the constraints of grammatical gender in the target sentences. This would benefit substitute-based WSDisc methods in languages other than English, for which most current approaches were developed.

4.3.2 Automatic Dataset

When applied to the noisier automatic dataset, the dropout-based method exhibited considerably reduced performance. The increased computational costs associated with this approach makes it challenging to fully explore its potential in this context. As such, further methodological improve-

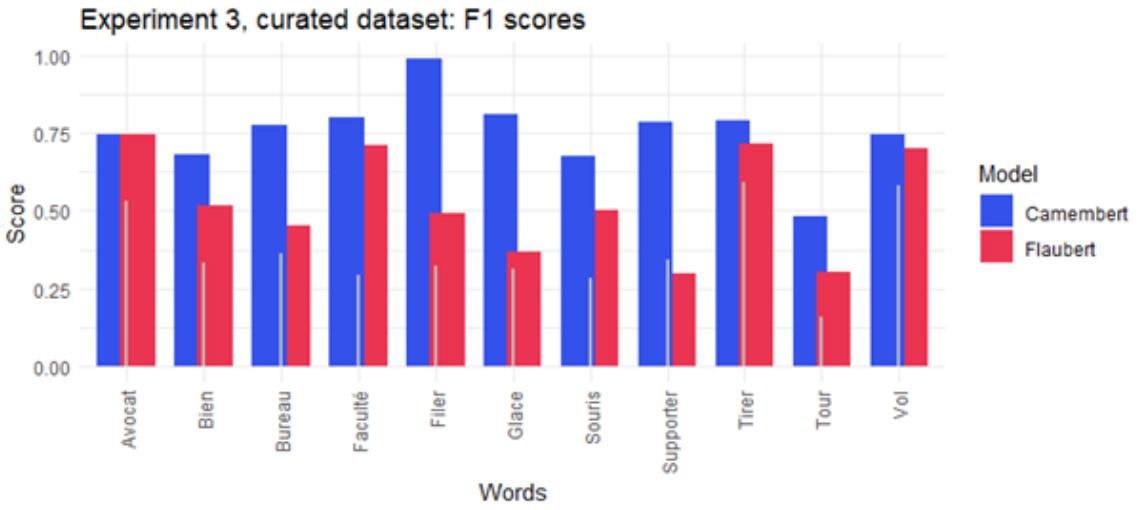

Figure 6: F1 scores for both models, obtained by clustering dropout-based substitute vectors (Method 3), curated dataset. The grey vertical line indicates the baseline of F1 scores when all points are assigned to the same cluster.

ments that could mitigate the effect of larger dataset sizes with more noise remain beyond the scope of the present study.

5. Discussion

This study demonstrates that contextualized embeddings from French encoder-only models contain the necessary information for successful Word Sense Discrimination. Clustering these embeddings into a variable number of clusters, determined unsupervisedly by a Bayesian Information Criterion, allows for effective grouping of word senses. On our curated, smaller dataset, the FlauBERT model achieved the highest scores, as shown in Figure 1. However, due to its greater robustness across methods and datasets, the CamemBERT model is better suited overall for Word Sense Discrimination tasks, as represented in Figure 7. Despite slightly lower results in Experiment 1.1, CamemBERT outperformed its competitor in all other experiments and shows a notably better robustness to noise present in "real-life" data. The third method, incorporating Zhou et al. (2019)'s approach into a substitute-based Word Sense Discrimination framework, yielded the highest performance on the curated dataset. This method outperforms our adaptation to French of the current state-of-the-art method for English significantly, but presents the drawback of increased computational costs due to the need to compute multiple embeddings for each target sentence. Additionally, its reliance on dropout methods introduces randomness, and requires multiple iterations to mitigate the effect of chance, making it less ideal for larger datasets or in settings where computing power is limited. While the dropout-based method achieves the highest results overall, we consider the embedding-based clustering method to be preferable for most WSDisc tasks due to its simplicity, reflected in the computational efficiency and superior performance on larger and noisier corpora. It can be seen in Figure 8 that the method demonstrates greater robustness across different datasets, making it a more scalable and practical solution that can be adapted easily to BERT-based models in other languages regardless of morphological and structural disparities.

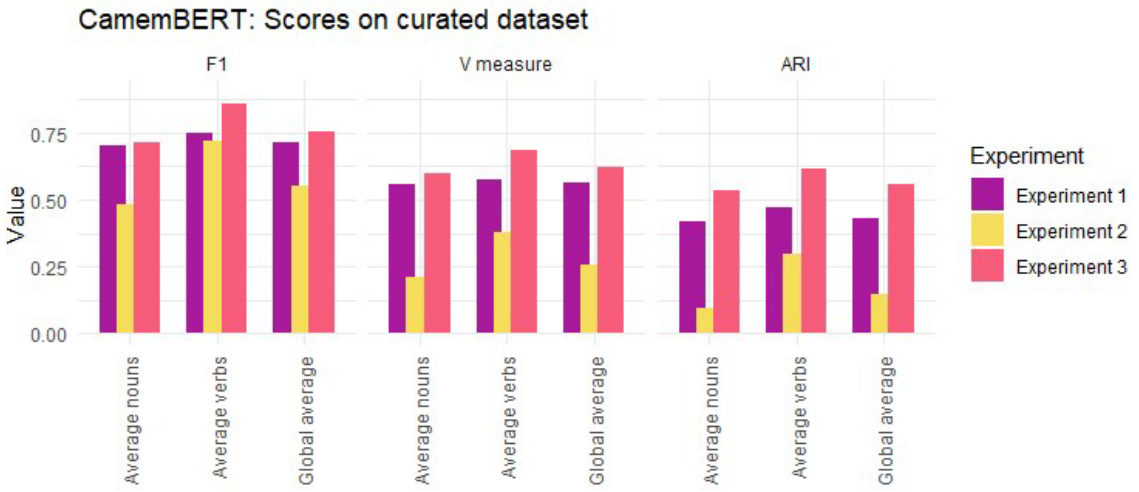

Figure 7: Comparison of experiment results for curated dataset

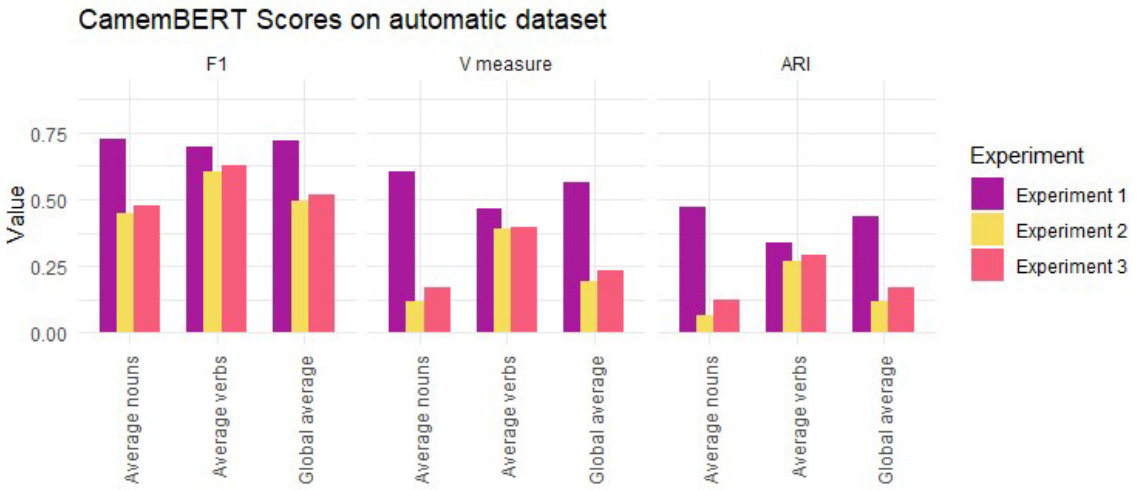

Figure 8: Comparison of experiment results for automatic dataset

6. Conclusion & Future Work

This study demonstrates the feasibility of unsupervised Word Sense Discrimination for the French language, using transformer-based contextualized embeddings and substitute-based clustering approaches. Our results show that while improved lexical substitution methods, such as dropout-based substitutes, yield higher performance on curated datasets, direct clustering of contextualized embeddings is a more practical and computationally efficient approach. This method demonstrates greater robustness across noisy datasets and is easily adaptable to other morphologically rich languages. Generalizing these findings to other languages and comparing them to the state-of-the-art in Word Sense Discrimination is not straightforward. On one hand, the lack of satisfactory French corpora means that the employed dataset and experimental design are unique, complicating direct comparisons with research focused mainly on English. On the other hand, existing methods of-

ten overlook the challenges posed by morphological and structural differences in other languages. Despite promising results, evaluating methods across diverse datasets remains challenging. Future work should apply the proposed methods to existing datasets, develop publicly available, multilingual Word Sense Discrimination benchmarks, and explore hybrid approaches that combine the computational efficiency of embedding-based clustering with the precision of substitute-based methods. Investigating nuanced evaluation metrics that capture the complexity of polysemy and homonymy in morphologically rich languages will also be critical for advancing Word Sense Discrimination research and promoting inclusivity in NLP.

References

Accou, Stef (2024), *Word Sense Discrimination with French Transformer models*, Master's thesis, KU Leuven. Faculteit Letteren, Leuven. Book Title: Word Sense Discrimination with French Transformer models.

Amrami, Asaf and Yoav Goldberg (2018), Word Sense Induction with Neural biLM and Symmetric Patterns, *in* Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, pp. 4860–4867. https://aclanthology.org/D18-1523.

Amrami, Asaf and Yoav Goldberg (2019), Towards better substitution-based word sense induction. arXiv:1905.12598 [cs]. http://arxiv.org/abs/1905.12598.

Arefyev, Nikolay, Boris Sheludko, Adis Davletov, Dmitry Kharchev, Alex Nevidomsky, and Alexander Panchenko (2019), Neural GRANNy at SemEval-2019 Task 2: A combined approach for better modeling of semantic relationships in semantic frame induction, *in* May, Jonathan, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 31–38. https://aclanthology.org/S19-2004.

Başkaya, Osman, Enis Sert, Volkan Cirik, and Deniz Yuret (2013), AI-KU: Using Substitute Vectors and Co-Occurrence Modeling For Word Sense Induction and Disambiguation, *in* Manandhar, Suresh and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 300–306. https://aclanthology.org/S13-2050.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Fellbaum, Christiane, editor (1998), *WordNet: An Electronic Lexical Database*, The MIT Press. https://direct.mit.edu/books/book/1928/WordNetAn-Electronic-Lexical-Database.

Greenacre, Michael, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina (2022), Principal component analysis, *Nature Reviews Methods Primers* **2** (1), pp. 1–21. Publisher: Nature Publishing Group. https://www.nature.com/articles/s43586-022-00184-w.

Hubert, Lawrence and Phipps Arabie (1985), Comparing partitions, *Journal of Classification* **2** (1), pp. 193–218. https://doi.org/10.1007/BF01908075.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli (2016), Embeddings for Word Sense Disambiguation: An Evaluation Study, *in* Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Berlin, Germany, pp. 897–907. https://aclanthology.org/P16-1085.

Jurgens, David and Ioannis Klapaftis (2013), SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses, *in* Manandhar, Suresh and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 290–299. https://aclanthology.org/S13-2049.

Kågebäck, Mikael, Fredrik Johansson, Richard Johansson, and Devdatt Dubhashi (2015), Neural context embeddings for automatic discovery of word senses, *in* Blunsom, Phil, Shay Cohen, Paramveer Dhillon, and Percy Liang, editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Association for Computational Linguistics, Denver, Colorado, pp. 25–32. https://aclanthology.org/W15-1504.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab (2020), FlauBERT: Unsupervised Language Model Pre-training for French.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. http://arxiv.org/abs/1907.11692.

Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados (2021), Analysis and Evaluation of Language Models for Word Sense Disambiguation, *Computational Linguistics* **47** (2), pp. 387–443.

Manandhar, Suresh and Ioannis P. Klapaftis (2009), SemEval-2010 task 14: evaluation setting for word sense induction & disambiguation systems, *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions - DEW '09*, Association for Computational Linguistics, Boulder, Colorado, p. 117. http://portal.acm.org/citation.cfm?doid=1621969.1621990.

Martelli, Federico and Roberto Navigli (2020), D4.6 Semantically annotated corpora, *ELEXIS - European Lexicographic infrastructure*.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot (2020), CamemBERT: a Tasty French Language Model, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219. arXiv:1911.03894 [cs]. http://arxiv.org/abs/1911.03894.

Melamud, Oren, Omer Levy, and Ido Dagan (2015), A Simple Word Embedding Model for Lexical Substitution, *in* Blunsom, Phil, Shay Cohen, Paramveer Dhillon, and Percy Liang, editors, *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Association for Computational Linguistics, Denver, Colorado, pp. 1–7. https://aclanthology.org/W15-1501.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018), Deep Contextualized Word Representations, *in* Walker, Marilyn,

Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237. https://aclanthology.org/N18-1202.

Raganato, Alessandro, Jose Camacho-Collados, and Roberto Navigli (2017), Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, pp. 99–110. http://aclweb.org/anthology/E17-1010.

Rosenberg, Andrew and Julia Hirschberg (2007), V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure, *in* Eisner, Jason, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Association for Computational Linguistics, Prague, Czech Republic, pp. 410–420. https://aclanthology.org/D07-1043/.

Scarlini, Bianca, Tommaso Pasini, and Roberto Navigli (2020), Sense-Annotated Corpora for Word Sense Disambiguation in Multiple Languages and Domains, *in* Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 5905–5911. https://aclanthology.org/2020.lrec-1.723/.

Schütze, Hinrich (1998), Automatic Word Sense Discrimination, *Computational Linguistics* **24** (1), pp. 97–123. Place: Cambridge, MA Publisher: MIT Press. https://aclanthology.org/J98-1004.

Segonne, Vincent, Marie Candito, and Benoît Crabbé (2019), Using Wiktionary as a resource for WSD : the case of French verbs, *in* Dobnik, Simon, Stergios Chatzikyriakidis, and Vera Demberg, editors, *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, Association for Computational Linguistics, Gothenburg, Sweden, pp. 259–270. https://aclanthology.org/W19-0422.

Sketch Engine (2020), Gutenberg Corpora 2020. https://www.sketchengine.eu/gutenberg-corpora-2020/.

Yamada, Kosuke, Ryohei Sasano, and Koichi Takeda (2021), Verb Sense Clustering using Contextualized Word Representations for Semantic Frame Induction, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Online, pp. 4353–4362. https://aclanthology.org/2021.findings-acl.381.

Zhong, Zhi and Hwee Tou Ng (2010), It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text, *in* Kübler, Sandra, editor, *Proceedings of the ACL 2010 System Demonstrations*, Association for Computational Linguistics, Uppsala, Sweden, pp. 78–83. https://aclanthology.org/P10-4014.

Zhou, Wangchunshu, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou (2019), BERT-based Lexical Substitution, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 3368–3373. https://www.aclweb.org/anthology/P19-1328.