Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Raffinage des représentations des tokens dans les modèles de langue pré-entraînés avec l'apprentissage contrastif : une étude entre modèles et entre langues

Anna Mosolova[1, 2] Marie Candito[1] Carlos Ramisch[2]
(1) Université Paris Cité, CNRS, LLF, Paris, France
(2) Aix Marseille Univ, CNRS, LIS, Marseille, France
first.last@u-paris.fr, first.last@lis-lab.fr

RÉSUMÉ

Les modèles de langue pré-entraînés ont apporté des avancées significatives dans les représentations contextuelles des phrases et des mots. Cependant, les tâches lexicales restent un défi pour ces représentations en raison des problèmes tels que la faible similarité des representations d'un même mot dans des contextes similaires (Ethayarajh, 2019). Mosolova *et al.* (2024) ont montré que l'apprentissage contrastif supervisé au niveau des tokens permettait d'améliorer les performances sur les tâches lexicales. Dans cet article, nous étudions la généralisabilité de leurs résultats obtenus en anglais au français, à d'autres modèles de langue et à plusieurs parties du discours. Nous démontrons que cette méthode d'apprentissage contrastif améliore systématiquement la performance sur les tâches de Word-in-Context et surpasse celle des modèles de langage pré-entraînés standards. L'analyse de l'espace des plongements lexicaux montre que l'affinage des modèles rapproche les exemples ayant le même sens et éloigne ceux avec des sens différents, ce qui indique une meilleure discrimination des sens dans l'espace vectoriel final.

ABSTRACT

Enhancing token-level representations in PLMs using contrastive learning : a cross-model and cross-language study

Pre-trained language models (PLMs) have significantly advanced the contextual representation of sentences and words. However, lexical-level tasks remain challenging with these representations due to issues such as low similarity of the same word in similar contexts (Ethayarajh, 2019). Mosolova *et al.* (2024) showed improvements on lexical tasks obtained by supervised token-level contrastive learning. In this paper, we study the generalizability of their findings for English on French, across other PLMs and parts of speech. We find that this contrastive learning method consistently improves performance on Word-in-Context tasks, surpassing that of standard PLMs. Further analysis reveals that fine-tuning results in increased similarity among instances sharing the same sense and greater dissimilarity among instances with different senses, suggesting improved semantic discrimination in the embedding space.

MOTS-CLÉS : sémantique lexicale, word-in-context, affinage.

KEYWORDS: lexical semantics, word-in-context, fine-tuning.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 667–682.

1 Introduction and Related Work

With the advent of large language models (LLMs), many natural language processing tasks (NLP) have come closer to being resolved (Grattafiori *et al.*, 2024). Yet, concerning lexical semantics tasks, LLMs do not offer as astonishing improvements as for text generation tasks. Having an automatic system to disambiguate the word occurrences of a large corpus remains a challenging task for most languages, although highly useful for linguistic corpus analysis. Anonymous (2025) report that in zero-shot setting, LLMs do not surpass supervised systems for Word Sense Disambiguation (WSD), indicating that high WSD performance still requires annotated data. State-of-the-art methods for WSD (Guzman Olivares *et al.*, 2025) rely both on labeled data (Babelnet (Navigli & Ponzetto, 2010)), and on pre-trained language models (PLMs) providing contextualized embeddings. Such embeddings are also a core part of current work in the unsupervised version of WSD, namely Word Sense Induction (WSI) (Abdine *et al.*, 2023; Mosolova *et al.*, 2025).

Despite their widespread use for WSD and WSI, contextualized embeddings have also been proven to be inappropriate per se for lexical semantics. In particular, studies on the Word-in-Context task (WiC), consisting in predicting whether two context sentences of a lemma share the same sense or not (Pilehvar & Camacho-Collados, 2019), have shown that original PLM representations fail to capture senses effectively and perform poorly (accuracy below 66% when using BERT-large). This inadequacy may stem from the known limitations of contextualized embeddings when it comes to represent lexical semantics : they show low similarity between the representations of the same word in different contexts and various biases such as frequency of the word or presence of punctuation in a sentence, which affect PLMs' ability to accurately represent word meaning (Ethayarajh, 2019; Timkey & van Schijndel, 2021; Jiang *et al.*, 2022).

For these reasons, approaches have been proposed to make contextualized embeddings more suitable for lexical tasks. A common technique is to fine-tune PLMs using contrastive learning, consisting in using specific loss functions to teach models to bring similar examples (positive examples) closer in the embeddings space while pushing away the dissimilar ones (negative examples) (Kaya & Bilge, 2019). Importantly, any WSD or WSI system that uses contextualized embeddings can benefit from such fine-tuned PLMs.

In NLP, contrastive learning was initially applied to enhance sentence representations (Yan *et al.*, 2021; Gao *et al.*, 2021; Fang *et al.*, 2020), and was later adapted for token-level representations. Previous works differ in how positive and negative examples are constructed : "self-supervised" approaches generate positive examples by automatically modifying the original instance, while negative examples are randomly selected from the rest of the dataset. In this vein, Liu *et al.* (2021) propose the MirrorWiC model, which uses random masking of words in the original context surrounding the target word to create self-supervised positive examples. Abdine *et al.* (2023) use the denoising autoencoder BART (Lewis *et al.*, 2020) to generate perturbed versions of the original example, serving as positive examples.

Supervised approaches, in contrast, create positive examples from instances with the same label, extracted from some classification dataset. Yamada *et al.* (2023) use supervised contrastive learning with the aim of performing semantic frame induction. The positive examples are extracted from the English Framenet (Baker *et al.*, 1998), as pairs of instances evoking the same semantic frame. Mosolova *et al.* (2024) propose to use examplar sentences of word senses from an electronic dictionary (the English Wiktionary) : positive examples are sets of exemplars of the same sense, while negative ones are exemplars of different senses of the same lemma. The authors emphasize that this approach

is portable to the many languages for which a large Wiktionary exists.

On top of the WiC task, these works have been evaluated on a range of lexical semantics tasks, including WSI (Manandhar *et al.*, 2010; Jurgens & Klapaftis, 2013), WSD (Raganato *et al.*, 2017), Usim (Erk *et al.*, 2013), CoSimLex (Armendariz *et al.*, 2020), and Frame Induction. All have shown that contrastive learning provides substantial improvements over the original PLMs.

In this paper, we focus on the approach proposed by Mosolova *et al.* (2024) and test its generalizability across different model architectures (RoBERTa and ModernBERT on top of BERT), parts of speech (English nouns and adjectives, on top of verbs), and languages (French on top of English). Additionally, we analyze changes in the embedding space after fine-tuning, an aspect that was not explored in the original study.

In the following sections, we present the approach of Mosolova *et al.* (2024) (Section 2), the methodology (Section 3), the experimental results (Section 4), and the qualitative analysis of the embedding space (Section 5), followed by our conclusions (Section 6).

2 Summary of Mosolova *et al.* (2024) supervised fine-tuning approach

The goal of study by Mosolova *et al.* (2024) was to evaluate whether contrastive fine-tuning produces more semantically meaningful token embeddings, better suited for lexical-level tasks than standard PLM embeddings. They expected that, during fine-tuning, representations of the same word used in the same sense will be pulled together, while those having different senses will be pushed further apart. For this purpose, they employed the multiple-positives contrastive loss for fine-tuning [1]. The loss for one lemma l is defined as :

$$\mathcal{L}(l) = \sum_{j \in E(l)} \frac{-1}{|S(j)|} \sum_{j' \in S(j)} \log \frac{e^{s(j,j')/\tau}}{\sum_{k \in E(l) \setminus j} e^{s(j,k)/\tau}} \tag{1}$$

Here $E(l)$ is the set of example sentences for lemma l, j is an example sentence within $E(l)$, and $S(j)$ is the subset of these examples, where the lemma l has the same sense as in example j, except for j itself. $E(l) \setminus j$ is the set of all examples of lemma l except j. $s(m, n)$ is the similarity measure between the embeddings of the target tokens in examples m and n (they use cosine similarity) and τ is a scalar temperature hyperparameter. Notice that, since the denominator sums over all examples of lemma l (not only those sharing the same sense), this will mechanically reduce the similarity between examples of different senses, while increasing the similarity between same-sense examples j and j'. The term $\frac{-1}{|S(j)|}$ is added to normalize the loss, ensuring that instances with many same-sense pairs do not contribute disproportionately. This prevents lemmas with a large number of examples from dominating the overall training objective.

For example, given the target sentence *Tu devrais frapper avant d'entrer* with the target lemma *entrer* and the following supplementary examples : (1) *Entrez votre nom d'utilisateur et votre mot de passe*, (2) *Il ne peut pas entrer dans le royaume de Dieu*, (3) *Elle avait prévu d'entrer dans la profession*

1. Mosolova *et al.* (2024) code and English Wiktionary dataset for verbs is available at https://github.com/anya-bel/contrastive_training

juridique, (4) *Des inconnus pourraient entrer dans la salle*, and (5) *Il est entré immédiatement*, the loss function is expected to bring the representation of the target lemma *entrer* in the target sentence closer to examples 2, 4, and 5, while pushing it away from examples 1 and 3.

Dataset For their experiments, the authors extracted examples from English Dbnary (Sérasset, 2015). The dataset includes all examples of verbal lemmas with at least 1 and at most 10 senses, excluding those with only a single sense and a single example, as well as multiword verbs. The dataset contains 13,118 verbs with a total of 68,271 examples having in total 26,398 senses. Mean number of examples per sense in the dataset is 2.59(±5.41), mean number of senses per verb is 2.01(±1.54) and mean number of examples per verb is 5.21(±12.68). The dataset is divided into 3 parts : 80% for fine-tuning, 10% for development and 10% for testing. We reuse their methodology to create datasets for English nouns and adjectives, as well as French verbs.

3 Experimental setup

This section details the creation of a dataset for English nouns and adjectives (to test the methodology on parts of speech other than verbs) and a dataset of French verbs (to test the method on a another language). Furthermore, we provide details regarding the training process, including the models evaluated and the hyperparameters used for fine-tuning. Lastly, we discuss the evaluation approach for the Word-in-Context datasets.

3.1 Fine-tuning datasets

	English		**French**
	Noun	**Adjective**	**Verb**
Lemmas	50,402	82,263	67,819
Examples	241,065	20,844	17,840
Senses	79,899	30,455	28,018
Avg nb of examples per lemma	3.02(±5.4)	2.28(±3.0)	3.49(±22.97)
Avg nb of examples per sense	2.19(±2.88)	1.81(±1.57)	2.26(±13.52)
Avg nb of senses per lemma	1.38(±1.26)	1.26(±0.96)	1.54(±1.78)

TABLE 1 – Statistics of the full Wiktionary datasets for English nouns and adjectives, and French verbs. Standard deviations are reported in parentheses.

We create datasets for English nouns, adjectives, and French verbs to fine-tune PLMs using the contrastive loss described above. Each dataset contains sets of lemma instances labeled with senses, which are necessary to form positive and negative pairs for the contrastive objective. We use Wiktionary as our source, as it provides freely available, sense-annotated sentences, grouped by meaning for

each word across numerous languages [2]. We use Dbnary (Sérasset, 2015) to access Wiktionary data [3] and follow the creation methodology described in the previous section. The statistics of the resulting datasets for English nouns and adjectives and French verbs are shown in Table 1.

In order to evaluate fine-tuning on all parts of speech, we also made experiments combining the fine-tuning sets for verbs, nouns and adjectives discussed above, keeping the development and test splits for each POS.

3.2 Fine-tuning details

We test four English PLMs : the previously evaluated BERT-base-uncased, a larger variant BERT-large-uncased (Devlin *et al.*, 2019), as well as two models with different pretraining configurations : RoBERTa-base (Liu *et al.*, 2019) and ModernBERT-base (Warner *et al.*, 2024). For French, we test CamemBERT-base (Martin *et al.*, 2020) and FlauBERT-base-cased [4] (Le *et al.*, 2020). All models are used through the *transformers* library [5] (Wolf *et al.*, 2020).

BERT-base-uncased was fine-tuned on English nouns, adjectives, and the combined dataset including all parts of speech (All POS), since results for verbs have already been reported. BERT-large, RoBERTa-base, and ModernBERT-base were fine-tuned on English verbs only, to test the impact of model size and alternative pretraining configurations. For French, both FlauBERT and CamemBERT were fine-tuned on verbs, allowing for a direct comparison with English results ; other parts of speech could be explored in future work. In total, we evaluate eight fine-tuning setups.

We adopt the hyperparameters optimized by Mosolova *et al.* (2024) for all our experiments : learning rate = $5e-6$, τ = 0.5, epochs = 2, each batch $E(l)$ contained at most 64 random examples of one lemma. We also applied Principal Component Analysis (PCA) with whitening to reduce the embeddings to 100 components, as this has been shown to considerably improve performance. We use target word embeddings from the PLM's last layer to compute the similarities $s(m, n)$. If a word was composed of several subwords, we averaged their embeddings. The average training time for one epoch with $10,000$ lemmas for a '-base' model was approximately 40 minutes on a single 4Gb GPU.

3.3 Evaluation method : the Word-in-Context task

To evaluate the impact of fine-tuning on token-level representations of PLMs, we employ Word-in-Context (WiC) task introduced by Pilehvar & Camacho-Collados (2019) to provide an intrinsic evaluation of this fine-tuning. The WiC task consists in predicting whether a target word used in two sentences has the same meaning in both contexts. For instance, given two French sentences : (1) *Les avions ne peuvent pas voler en ce moment* and (2) *Quelqu'un a volé mon sac dans le métro*, the

2. We acknowledge that Wiktionary senses can be overly fine-grained, and thus that trying to set apart instances of very close senses might seem counter-intuitive. Nevertheless, Mosolova *et al.* (2024) showed that fine-tuning still improves the resulting embedding space. Future work may include a more extensive use of the sense hierarchy in Wiktionary (existing in at least English and French versions) to assign different weights to negative examples, coming from homonyms (namely different Dbnary "lexical entries") or from senses of the same entry.

3. For English nouns and adjectives, we use English Dbnary dump of 06/12/2024. For French verbs, we use French Dbnary dump of 01/03/2024, https://kaiko.getalp.org/about-dbnary/.

4. We tested the '-cased' version of FlauBERT to ensure a fair comparison with the CamemBERT-base model, which does not have an uncased option.

5. For FlauBERT, we implemented target token search, as its tokenizer does not have a 'fast' version.

model should predict *False*. We only use the development and test parts of the dataset, as we aim to evaluate the quality of the embeddings directly, without any additional training on the target task. To evaluate our models, we adopt an unsupervised approach, using a threshold-based classifier with cosine similarity measured between the target word embeddings extracted from the PLM's last layer. The threshold is tuned on the development set with a step size of 0.02 and reused on the test set.

For the evaluation of **English** PLMs, we use the original WiC dataset introduced by Pilehvar & Camacho-Collados (2019), which contains examples of nouns and verbs (hereafter referred to as En-Orig-WiC). We also reuse the Wikt-WiC and Framenet-WiC datasets from Mosolova *et al.* (2024) (referred to as En-Wikt-WiC$_{verb}$ and En-Framenet-WiC$_{verb}$, respectively). Additionally, we create two WiC-like datasets from the development and test parts of the nouns and adjectives datasets mentioned above (En-Wikt-WiC$_{noun}$ and En-Wikt-WiC$_{adjective}$, respectively). Both En-Wikt-WiC$_{noun}$ and En-Wikt-WiC$_{adjective}$ datasets comprise 2000 entries with 1000 positive and 1000 negative pairs.

For **French** PLMs, we use the French part of XL-WiC dataset [6] by Raganato *et al.* (2020) (hereafter referred to as Fr-XL-WiC) [7]. Note we fine-tuned on verbs only, but XL-WiC contains instances of verbs and also nouns. So we also extracted a WiC verbal dataset from the development and test parts of the fine-tuning dataset, each containing 1,200 examples with an equal number of positive and negative examples (Fr-Wikt-WiC$_{verb}$).

As all the datasets are balanced, we use accuracy as the evaluation metric, following the standard practice in most WiC tasks. In all figures and tables, we report the average of 5 runs of fine-tuning along with the standard deviation.

4 WiC Results

For each experiment, we report two baselines : the performance of PLM's embeddings before fine-tuning both without and with PCA. We also present the fine-tuning results under these same conditions.

Generalisation across models : The results of fine-tuning of BERT-large, RoBERTa-base and ModernBERT-base models on the verbs dataset are presented in Figure 1, together with the results of 'bert-base' model from Mosolova *et al.* (2024) and MirrorWiC model from Liu *et al.* (2021) for comparison. Dimensionality reduction using PCA proves beneficial for all models on this task, even without fine-tuning, and its positive effect is largely maintained after fine-tuning. Fine-tuning itself improves performance across all models. On the En-Wikt-WiC$_{verb}$ dataset, BERT-base achieves the best performance among all models before fine-tuning (59.6%). However, after fine-tuning, other models, particularly BERT-large and ModernBERT-base, show substantial improvements and outperform BERT-base, with BERT-large reaching the highest accuracy of 75.0%. This is particularly remarkable considering that all models were fine-tuned using the best hyperparameters reported for the BERT-base model. On Orig-WiC, RoBERTa-base achieves a new best result in the unsupervised setting after fine-tuning with PCA. On FrameNet-WiC$_v$, it also obtains a surprisingly strong score before fine-tuning with PCA (73.9%), which slightly decreases after fine-tuning. In contrast, the other

6. We automatically corrected several span annotation errors in the development and test sets using SpaCy, as some target words were incorrectly annotated as the last character of the sequence.

7. As this dataset is composed of Wiktionary examples, we removed the overlapping examples from the fine-tuning dataset.

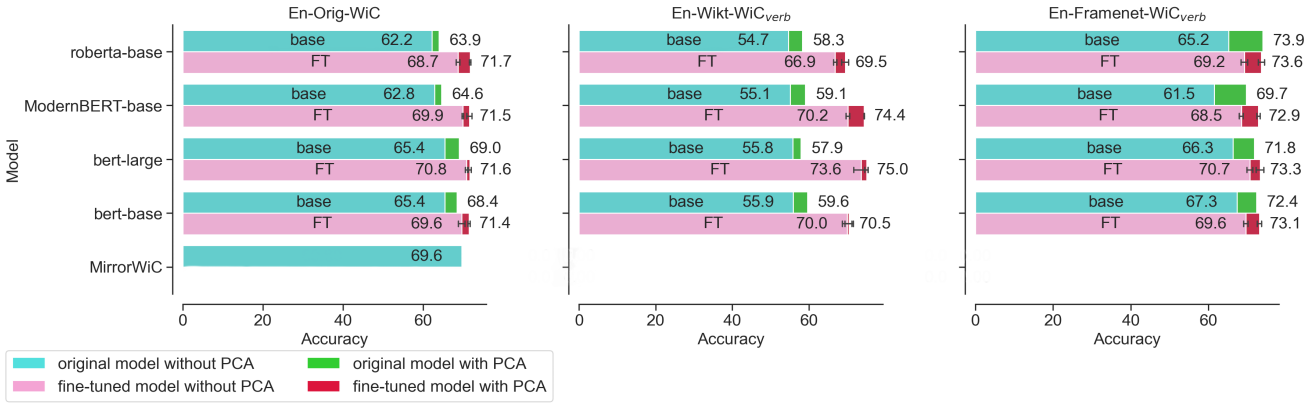

FIGURE 1 – WiC test sets results of fine-tuning on the **English verb** dataset. 'bert-base' results are taken from Mosolova *et al.* (2024), MirrorWiC results from Liu *et al.* (2021). **base** lines are baseline results before fine-tuning, **FT** lines are averages of 5 runs (errors bars are std. dev.).

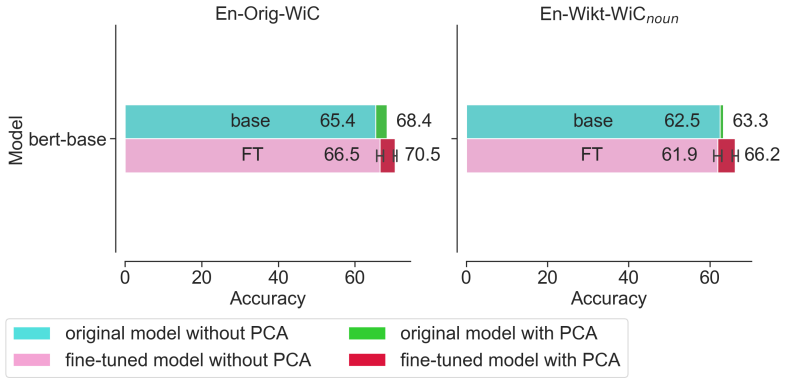

FIGURE 2 – WiC test sets results of **bert-base-uncased** fine-tuning on the **English noun** dataset. **base** lines are baseline results before fine-tuning, **FT** lines are averages of 5 runs (errors bars are std. dev.).

models benefit more consistently from fine-tuning on this dataset. Overall, fine-tuning positively influences all models, improving their performance on the WiC task by enhancing their ability to distinguish between same sense and different sense examples. PCA application consistently improves the models' performance as well.

Generalisation across POS : Regarding fine-tuning on different parts of speech ('bert-base' part of Figure 1 for verbs, Figure 2 for nouns and Figure 3 for adjectives), on the En-Orig-WiC dataset, the highest improvement is achieved when using only verb data. Results after fine-tuning on nouns also improve, but are 1% worse than those obtained from verb-only and all-POS fine-tuning (Figure 4). Interestingly, results on the En-Orig-WiC improve after fine-tuning on adjectives as well, despite adjectives not being present in the development and test sets. This suggests that during fine-tuning, the entire embedding space changes, not just the targeted lemmas. For the En-Wikt-WiC per POS test sets, fine-tuning consistently brings clear improvements, achieving the best results when combined with PCA in all settings. In summary, fine-tuning on verbal examples is more effective, likely due to the higher polysemy rate of verbs compared to other parts of speech (see Section 3 for details). When fine-tuning is combined with PCA, it consistently outperforms the original model for all POS.

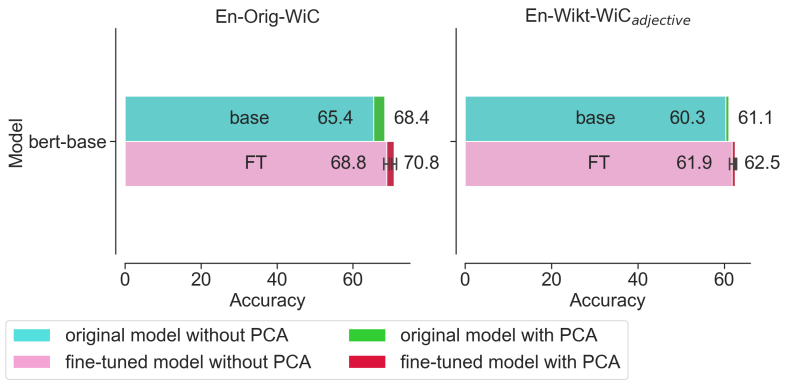

FIGURE 3 – WiC test sets results of **bert-base-uncased** fine-tuning on the **English adjective** dataset. **base** lines are baseline results before fine-tuning, **FT** lines are averages of 5 runs (errors bars are std. dev.).

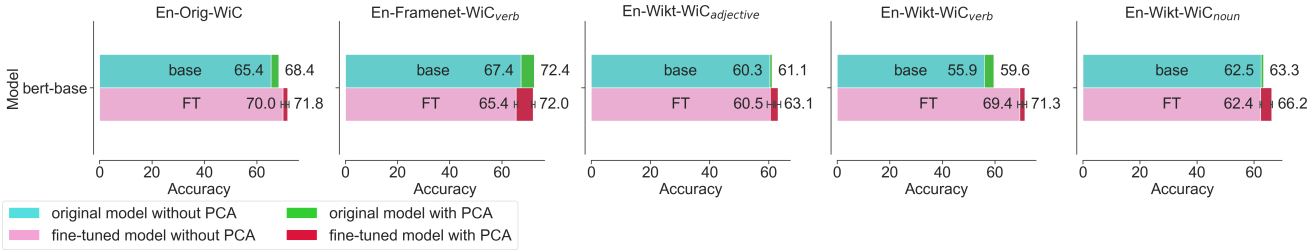

FIGURE 4 – WiC test sets results of **bert-base-uncased** fine-tuning on the **English all-POS** dataset. **base** lines are baseline results before fine-tuning, **FT** lines are averages of 5 runs (errors bars are std. dev.).

Additionally, even POS-specific fine-tuning changes the overall structure of the embedding space, influencing the representations of words beyond the targeted part of speech.

Figure 4 shows the results of fine-tuning on examples from all parts of speech. This fine-tuning further improves the results on the En-Orig-WiC dataset, pushing the highest accuracy on this dataset to 71.8%. This model also shows the best results on the En-Wikt-WiC$_{noun}$ and En-Wikt-WiC$_{adjective}$ test sets, but lower results on the verb-related test set, namely En-Framenet-WiC$_{verb}$ and En-Wikt-WiC$_{verb}$. Taking all results into account, fine-tuning on all parts of speech appears to be the best solution for tasks involving several POS as well as those using a single POS.

Generalization across languages : For the French PLMs (Figure 5), the fine-tuning process showed an astonishing 10% improvement on both models, with CamemBERT-base performing better on the Fr-Wikt-WiC$_{verb}$ dataset and FlauBERT on the Fr-XL-WiC dataset. However, the application of PCA shows inconsistent results, with 5% gain on Fr-Wikt-WiC$_{verb}$ for both models, but decreases of 0.4% and 1.8% on Fr-XL-WiC for FlauBERT and CamemBERT, respectively. These findings show that fine-tuning on French is as beneficial for this task as fine-tuning on English. Additionally, the previously observed improvement in the overall embedding space across all parts of speech in English is present in French as well, given that Fr-XL-WiC includes both verbs and nouns, while fine-tuning was done using only verbal examples.

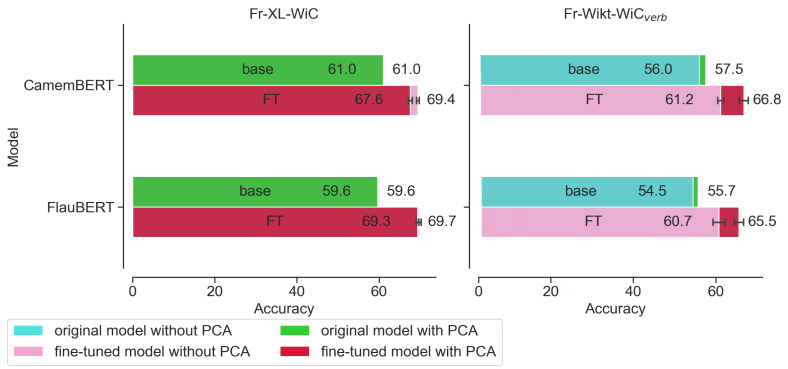

FIGURE 5 – WiC test sets results of fine-tuning on the **French verb** dataset. CamemBERT is camembert-base, FlauBERT is flaubert-base-cased. **base** lines are baseline results before fine-tuning, **FT** lines are averages of 5 runs (errors bars are std. dev.).

Influence of dimensionality reduction : PCA application is beneficial in all setups for English, both with and without fine-tuning. However, its effectiveness for French requires additional analysis, as it demonstrated opposite trends on the Fr-Wikt-WiC$_{verb}$ and Fr-XL-WiC datasets with results improvements observed only on the former.

Overall, the experiments show that both contrastive fine-tuning and dimensionality reduction enhance PLMs' lexical semantics knowledge which generalizes on the entire embedding space, even when fine-tuning is done on a single part of speech.

5 Modified embedding space analysis

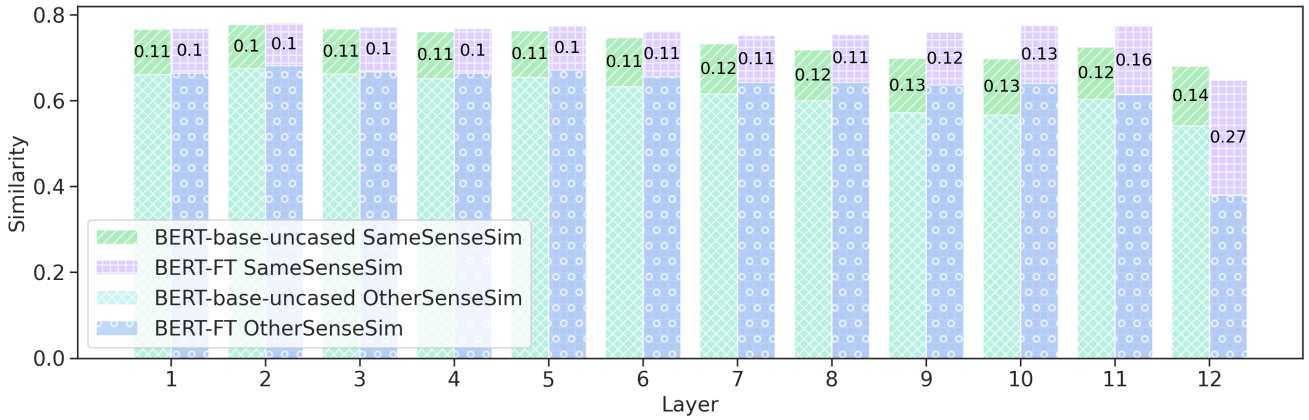

FIGURE 6 – Same-Sense Similarity and Other-Sense Similarity computed before and after fine-tuning on the bert-base-uncased model on the English verb dataset. The metrics are computed on English Wiktionary development dataset. Left bars with diagonal and crossed diagonal patterns correspond to the original model, bars with circles and grid patterns – to the fine-tuned one. The value inside bars corresponds to the difference between SameSenseSim and OtherSenseSim.

After fine-tuning, we analyzed the qualitative differences between the original and fine-tuned models using a modified version of the self-similarity metric proposed by Ethayarajh (2019). This metric

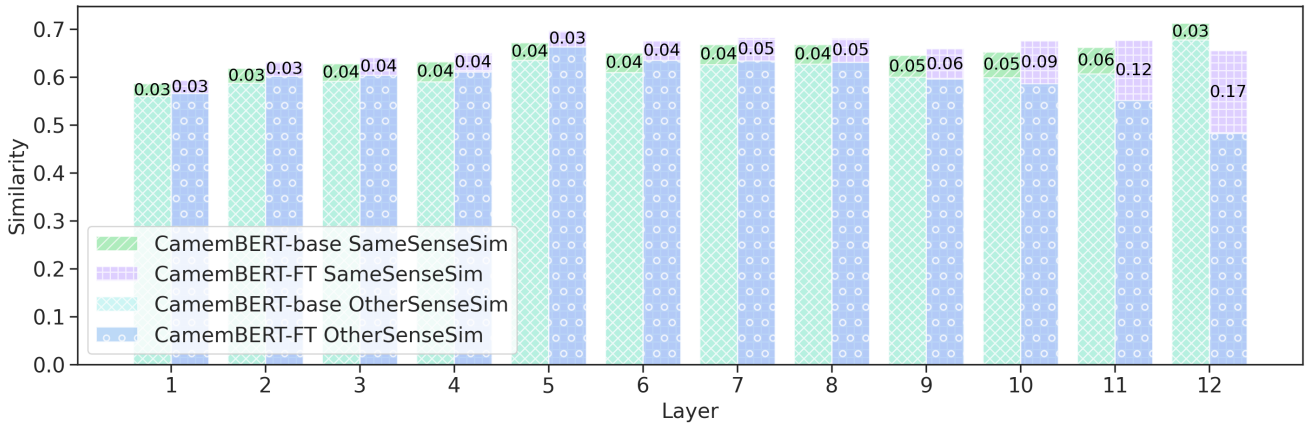

FIGURE 7 – Same-Sense Similarity and Other-Sense Similarity computed before and after fine-tuning on the CamemBERT-base model on the French verb dataset. The metrics are computed on French Wiktionary development dataset. Left bars with diagonal and crossed diagonal patterns correspond to the original model, bars with circles and grid patterns – to the fine-tuned one. The value inside bars corresponds to the difference between SameSenseSim and OtherSenseSim.

assesses the change in target word representations by measuring its similarity in contexts with the same and different senses :

$$SenseSim(l) = \frac{1}{|LL|} \sum_{j} \sum_{k \neq j}^{LL} cos(f_l(s_j, i_j), f_l(s_k, i_k)) \tag{2}$$

where s is a sentence where lemma l appears in position i and f is function that maps $s[i]$ to its representation of the layer l of the model f. LL represents lists of occurrences of the lemma l with the same and different senses depending if Same Sense similarity (SameSenseSim) or Other Sense similarity (OtherSenseSim) is measured. We expect the SameSenseSim to increase after fine-tuning, while the OtherSenseSim should decrease.

We computed both metrics on the English and French Wiktionary development sets for verbs for each layer of the bert-base-uncased, roberta-base and camembert-base models before and after fine-tuning. The results for BERT and CamemBERT are presented in Figures 6 and 7 (we put the corresponding figure and results for RoBERTa in Appendix B).

All figures show considerable changes in the similarities at the PLM's last layer. In Figure 6, the difference between the last layer's SameSenseSim and OtherSenseSim scores reached 0.27 points after fine-tuning (0.17 in Figure 7 for French). For English, before the 12th layer, both SameSenseSim and OtherSenseSim increase after fine-tuning, yet the gap between them widens, suggesting that while fine-tuning brings all representations closer, it also increases the distance between representations of words used in different senses. The 12th layer shows the largest gap between these two scores as it is directly involved in loss computation during fine-tuning. As for French, starting from the 9th layer, SameSenseSim increases after fine-tuning, while OtherSenseSim decreases, and the gap between the two widens similarly to the BERT-base model.

These observations confirm that fine-tuning has a considerable effect both on French and English models and reorganizes the last layers of PLMs to better encode sense information.

6 Conclusion

In this paper, we investigated whether the supervised contrastive learning technique proposed by Mosolova *et al.* (2024) is extendable across different model sizes, architectures, parts of speech and languages. Our results demonstrate that fine-tuning with contrastive learning consistently improves performance across all tested configurations, including BERT, RoBERTa, ModernBERT, CamemBERT, and FlauBERT models, as well as multiple parts of speech [8].

Additionally, we analyzed changes in the embedding space before and after fine-tuning. Our findings indicate an increase of the difference between same-sense and other-sense similarities, confirming that fine-tuning helps models better distinguish between different word senses in the embedding space.

Future work could include applying these embeddings to downstream tasks such as Word Sense Induction (Manandhar *et al.*, 2010), which requires embeddings of the same sense of the same lemma to be close, as well as Concept Induction (Liétard *et al.*, 2024) and Frame Induction (Yamada *et al.*, 2023) tasks, which involve grouping similar senses across different lemmas. Another promising direction would be to exploit additional information from Wiktionary, such as synonyms, antonyms, and cross-definition links, to construct more positive and negative pairs for contrastive learning.

Acknowledgements

We thank the reviewers for their valuable feedback on our work.

This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

Références

ABDINE H., KAMAL EDDINE M., BUSCALDI D. & VAZIRGIANNIS M. (2023). Word sense induction with agglomerative clustering and mutual information maximization. *AI Open*, **4**, 193–201. DOI : 10.1016/j.aiopen.2023.12.001.

ANONYMOUS (2025). Are large language models good word sense disambiguators ? In *Submitted to ACL Rolling Review - December 2024*. under review.

ARMENDARIZ C. S., PURVER M., ULČAR M., POLLAK S., LJUBEŠIĆ N. & GRANROTH-WILDING M. (2020). CoSimLex : A resource for evaluating graded word similarity in context. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5878–5886, Marseille, France : European Language Resources Association.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference*

8. We publish the best model for each setup on the Hugging Face Hub at : https://huggingface.co/annamos. The datasets are also available on GitHub under the Creative Commons Attribution-ShareAlike 3.0 license : https://github.com/anya-bel/contrastive_learning_transfer

on Computational Linguistics, Volume 1, p. 86–90, Montreal, Quebec, Canada : Association for Computational Linguistics. DOI : 10.3115/980845.980860.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

ERK K., MCCARTHY D. & GAYLORD N. (2013). Measuring word meaning in context. *Computational Linguistics*, **39**(3), 511–554. DOI : 10.1162/COLI_a_00142.

ETHAYARAJH K. (2019). How contextual are contextualized word representations ? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 55–65, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1006.

FANG H., WANG S., ZHOU M., DING J. & XIE P. (2020). CERT : Contrastive Self-supervised Learning for Language Understanding. arXiv : 2005.12766.

GAO T., YAO X. & CHEN D. (2021). SimCSE : Simple contrastive learning of sentence embeddings. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6894–6910, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.552.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A. *et al.* (2024). The llama 3 herd of models. arXiv : 2407.21783.

GUZMAN OLIVARES D., QUIJANO L. & LIBERATORE F. (2025). SANDWiCH : Semantical analysis of neighbours for disambiguating words in context ad hoc. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 7019–7033, Albuquerque, New Mexico : Association for Computational Linguistics.

JIANG T., JIAO J., HUANG S., ZHANG Z., WANG D., ZHUANG F., WEI F., HUANG H., DENG D. & ZHANG Q. (2022). PromptBERT : Improving BERT sentence embeddings with prompts. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 8826–8837, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.603.

JURGENS D. & KLAPAFTIS I. (2013). SemEval-2013 task 13 : Word sense induction for graded and non-graded senses. In S. MANANDHAR & D. YURET, Éds., *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, p. 290–299, Atlanta, Georgia, USA : Association for Computational Linguistics.

KAYA M. & BILGE H. Ş. (2019). Deep metric learning : A survey. *Symmetry*, **11**(9), 1066.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for French. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIÉTARD B., DENIS P. & KELLER M. (2024). To word senses and beyond : Inducing concepts with contextualized language models. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 2684–2696, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.156.

LIU Q., LIU F., COLLIER N., KORHONEN A. & VULIĆ I. (2021). MirrorWiC : On eliciting word-in-context representations from pretrained language models. In A. BISAZZA & O. ABEND, Éds., *Proceedings of the 25th Conference on Computational Natural Language Learning*, p. 562–574, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.conll-1.44.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. arXiv : 1907.11692.

MANANDHAR S., KLAPAFTIS I., DLIGACH D. & PRADHAN S. (2010). SemEval-2010 task 14 : Word sense induction &disambiguation. In K. ERK & C. STRAPPARAVA, Éds., *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 63–68, Uppsala, Sweden : Association for Computational Linguistics.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

MOSOLOVA A., CANDITO M. & RAMISCH C. (2024). Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In Y. GRAHAM & M. PURVER, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 34–41, St. Julian's, Malta : Association for Computational Linguistics.

MOSOLOVA A., CANDITO M. & RAMISCH C. (2025). In the LLM era, Word Sense Induction remains unsolved. *To appear in Findings of the Association for Computational Linguistics : ACL 2025*.

NAVIGLI R. & PONZETTO S. P. (2010). BabelNet : Building a very large multilingual semantic network. In J. HAJIČ, S. CARBERRY, S. CLARK & J. NIVRE, Éds., *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 216–225, Uppsala, Sweden : Association for Computational Linguistics.

PILEHVAR M. T. & CAMACHO-COLLADOS J. (2019). WiC : the word-in-context dataset for evaluating context-sensitive meaning representations. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *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)*, p. 1267–1273, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1128.

RAGANATO A., CAMACHO-COLLADOS J. & NAVIGLI R. (2017). Word sense disambiguation : A unified evaluation framework and empirical comparison. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for*

Computational Linguistics : Volume 1, Long Papers, p. 99–110, Valencia, Spain : Association for Computational Linguistics.

RAGANATO A., PASINI T., CAMACHO-COLLADOS J. & PILEHVAR M. T. (2020). XL-WiC : A multilingual benchmark for evaluating semantic contextualization. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7193–7206, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.584.

SÉRASSET G. (2015). Dbnary : Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, **6**(4), 355–361.

TIMKEY W. & VAN SCHIJNDEL M. (2021). All bark and no bite : Rogue dimensions in transformer language models obscure representational quality. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4527–4546, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.372.

WARNER B., CHAFFIN A., CLAVIÉ B., WELLER O., HALLSTRÖM O., TAGHADOUINI S., GALLAGHER A., BISWAS R., LADHAK F., AARSEN T., COOPER N., ADAMS G., HOWARD J. & POLI I. (2024). Smarter, better, faster, longer : A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv : 2412.13663.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

YAMADA K., SASANO R. & TAKEDA K. (2023). Semantic frame induction with deep metric learning. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1833–1845, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.134.

YAN Y., LI R., WANG S., ZHANG F., WU W. & XU W. (2021). ConSERT : A contrastive framework for self-supervised sentence representation transfer. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 5065–5075, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.393.

A Fine-tuning results on the WiC development sets

In Tables 2, 3, 4, 5 and 6, we show models' results on the development set of WiC datasets.

B RoBERTa : Embedding space analysis figures

Figure 8 shows the changes of the Same-Sense and Other-Sense similarity scores before and after fine-tuning of the roberta-base model. In Figure 8, we see that before fine-tuning, RoBERTa's

Model	FT	PCA	Orig-WiC	Framenet-WiC$_v$	Wikt-WiC$_v$
bert-base	-	-	67.9	70.9	58.0
bert-base	-	+	69.7	73.9	58.9
bert-base	+	+	73.5(±0.5)	76.0(±0.2)	64.8(±0.5)
bert-large	-	-	67.1	69.1	57.2
bert-large	-	+	69.3	74.5	58.3
bert-large	+	-	71.8(±0.7)	73.4(±0.4)	66.5(±0.5)
bert-large	+	+	73.4(±0.5)	75.9(±0.2)	66.3(±0.5)
roberta-base	-	-	65.0	67.4	55.1
roberta-base	-	+	68.2	75.4	57.6
roberta-base	+	-	70.6(±0.5)	72.2(±0.5)	65.9(±0.7)
roberta-base	+	+	72.8(±0.8)	75.7(±0.2)	66.3(±0.2)
ModernBERT-base	-	-	64.3	65.2	54.6
ModernBERT-base	-	+	66.1	73.4	58.4
ModernBERT-base	+	-	70.7(±0.8)	72.0(±0.5)	66.0(±0.2)
ModernBERT-base	+	+	72.8(±0.7)	75.4(±0.3)	65.9(±0.3)
MirrorWiC	-	-	71.9	-	-

TABLE 2 – WiC development sets results of fine-tuning on the **English verb** dataset. 'bert-base' results are taken from Mosolova *et al.* (2024), MirrorWiC results from Liu *et al.* (2021). **FT** : with or without fine-tuning, **PCA** : with or without PCA dimensionality reduction (100 components, with whitening). The first two lines of each model are baseline results before fine-tuning, last two lines are averages of 5 runs (std. dev. in parentheses). v subscript indicates a verbs-only dataset.

FT	PCA	Orig-WiC	Wikt-WiC$_n$
-	-	67.9	62.2
-	+	69.7	62.6
+	-	68.3(±0.8)	64.1(±0.6)
+	+	72.9(±0.7)	67.0(±0.3)

TABLE 3 – WiC development sets results of **bert-base-uncased** fine-tuning on the **English noun** dataset. **FT** : with or without fine-tuning, **PCA** : with or without PCA dimensionality reduction (100 components, with whitening). The first two lines are baseline results before fine-tuning, last two lines are averages of 5 runs (std. dev. in parentheses). n subscript indicates a nouns-only dataset.

FT	PCA	Orig-WiC	Wikt-WiC$_a$
-	-	67.9	60.7
-	+	69.7	62.3
+	-	69.4(±0.9)	61.8(±0.8)
+	+	71.6(±0.4)	64.8(±0.2)

TABLE 4 – WiC development sets results of **bert-base-uncased** fine-tuning on the **English adjective** dataset. **FT** : with or without fine-tuning, **PCA** : with or without PCA dimensionality reduction (100 components, with whitening). The first two lines are baseline results before fine-tuning, last two lines are averages of 5 runs (std. dev. in parentheses). a subscript indicates an adjectives-only dataset.

SameSenseSim and OtherSenseSim scores are nearly identical, which explains why this model performed the poorest on the En-Wikt-WiC$_{verb}$ dataset (see Figure 1). After fine-tuning, the gap between SameSenseSim and OtherSenseSim widens in the last four layers (0.26 points for the last layer), confirming the pattern previously observed with the BERT and CamemBERT models.

FT	PCA	Orig-WiC	Framenet-WiC$_v$	Wikt-WiC$_v$	Wikt-WiC$_n$	Wikt-WiC$_a$
-	-	67.9	70.9	58.0	62.2	60.7
-	+	69.7	73.9	58.9	62.6	62.3
+	-	72.1(±0.5)	69.5(±0.3)	63.9(±0.5)	63.9(±0.5)	61.9(±0.8)
+	+	73.1(±0.6)	75.7(±0.5)	64.9(±0.7)	67.2(±0.3)	65.0(±0.5)

TABLE 5 – WiC development sets results of **bert-base-uncased** fine-tuning on the **English all-POS** dataset. **FT** : with or without fine-tuning, **PCA** : with or without PCA dimensionality reduction (100 components, with whitening). The first two lines are baseline results before fine-tuning, last two lines are averages of 5 runs (std. dev. in parentheses). n, v and a subscripts indicate nouns-only, verbs-only and adjectives-only datasets, respectively.

Model	FT	PCA	Orig-WiC	Wikt-WiC$_v$
CamemBERT	-	-	62.0	54.8
CamemBERT	-	+	64.7	56.3
CamemBERT	+	-	71.6(±0.2)	63.2(±0.5)
CamemBERT	+	+	71.1(±0.1)	68.2(±0.3)

Model	FT	PCA	Orig-WiC	Wikt-WiC$_v$
FlauBERT	-	-	61.3	54.4
FlauBERT	-	+	61.7	55.9
FlauBERT	+	-	71.1(±0.3)	64.4(±1.1)
FlauBERT	+	+	71.6(±0.2)	67.9(±0.4)

TABLE 6 – WiC development sets results of fine-tuning on the **French verb** dataset. CamemBERT is camembert-base, FlauBERT is flaubert-base-cased. **FT** : with or without fine-tuning, **PCA** : with or without PCA dimensionality reduction (100 components, with whitening). The first two lines are baseline results before fine-tuning, last two lines are averages of 5 runs (std. dev. in parentheses). v subscript indicates a verbs-only dataset.

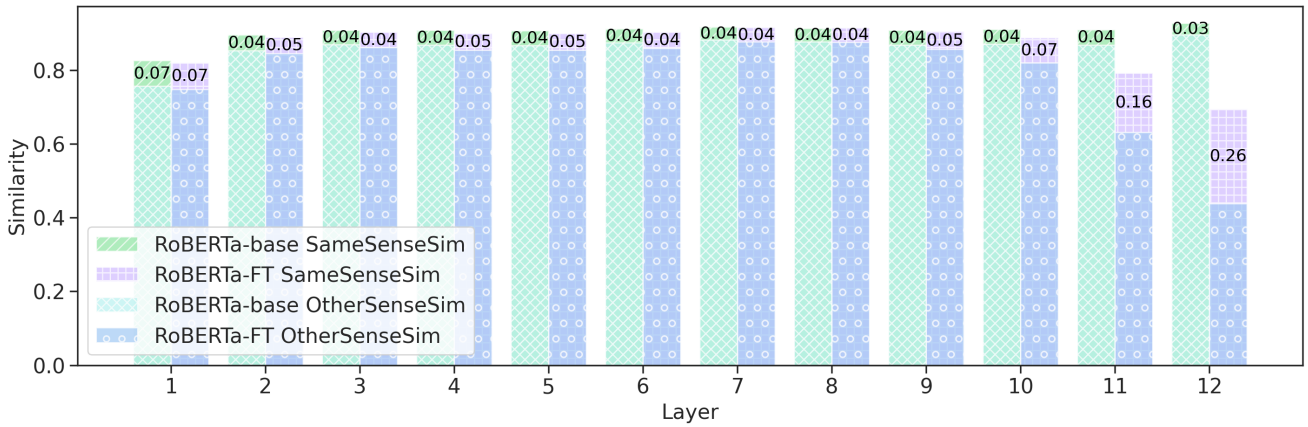

FIGURE 8 – Same-Sense Similarity and Other-Sense Similarity computed before and after fine-tuning on the roberta-base model on the English verb dataset. The metrics are computed on Wiktionary development dataset. The value inside column corresponds to the difference between SameSenseSim and OtherSenseSim.