

Evaluating and Debiasing Multilingual Backpack Language Model Representations

Jennifer Jane Ries
Columbia University
jr3954@columbia.edu

Kavya Venkatesh
Columbia University
kv2458@columbia.edu

Abstract

Backpack language model embeddings represent each word as a weighted sum of sense vectors, offering more interpretable and intervenable lexical representations than standard contextual embeddings. Backpack sense vector representations show strong performance on lexical similarity tasks in a monolingual setting, as well as cross-lingual transfer through finetuning on aligned parallel multilingual data. Our work investigates whether representations from Backpack models trained from scratch on parallel aligned multilingual data allow for interpretable control without sacrificing translational and cross-lingual representation learning ability. We train matched 131M parameter Transformer and 132M parameter Backpack models from scratch on parallel aligned English-French EuroParl and evaluate their expressivity using validation perplexity, their learned monolingual and cross-lingual lexical relationships using Multi-SimLex and their translational ability using sentence-level cosine similarity. While the Transformer attains higher English and cross-lingual Multi-SimLex, the Backpack outperforms the Transformer on French Multi-SimLex and shows stronger translated-vs-random sentence separation. We finally compare sense-level gender bias ablations on the Backpack model with nullspace projection debiasing on the Transformer, demonstrating the effectiveness of sense vector control for debiasing multilingual representations.

1 Introduction

Cross-lingual lexical similarity evaluation assesses whether language model embeddings place translationally related words in similar regions of representation space while still maintaining fine-grained semantic distinctions (Vulić et al., 2020). This question is central to multilingual NLP, as multilingual encoders must align meaning across languages

with different polysemy patterns, semantic structures, and cultural contexts (Libovický et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). However, the opaque nature of contextual embeddings makes it difficult to systematically evaluate, let alone manipulate, the organization of their cross-lingual lexical representations.

Hewitt et al. (2023)’s Backpack model addresses the lack of control inherent in contextual embeddings, without sacrificing embedding expressivity, in a monolingual context. Backpack language models address this interpretability gap by decomposing each token into a weighted combination of shared sense vectors (Hewitt et al., 2023).

For monolingual English data, Hewitt et al. (2023) find the Backpack’s learned word embeddings’ performance on lexical similarity tasks to be better than those of a Transformer without Backpack embedding architecture. Sun and Hewitt (2023) likewise find improved performance on lexical similarity tasks relative to a matched Transformer baseline for monolingual character-level Mandarin Chinese data.

Training on parallel aligned multilingual data allows for strong cross-lingual lexical transfer in language models such as multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Moreover, Cruz et al. (2026) show the occurrence of cross-lingual transfer in aligned multilingual Backpack representations produced through fine-tuning English Backpack models on parallel translated sentence pairs.

Beyond evaluating expressivity and performance on lexical similarity tasks, Hewitt et al. (2023) and Sun and Hewitt (2023) demonstrate the effectiveness of using the Backpack structure for interpretable intervention in the form of sense-level debiasing. Each show that using sense vector ablations to reduce learned gender bias in Backpack representations of profession words is more effective than the Transformer nullspace projection debi-

asing method (Bolukbasi et al., 2016) under similar conditions.

Whereas Hewitt et al. (2023) and Sun and Hewitt (2023) mitigate gender bias by fully ablating a given sense identified as having a strong contribution to gender bias, Afzali et al. (2026) show the effectiveness of sense-reweighting, where they identify two such senses and reduce their weights in the calculation of the full representation, so as to simultaneously reduce gender bias and retain information from all sense vectors.

In this work, we explore Backpack model expressivity, performance on lexical similarity tasks, and interpretable gender debiasing in a multilingual context. To examine cross-lingual transfer in Backpack sense vectors, we train a 131M parameter Transformer and a 132M parameter Backpack from scratch on parallel interleaved English-French EuroParl (Koehn, 2005) for 100,000 iterations with matched hyperparameters. We evaluate both models’ expressivity using validation perplexity and cross-lingual representation learning using monolingual and cross-lingual Multi-SimLex (Vulić et al., 2020) and comparisons between sentence-level cosine similarity between translation pairs and randomly paired sentences, adapted from Hewitt et al. (2023). In addition, we perform sense-level debiasing ablations in the style of Hewitt et al. (2023) and Sun and Hewitt (2023), adapted to our multilingual Backpack model.

The Transformer achieves a lower perplexity on a held-out validation set of Europarl, 29.8 compared to 35.5 for the Backpack, and performs better on the cross-lingual Multi-SimLex lexical similarity task (for EN-FR: $\rho=0.109$, $p < 0.001$ compared to the Backpack’s $\rho=0.066$, $p = 0.004$, and for FR-EN: $\rho=0.143$, $p < 0.001$ compared to the Backpack’s $\rho=0.123$, $p < 0.001$). However, the Backpack shows slightly stronger translated-vs-random sentence separation, $\Delta\mu=0.045$ compared to $\Delta\mu=0.034$ for the Transformer.

Moreover, our experiments find Backpack sense vector ablations to be more effective at gender debiasing than Transformer embedding nullspace projection, suggesting that the fine-grained control of Backpack representations Hewitt et al. (2023), Sun and Hewitt (2023), and Afzali et al. (2026) demonstrate in monolingual contexts is also present in a multilingual context.

2 Related Work

Past work has developed methods, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), for generating non-contextual word embeddings. In these non-contextual embeddings, a word is represented by a singular vector, regardless of whether the word’s meaning can change based on its context.

Contextual word embeddings, in which a given word can be associated with multiple vectors, each representing its meaning in a different context, are central to Vaswani et al. (2017)’s Transformer architecture, though they predate it (e.g., in recurrent language models (Elman, 1990)). Vaswani et al. (2017)’s highly expressive Transformer architecture builds contextual representations by applying nonlinear transformations to input representations. However, though the expressivity of Transformers is desirable, the repeated application of nonlinearities to the input representation means that their contextual word embeddings do not respond predictably to intervention, for example, for debiasing a model’s generated text (Hewitt et al., 2023).

Hewitt et al. (2023) propose the Backpack architecture, in which each word is represented as a weighted linear combination of non-contextual sense vectors, allowing for more fine-grained control of meaning representations. In terms of maintaining expressivity, they find its learned word embeddings’ performance on lexical similarity tasks to be better than those of a Transformer without Backpack embedding architecture (Hewitt et al., 2023).

Hewitt et al. (2023)’s work is with English training data. Sun and Hewitt (2023) extend their work, implementing a character-tokenized Chinese Backpack Model, finding that Backpack character embeddings perform better on lexical similarity tasks than character embeddings produced by a Transformer without Backpack embedding architecture.

Multilingual language models like multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) learn cross-lingual representations through training on parallel multilingual data, and have shown strong cross-lingual lexical transfer in their representations (Libovický et al., 2019). However, they lack the interpretable sense-level structure of Backpack models. Cruz et al. (2026) induce cross-lingual transfer in multilingual Backpack representations through finetuning on parallel translation data.

Lexical similarity benchmarks such as SimLex-999 (Hill et al., 2015) and Multi-SimLex (Vulić et al., 2020) can be used to measure learned lexical similarities as a proxy for the caliber of a model’s within-language and cross-lingual lexical representations. On monolingual Multi-SimLex, strong static embeddings reach $\rho \approx 0.55\text{--}0.61$ (fastText), whereas on cross-lingual EN-FR, pretrained M-BERT reaches $\rho \approx 0.40$ and aligned Cross Lingual Word Embedding methods reach $\rho \approx 0.66\text{--}0.71$ (Vulić et al., 2020). We use these numbers only as external reference points given that our models are trained on EuroParl, which, as parliamentary data, is limited in scope, so direct comparison to fully pretrained or post-processed models is intentionally limited.

In addition to their results supporting the maintenance of expressivity in Backpack language models, Hewitt et al. (2023), Sun and Hewitt (2023), and Afzali et al. (2026) each show the efficacy of leveraging the sense structure of the Backpack architecture for debiasing, specifically through fully or partially ablating senses showing a strong contribution to the overall gender bias of a given representation.

3 Approach

3.1 Backpack Model Architecture

3.1.1 Probabilistic Backpack Model

Hewitt et al. (2023) modify the standard Transformer architecture at the embedding layer. For the sequence $x_{1:i}$ in vocabulary V with the Backpack representation o_{i-1} and $E \in \mathbb{R}^{d \times |V|}$, the probabilistic Backpack model is defined as follows:

$$p(x_i | x_{<i}) = \text{softmax}(E^\top o_{i-1})$$

Instead of assigning each word $x \in V$ a single embedding vector $e_x \in \mathbb{R}^d$, Backpack embeddings break word representations down into a combination of sense vectors. This makes Backpack representations more interpretable than contextual embeddings, as each sense can be analyzed individually. Moreover, each sense might encode information about a particular semantic feature or features.

In addition, we hypothesize that, because senses are shared across the vocabulary, in a multilingual setting, cross-lingual transfer may occur in sense vectors. This is particularly useful in multilingual language modelling, where we want lexical representations of words from different languages with

shared meaning (e.g., *parliament/parlement*) to be close to each other in representation space.

3.1.2 Sense vectors

As opposed to single-vector word representations like word2vec (Mikolov et al., 2013), given a vocabulary V , Backpack models learn k sense vectors

$$C(x)_1, \dots, C(x)_k$$

with

$$C : V \rightarrow \mathbb{R}^{k \times d}$$

for every word $x \in V$.

3.1.3 Weighted sum

The representation for word x_i in $x_{1:i}$ is the weighted sum of the sense vectors of the words in the sequence.

Given contextualization weights $\alpha \in \mathbb{R}^{k \times n \times n}$ and the contextualization function for sequence $x_{1:n}$

$$\alpha = A(x_{1:n}),$$

where

$$A : V^n \rightarrow \mathbb{R}^{k \times n \times n},$$

Element x_i in sequence $x_{1:n}$ is represented by the following expression:

$$o_i = \sum_{j=1}^n \sum_{\ell=1}^k \alpha_{\ell ij} C(x_j)_\ell.$$

3.2 Parameterization

3.2.1 Senses

We use the following sense function C , based on Hewitt et al. (2023) and Sun and Hewitt (2023):

$$C(x) = FF(E_x)$$

3.2.2 Context weights

Following the example of Hewitt et al. (2023) and Sun and Hewitt (2023), we use a Transformer to parameterize our contextualization weights:

$$h_{1:n} = \text{Transformer}(E_{x_{1:n}})$$

$$\alpha_\ell = \text{softmax}\left(h_{1:n} K^{(\ell)\top} Q^{(\ell)} h_{1:n}^\top\right)$$

for senses ℓ , and matrices $K^{(\ell)}, Q^{(\ell)} \in \mathbb{R}^{d \times d/k}$

3.3 Baselines

3.3.1 Transformer baseline

We adapt [Sun and Hewitt \(2023\)](#)’s baseline methodology. Our baseline is a StandardTransformerLM with the same depth, width, optimizer, data, and tokenizer as the Backpack model. The only architectural difference is the embedding layer, where the Transformer uses standard token embeddings for lexical representation whereas the Backpack embeddings represent the weighted sum of sense vectors. Both models use the XLM-RoBERTa-base tokenizer ([Conneau et al., 2020](#)) with vocabulary size 250,002, and are trained from scratch on interleaved EN-FR EuroParl ([Koehn, 2005](#)).

3.3.2 Training parity

We use identical hyperparameters where possible (Table 1): 6 layers, 6 heads, embedding dimension 384, context length 512, batch size 32, AdamW with learning rate 3×10^{-4} , and 100,000 training iterations for both models. Checkpoints are selected by best validation loss and evaluated using the same evaluation pipeline. Results in Table 4 are reported from this shared evaluation code. We do not claim parity with pretrained multilingual BERT ([Devlin et al., 2019](#)), XLM-RoBERTa ([Conneau et al., 2020](#)) or aligned fastText Cross Lingual Word Embedding systems ([Vulić et al., 2020](#)). These models are cited only to contextualize the absolute scale of our Multi-SimLex scores.

4 Experiments

4.1 Data

Informed by the use by [Cruz et al. \(2026\)](#) of datasets of aligned sentence translation pairs to facilitate cross-lingual transfer in Backpack representations, we use the EuroParl French-English parallel corpus ([Koehn, 2005](#)), which contains aligned sentences from European Parliament proceedings in multiple languages. We focus on the English-French language pair, using approximately 50,000 parallel sentence pairs. The data is split 90/10 for training and validation.

To create cross-lingual alignment, we interleave parallel sentences with language separators (e.g., <EN> and <FR> tags), creating sequences that look like <EN> English sentence <FR> French sentence. We also include reverse-order pairs to encourage bidirectional alignment learning. All text is tokenized using XLM-RoBERTa tokenizer ([Conneau et al., 2020](#)) with a vocabulary

size of 250,002 tokens, including 250,000 vocab tokens, ‘<EN>’ and ‘<FR>’.

4.2 Experimental details

Implementation is in PyTorch using XLM-RoBERTa tokenization ([Conneau et al., 2020](#)).¹

Both the Backpack and Transformer are trained from scratch on the EuroParl English–French parallel corpus ([Koehn, 2005](#)) for 100,000 iterations with matched hyperparameters (Table 1). Reported results use the best validation-loss checkpoint from each run. All models are trained with gradient clipping (max norm 1.0), dropout (0.1), and mixed precision training (float16) for efficiency.

Setting	Transformer	Backpack
Parameters	131.18M	132.43M
Layers	6	6
Heads	6	6
Embedding dim	384	384
Senses (K)	–	16
Context length	512	512
Batch size	32	32
Learning rate	3×10^{-4}	3×10^{-4}
Optimizer	AdamW	AdamW
β_1	0.9	0.9
β_2	0.95	0.95
Weight decay	1×10^{-1}	1×10^{-1}
Max iterations	100,000	100,000

Table 1: Model configurations and training hyperparameters

4.3 Evaluation methods

4.3.1 Multi-SimLex word-level similarity

Adapted from the monolingual lexical similarity evaluation using SimLex999 ([Hill et al., 2015](#)) by [Hewitt et al. \(2023\)](#), we compute cosine similarity between the embedding representations of word pairs, then compute the Spearman rank correlation ρ between the embedding cosine similarity scores and human similarity scores in Multi-SimLex ([Vulić et al., 2020](#)).

In the Backpack embedding representation for word x , the embedding information is spread over k sense vectors $C(x)_1, \dots, C(x)_k \in \mathbb{R}^d$. We form

¹Our code is available at <https://github.com/kavyavenk/multilingual-backpacks/>

a single word embedding by averaging the senses:

$$\tilde{e}_w = \frac{1}{k} \sum_{\ell=1}^k C(w)_\ell.$$

We center and ℓ_2 -normalize embedding vectors before computing cosine similarity. We report Spearman rank correlations between embedding cosine similarity scores and human similarity scores for monolingual English and French Multi-SimLex pairs, as well as cross-lingual EN-FR and FR-EN Multi-SimLex pairs.

4.3.2 Sentence-level similarity

We construct sentence representations using mean pooling, then compute cosine similarity between translation sentence pairs.

For a sentence $x = (x_1, \dots, x_T)$, we define a sentence representation $g(x)$ by mean pooling over word embeddings:

$$g(x) = \frac{1}{T} \sum_{t=1}^T h_t,$$

where h_t is the final-layer hidden state for token x_t . Given M parallel sentence pairs $\{(x_i^{en}, x_i^{fr})\}_{i=1}^M$ from EuroParl, we compute

$$s_i^{(trans)} = \cos(g(x_i^{en}), g(x_i^{fr})).$$

To interpret these scores, we also construct a random-sentence baseline, in which we randomly pair each English sentence x_i^{en} with a mismatched French sentence $x_{\pi(i)}^{fr}$ (where π is a random permutation), and compute

$$s_i^{(rand)} = \cos(g(x_i^{en}), g(x_{\pi(i)}^{fr})).$$

We then report the mean and standard deviation of both distributions, $\mu_{trans}, \sigma_{trans}$ and $\mu_{rand}, \sigma_{rand}$, and compare them.

4.3.3 Sense vector ablations

In order to determine top contributing Backpack sense vectors to bias outcomes, we run sense vector ablation experiments in the style of [Hewitt et al. \(2023\)](#). We perform two separate sense vector ablation experiments: one for English and one for French.

We follow [Hewitt et al. \(2023\)](#)’s example, using:

$$\text{Bias}(x) = \max\left(\frac{P(\text{he} | x)}{P(\text{she} | x)}, \frac{P(\text{she} | x)}{P(\text{he} | x)}\right)$$

to determine whether the Backpack has learned a gender bias towards either male or female for a given occupation ($\text{Bias}(x) = 1$ indicates completely gender-balanced predictions).

In order to account for changes in bias score as a result of the prompt template, we include 5 different templates which we use for each profession. Thus each profession’s individual bias score is the average over the bias scores produced using each template. These English prompts are shown in Table 2.

English Prompts
The [PROFESSION] said that
The [PROFESSION] walked in and said that
I spoke to the [PROFESSION] because
Yesterday, the [PROFESSION] said that
The [PROFESSION] told me that

Table 2: English prompts for evaluating gender bias of professions

We modify our list of templates for the French experiments to match the gendered nature of French grammar. While the templates are direct translations of the English templates, for each we include the construction using ‘le’ and the construction using ‘la’ before each profession name. These French prompts are shown in Table 3.

French Prompts
Le [PROFESSION] a dit qu’
Le [PROFESSION] est entré et a dit qu’
J’ai discuté avec le [PROFESSION] parce qu’
Hier, le [PROFESSION] a dit qu’
Le [PROFESSION] m’a dit qu’
La [PROFESSION] a dit qu’
La [PROFESSION] est entrée et a dit qu’
J’ai discuté avec la [PROFESSION] parce qu’
Hier, la [PROFESSION] a dit qu’
La [PROFESSION] m’a dit qu’

Table 3: French prompts for evaluating gender bias of professions

For the English profession words themselves, we use the set of 40 occupation names from [Hewitt et al. \(2023\)](#). These English profession words are as follows: mechanic, accountant, farmer, baker, assistant, construction, guard, carpenter, analyst, physician, cook, clerk, manager, developer, librar-

ian, salesperson, mover, hairdresser, auditor, sheriff, janitor, driver, chief, teacher, writer, receptionist, CEO, nurse, housekeeper, secretary, counselor, attendant, supervisor, designer, lawyer, editor, cashier, laborer, tailor, cleaner.

However, many professions are gendered in French (for example, the masculine form of the French word for baker is *boulangier* whereas the feminine form is *boulangère*), including many on the list we employ for the English sense vector evaluations. To mitigate this while maintaining comparability, for French debiasing we use only the subset of the list of professions where the male and female forms of the profession are identical.

Thus, the French profession words are as follows: *médecin* (physician), *analyste* (analyst), *bibliothécaire* (librarian), *comptable* (accountant), *designer* (designer), *manager* (manager), *réceptionniste* (receptionist), and *secrétaire* (secretary).

For each English and French, we run ablations in which we iteratively zero out each of the 16 sense vectors and re-compute the average bias score over all of the professions. In this way, we identify the sense vector that contributes the most on-average to gender bias in representations of all of the profession words for each language. We then report the contribution ablating that sense vector makes to gender bias in each profession’s representation.

4.3.4 Transformer nullspace projection debiasing

Following [Hewitt et al. \(2023\)](#), and based on [Bolukbasi et al. \(2016\)](#), in order to debias our Transformer baseline model in a comparable way for comparison with the Backpack sense vector ablations, we project the embedding for each English profession word from the same list onto the nullspace of:

$$E\mathbf{x}_{he} - E\mathbf{x}_{she}$$

We do the equivalent for the French using ‘il’ and ‘elle.’ We treat these as approximations of the profession words’ learned gender bias direction.

5 Results

We evaluate both models using validation perplexity, monolingual Multi-SimLex, cross-lingual Multi-SimLex, sentence similarity with a random baseline, and sense-level gender bias ablations.

On a held-out validation set of EuroParl, the Transformer attains lower perplexity than the Back-

pack, 29.8 compared to the Backpack’s 35.5, reflecting its more confident next-token predictions.

5.1 Lexical similarity

Monolingual Multi-SimLex Spearman correlations are low ($\rho < 0.13$) but statistically significant for both models, which is as expected for language models trained on parliamentary data with limited topical scope. The Transformer is stronger on English (Backpack $\rho=0.052$, $p=0.024$; Transformer $\rho=0.122$, $p<0.001$), while the Backpack is stronger on French (Backpack $\rho=0.120$, $p<0.001$; Transformer $\rho=0.105$, $p<0.001$).

Cross-lingual Multi-SimLex Spearman correlations are likewise low but statistically significant for both models. The Transformer is higher in both directions (EN-FR: Backpack 0.066, $p=0.004$ vs. Transformer 0.114, $p<0.001$; FR-EN: Backpack 0.123, $p<0.001$ vs. Transformer 0.132, $p<0.001$).

Our correlations are below strong pretrained or aligned static baselines (fastText $\rho \approx 0.55$ – 0.61 monolingually, multilingual BERT $\rho \approx 0.40$ and aligned Cross Lingual Word Embedding $\rho \approx 0.66$ – 0.71 on EN-FR ([Vulić et al., 2020](#))), reflecting our models’ training on a narrow domain with computational limits. Our primary comparison is therefore between the Backpack and the matched Transformer, rather than between the Backpack and pretrained state-of-the-art models such as multilingual BERT.

5.2 Sentence-level similarity

Parallel sentence embeddings are only slightly more similar than randomly paired sentences, with Backpack $\Delta\mu=0.045$ and Transformer $\Delta\mu=0.034$ over 200 EuroParl validation pairs, indicating that both models learn only some translational information. The fact that sentence-level translation embedding similarity is low may also be attributed to the fact that both our Transformer baseline and our Backpack model are trained on parallel French-English data, but, unlike pretrained multilingual BERT/XLM-RoBERTa or aligned fastText Cross Lingual Word Embedding systems ([Vulić et al., 2020](#)), are not specifically designed to produce cross-lingually aligned representations.

5.3 Sense vector ablation for debiasing

5.3.1 English debiasing

Across the 40 professions, we find an average bias score of 24.659, indicating that the Backpack has learned significant gender bias with respect to

Metric	Backpack	Transformer
Validation perplexity	35.5	29.8
<i>Monolingual Multi-SimLex</i>		
English ρ	0.052 ($p=0.024$)	0.122 ($p<0.001$)
French ρ	0.120 ($p<0.001$)	0.105 ($p<0.001$)
<i>Cross-lingual Multi-SimLex</i>		
EN-FR ρ	0.066 ($p=0.004$)	0.114 ($p<0.001$)
FR-EN ρ	0.123 ($p<0.001$)	0.132 ($p<0.001$)
<i>Sentence similarity ($N=200$, mean pooling)</i>		
$\Delta\mu = \mu_{trans} - \mu_{rand}$	0.045	0.034

Table 4: Lexical, sentence-level similarity results

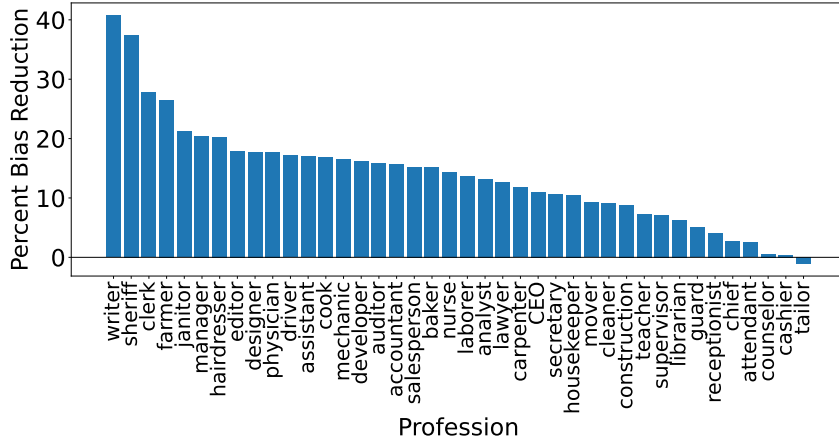


Figure 1: English percent reduction in bias score by profession after ablation of only sense 1.

occupation. For example, this includes the bias score 27.207 towards male for ‘physician.’ In the same experimental setup, our Transformer baseline achieves an average bias score of 18.635. Through nullspace projection debiasing, the Transformer baseline’s average bias score decreases to 17.527, about a 6% reduction in bias score overall.

Through the sense vector ablation experiment, some senses stand out as particularly impactful to bias. On either end of the spectrum, ablating sense 1 reduces the average bias by about 14%, whereas ablating sense 2 increases the average bias score by about 2%.

In order to probe this result, we examine the impact of ablating sense 1 on each profession’s individual bias score, shown in Figure 1. Though sense 1 reduces bias the most of all senses across the dataset as a whole, on an individual level it contributes varying amounts to bias. For example, ablating sense 1 reduces bias for the term ‘writer’ by about 41%, had almost no effect on bias for ‘cashier’ with only a 0.35% reduction, and increases bias for ‘tailor’ by about 1%. This sug-

gests that gender bias representation is distributed over multiple sense vectors such that the ablation of a single sense vector cannot uniformly debias a set of representations. That said, a 14% average reduction in gender bias with the ablation of only sense 1 is promising evidence for the usefulness of sense vectors in gender debiasing, especially given that nullspace projection debiasing only reduces gender bias by 6% on average for the same profession word list.

5.3.2 French debiasing

Our Backpack’s average bias score for the French professions list is 6.900, and our Transformer baseline’s average bias score is 3.252. Rather than sense 1, for French sense 7 has the highest contribution to gender bias: ablating this sense decreases the average bias score by about 16%. On the opposite end of the spectrum, ablating sense 12 increases the average bias score by about 11%. We find that the sense vector with the highest contribution to overall bias differs between languages, suggesting that there is not a universal “gender bias sense,”

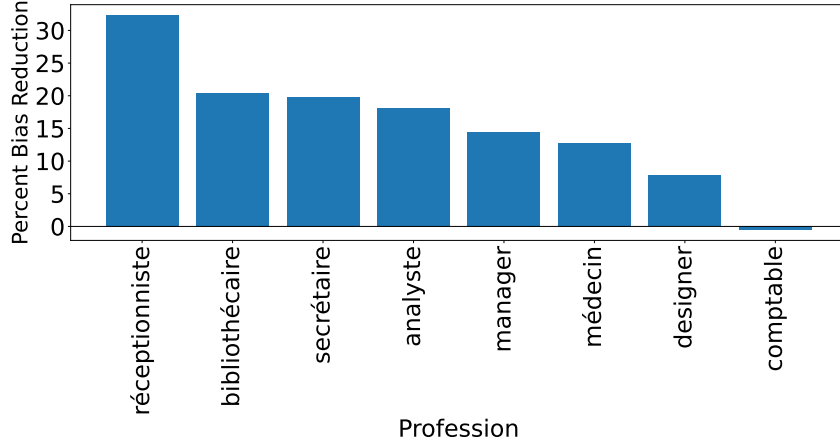


Figure 2: French percent reduction in bias score by profession after ablation of only sense 7.

rather that it is language-dependent. Through the nullspace projection method described above for the Transformer, the bias score has only a small reduction, from 3.252 to 3.228, or about 1%.

Figure 2 shows the results on each profession from ablating sense 7, which reduces bias the most of all senses across the dataset. As with the sense 1 ablation for English, ablating affects different occupation words’ bias scores differently. For example, ablating sense 7 reduces the gender bias for ‘réceptionniste’ by about 33%, but increases gender bias for ‘comptable’ by about 1%. This shows distribution of gender bias information across senses, but also a non-trivial overall debiasing effect from ablating a single sense vector. Similar to English, French sense vector ablation debiasing outperforms the Transformer nullspace projection debiasing, which only achieves about a 1% reduction in gender bias for the same profession words list.

5.3.3 Performance cost

While promising, our debiasing results must be considered alongside cost to performance. When interpreting the following performance-based metrics, it must be considered that the Backpack sense ablations involve the elimination of a significant portion of representation space, whereas Transformer nullspace debiasing only eliminates the English or French gender direction from representation space.

In order to evaluate the performance cost of Transformer nullspace projection in a comparable way to Backpack sense ablations, we project each word of the full vocabulary onto the nullspace of the English and French gender bias directions.

For the Backpack, ablating sense 1 leads to a jump in validation perplexity score to 48.1, while

ablating sense 7 leads to a jump to 47.3, representing a nontrivial reduction in predictive ability that is not matched by the Transformer with English or French nullspace projections, which for each maintains the same perplexity score as the original Transformer for the English projection, and sees an increase of 0.1 in perplexity score for the French projection. This is expected given the amount of representation space removed by our sense ablations compared to our nullspace projections.

On our sentence-level similarity evaluation, sense ablations cause slight increases in $\Delta\mu$, meaning that, as for the original Backpack model, parallel sentence embeddings are slightly more similar than randomly paired sentences. Compared to the original Backpack model’s $\Delta\mu$ of 0.045, ablating sense 1 results in a $\Delta\mu$ of 0.0617, and ablating sense 7 results in a $\Delta\mu$ of 0.0463. For the Transformer nullspace projections, we see no change in $\Delta\mu$ for either the English or French nullspace projections. Thus, neither sense ablations nor nullspace projection detract from either model’s modest translational ability.

We likewise run our monolingual and multilingual MultiSimLex evaluations on the Backpack with each of senses 1 and 7 ablated, as well as the Transformer with each of the English and French nullspace projections. Overall, we see both the Backpack model with sense ablations and the Transformer with nullspace projections retain the majority of their monolingual and multilingual representation learning ability.

For monolingual MultiSimLex Spearman correlations, the sense 1 ablation achieves $\rho = 0.052$, $p = 0.023$ for English and $\rho = 0.114$, $p < 0.001$ for French. On the same evaluation metric, ablation

of sense 7 has $\rho = 0.046$, $p = 0.045$ for English and $\rho = 0.112$, $p < 0.001$ for French. This represents only a small decrease in monolingual representation learning ability compared to the Backpack model’s original $\rho = 0.052$, $p = 0.024$ for English and $\rho = 0.120$, $p < 0.001$ for French. For monolingual MultiSimLex, the English nullspace projection had $\rho = 0.121$, $p < 0.001$ for English and $\rho = 0.104$, $p < 0.001$ for French. Meanwhile, the French nullspace projection had monolingual English $\rho = 0.123$, $p < 0.001$ and French $\rho = 0.107$, $p < 0.001$. Both are nearly identical to the Spearman correlation of the original Transformer, $\rho = 0.122$, $p < 0.001$ for English and $\rho = 0.105$, $p < 0.001$ for French, reflecting essentially the same monolingual representation learning ability after nullspace projection.

For multilingual MultiSimLex, sense 1 ablation leads to $\rho = 0.057$, $p = 0.013$ for EN-FR, compared to the original Backpack model’s $\rho = 0.066$, $p = 0.004$, and $\rho = 0.111$, $p < 0.001$ for FR-EN, compared to the original Backpack model’s $\rho = 0.123$, $p < 0.001$. Ablation of sense 7 results in $\rho = 0.057$, $p = 0.014$ for EN-FR and $\rho = 0.108$, $p < 0.001$ for FR-EN. These represent only slight decreases in multilingual representation learning ability after sense ablation. English nullspace projection is again nearly identical to the original Transformer for multilingual MultiSimLex, with $\rho = 0.115$, $p < 0.001$ for EN-FR, compared to $\rho = 0.114$, $p < 0.001$ for the original Transformer, and $\rho = 0.131$, $p < 0.001$ for FR-EN, compared to $\rho = 0.132$, $p < 0.001$ for the original Transformer. The same was true of French nullspace projection, giving $\rho = 0.114$, $p < 0.001$ for EN-FR and of $\rho = 0.134$, $p < 0.001$ for FR-EN.

Despite the reduction in predictive ability consistent with the ablation of a full sense, we see that the Backpack model with sense ablations maintains translational and monolingual and multilingual representation learning ability on par with the Transformer with nullspace projections.

6 Discussion

Both the Transformer baseline and the Backpack models demonstrate cross-lingual representation learning ability: while the Transformer attains higher cross-lingual Multi-SimLex (0.114 ($p < 0.001$) versus 0.066 ($p = 0.004$) for EN-FR and 0.132 ($p < 0.001$) versus 0.123 ($p < 0.001$) for

FR-EN), the Backpack shows stronger translated-vs-random sentence separation, with $\Delta\mu=0.045$ compared to $\Delta\mu=0.034$ for the Transformer.

Moreover, the Backpack’s main advantage over the Transformer baseline is the result of interpretable intervention. For both English and French experiments, sense vector debiasing is significantly more effective than the nullspace projection method used for Transformer debiasing. For the English sense vector ablations, ablating sense 1 reduces bias by about 14% compared to 6% for Transformer nullspace projection debiasing. On the French gender-neutral subset, sense 7 ablation reduces bias by about 16% versus about 1% for the Transformer nullspace projection debiasing. While the sense vector with the highest contribution to overall bias is language dependent, and even within each language its contribution to individual profession words’ bias scores varies significantly (see Figures 1 and 2), ablation of a single sense vector still significantly reduces the average bias score for both languages. Thus, in the multilingual context of our experiments, the Backpack architecture allows for sense-level debiasing, which we find to be more effective than nullspace projection debiasing, without sacrificing a significant amount of translational and cross-lingual representation learning ability.

7 Conclusion

We present an analysis of multilingual Backpack representations and the potential for leveraging their sense structure for interpretable debiasing. We find that Backpack models show comparable cross-lingual representation learning ability to Transformer models of the same size, and also that Backpack sense-level debiasing is more effective for gender debiasing than Transformer nullspace projection. These results are particularly significant given that, as multilingual models increase in scale and power, interpretable debiasing methods which are effective in multilingual contexts become increasingly essential.

Limitations

The size and training of both our Transformer baseline and Backpack model were limited by our access to computational resources, meaning that future work might replicate our findings on larger models trained on larger datasets for more iterations. In addition, our contributions with respect

to multilingual Backpack language models focus on the English-French language pair. To this end, future work should scale to more language pairs, including languages using non-Roman scripts, and language pairs themselves using different scripts. Finally, future work might examine the efficacy of sense-level debiasing methods on Backpack models trained on more than two languages, and on Backpack models trained on lower-resource languages than English and French.

Acknowledgments

We thank John Hewitt for his advisement on this project.

References

- Amirabbas Afzali, Amirreza Velae, Iman Ahmadi, and Mohammad Aliannejadi. 2026. [Controlling gender bias in retrieval via a backpack architecture](#). In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 557–565. Springer.
- Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32.
- Jan Cruz, David Adelani, and Alham Aji. 2026. [Multilinguality as sense adaptation](#). *Preprint*, arXiv:2601.10310.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*.
- Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–213.
- John Hewitt, Christopher Thickstun, Christopher D. Manning, and Percy Liang. 2023. [Backpack language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–556, Toronto, Canada. Association for Computational Linguistics.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. [SimLex-999: Evaluating semantic models with \(genuine\) similarity estimation](#). *Computational Linguistics*, 41(4):665–695.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.
- Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual BERT? *arXiv preprint arXiv:1911.03310*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *International Conference on Learning Representations (Workshop Track)*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Hao Sun and John Hewitt. 2023. [Character-level Chinese backpack language models](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 237–252, Singapore. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.
- Ivan Vulić, Edoardo Maria Ponti, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2020. [Multi-SimLex: A large-scale evaluation and study of monolingual and cross-lingual lexical semantic similarity](#). *Computational Linguistics*, 46(3):645–692.