Disentangling lexical and grammatical information in word embeddings

Li Liu
OLST, Université de Montréal
Montréal, Canada
li.liu.2@umontreal.ca

François Lareau
OLST, Université de Montréal
Montréal, Canada
francois.lareau@umontreal.ca

Abstract

To enable finer-grained linguistic analysis, we propose a method for the separation of lexical and grammatical information within contextualized word embeddings. Using CamemBERT embeddings for French, we apply our method to 14,472 inflected word forms extracted from the French Lexical Network (LN-fr), covering 1,468 nouns, 202 adjectives, and 299 verbs inflected via 14 distinct grammatical feature values. Our iterative distillation process alternates two steps until convergence: (i) estimating lexical or grammatical vectors by averaging the embeddings of words that share the same lexeme or grammatical feature value, and (ii) isolating the complementary component of each word embedding by subtracting the estimated vector. To assess the quality of the decomposition, we measure whether the resulting lexical and grammatical vectors form more compact clusters within their respective groups and whether their sum better reconstructs the original word embeddings. All evaluations rely on euclidean (L2) distance. The observed improvements in both clustering and reconstruction accuracy demonstrate the effectiveness of our approach.

1 Introduction

Static word embeddings, such as those generated by *word2vec* (Mikolov et al., 2013b,a), assign a single, fixed vector to each word form based on its general contextual usage. This approach conflates distinct meanings of polysemous words or homonyms and fails to capture morphological compositionality, as it does not model how word forms may share a common core lexical meaning or how affixes encode grammatical features. For morphologically rich languages, this entanglement can hinder fine-grained linguistic analysis. Previous work, such as Lareau et al. (2015), addressed this issue by proposing a method to decompose static embeddings into lexical and inflectional components, aiming to obtain semantically purer representations.

Contextualized embeddings from pretrained language models such as BERT (Devlin et al., 2019) produce dynamic, context-sensitive vectors that implicitly encode a range of linguistic information, including morphology and syntax. This has substantially improved the modeling of polysemy, homonymy, and morphosyntactic variation compared to static embeddings. However, it remains unclear how lexical and grammatical features are represented within these embeddings and whether they can be meaningfully disentangled. In this paper, we revisit the problem of separating lexical and grammatical information in word embeddings, focusing on embeddings produced by CamemBERT (Martin et al., 2020) for French, a language with a relatively rich morphology.

This work was originally motivated by a separate study where we aimed to measure the semantic idiomaticity of French idioms. Semantic idiomaticity refers to the extent to which the meaning of an idiom cannot be inferred from its component words. While CamemBERT is able to distinguish free simple lexemes from words within idioms, it struggles with component words within idioms of different levels of semantic idiomaticity (Liu and Lareau, 2024). This suggests that the model captures idiomaticity at a superficial lexical level, but is not sensitive to the internal semantic structure of idioms. We hypothesized that this limitation is due to the entanglement of multiple types of idiomaticity, not only semantic, but also morphological and syntactic. In order to isolate purely semantic meaning from grammatical interference, we turned to the problem of disentangling lexical and grammatical components in contextual embeddings. The current study develops and evaluates a method for this task, inspired by the methodology proposed by Lareau et al. (2015).

We assume that a word embedding can be modeled as the linear combination of two components, a lexical vector capturing its core lexical meaning,

and a grammatical vector encoding morphosyntactic information. Therefore, we should be able to isolate one component by subtracting the other. Our method relies on two assumptions:

1. All inflected forms of a lexeme share a common core lexical meaning.
2. All words inflected via the same grammatical feature value (i.e., all plural nouns, or all feminine adjectives) share a common grammatical meaning, regardless of allomorphy.

Under this framework, the lexical vector of a lexeme can be estimated either directly, by averaging the embeddings of its inflected forms, or indirectly, by subtracting a shared grammatical vector from each word embedding. Likewise, grammatical vectors associated with specific feature values can be derived by averaging over relevant word embeddings or by removing lexical content.

To obtain more accurate and disentangled representations, we develop an iterative distillation process that integrates both estimation strategies. At each step, one component is isolated by subtracting the current estimate of the other, then refined by averaging over the pertinent group of words (e.g., all inflected words of a lexeme, or all words sharing a grammatical feature value). This process incrementally improves both components over successive iterations.

We hypothesize that, after distillation, lexical vectors belonging to the same lexeme on the one hand, and grammatical vectors sharing the same feature value on the other, get closer in the vector space. We evaluate this by comparing the average pairwise L2 distances within each group before and after distillation. We also assess the reconstruction accuracy of the original embeddings by measuring the difference between each embedding and the sum of its distilled lexical and grammatical components.

In this study, we focus specifically on inflection and leave aside derivation, as it is often non-compositional. We worked on French because it has a sufficiently rich morphology for it to be non-trivial, and we had access to the data we needed. However, our method is language-agnostic, and such data is relatively easy to come by for a variety of languages.

2 Related work

Recent studies have highlighted that contextualized word embeddings encode various types of linguistic information in a high entangled form (López-Otal et al., 2025; Ravfogel et al., 2020). This has sparked growing interest in disentangling grammatical information. However, most existing work addresses this challenge in the context of downstream tasks or model performance, rather than focusing on extracting grammatically meaningful representations for linguistic analysis (Huang et al., 2021; Li et al., 2021; Chen et al., 2019; Ravfogel et al., 2020; Omrani Sabbaghi and Caliskan, 2022).

To our knowledge, few studies have explicitly addressed this question from the perspective of linguistic analysis. The work most closely related to ours that we know of is by Lareau et al. (2015), who developed a method applied to decompose static *word2vec* embeddings in English. Their approach, based on averaging and subtraction, was tested on a small-scale dataset of around 20 verbs, with a primary focus on lexical vectors. Their method struggled with homonyms due to the static nature of *word2vec* embeddings. In contrast, our approach leverages contextualized embeddings, which mitigate this issue. It is also applied to a much larger and more diverse natural corpus. While inspired by their methodology, we extend it with an iterative refinement process and expand the analysis to include grammatical vectors as well. In addition, we introduce a broader set of evaluation metrics.

3 Experiment

3.1 Data

For our experiment, we used data from French Lexical Network (LN-fr) v3 (Polguère, 2009; Lux-Pogodalla and Polguère, 2011; Polguère, 2014; ATILF, 2023), an open-access lexical database manually developed according to the methodological principles of explanatory combinatorial lexicology (Mel'čuk, 2006). Each entry in LN-fr represents a disambiguated lexical unit in French, corresponding to a distinct and well-defined sense of a simple lexeme or an idiom. In our study, we focused exclusively on simple lexemes (hereafter referred to as lexemes). Each lexeme has a part of speech (POS) tag; since we studied inflectional types in French, we extracted only the nouns, adjectives, and verbs, other classes being invariant.

Each lexeme is associated with one or more lexicographic examples sourced from corpora. These examples were carefully selected by lexicographers to reflect real-world usage, showcasing the syntax, semantics, and combinatorial properties of the

lexemes (Lux-Pogodalla, 2014). Furthermore, the annotation explicitly identifies the position of the words corresponding to the lexeme within each example. These words represent inflected forms of the lexeme in the sentence. A single lexeme may be associated with multiple words within an example. This can occur through repetitions or analytic forms (such as past tense, e.g., *ai mangé* '(I) have eaten'). To simplify the analysis, such cases were excluded. Only examples containing a single word corresponding to a lexeme were retained.

Most grammatical features are not annotated in LN-fr, with the exception of number and gender. We therefore used Stanza (Qi et al., 2020) to analyze the examples of the lexemes and complete the annotation of their remaining grammatical features. For number and gender, we compared the LN-fr annotations with those produced by Stanza and found them to be fully consistent. Given that nouns and adjectives, the main categories marked for these features, account for over 86% of our data, this consistency supports the reliability of Stanza for morphological annotation and indicates that our method should work even without annotated data. To further reduce potential errors, we compared the POS tags and lemmas returned by Stanza with the manual annotations in LN-fr, removing 230 lexemes where the POS assignments did not match.

We generated word embeddings for lexemes in our data using CamemBERT (Martin et al., 2020) for our experiment. CamemBERT is a pretrained contextualized language model for French, where each token in the sentence is represented differently depending on the other tokens in the context. We used the representations from the last layer. For words tokenized into sub-word tokens, we sum all sub-word embeddings to get the word's embedding. We used example sentences retrieved from LN-fr as context and generated vectors that represent the inflected forms of lexemes. We considered only lexemes with at least four examples in the database. This ensures more stable and representative lexical embeddings by averaging over multiple contexts and helps achieve better coverage of a lexeme's inflectional paradigm. Moreover, to reduce model-internal bias, we applied mean-centering to all embeddings, removing common components unrelated to lexical or grammatical distinctions.

In total, we extracted from LN-fr nearly 2,000 lexemes, with significantly more nouns than adjectives or verbs. Each lexeme is accompanied by its word forms, example sentences, along with corresponding word embeddings and annotated grammatical information. Table 1 summarizes the number of words associated with each feature value within each grammatical category. We group nouns, adjectives, and verbs according to the grammatical categories they express: nouns by number, adjectives by number and gender, and verbs by number, tense, mood, finiteness, voice, gender and person. Words lacking relevant annotation are excluded from the count, and only feature values with at least 200 words across lexemes are retained to ensure sufficient data for reliable analysis.

	Lexemes	**Words**
Noun	**1468**	**11159**
sing		8716
plur		2443
Adjective	**202**	**1344**
sing		1038
plur		306
masc		837
fem		507
Verb	**299**	**1969**
sing		834
plur		359
pres		903
imp		245
ind		1159
inf		715
fin		1182
per-3		1056
Total	**1969**	**14472**

Table 1: Lexemes and words (counted by grammatical feature values) in our dataset. Abbreviations for grammatical feature values: *sing*=singular, *plur*=plural, *masc*=masculine, *fem*=feminine, *pres*=present, *imp*=imperfect, *ind*=indicative, *inf*=infinitive, *fin*=finite, *per-3*=third person.

3.2 Methodology

Relying on the assumptions outlined in §1, we propose an iterative distillation method to decompose word embeddings into two components: a **lexical vector** representing its core lexical meaning and a **grammatical vector** encoding morphosyntactic information.

Let L denote a lexeme with n observed inflected forms $\{w_1, w_2, \ldots, w_n\}$. These forms share a common lexeme vector, estimated by the average of their word embeddings:

$$\mathbf{L} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}_i$$

Similarly, for a grammatical feature value G that appears in m words $\{w_1, w_2, \ldots, w_m\}$ in our dataset, we define a shared grammatical vector as:

$$\mathbf{G} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{w}_i$$

For example, the lexeme *eat* includes words such as *eat*, *eats*, *ate*, etc., whose embeddings share a lexical component $\overrightarrow{eat}$. The nominal number feature PLUR applies to words like *apples*, *bikes* and *houses*, whose embeddings share a grammatical component $\overrightarrow{\text{PLUR}_\text{N}}$.

Let w be a word comprising a lexical base l and a grammatical feature value g, its word embedding can be approximated as the sum of two components:

$$\mathbf{w} \approx \mathbf{l}_w + \mathbf{g}_w$$

To obtain purer lexical and grammatical embeddings, we apply two update steps:

Lexical vector update For each w of a lexeme L, we initialize its local grammatical vector g_w using its feature value vector $\mathbf{G}$. Then we subtract this grammatical vector to isolate its current lexical vector l_w. For each lexeme L, we average all its words' l_w to update the lexeme vector $\mathbf{L}$.

$$\mathbf{l}_w = \mathbf{w} - \mathbf{G}, \quad \mathbf{L} \leftarrow \frac{1}{|L|} \sum_{w \in L} \mathbf{l}_w$$

Grammatical vector update Likewise, for each w, we can initialize its local lexical vector l_w using its lexeme vector $\mathbf{L}$. Then by subtracting this l_w from the word embedding, we get the word's current grammatical vector g_w. The average of g_w of all words inflected via G is calculated to update $\mathbf{G}$.

$$\mathbf{g}_w = \mathbf{w} - \mathbf{L}, \quad \mathbf{G} \leftarrow \frac{1}{|G|} \sum_{w \in G} \mathbf{g}_w$$

Our approach is an iterative process that alternates between the two updates, where the output of one step serves as the input for the next. The process continues until the difference between successive updates becomes negligible—typically after just five or six iterations.

To distill lexical or grammatical vectors of a word, we can either start with the lexical vector update by estimating $\mathbf{G}$, or with the grammatical vector update by estimating $\mathbf{L}$. As subtraction is central to both steps, the quality of the initial estimate is critical: any noise in the subtracted vector propagates into the result. Thus, the more accurate the initial estimate, the better the decomposition. It should be stressed that this initial estimate is not random; it is the mean of a set of vectors, and thus yields the same result every time.

We find that initializing with grammatical vectors is more robust. When estimating a grammatical vector $\mathbf{G}$ for a feature value (e.g., PLUR_N), we average the embeddings of all words (e.g., *apples*, *bikes*, *houses*, etc.) that share this feature. Since these words are typically lexically diverse, their lexical components tend to cancel each other out, resulting in a relatively clean approximation of the grammatical meaning. While estimating a lexeme vector $\mathbf{L}$ (e.g., for lexeme *eat*), we average a small number of words inflected from the lexeme (e.g., *eat*, *eats*, *ate*). These words often differ in grammatical properties and appear in different contexts, introducing noise that can distort the estimate of their core lexical meaning.

Furthermore, we extend this idea to handle multi-feature grammatical information, which is common in French. While nouns typically carry only one grammatical feature NUMBER (SING or PLUR), adjectives and verbs express multiple features simultaneously. Specifically, adjectives reflect both number and gender, while verbs can encode tense, mood, person, number, etc. When extracting a word's lexical vector, we remove a composite grammatical vector that corresponds to the full set of feature values it carries. This vector is estimated by averaging the embeddings of all words sharing the exact same feature combination. Conversely, in extracting grammatical vectors, we isolate each feature value independently. For example, to estimate the vector for present tense, we use all verb forms that express the present tense, regardless of their other grammatical properties. This targeted averaging provides a clearer estimate of the intended grammatical dimension.

In summary, our method alternates between subtracting a full grammatical vector to refine lexical vectors, and subtracting the current lexical vector to isolate individual grammatical components. We apply this procedure to verbs, nouns, and adjectives grouped by their feature values, iterating until convergence.

3.3 Evaluation metrics

To assess the effectiveness of this method, we adopt two complementary evaluation metrics.

Internal distance Our evaluation is grounded in the assumptions outlined earlier: (1) words belonging to the same lexeme differ only in their grammatical components, and (2) words that share the same grammatical components differ primarily in their lexical meaning. From this, we hypothesize that once the grammatical components are removed from the original embeddings, the remaining lexical vectors should form tighter clusters within each lexeme group, because what is left should be close to the naked lexical information. Similarly, if the lexical component is subtracted, the remaining grammatical vectors should show greater internal consistency within each feature value group. To verify this, we measure the internal compactness of each group before and after distillation.

For each lexeme, we calculate the average pairwise distance among the embeddings of its inflected words prior to distillation. We then repeat the measurement using only the lexical components l_w extracted from these words after distillation. A reduction in distance suggests that the lexical content has been more effectively isolated.

Similarly, for each feature value, we first calculate the average pairwise distance among the original embeddings of all words marked with that value. We then calculate the same measure using only the grammatical vectors g_w corresponding to the feature value isolated from those words. A tighter clustering in this space would indicate that the shared morphosyntactic property has been captured more clearly.

As a baseline, we compute the average pairwise distance between random word pairs that do not share either a lexeme or any grammatical feature values, applying the same subtraction procedure. For each run, we sample up to 10,000 such random pairs; if the total number of admissible pairs is smaller, we use all available pairs. We repeat this process 10 times and report the mean across runs. Since these words are unrelated in both lexical and grammatical dimensions, their vectors should not become closer after distillation. This allows us to verify that any observed distance reduction in groups defined by shared lexemes or feature values is not merely an artifact of the subtraction process, but reflects meaningful linguistic structure.

Reconstruction accuracy We evaluate whether the lexical and grammatical components can faithfully reconstruct the original word embeddings. For each word, we compare its original embedding with two reconstructed versions: one using the initial estimates of its lexical vector and the grammatical vectors corresponding to its set of feature values, and another using the distilled vectors obtained. A lower reconstruction error in the latter case implies improved preservation of the original embeddings' structure.

3.4 Distance metric

Both the distillation process and the subsequent evaluation require a way to quantify how the vectors change under our distillation method. To compare these vectors, we initially calculated both cosine similarity and L2 distance.

Cosine similarity is commonly adopted as a metric for semantic similarity in natural language processing (NLP), as it captures the angular relationship between vectors while ignoring their magnitude. However, our method involves vector subtraction, which can substantially alter both direction and length. This makes cosine similarity potentially misleading: in extreme cases, two vectors may retain the same angle (i.e., yield a high cosine similarity) while differing greatly in magnitude, making them appear semantically close even when they are not. Previous studies have also shown that cosine similarity can be distorted in contextualized embedding models due to anisotropy and frequency effects (Ethayarajh, 2019; Timkey and van Schijndel, 2021; Zhou et al., 2022).

In our evaluation, cosine similarity and L2 distance often led to divergent conclusions. Since L2 distance captures both angular and magnitude-related differences, we consider it to be a more reliable indicator of the structural changes introduced by our method. Furthermore, we found no strong theoretical reason to prefer cosine similarity in our setting beyond its popularity in previous work. Given these considerations, we focus exclusively on L2 distance in the results reported below.

4 Results and Discussion

In this section, we evaluate whether the lexical and grammatical components extracted from word embeddings display greater internal consistency after distillation. In each comparison, we measure the target evaluation metric before and after the procedure. In all result tables, the *before* column reports results calculated using the original embeddings, while the *after* column shows results based on distilled vectors. The relative change (Δ) is calculated

as $\frac{after-before}{before}$, reflecting the proportion of the resulting change. All reported results are rounded to two decimals.

4.1 Lexical vectors become more consistent after grammatical removal

Table 2 reports the results of the evaluation of the lexical vector distilled by removing the grammatical component(s).

For each lexeme in our dataset, we select inflected words that differ in grammatical features and calculate the average pairwise L2 distance between their original embeddings. This distance serves as a measure of internal lexical dispersion prior to distillation. We use the same measure on lexical vectors derived after removing grammatical components. If the grammatical information has been successfully removed, the resulting lexical vectors should exhibit lower internal dispersion. We exclude lexemes with identical feature values, as their embeddings are already highly similar in the original space; the procedure would yield negligible effect.

As shown in Table 2, we observe a consistent decrease in distance across all lexical categories in the range of around 8% to 14%. This indicates that the distilled lexical vectors are more tightly clustered, supporting the effectiveness of our distillation method across different parts of speech.

For baseline comparison, we evaluate random word pairs drawn from different lexemes that differ in feature values. These words are not expected to share a semantic content, so removing grammatical information should not significantly reduce their distance. Indeed, the random groupings of the same part of speech exhibit notably smaller reductions, with average decreases reaching only about half of those observed in lexeme-aligned groups. This confirms that the increased compactness observed in structured lexeme groups reflects meaningful decomposition rather than trivial consequence of mean subtraction or vector manipulation.

4.2 Grammatical vectors get closer after distillation

In addition to lexical coherence, we also evaluate the internal consistency of the grammatical vectors, with results presented in Table 3.

For each feature value, we find words from distinct lexemes that share this value. The average pairwise L2 distance between their original embeddings measures how dispersed these words are before distillation. We then compute that distance using the grammatical vectors extracted after removing lexical components. If the subtraction is effective, these vectors should converge toward a representation of the shared grammatical property, lowering the average distance. Again, we omit tokens from the same lexeme to avoid trivial reductions stemming from shared lexical information.

As shown in Table 3, the grammatical vectors exhibit a substantial reduction in pairwise distance across all feature values within all grammatical categories, ranging from approximately 31% to 45%. This level of reduction is markedly higher than what we observed for lexical vectors. This stark contrast suggests that grammatical information is more effectively disentangled. A possible explanation lies in the structural difference between the comparison groups in each evaluation. In the lexical vector analysis, we compare tokens from the same lexeme that differ only in grammatical features. Such tokens already occupy relatively close positions in the embedding space even before distillation, leaving limited room for further convergence. By contrast, in the grammatical vector analysis, the compared tokens share a grammatical feature but are from different lexemes, are therefore initially more widely dispersed. After the lexical component is removed, this dispersion is greatly reduced, as the remaining grammatical vectors align more closely around the shared grammatical property. Moreover, grammatical features are often shared by a larger number of tokens than individual lexemes, making the averaged estimates for grammatical vectors more robust.

To establish a control, we measure distances between randomly sampled words that differ in both lexeme and feature value. Since such pairs are not expected to encode common grammatical information, their grammatical vectors should remain dispersed. This comparison ensures that the observed distance reductions in feature-based groups cannot be explained by vector subtraction alone. Table 4 shows reductions in distance consistently small across all grammatical categories, less than 10%. These values are markedly lower than those observed in structured feature-based groups (cf. Table 3), confirming that the substantial convergence seen reflects the extraction of meaningful shared grammatical information.

	Lexemes			Random lexemes		
	before	**after**	**Δ**	**before**	**after**	**Δ**
Noun	5.30	4.71	-10.32%	6.70	6.41	-4.24%
Adjective	4.72	4.34	-7.97%	6.78	6.50	-4.03%
Verb	5.66	4.82	-14.07%	7.14	6.48	-9.22%

Table 2: Pairwise distance of lexical vectors before and after distillation. Results are computed over word pairs from the same lexeme, as well as random word pairs from different lexemes, all inflected with different grammatical feature values.

	before	**after**	**Δ**
Noun			
sing	6.08	3.33	-45.18%
plur	7.34	4.23	-42.31%
Adjective			
sing	6.33	3.62	-42.85%
plur	7.15	4.35	-39.12%
masc	6.63	3.83	-42.19%
fem	6.54	3.88	-40.70%
Verb			
ind	7.09	4.40	-37.89%
per-3	7.02	4.36	-37.88%
sing	6.71	4.13	-38.50%
plur	7.57	4.70	-37.91%
pres	6.66	4.09	-38.51%
imp	7.81	4.77	-38.86%
inf	5.93	3.65	-38.36%
fin	7.06	4.81	-31.94%

Table 3: Pairwise distance of grammatical vectors before and after the distillation, calculated over word pairs that share the same grammatical feature value but originate from distinct lexemes.

	before	**after**	**Δ**
N-number	7.10	6.73	-5.20%
Adj-number	7.06	6.77	-4.14%
Adj-gender	6.67	6.55	-1.76%
V-mode	8.07	7.29	-9.66%
V-person	7.36	7.25	-1.42%
V-number	7.45	7.15	-4.08%
V-tense	7.75	7.34	-5.27%
V-finiteness	6.99	6.61	-5.42%

Table 4: Pairwise distance of random grammatical vectors before and after distillation, calculated between random word pairs differing in grammatical feature value and lexeme.

4.3 Word embedding reconstruction

Extending the above evaluations, to assess whether the distilled components provide more accurate representations of lexical and grammatical information, we evaluate how well they can reconstruct the original word embeddings. Specifically, we measure the L2 distance between the original embedding of each token and its reconstructed form, obtained by summing its lexical vector and the average of its grammatical vectors corresponding to each of its feature values.

	before	**after**	**Δ**
Noun	2.58	2.52	-2.38%
Adjective	2.89	2.58	-11.40%
Verb	3.50	2.27	-35.28%

Table 5: Reconstruction error of word embeddings from their lexical and grammatical components before and after the distillation. Results are averaged over part of speech.

As a baseline, we first perform reconstruction using the initial, undistilled estimates of lexical and grammatical vectors. These initial estimates are expected to contain overlapping or entangled information, resulting in higher reconstruction error. After distillation, however, the components are refined to better isolate the intended dimensions of meaning, which should lead to more faithful reconstructions.

Our results are reported in Table 5. We observe consistent reductions across all parts of speech. The improvement is most substantial for verbs, with an average reduction of over 35%, while adjectives and nouns show smaller but still meaningful improvements (11.4% and 2.4%, respectively). This pattern may be attributed to differences in morphological complexity and the way grammatical information is distributed across parts of speech. In French, both nouns and adjectives typically mark number and gender using the same surface morphemes (e.g., *-s* and *-x* for PLUR; *-e*,*-euse* and *-trice* for FEM), which are shared between lexemes. While such suffixes are consistent and formally simple, they express only a limited set of grammatical features, and the shared form across categories may blur the information, making it more difficult for the model to disentangle the lexical and grammatical components precisely.

In contrast, French verbs undergo more complex

inflection, where a single suffix often encodes multiple feature values simultaneously. For instance, the ending *-ent* in *ils parlent* ('they speak') marks third person, plural, present tense, and indicative mood all within a single affix. Despite the greater surface complexity, the richness and density of grammatical encoding in verbal morphology may provide a stronger signal, allowing the model to better isolate and represent grammatical content. The more pronounced improvement observed in verbs thus likely reflects this concentrated grammatical structure, which becomes more salient and recoverable after distillation.

5 Conclusion

We aimed to disentangle contextualized word embeddings in CamemBERT into lexical and grammatical parts. We proposed an iterative distillation method based on the complementarity of averaging and subtraction. A word's lexical vector can be approximated either by averaging the vectors of all words that share the same lexical meaning, or by subtracting the vectors corresponding to its grammatical features. Similarly, its grammatical vector can be obtained either by subtracting the lexical part from the original embedding, or by averaging the embeddings of all words that share the same grammatical feature.

If effectively separated, the lexical and grammatical vectors should be more distinct, with minimal overlap between their respective contents compared to their initial estimates. Each vector should convey more clearly the structural regularities shared with similar words, resulting in tighter alignment within their lexical or grammatical groups. As such, they are better suited to jointly approximate the original word embedding. As expected, in our evaluation, the final lexical and grammatical vectors that we extracted are more clearly clustered with their structurally similar counterparts, when combined, reproduce original word embeddings with minimal loss.

Notably, the reduction in distance is much more pronounced for grammatical vectors than for lexical vectors—around 40% versus 10% on average. Since the initial grammatical vectors are averaged over a large set of lexemes and contain great lexical noise, which is removed during distillation, leading to tighter alignment. The extent of this convergence is relatively stable across different feature values, but varies across parts of speech. Verbs, especially, show a stronger reduction in distance and a larger drop in reconstruction error compared to nouns and adjectives. This may be due to the richness of verbal morphology in French, where suffixes often encode several grammatical features at once, making the grammatical signal more prominent and its removal more effective. Nouns show only minor reconstruction gains, likely due to limited grammatical variation from number inflection alone. Also, large number of nouns in our dataset may stabilize their initial estimates, leaving less room for improvement. Adjectives fall in between, showing moderate gains.

Another important factor behind these observations concerns tokenization and the model's sensitivity to morphological markers. In CamemBERT, frequent, short, and morphologically informative tokens are more likely to be consistently encoded or even assigned special status during training (Rogers et al., 2021; Clark et al., 2019; Mohebbi et al., 2021). In contrast, lexical roots often span longer or rarer sub-word tokens and are more prone to being split or distorted, especially in low-frequency contexts. As a result, grammatical information is already more cleanly separated and clearly encoded in the model's internal representations, making it easier to distill effectively. This also explains why verbs, whose suffixes encode multiple features in compact forms, benefit the most from the process.

Future work will further explore how factors such as word frequency and tokenization affect the separation of lexical and grammatical vectors. Our method assumes a linear relationship between lexical and grammatical vectors; in a follow-up study, we plan to explore non-linear relationships. Given that our method does not rely on language-specific morphological rules, we will apply and evaluate it across languages. In addition, we are interested in extending the approach using learning-based methods, and in incorporating morphology-aware tokenizers to improve grammatical representation. Finally, we aim to assess the practical value of our decomposition through downstream tasks.

Limitations

One limitation of our study is data imbalance, which may affect result comparability and robustness. The number of words varies widely across parts of speech and grammatical features: nouns are far more numerous, yet have fewer grammatical features. Some feature values are sparsely rep-

resented, leading to less reliable vector estimates. Lexemes also vary in the number of inflected forms. Due to limited data, certain features such as verbal voice and gender were excluded, making the evaluation less complete.

Source code

This experiment can be reproduced by downloading the data we used and our source code from `https://github.com/liliulng/disentangle-wemb`.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable suggestions. We are also grateful for the financial support of the China Scholarship Council (#202008310177) and the Fonds de recherche du Québec (#366841).

References

ATILF. 2023. Réseau lexical du français (rl-fr). ORTOLANG (Open Resources and TOols for LANGuage)—www.ortolang.fr.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. *arXiv preprint arXiv:1904.01173*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL'19: Human Language Technologies*, volume 1, pages 4171–4186, Minneapolis, MN, USA. ACL.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

James Y Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. *arXiv preprint arXiv:2104.05115*.

François Lareau, Gabriel Bernier-Colborne, and Patrick Drouin. 2015. La séparation des composantes lexicale et flexionnelle des vecteurs de mots. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 242–248, Caen, France. ATALA.

Dingcheng Li, Hongliang Fei, Shaogang Ren, and Ping Li. 2021. A deep decomposable model for disentangling syntax and semantics in sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4300–4310.

Li Liu and Francois Lareau. 2024. Assessing BERT's sensitivity to idiomaticity. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 14–23.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *arXiv preprint arXiv:2504.08001*.

Veronika Lux-Pogodalla. 2014. Integrating lexicographic examples in a lexical network (intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014*, volume 2, pages 586–591, Marseille, France. ATALA.

Veronika Lux-Pogodalla and Alain Polguère. 2011. Construction of a French Lexical Network: Methodological Issues. In *First International Workshop on Lexical Resources (WoLeR 2011)*, pages 54–61, Ljubljana, Slovenia.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219, Online. ACL.

Igor A. Mel'čuk. 2006. Explanatory combinatorial dictionary. In Giandomenico Sica, editor, *Open Problems in Linguistics and Lexicography*, pages 225–355. Polimetrica, Monza, Italy.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems (NIPS 2013)*, 26.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806.

Shiva Omrani Sabbaghi and Aylin Caliskan. 2022. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531.

Alain Polguère. 2009. Lexical systems: graph models of natural language lexicons. *Language Resources and Evaluation*, 43:41–55.

Alain Polguère. 2014. From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, 27(4):396–418.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Jacob Goldberger, and Yoav Goldberg. 2020. Unsupervised distillation of syntactic information from contextualized word representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 91–106, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.