# Multilingual Backpack Language Models

Clément ADANDE (clementa@aims.ac.za)

African Institute for Mathematical Sciences (AIMS)

Supervised by: Francois MEYER

University of Cape Town, South Africa

24 October 2024

*Submitted in partial fulfillment of a AI for Science masters degree at AIMS South Africa*

# Abstract

Backpack language models (LMs) have been proposed as a language modelling architecture which offers a flexible interface for interpretability and control. They learn multiple sense vectors per word, explicitly encoding polysemy in natural languages. So far, Backpack LMs have only been trained as monolingual LMs, for respectively English and Chinese. For each of these languages, after training the Backpack LMs, sense vectors specialise and encode various aspects of a word allowing convenient handling of polysemous words as they convey different senses. However, they have not been trained on multilingual modelling, which has become the standard approach for extending the coverage of language models to more languages, particularly low-resource languages. The multi-sense modelling of the Backpack LMs seems particularly relevant to multilingual modelling, where many words and subwords have different meanings in different languages. This work explores the effectiveness of Backpack LMs in multilingual settings by training and evaluating Backpack LMs on two languages simultaneously. We found that the Backpack LMs learn meanings of words efficiently and do not encode language-specific sense vectors. Our Backpack LM (112M parameters) has lower perplexity than the baseline Transformer (93M parameters). Based on a cloze task, the Backpack LM (112M parameters) also slightly outperforms the baseline Transformer (93M parameters).

## Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Clément Adandé, 24 October 2024

# Contents

# 1. Introduction

## 1.1 Overview

Large language models (LLMs) have made significant progress in generating coherent and contextually appropriate text. Standard LLM architectures, as seen in Transformers (Vaswani et al., 2017), consist of learning a single embedding of each word and then passing these embeddings through a neural network. This means that multiple senses of a polysemous word are not modelled separately, and can lead to ambiguities in understanding and generating text. For instance, the word *bank* could mean either *the side of a river* or *a financial institution*, and instead of encoding these senses with different vectors, the model combines them into a single representation and selects the most appropriate meaning based on the surrounding context. Known as Word Sense Disambiguation (WSD) (Navigli, 2009), this conflation of word meanings is a popular limitation for many language models.

Addressing these ambiguities requires intervention in the model's architecture to better separate word senses. However, contemporary LLMs are difficult to intervene in because their underlying structures often hide a non-linear structures in their contextual representations, for example transformers' architecture results in a composition of sequence of linear and non-linear functions making them non-linear and this complicates the intervention in transformers. This issue is addressed by the recently introduced Backpack Language Model (Hewitt et al., 2023). The core idea is to model a set of non-contextual sense vectors for each word and represent a word in a sequence as a weighted sum of sense vectors of the constituent words in that sequence, where the weights are determined by the context. These models have demonstrated considerable efficiency for interpretability and ability to separately encode different meanings of words in English (Hewitt et al., 2023) and Chinese (Sun and Hewitt, 2023), giving a more flexible and convenient way of dealing with linguistic ambiguity and many other issues, for example gender bias. For instance, Hewitt et al. (2023) found that a specific sense of Backpack LM encapsulates the gender bias for stereotypically gendered profession nouns, and they adjusted or downscaled this sense vector to mitigate the bias.

## 1.2 Objective

Backpack LMs have so far been tested as monolingual language models in two different works: first on English (Hewitt et al., 2023) and second on Chinese (Sun and Hewitt, 2023). Despite their success in these languages, they have not yet been extended to multilingual language modelling.

Multilingual language models have become popular due to cross-lingual transfer[1]which is particularly beneficial for low-resource languages. In a multilingual context, words or tokens often carry different meanings across languages exacerbating the challenge of modelling the different meanings of words. This project aims to test Backpack LMs on multilingual settings[2] for the first time and explore the possibilities of the Backpack LMs in preserving their capabilities on multilingual settings, particularly focusing on the situation where certain words can have various meanings depending on the language. For simplified experiments and analysis, only two languages are considered in this work. By training these models on corpora from two different languages simultaneously, namely English and French, the objective of this project is twofold.

[1] Cross-lingual transfer refers to transfer learning using data and models available for one language for which ample such resources are available (e.g., English) to solve tasks in another, commonly more low-resource, language.

[2] Our code is available on GitHub at https://github.com/clemsadand/multilingual-backpack-lm.

1. Compare the performance of Backpack LMs in vanilla LMs in multilingual settings. We aim to assess whether the performance of Backpack LMs matches that of vanilla LMs and to determine if the benefits of Backpack LMs affect performance or not. This is achieved through the metric perplexity to access the model confidence in text generation and the metric accuracy on cloze task to assess the context-dependent generation.
2. Analyse the characteristics of the senses learned by a multilingual Backpack LM. We aim to examine whether the Backpack LM encodes any language-specific senses, in other words, verify if the sense distributions vary across languages. This involves inspecting the senses of a sample of words and their lexical relationship, and analysis of senses based on language-specific contexts.

## 1.3 Research Questions

In this work, we explore the ability of the multilingual Backpack LMs focusing on specific research axes. We seek to answer the following questions:

1. Do multilingual Backpack LMs outperform standard GPT-2 on intrinsic evaluation?
2. Do multilingual Backpack LMs encode language-specific sense vectors, or are the sense vectors shared across languages?
3. Do multilingual Backpack LMs handle shared words and cognates in language modelling?

## 1.4 Summary of Results

After training the Backpack LMs on multilingual corpora, we found that the Backpack LMs learn meanings of words efficiently. By inspecting the senses of some words, we found that, depending on the language of the context, each sense encoded by the Backpack LM highly weights words or subwords in that language in general. This demonstrates that the multilingual Backpack LMs do not encode language-specific sense vectors. Our Backpack LM (112M parameters) has lower perplexity than the baseline Transformer (93M parameters). Based on a cloze task, the Backpack LM (112M parameters) also outperforms the baseline Transformer (93M parameters).

## 1.5 Report Layout

The remainder of this report is presented as follows:

- **Chapter 2: Background**
  - We review relevant literature and foundational concepts that support our research, such as word embeddings, word sense representations, intervention in transformers, and Backpack LMs.
- **Chapter 3: Multilingual Backpack Language Modelling**
  - **Backpack Architecture:** We explain the architecture and key features of Backpack LMs and how they differ from standard transformer-based language models.
  - **Multilingual Backpacks:** We describe the extension of Backpack LMs to multilingual settings and analyse their behavior in capturing shared and language-specific meanings.
- **Chapter 4: Methodology**
  - **Data:** We describe the datasets used for training and evaluation, specifically Europarl and MultiUN.
  - **Tokenization:** This section covers the customized tokenization process using Byte Pair Encoding (BPE) applied to the training corpora.
  - **Data Preprocessing Workflow:** We outline the steps for merging, tokenizing, and splitting the datasets for training.
  - **Training Setup:** We detail the training configurations for Backpack LMs and baseline GPT-2 models.
- **Chapter 5: Experiments**
  - **Performance Evaluation:** We explain the metrics and methods used to assess model performance, such as perplexity and cloze task accuracy.
  - **Word Representation Analysis:** We analyse how words are represented and the sense vectors learned by Backpack LMs.
- **Chapter 6: Results and Discussion**
  - **Perplexity-Based Evaluation:** We present the perplexity scores and compare the performance of Backpack LMs and baseline transformers across French and English.
  - **Cloze Task:** We evaluate the models using a cloze task to assess their ability to predict missing words.
  - **Sense Visualization and Distribution:** We explore the semantic relationships captured by the sense vectors in both languages.
  - **Sense Distribution Analysis:** We examine the contextual ratio of sense vectors for various word categories, such as cognates and shared words.
- **Chapter 7: Conclusion and Perspective**
  - We summarize the findings and suggest directions for future research in multilingual Backpack language models.

- **Appendices**
  - **Basic Concepts:** This section provides definitions and explanations of fundamental concepts used throughout the report.
  - **Validation Loss:** Details on the validation loss during the training of the Backpack LM and baseline model.
  - **Sense Visualizations:** Additional examples and visualizations of the senses learned by the models.
  - **Sense Distribution:** Detailed results on the distribution of senses across different word categories.

# 2. Background

## 2.1 Word Embeddings

Modern NLP systems represent words as vectors of real numbers called embeddings. This has a long history (Weaver, 1955), originating in the field of document retrieval (Salton et al., 1975). Following by the paper Salton et al. (1975), several works in Language Modeling were based on this concept (Rumelhart et al., 1986; Elman, 1990; Bengio et al., 2003; Turian et al., 2010; Mikolov et al., 2013; Pennington et al., 2014; Vaswani et al., 2017; Radford et al., 2019; Devlin et al., 2019) which led to language models using neural networks to learn the vector representation before feeding it to the models.

Static word embedding represents each word as a single vector regardless of its context (Mikolov et al., 2013; Pennington et al., 2014). Mikolov et al. introduced Word2Vec in 2013, which learns words based on the surrounding context in which their appear. It has two variants: Continuous Bag of Words (CBOW) which, predicts a target word based on its context, and Skip-Gram, which predicts surrounding words given a target word (Mikolov et al., 2013). Following this, Pennington et al. developed GloVe, another word embedding model in 2014. Standing for Global Vectors for Word Representation, GloVe is a count-based method which constructs a matrix of word co-occurrences and factorizes the entire corpus, resulting in rich semantic meanings (Pennington et al., 2014). Despite their contribution to word representation, these models encode fixed and context-independent senses.

Contextual word embeddings represent words as vectors that vary depending on the context in which they appear (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019). Meaning Bidirectional Encoder Representations from Transformer, BERT was introduced by Devlin et al. (2019) and uses a bidirectional transformer (Vaswani et al., 2017) architecture[1] to generate embeddings for each word in a sentence. Unlike BERT, Generative Pre-trained Transformer 2 (GPT-2) uses a unidirectional transformer to learn word representation (Radford et al., 2019). BERT and GPT-2 are two popular models used in several studies to learn contextual representation of words.

While both static and contextual word embeddings capture valuable semantic information, they still face challenges in correctly representing words with multiple meanings. To address this issue, word sense representation techniques have been developed.

## 2.2 Word Sense Representations

Traditional word embedding models represent each word as a single vector which is inconvenient when encoding polysemous words (words with multiple meanings). In fact, as different senses of a polysemous word are condensed in one single vector, it may be problematic for language models to predict the suitable meaning on downstream tasks such as text generation or machine translation. This limitation is known as the meaning conflation deficiency (Camacho-Collados and Pilevar, 2018) and several works have proposed different approaches to address it using sense embedding models (Miller et al., 1990; Song et al., 2016; Navigli and Ponzetto, 2012; Camacho-Collados and Pilevar, 2018; Roh et al., 2021; Rodrigues da Silva and Caseli, 2021). In 2018, Camacho-Collados and Pilevar conducted an inventory of word meaning representations and noticed two main approaches used in the literature for encoding word senses: unsupervised models, which deduce meanings from text corpora (Song et al., 2016; Roh et al., 2021; Rodrigues da Silva and Caseli, 2021), and knowledge-based models, which rely on sense

[1] A bidirectional transformer processes text in both (left and right context) directions.

inventories to embed different meanings of words (Miller et al., 1990; Navigli and Ponzetto, 2012). For example, Roh et al. (2021) propose an unsupervised framework that captures multi-sense representations of polysemous words using a model including unsupervised clustering for sense labeling.

## 2.3 Intervention in Transformers

As mentioned in Section 2.1, contextual word embeddings models incorporate the transformers (Vaswani et al., 2017) into their architecture. By design, transformers have a complex architecture, and their omnipresence raises important questions about their interpretability and possibilities for intervention, both in transformers and LLMs in general. Roughly, interpretability seeks to understand and explain how the models works (Lipton, 2017; Voita et al., 2021; Tan et al., 2023) and intervention attempts to control the model behavior after training (Tan et al., 2024). Recent works have made notable contributions to these questions. Backpack language models presented in Hewitt et al. (2023) represent a valuable effort in this regard.

## 2.4 Backpack LMs

The Backpack LM (Hewitt et al., 2023) is a new architecture, encoding multiple meanings of words and offering a flexible interface for interpretability and control. It embeds various meanings of each word as non-contextual sense vectors, and based on the context, each word is represented as a weighted sum of its sense vectors, where the weights are computed as a function of the context. In subsequent sections, the Backpack architecture is described (see Section 3.1 in this report). Originally tested on English and then on Chinese, the Backpack LMs have demonstrated their performance on English (Hewitt et al., 2023) as well as Chinese (Sun and Hewitt, 2023) on several NLP tasks, namely lexical relationships and mitigating gender biases (Hewitt et al., 2023; Sun and Hewitt, 2023). In Hewitt et al. (2023), Backpack models achieve similar perplexity to Transformer models on language modeling tasks and outperform GPT-J-6B on lexical similarity benchmarks like SimLex999. In addition, the Backpack architecture enables controllable interventions, such as topic-guided generation and gender bias reduction, by modifying sense vectors. Similarly, in Sun and Hewitt (2023), Backpack model performs comparably to GPT-2 on perplexity and word prediction tasks. Moreover, it demonstrates the ability to intervene in gender-biased representations at the character level and allows control over word meanings by manipulating sense vectors.

In summary, the evolution of word embeddings from single to multiple and contextualized representations constitutes a significant step forward in NLP. While traditional language models struggled with polysemy, recent models like Transformers have shown their performance on a wide range of tasks and are limited in terms of interpretability and control. However, encoding multiple non-contextual senses of each word, Backpack LMs hold promise for improving interpretability and control.

# 3. Multilingual Backpack Language Modelling

Unlike a monolingual LMs, multilingual LMs are described to handle multiple languages simultaneously. Multilingual models have exactly the same architecture as monolingual models. The main difference lies not in their architecture, but in their training data and objectives. In this chapter, we present the general architecture of Backpack LMs, describe how they differ from standard language models, and extend their application to multilingual settings.

## 3.1 Backpack Architecture

The main difference between backpack LMs and vanilla Transform-based LMs lies in how they encode words. In fact, in a sequence, transformers contextualize the meaning of each word based on their relevance with the previous words (see the appendix A.4 to learn a relevance is incorporated in transformers), while the Backpack language models represent word as weighted sum of non-contextual sense vectors of the previous words within this sequence. Figure 3.1 illustrates these distinct characteristics, and shows how Transformers treat the sequence as a whole, whereas Backpacks combine individual learned word aspects non-contextually. In other words, in Backpack, the word sense embeddings are non-contextual, but they are combined as a weighted sum, and the weights in this sum are contextual.

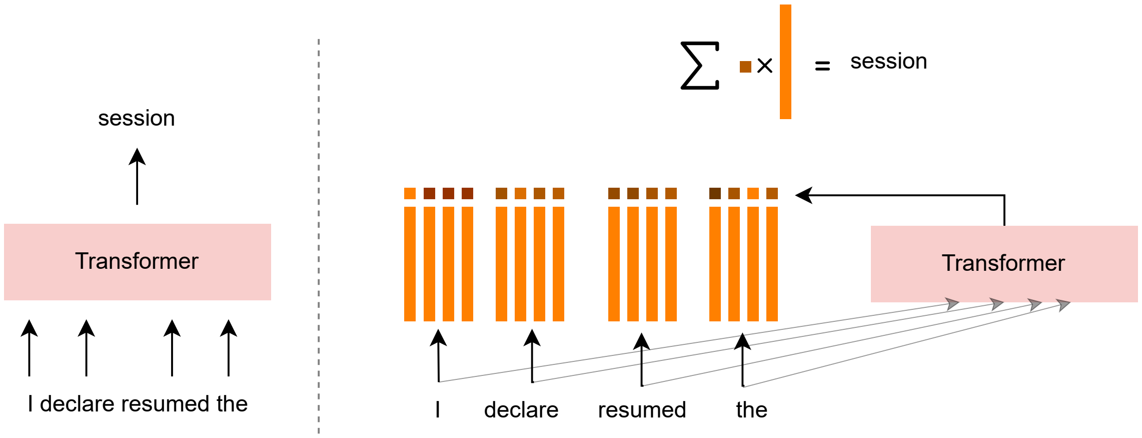


Figure 3.1: Transformers are monolithic functions of sequences. In Backpacks, the output is a weighted sum of non-contextual, learned word aspects (Hewitt et al., 2023). To predict the next word *session*, the Backpack uses a linear combination of the previous words in the sequence *I declare resumed the*.

In the remaining of this report, unless explicitly mentioned, $n$, $d$ and $k$ are non-negative integers, where $n$ represent size of a sequence of words, $d$ the dimension of an embedding vector and $k$ the number of sense vectors learned by the Backpack LMs.

The core components of Backpack LM architectures are sense vectors and the Backpack representation.

Given a subword vocabulary $\mathcal{V}$ and a subword $w \in \mathcal{V}$, the Backpack LM learns $k$ **sense vectors** $C(w)_1$, $C(w)_2$, $\ldots C(w)_k$ which belong to $\mathbb{R}^d$ where $C$ is the **sense function** mapping $\mathcal{V}$ to $\mathbb{R}^{d\times k}$ and is parameterized by:

$$C(w) = FF(Ew)$$

with $FF : \mathbb{R}^d \to \mathbb{R}^{d \times k}$ a feedforward layer network and $E \in \mathbb{R}^{d \times |\mathcal{V}|}$ is the embedding matrix.

The Backpack LMs encode contextual word meanings. Given a sequence of subwords $w_1$, $w_2$, ..., $w_n$ of $\mathcal{V}$, the Backpack representation $o_i$ of the subword $w_i$ is the weighted sum of the sense vectors of words within this sequence (Hewitt et al., 2023). Formally, the backpack representation of the word $w_i$ is defined by:

$$o_i = \sum_{j=1}^{n} \sum_{\ell=1}^{k} \alpha_{ij\ell} C(w_j)_\ell, \tag{3.1.1}$$

where $\alpha_{ij\ell}$ is the **contextualization weight** and is parameterized by:

$$\alpha_\ell = \text{softmax}\left(h_{1:n}^T K^{(\ell)^T} Q^{(\ell)} h_{1:n}\right) \tag{3.1.2}$$

for each predictive sense $\ell$ with matrices $K^{(\ell)}$, $Q^{(\ell)}$ and $h_{1:n}$ calculated by a Transformer (Vaswani et al., 2017) with autoregressive masking, i.e

$$h_{1:n} = \text{Transformer}(Ex_{1:n}). \tag{3.1.3}$$

Defined in Hewitt et al. (2023), a **Backpack LM** is an autoregressive probabilistic model which defines a probability distribution over the next token $w_i$ of a sequence $w_1$, $w_2$, ..., $w_{i-1}$ of subwords such as:

$$P(w_i | w_1, w_2, \ldots, w_{i-1}) = \text{softmax}(Eo_i), \tag{3.1.4}$$

where $o_i$ represents the backpack representation of the subword $o_i$.

## 3.2 Multilingual Backpacks

Previous works involving the Backpack LMs have made a monolingual usage. First, Hewitt et al. (2023) tested Backpack on English and second, Sun and Hewitt (2023) tested them on Chinese. Consequently, they both attest that the Backpack LMs learn lexical structure in each of these languages. This project extends the Backpack LMs to multilingual settings for the first time. We train the Backpack LMs on French and English parallel corpus. There is a reason for this choice. In addition to the rich structure of language-specific words in English and French, these languages share a significant number of words, partially attributed to the Normal Conquest of England in 1066 (Bandrivska, 2022). Testing Backpack LMs on these languages simultaneously allows us to take advantage of their diversity and to observe whether the models can capture shared and language-specific meanings.

Hewitt et al. (2023) mention that senses capture different aspect of words such as relatedness, verb objects, wordpiece, proper noun associations,... We hypothesise that the multilingual Backpack LM might present similar patterns as well and further, reserve some senses to identify languages.

# 4. Methodology

In this chapter, we describe the methodology used for training our models on parallel corpora. We begin by introducing the dataset, followed by a description of the tokenization process used to prepare the data for training. Finally, we outline the model training process.

## 4.1 Data

This project made use of two principal datasets: Europarl (Koehn, 2005) and MultiUN (Eisele and Chen, 2010; Tiedemann, 2012; Ziemski et al., 2016) parallel corpora.

### Europarl parallel corpus

The Europarl corpus is a domain-specific dataset from the proceedings of the European Parliament from 1996 to 2011. It is available 21 European languages including French and English. The the French-English subset is described in Table 4.3. The Europarl corpus is cleaned, and aligned with a tool based on the Church and Gale algorithm, a method for aligning corresponding sentences in a parallel corpus. The dataset is openly accessible for research on many websites including www.statmt.org[1]. It was first introduced by Koehn (2005).

### MultiUN parallel corpus

The MultiUN corpus is extracted from the official records and other parliamentary documents of the United Nations. It is available in 6 languages including English and French and has been manually translated between 1990 and 2014. The French-English subset is described in Table 4.3. The dataset is available on a dedicated the official United Nations website [2] and also on the OPUS website[3], and its use is presented in works such as Eisele and Chen (2010); Tiedemann (2012).

| | Sentences | French words | English words |
|---|---|---|---|
| Europarl | $2,007,723$ | $51,388,643$ | $50,196,035$ |
| MultiUN | $13,172,019$ | $362,171,080$ | $320,100,029$ |

Table 4.1: Number of sentences and corresponding word counts for French and English in the Europarl and MultiUN datasets

In the last decade, the Europarl and MultiUN datasets have been widely used in NLP research, specifically for machine translation. Although these datasets are limited in terms of domain, they are still suitable for our purposes of testing multilingual Backpack LMs. In this study, we use them the efficiency of the Backpack LM in handling the polysemous and language-based meanings of words. We acknowledge their limitations and take this into account in designing our experiments. These datasets provide Backpack LMs access to a wide range of sentence patterns and linguistic traits, which is essential for our analysis.

[1] https://www.statmt.org/europarl/

[2] https://conferences.unite.un.org/uncorpus

[3] https://opus.nlpl.eu/MultiUN/en&fr/v1/MultiUN

## 4.2 Tokenization

### 4.2.1 General Overview

In NLP, **tokenization** is an important step for data preprocessing. It converts raw text into a format that language models can process for sequence modelling. The tokenization consists in splitting corpus or text into small *units*, called **tokens**, which could be words, subwords or characters. There are different types of tokenization: **word-level**, **subword-level** and **character-level** tokenizations. Table 4.2 describe each of these types with examples.

| Type | Description | Example of text | Tokenized text |
|---|---|---|---|
| Word-level | Splits the text into words. | "The report was well received." | `["The", "report", "was", "well", "received."]` |
| Subword-level | Splits the text into smaller units. | "The report was well received." | `["The", " re", "was", "well", " re", "ceived."]` |
| Character-level | Splits text into characters. | "received" | `["r", "e", "c", "e", "i", "v", "e", "d"]` |

Table 4.2: Different Types of Tokenization

A **vocabulary** is the set of tokens resulting in the tokenization of a text.

The tool or library implementing the tokenization process is commonly named the **tokenizer**. Different tokenizers exists for each type of tokenization. Among these, we can mention the **Byte Pair Encoding (BPE) tokenizer** (Sennrich et al., 2016). BPE tokenizer is the one used in this project to obtain a wide range of tokens shared between French and English on the parallel corpora. An advantage of subword tokenizer is that it can easily handle new or out-of-vocabulary words.

**Algorithm 1** BPE Algorithm (Sennrich et al., 2016).

```
1. Initialize the vocabulary with all the bytes or characters in the text corpus.
2. Calculate the frequency of each byte or character in the text corpus.
while the desired vocabulary size is not reached do
    a. Find the most frequent pair of consecutive bytes or characters in the text corpus
    b. Merge the pair to create a new subword unit.
    c. Update the frequency counts of all the bytes or characters that contain the merged pair.
    d. Add the new subword unit to the vocabulary.
end while
3. Represent the text corpus using the subword units in the vocabulary.
```

### 4.2.2 Customized Tokenizer

We used a HuggingFace[4] library to train a customized BPE tokenizer which was trained on the same corpora used for model training. The resulting tokens, which will be learnt by Backpack LMs, consists of French and English subwords.

[4] https://huggingface.co/learn/nlp-course/en/chapter6/8

We trained two BPE tokenizers: one exclusively on the Europarl dataset and the another on the combination of Europarl and MultiUN corpora with a maximum vocabulary size set to 50,256 and 10,000 respectively.

- We train one set of models on Europarl using 50,000 subwords. This is for analysis, since many words will be included. We will refer to this tokenizer as *50k-tokenizer*.
- We train another set of models on the combined corpus to see how performance scales. Since this takes longer to train, and larger vocabularies increase training (the softmax increases) we use a smaller vocabulary of 10,000 to be computationally feasible. We will refer to this tokenizer as *10k-tokenizer*.

We did not perform preprocessing techniques on the dataset.

### 4.2.3 Corpora Tokenization

The corpora Europarl (Koehn, 2005) and MultiUN (Ziemski et al., 2016) are tokenized using these customized tokenizers.

For the 10k-tokenizer, Table 4.3 presents the number of French and English tokens used in the training files of two datasets, Europarl and MultiUN, along with the running time in minutes for tokenizing these files. We used eight 8 CPUs of RAM with a machine type `e2-standard-16` the for task.

| | **French tokens** | **English tokens** | **Running time** (minutes) |
|---|---|---|---|
| Europarl | 1,987,645 | 1,987,645 | 12 |
| MultiUN | 7,950,585 | 7,948,132 | 55 |
| Total | 9,938,230 | 9,935,777 | 67 |

Table 4.3: Token counts for the French and English corpora from Europarl and MultiUN, along with the running time required for tokenizing each dataset.

For the 50,256-token tokenizer, Table 4.4 presents the number of French and English tokens used in the training files of Europarl, along with the running time in minutes for tokenizing these files. We used two CPUs of RAMs for task.

| | **French tokens** | **English tokens** | **Running time** (minutes) |
|---|---|---|---|
| Europarl | 2,007,723 | 2,007,723 | 35 |

Table 4.4: Token counts for the French and English corpora from Europarl, along with the running time required for tokenizing each dataset.

## 4.3 Data Preprocessing Workflow

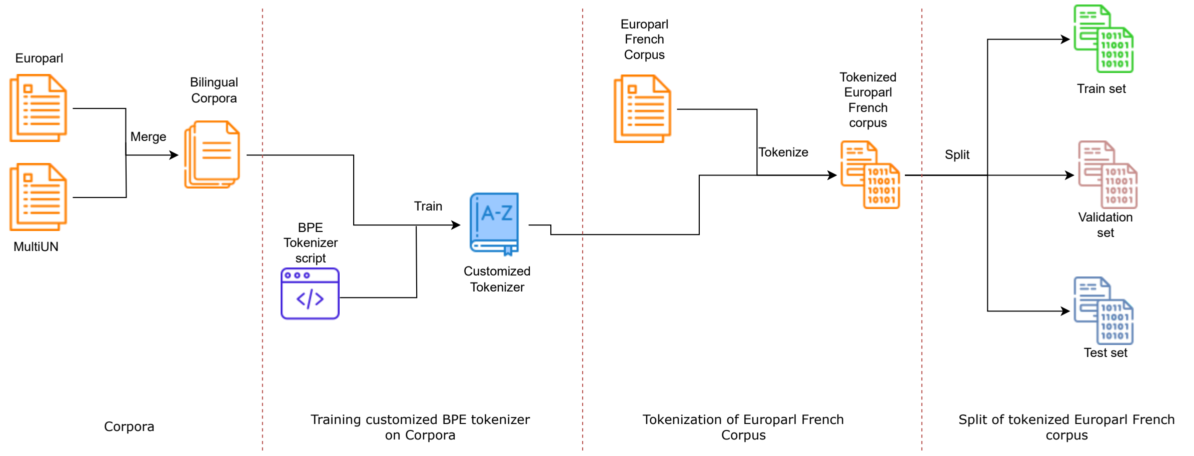

Europarl
MultiUN
Merge
Bilingual Corpora
BPE Tokenizer script
Train
Customized Tokenizer
Europarl French Corpus
Tokenize
Tokenized Europarl French corpus
Split
Train set
Validation set
Test set
Corpora
Training customized BPE tokenizer on Corpora
Tokenization of Europarl French Corpus
Split of tokenized Europarl French corpus

Figure 4.1: Workflow for data processing. The diagram illustrates the steps from merging bilingual corpora (Europarl and MultiUN) to training a customized BPE tokenizer. It includes processes for tokenizing the Europarl French corpus and splitting the tokenized data into train, validation, and test sets for further processing.

Figure 4.1 shows the workflow of the data processing. We begin by merging the Europarl and MultiUN corpora to create a bilingual dataset. Next, we train a customized BPE tokenizer on the merged corpora, to allow the tokenizer to learn subword representations suited to both languages. As both Europarl and MultiUN include French and English corpora, we made use of the trained tokenizer to proceed to the tokenization of every corpora and then split it into three subsets: training, validation and test sets. Figure 4.1 illustates this last step for the Europarl French corpus. We tokenize the Europarl French corpus. After tokenization, we split the processed corpus into training, validation, and test sets. This workflow is important as it ensures that the data is well-prepared for training based on our customised tokenizer. The same process was applied to the Europarl English corpus as well as MultiUN data.

## 4.4 Training Setup

Our models are trained on the parallel French-English corpora outlined in SubSection 4.4.2. While the parallel corpora are usually used for translation tasks, we use them for language modelling instead. Thus, we do not sample the parallel sentences together. To equip these models with multilingual capabilities, they are trained one language at a time, on either the English corpus or the French corpus. At each gradient step, the models learn from a batch of the corpus consisting of only one of the two languages. This way, the models focus on the meanings of words separately for each language, allowing the model to enhance shared words accurately depending on the context. We trained Backpack LMs and GPT-2s as baselines.

### 4.4.1 Baseline Transformers

We trained two GPT-2 baseline transformers and named them, respectively, mini and small, according to the number of parameters. The small GPT-2 was trained solely on the Europarl dataset (50k-tokenizer), while both the small and mini GPT-2 were trained on the Europarl and MultiUN datasets (with 10k-tokenizer). Table 4.5 presents the characteristics of each GPT-2 variant. The mini GPT-2 transformer has $6$ transformer layers, each with 6 attention heads, an embedding dimension of 384, and a block size of 256. We handled the memory constraint by using a mini-batch of 4096 tokens and gradient accumulation over 64 steps. The learning rate was set to $6 \times 10^{-5}$, with a linear decay followed by a cosine decay strategy across 6,000 iterations and a minimum learning rate of $6 \times 10^{-6}$ to prevent excessive drops in learning progress. We included a warm-up period for the first 6,000 steps to gradually ramp up the learning rate, avoiding abrupt changes early in training and preventing divergences.

| GPT-2 | Layers | Heads | Embedding |
|---|---|---|---|
| Mini | 6 | 6 | 384 |
| Small | 12 | 12 | 768 |

Table 4.5: Different configuration of GPT-2 model considered as baseline: mini and Small, showing the number of layers, attention heads, and embedding size for each model. The number of parameters of mini GPT-2 is less than the one of small GPT-2.

We used the same configuration to train all the models, GPT-2 and Backpack LMs.

### 4.4.2 Backpack Training

We initially trained two small Backpack LMs on only Europarl with respectively 8 sense vectors and 16 sense vectors using 50,256 subwords. Following this, we extended the training to include both the Europarl and MultiUN datasets, where we trained one small and one mini Backpack LMs with 16 sense vectors using 10,000 subwords. We called them small or mini Backpack LMs to highlight the underlying GPT-2 transformers.

In total, we trained seven language models. Table 4.6 presents each of these models with their the number of parameters.

| Model | Dataset | Parameters |
|---|---|---|
| Small GPT-2 | Europarl | 124M |
| Small Backpack 8 | | 143M |
| Small Backpack 16 | | 162M |
| Mini GPT-2 | Europarl & MultiUN | 14M |
| Mini Backpack | | 19M |
| Small GPT-2 | | 93M |
| Small Backpack | | 112M |

Table 4.6: Different language models with their dataset and number of parameters. The number 8 and 16 indicate the number of sense vectors.

For the next steps and for the sake of simplicity, we added *-eu* to the names of all models trained only on the Europarl dataset. For example, we will use Small GPT-2-eu to refer to the one trained exclusively on Europarl.

During the train, the validation loss of the small Backpack-eu-8 consistently remains low than the one of small GPT2-eu.

# 5. Experiments

## 5.1 Performance Evaluation

There are many ways of evaluating language models. Different methods used in NLP to assess model performances can be clustered into two groups (Shi et al., 2018).

One, named **intrinsic evaluations**, focus on assessing the ability of the model to capture linguistic structures and information directly from the data it was trained on. These are useful to measure how well a model captures the statistical patterns of language. They use metrics such as accuracy, precision, recall, F1 score, and perplexity to quantify the model performance. Intrinsic evaluations are commonly conducted on standardized test dataset or benchmarks (Shi et al., 2018).

The second group, called **extrinsic evaluations**, focus on downstream task performance, i.e real-world applications such as text generation, machine translation, question-answering. They test how well the model can behave in tangible real situations.

The evaluation of any model can be done via both of these methods. However, depending on the goal of the model is designed for, one method might be more appropriate than another.

To measure the Backpack LM performances, we made use of intrinsic and extrinsec evaluations. We evaluate our models on the intrinsic evaluation metric on perplexity, and on the extrinsic task of cloze prediction.In addition, we will implement another intrinsic evaluation method. We will evaluate the models on semantic similarity.

### 5.1.1 Perplexity

Widely used, perplexity intuitively informs about the number of tokens. It measures the uncertainty associated with the next-token prediction. In other terms, it measures how well a language model (a probability distribution) predicts the next tokens. Mathematically, perplexity is defined as the exponential of the cross-entropy loss, and it can be interpreted as the inverse probability of the test set normalized by the number of words. The lower the perplexity score is the better the model predicts the next word or subword in a sequence.

Given a sequence of words $w_1$, $w_2$, ..., $w_T$, the perplexity $PPL$ is defined as:

$$PPL = 2^{-\frac{1}{T}\sum_{t=1}^{T}\log_2 P(w_t|w_1,w_2,\ldots,w_{t-1})}$$

where $P(w_t|w_1, w_2, \ldots, w_{t-1})$ is the conditional probability of word $w_t$ given the previous words predicted $w_1$, $w_2$, ..., $w_{t-1}$.

We utilise a held-out test set drawn from the Europarl (Koehn, 2005) and MultiUN (Ziemski et al., 2016) datasets, which were not seen during training, to compute the perplexity of our models. This allows us to assess how well the model's predictions generalize to new, unseen corpora and language patterns.

### 5.1.2 Cloze task

A cloze task is a type of tasks in which words are removed from a sentence, and then the model predicts the missing words Hu et al. (2021). This task provides insights into the model's ability to understand

| Id | Number of tokens | Cloze passage | Answer |
|---|---|---|---|
| 0 | 74 | As to how we should deal with this serious matter here in this Chamber, I think it would be appropriate if the problems in the Middle East were to be raised tomorrow in the debate with the President - in - Office of the Council and the President of the Commission, either by individual Members or by political group chairmen or whoever speaks on behalf of the `<mask>`. | groups |
| 1 | 96 | Yet, while our own institution has done so much to ensure the successful completion of the work of the Convention, particularly by, on many occasions, making its premises available to the Convention, I am sorry that today, as our plenary part - session in Strasbourg is opening, it was not possible to organize the formal hand over of this Charter here in Strasbourg, in conditions worthy of it, now that the work of drafting it has been `<mask>`. | completed |
| | | | |

Table 5.1: Examples from the test set of the cloze task, showing incomplete passages where the missing word is replaced by `<mask>`. The task is to predict the correct word to fill the gap, with answers provided in the final column

context. In this report, the Backpack LMs are also evaluated on customized dataset designed for this task.

The test set is drawn from the Europarl (Koehn, 2005) test set[1] and MultiUN (Ziemski et al., 2016) held-out test set. We consider the sentences with number of tokens between $20$ and $256$ which last words are not numbers or punctuation. Based on these criteria, we sample 1000 examples from each dataset. Every example is a dictionary saved with the keys:

- `id`: a number between $0$ and $999$;
- `original_passage`: a line in the test set;
- `length_of_passage`: the length of the tokenized sequence of the original passage;
- `cloze_passage`: the original sequence with the last word removed;
- `answer`: the last word in the original passage;
- `language`: the language of the original passage (French or English).

The cloze task performed consists of predicting the masked word and checking whether the model's prediction is correct. Additionally, we measure the top-$k$ accuracy, which assesses whether the correct answer is among the top-$k$ predictions made by the model, where $k$ is $1$ or $3$.

For each example, we use the `cloze_passage` as prompt and generate the next tokens until sequence `length_of_passage` with the beam search method (Sun and Hewitt, 2023). As described in Sun and

[1] https://www.statmt.org/europarl/v1/common-test2.tgz

Hewitt (2023), we kept ten generations from the beam at each step.

We evaluate the Backpack LMs performance through the accuracy defined as the ratio of correctly predicted words to the total number of words in the task. Thus, recall that the number of words in the task is the equal to the size of the evaluation set, the accuracy is given by:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}.$$

## 5.2 Analyzing Word Representation

Previous experiments with Backpack LMs have shown that different senses encode different grammatical roles and different aspects of word meaning (Hewitt et al., 2023; Sun and Hewitt, 2023). For example, in the original paper Hewitt et al. (2023), sense 12 seems to capture related word while the sense 14 seems to capture associated objects for verbs (see Table 5.3).

| Sense 12 (relatedness) | | | Sense 14 (Verb objects, nmod nouns) | | |
|---|---|---|---|---|---|
| tasty | quickly | believe | build | attest | appreaciate |
| tasty | quick | belief | bridges | worthiness | finer |
| culinary | quickest | Belief | wall | Published | nuance |
| tasted | quick | beliefs | lasting | superiority | beauty |
| delicious | quicker | believing | ig | accuracy | irony |
| taste | fast | believe | rapport | validity | simplicity |

Table 5.3: Visualisation of senses 12 and 14 of some word drawn from original paper Hewitt et al. (2023). Each column presents different word forms associated with the specified senses.

When trained on multiple languages, it is possible that this finding will still hold; for example, different senses may capture distinct parts-of-speech or aspects of word meaning. Moreover, in a multilingual setting, senses might also specialise for individual languages. For instance, consider the false cognate[2] *sensible*, which means *sensitive* in French but *reasonable* in English. In this case, a sense vector could capture its meaning as *reasonable*, while another sense vector would encode its meaning as *sensitive*. That is to say that the multilingual Backpack LM could associate some senses with English aspects of a false cognate and other senses with French aspects of the same false cognate. This specialisation could enhance the model's ability to disambiguate meanings based on the context of the language being processed. This Section explores which senses are distributed across languages and what they capture through some experiments. We are interested in the following question: is there any evidence that the senses learned by multilingual Backpack are tied to particular languages, or do they rather encode language-independent aspects of meaning?

[2] A false cognate is a word with a similar form in both languages but has different meanings. For example, *librairie* is in French and means *bookstore*, while *library* is in English and refers to *the place with books to borrow*.

### 5.2.1 Word representation

The Backpack language models learn sense vectors of words or subwords in the vocabulary. In production or development, the language models might have to deal with words out-of-vocabulary. BPE tokenizers address this by representing any word as concatenation of subwords or tokens within the vocabulary. To handle sense vectors words, we follow the same representation as in Hewitt et al. (2023).

We represent the sense vectors of a word by averaging the sense vectors of its individual subwords (Hewitt et al., 2023; Sun and Hewitt, 2023). This means that each subword contributes to the overall meaning of the word. For example, for the word "gouvernements" we would compute its sense vector $\ell$ as the average of the sense vector $\ell$ for the subwords "gou", "verne", "ment" and "s". Precisely, let us consider a word $w$ made of the subwords $w_1$, $w_2$, $\ldots$, $w_m$. The sense vector $\ell$ of the word $w$ is computed as follow:

$$C(w)_\ell = \frac{1}{m}\sum_{i=1}^{m} C(x_i)_\ell,$$

for any $\ell \in \{1, 2, \ldots, k\}$.

### 5.2.2 Sense visualization

Inspired by the original paper (Hewitt et al., 2023), we expect the multilingual Backpack LMs to capture different aspects of meaning that contribute separately to the next word generation. To achieve this, we need to inspect what the senses are encoding. Thus, for a fixed sense $\ell$, we project this sense of the word $w \in \mathcal{V}$ onto the embedding space $E^T C(w)_\ell$ and filter out the ones with the highest scores resulting from this projection. Once we have identified the top-scoring words, we analyse the possible similarities and differences in meaning for some words. Throughout this analysis, we will gain insights into what the multilingual Backpack LM learned for each sense across different languages and contexts. Furthermore, we can deduce whether certain senses are more sensitive to a specific language or not.

### 5.2.3 Sense distribution

The Backpack LMs learn contextual meanings of words and represent them as multiple sense vectors. In multilingual settings, senses of word become more complex as the models are exposed to multiple languages, each having its own linguistic structures. This raises the question: how does the sense distribution for the same word vary across different contexts? To answer this question, we focus our analysis on four different categories of words: French-specific words, English-specific words, shared words and cognates[3]. Then, we examine the variation of these words across a sample of contexts. In order to access the contribution of a word or subword to a context, we define the contextual ratio.

#### Contextual Ratio

Let us consider a word $w$; and a context $c$, where tokenized sequence is $w_1$, $w_2$,. . . ,$w_n$. Suppose that this context contains the word $w$ and that the constituent tokens of $w$ are part of the sequence $w_1$, $w_2$,. . . ,$w_n$ with their indices in this sequence are denoted $j_1$, $j_2$, $\ldots$, $j_m$.

The Backpack representation of any subword $w_i$ which is defined by:

$$o_i = \sum_{\ell=1}^{k}\sum_{j=1}^{n} \alpha_{\ell ij} C(w_j)_k,$$

[3] A cognate is word with a similar form and meaning in both languages due to shared Latin or Greek roots. For example, nation and important are both in French and English.

encapsulates the contribution of each subword within this context. For a fixed sense vector $\ell$, the contribution of a subword $w_i$ to the representation of $w_j$ is $\alpha_{\ell ij}$. Thus, given the sense vector $\ell$, we define the **contextual ratio of the subword** $j_{w_s}$ by:

$$\lambda_{\ell j_s} = \frac{\alpha_{\ell j_s j_s}}{\sum\limits_{i=1}^{n} \alpha_{\ell i j_s}}.$$

Consequently, given the sense vector $\ell$, we define the **contextual ratio of the word** $w$ by the average value of the contextual ratio of its constituent tokens.

### Contextual Variation of Word

We drew a sample of 1,000 contexts of most frequent words in the aforementioned categories: French-specific words, English-specific words, cognates and shared words.

| | Number of words |
|---|---|
| **French words** | 173 |
| **English words** | 143 |
| **Cognates** | 21 |
| **Shared words** | 37 |

Table 5.4: Number of words sampled from different lexical categories, including French-specific words, English-specific words, cognates, and shared words.

The selection of these words is done as follows:

1. Start with a list of the 10,000 most frequent words from the French and English corpora of Europarl;
2. Manually select words from this list based on the desired categories (French-specific, English-specific, cognates, and shared words).

Table 5.4 presents the number of words considered for each category. For a given category, we choose a word within that category, compute the contextual ratio for the 1,000 contexts[4] in which the word occurs, obtain a statistical series of 1,000 real values, and calculate the variance of this series. To illustrate, following these steps, we obtain a series of 173 real values, which represent the variances of contextual ratios of different words in the category of French-specific words. For a given word, a small variance indicates a stable contribution of that word in different contexts. This section aims to measure the proportion of words whose contextual contributions remain the same across languages. This will help to gain insights into any language-specific senses.

[4] We selected *frequent words* to ensure that these words appear in at least 1,000 distinct contexts.

# 6. Results and Discussion

## 6.1 Perplexity-based Evaluation

The perplexity metric served to measure how well a model can predict a held-out corpus. We evaluate the perplexity on French and English separately for both models to quantify their ability in generating the right next tokens. Table 6.1 present the perplexity of different models across French and English, with two different tokenizers. Knowing that the lower the perplexity the more accurately the model is predicting the corpus, we can draw the following observations.

For the 10k-tokenizer, Small Backpack shows the best performance in both English and French. It achieves the lowest perplexity among all 4 models: 14.40 for French and 19.60 for English, and outperforms its baseline GPT-2 on perplexity. However, this does not hold for all, as Mini GPT-2 perplexity is smaller than the ones of Mini Backpack LM. As the difference is small, we can deduce that multilingual Backpack LMs do not lose any predictive performance compared to vanilla architectures for the 10k-tokenizer. Similarly, for the 50k-tokenizer, the Small Backpack-eu-8 has the better perplexity. It slightly outperforms GPT-2 in both languages. The Small Backpack-eu-16 achieves a higher perplexity among all 3 models, and this indicates its difficulties in handling larger vocabularies.

The Mini Backpack and Small Backpack have encoded the 16 sense vectors for each word or subword in the vocabulary. But the Small Backpack LMs achieves the best perplexity. This suggests that, for a fixed size of sense vectors, when the dataset is large, a higher number of parameters is suitable to ensure the predictive performance of the Backpack LMs.

| | Perplexity | |
|---|---|---|
| **Models** | **French** | **English** |
| **Mini Backpack** | 19.74 | 26.62 |
| **Mini GPT-2** | 18.59 | 25.02 |
| **Small Backpack** | **14.40** | **19.60** |
| **Small GPT-2** | 17.28 | 23.75 |
| **Small Backpack-eu-8** | **17.82** | **27.44** |
| **Small GPT-2-eu** | 17.92 | 28.24 |
| **Small Backpack-eu-16** | 23.68 | 38.48 |
| **Small GPT-2-eu** | 17.92 | 28.24 |

Table 6.1: Perplexity scores for different models across French and English languages. Lower perplexity indicates better model performance.

Comparatively to the monolingual setting shown by the previous papers (Hewitt et al., 2023; Sun and Hewitt, 2023) on Backpack LMs, we also found that multilingual Backpack models performed similarly or better than vanilla LMs.

## 6.2 Cloze Task

The cloze task consists of predicting missing words in a sentence, assessing a model's ability to understand context and generate coherent text based on partial information. We implemented the cloze task on Europarl and MultiUN held-out test sets. The top-1 accuracy measures the proportion of tokens where the model's prediction is correct on the first attempt, while the top-3 accuracy measures whether the correct token is among the model's top three predictions (first or second or third guesses). Table 6.2 shows accuracies for the cloze task of Backpack LMs and their baseline transformers.

Table 6.2 shows that the top-1 accuracy is lower than the top-3 accuracy across all models. For example, the Small GPT-2 achieves a top-1 accuracy of 15% which is smaller than its top-3 accuracy of 21%. This trend is consistent across different models. This means that while the models can generate potential responses, their may struggle to identify the most accurate next token on a first attempt.

| | **Accuracy (%)** | |
|---|---|---|
| **Models** | **Top-1** | **Top-3** |
| **Mini Backpack** | 12.5 | 19 |
| **Mini GPT-2** | 15 | 21 |
| **Small Backpack** | **17.8** | **25** |
| **Small GPT-2** | 17.5 | 20 |
| **Small Backpack-eu-8** | **25** | **26** |
| **Small GPT-2-eu** | 22.5 | 23.5 |
| **Small Backpack-eu-16** | 17.5 | 18 |
| **Small GPT-2-eu** | 22.5 | 23.5 |

Table 6.2: Top-1 and Top-3 accuracies results for different models to access the understanding of the model.

Table 6.2 suggests that the Backpack models demonstrate higher accuracy than their corresponding GPT-2 in most instances. For example, the Small Backpack eu 8 achieves a top-1 accuracy of 25% and a top-3 accuracy of 26%, which surpasses the Small GPT-2 eu, with 22.5% for top-1 and 23.5% for top-3 accuracy. The observation remains steady in various Backpack LMs and their baselines transformers. These results suggest that while GPT-2 models perform reasonably well, the Backpack models, even in more complex settings, tend to better capture context and provide more accurate predictions on cloze task.

In most cases, the results reveal that the Backpack LMs perform better than GPT-2 transformers on the cloze task. Therefore, based on extrinsic evaluation, the Backpack LMs attempt to understand easily context than vanilla Transformer LMs in general. Again, this is not surprising. In fact, by construction, the Backpack LM learns contextual meaning of words using GPT-2 transformer to encode the relevance. This contributes the Backpack ability to capture nuanced relationship between word meanings based on their prior context.

## 6.3 Sense Visualisation

A wide range of words or subwords in the vocabulary of the 50k-tokenizer is convenient to clearly inspect what the Backpack LMs encode as senses of a word. We conduct sense visualisation experiment with the Small Backpack-eu-8. Tables 6.4, 6.6 and 6.8 present the results and highlight senses 2, 4 and 7 (refer to Section C in the appendix for more sense visualisations).

In both French and English, for verbs, sense 2 focuses on related verbs that share similar semantic fields (see Table 6.4). For example, in English, for the verb *consider*, the multilingual Backpack associates it with such as the verbs *think*, *examine*, *realize*; and in French, for the verb *investir* (invest), the model links it to such as the verbs *moderniser* (modernise), *investir* (invest), *diversifier* (diversify).

| Sense 2 | | |
|---|---|---|
| consider | express | prevent |
| think | empower | detect |
| examine | explain | deter |
| realize | mind | manage |
| investigate | learn | compensate |
| learn | publicise | éradiquer |

| Sense 2 | | |
|---|---|---|
| investir<br>invest | garantir<br>guarantee | soutenir<br>support |
| moderniser<br>modernize | veillons<br>watch over | veillons<br>watch over |
| investir<br>invest | recherchons<br>research | apportons<br>bring |
| diversifier<br>diversify | sauvegarder<br>safeguard | soutenir<br>support |
| know | travaillons<br>work | educating |
| restructurer<br>restructure | protégeons<br>protect | parvenons<br>achieve |

Table 6.4: Sense 2 - Synonym and Related Verbs in English and French. Left, words in English; right, words in French.

Sense 4 encodes different grammatical forms, with related nouns and adverbs in both languages for almost all words (see Table 6.6). For example, in English, for *law*, the model learns associated terms such as *Constitutions*, *Arrest*, *juges* (judges); and in French, for *égalité* (equality), the model encodes masculins (males), discriminations (discriminations), inégalité (inequality).

Sense 7 learns next possible word in both French and English text generation (see Table 6.8). In English, *first* is connected with *time*, *consecutive*, *contact*; and in French, *nouveau* (new) associated with *cycle* (cycle), *vaccin* (vaccine).

The senses encoded by the multilingual Backpack LMs are generally specific to the language of the target words (see Tables 6.4, 6.6, and 6.8). Any sense does not specialise for one given language between French and English. This demonstrates the presence of language-independent senses. Sense 2 encodes related verbs, highlighting their semantic relationships in both English and French. Sense 4 encodes semantics, or the meaning of words, in both languages. Sense 7 encodes the next word prediction, again for English and French.

To conclude, the experiments shows that these Backpack LMs are encoding different aspects of word meaning in different senses and these senses appear to serve the same function for both languages most of the time proving a language-independent senses.

| Sense 4 | | |
|---|---|---|
| rights | law | quick |
| rights | Constitutions | quickly |
| Universal | law | quick |
| constitutions | jur | faster |
| right | Arrest | fast |
| Covenant | juges<br>judges | quickest |

| Sense 4 | | |
|---|---|---|
| égalité<br>equality | emploi<br>job | nécessaire<br>necessary |
| égalité<br>equality | employabilité<br>employability | efficacement<br>effectively |
| masculins<br>males | emploi<br>job | indispensables<br>indispensable |
| discriminations<br>discriminations | Employment<br>employment | nécessaire<br>necessary |
| inégalité<br>inequality | chômeurs<br>unemployed | indispensable<br>indispensable |
| féminin<br>feminine | emploi<br>job | primordiales<br>essential |

Table 6.6: Sense 4: Relatedness in English and French. Left, words in English; right, words in French. Under each French word its possible translation.

| Sense 7 | | |
|---|---|---|
| first | good | how |
| time | at | much |
| times | our | worked |
| these | the | things |
| consecutive | how | many |
| contact | its | the |

| Sense 7 | | |
|---|---|---|
| moins<br>less | nouveau<br>new | où<br>where |
| insurmont<br>insurmount | départ<br>departure | règne<br>reign |
| équivalente<br>equivalent | cycle<br>cycle | sévit<br>prevails |
| équivalent<br>equivalent | paradigme<br>paradigm | prévaut<br>prevails |
| partielle<br>partial | vaccin<br>vaccine | siègent<br>sits |
| paradoxal<br>paradoxical | raisonnement<br>reasoning | dominant<br>dominate |

Table 6.8: Sense 7: Next Tokens in English and French. Left, words in English; right, words in French. Moins (less), nouveau (new), où (where). Under each French word its possible translation.

## 6.4 Sense Distribution

### 6.4.1 Case-Study: Cognates and Shared-Words

We investigate sense distributions with cognate words and shared words (words that are identical in French and English). Additional examples can be found in Section D of the appendix. Figure 6.1 shows two boxplots of the contextual ratio of the cognates *democracy* (top) and *démocratie* (below) across eight different senses. Both boxplots offer insight into the behavior of these senses within English contexts (for democracy) and French contexts (for démocratie).

The boxplots show that sense 8 has the largest interquartile range (IQR) and the highest median value for each word. Thus, it contributes the most contextually and is more varied in how much it contributes compared to other senses. Below sense 8, sense 3 also has an important contribution to the word meaning across different contexts. However, sense 1, 2, and 4 have lower IQR and seem to contribute consistently to the word meaning as their IQR is typically constant. Furthermore, the plots show a large number of outliers across most senses, in particular for senses 2, 4, 5, and 6, indicating contexts in which these senses contribute disproportionately higher than their typical range. This trend remains similar for other cognates (see Figures 4-6 in the appendix). Overall, the analysis suggests that while certain senses (like 8 and 3) have an important and varied effect, other senses contribute often in a more consistent way, with occasional outliers reflecting context-specific shifts in meaning both in French and English.

As in the case of cognates, Figure 6.2 reveals the same trend for sense 3 and 8 across language-specific contexts of the word "*impact*". Moreover, the IQR is high in English than French across all senses except for some rare cases (see Figure 9 in the appendix). This indicates that, for shared words, their contextual ratios is greater in English contexts compared to French contexts. However, this pattern is not uniform across different words and does not distinctly favor either language.

#### Quantifying Sense Distribution Variance

Table 6.9 presents the proportion of words exhibiting different amounts of variance across their occurrences in different contexts for the sense 4. For the full set of 8 senses, refer to Table 6.10. The following observation drawn for sense 4 is still similar to other senses. Table 6.9 distinguishes between proportion variances less than 0.01 and those between 0.01 and 0.1. The experiment reveals that all variances are less than 0.1.

| | | **Proportion of Variances** | |
|---|---|---|---|
| | **Contexts** | $< 0.01$ | $\geq 0.01$ and $\leq 0.1$ |
| French words | French | 24.3 | 75.7 |
| French Cognates | | 9.5 | 90.5 |
| Shared words | | 10 | 90 |
| English words | English | 23.8 | 76.2 |
| English Cognates | | 38 | 62 |
| Shared words | | 37.8 | 62.2 |
| Shared words | Combined Contexts | 16.2 | 83.8 |

Table 6.9: Sense 4: Proportion (in %) of variances for different categories of words. Lower values indicate more consistent contextual contributions.

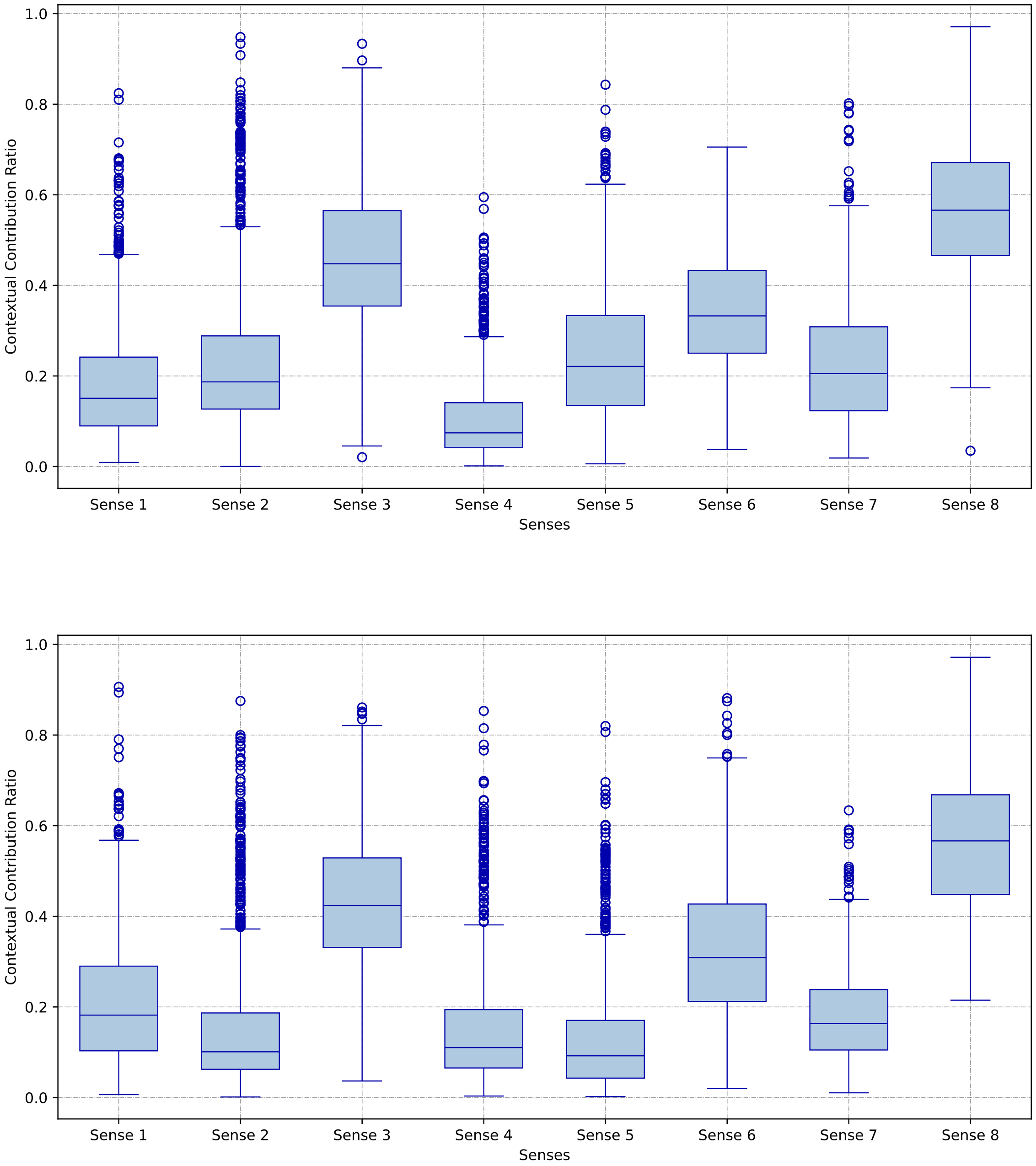


Figure 6.1: Box plots of the contextual ratios of words democraty (top) and démocratie per senses.

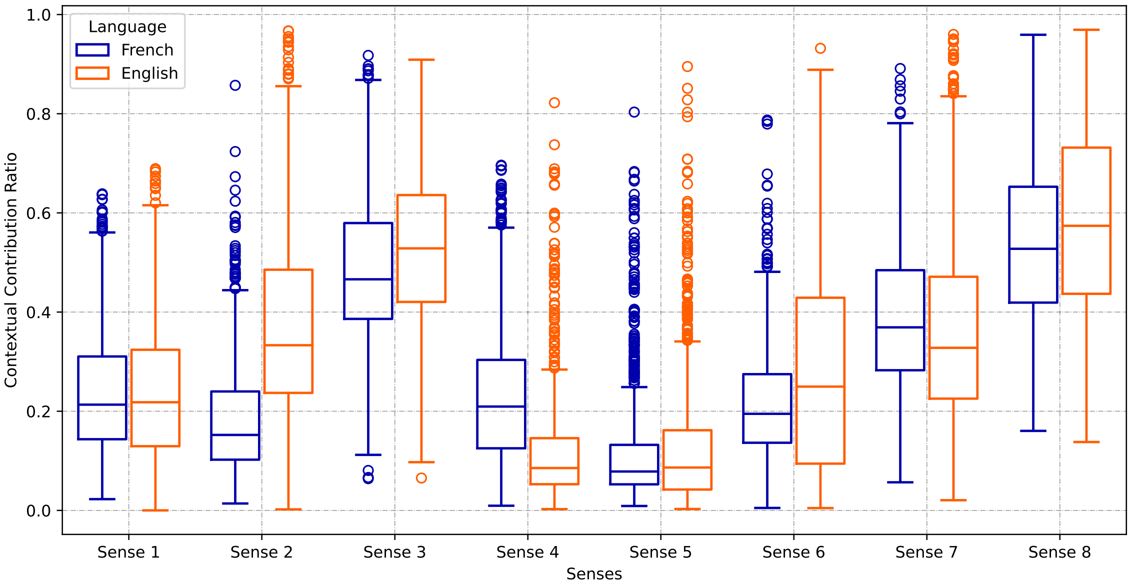


Figure 6.2: Box plots of the contextual ratios of word impact in French and English contexts. Blue is associated to the French contexts and Orange to English contexts.

The majority of both French and English words maintain a stable contribution across different contexts, with an important proportion having their variances between 0.01 and 0.1. English cognates stand out with a higher percentage (38%) showing less than 0.01 variance. This suggests a greater consistency in the contextual contribution of English. Overall, all variances are less than 0.1, indicating a lack of clear differentiation between French and English words. As a result, sense 4 is more stable in English contexts than in French contexts. Given the small difference, we conclude that sense 4 does not encode a specific language but rather favors English contexts in most contexts.

| | | **Proportion of Variances** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Contexts** | Sense 1 | | Sense 2 | | Sense 3 | | Sense 4 | | Sense 5 | | Sense 6 | | Sense 7 | | Sense 8 | |
| French words | French | 7.5 | 92.5 | 12.1 | 87.9 | 2.9 | 97.1 | 24.3 | 75.7 | 16.8 | 83.2 | 5.2 | 94.8 | 1.7 | 98.3 | 0.6 | 99.4 |
| French cognates | | 4.8 | 95.2 | 4.8 | 95.2 | 4.8 | 95.2 | 9.5 | 90.5 | 9.5 | 90.5 | 4.8 | 95.2 | 9.5 | 90.5 | 4.8 | 95.2 |
| Shared words | | 3.3 | 96.7 | 3.3 | 96.7 | 0.0 | 100.0 | 10.0 | 90.0 | 20.0 | 80.0 | 0.0 | 100.0 | 3.3 | 96.7 | 0.0 | 100.0 |
| English words | English | 13.3 | 86.7 | 3.5 | 96.5 | 2.8 | 97.2 | 23.8 | 76.2 | 6.3 | 93.7 | 4.2 | 95.8 | 2.8 | 97.2 | 2.8 | 97.2 |
| English Cognates | | 19.0 | 81.0 | 0.0 | 100.0 | 0.0 | 100.0 | 38.1 | 61.9 | 4.8 | 95.2 | 0.0 | 100.0 | 0.0 | 100.0 | 0.0 | 100.0 |
| Shared words | | 26.9 | 73.1 | 0.0 | 100.0 | 0.0 | 100.0 | 42.3 | 57.7 | 3.8 | 96.2 | 0.0 | 100.0 | 0.0 | 100.0 | 0.0 | 100.0 |
| Shared words | Combined Contexts | 5.4 | 94.6 | 0.0 | 100.0 | 0.0 | 100.0 | 16.2 | 83.8 | 2.7 | 97.3 | 0.0 | 100.0 | 0.0 | 100.0 | 0.0 | 100.0 |

Table 6.10: Proportion (in %) of variances for different senses and categories of words. Lower values indicate more consistent contextual contributions. For each sense the first column present the proportion of variances less than 0.01 and the proportion of variances between 0.01 and 0.1.

## 6.5 Do multilingual Backpack LMs learn language-specific senses?

Similar to the monolingual Backpack LM (Hewitt et al., 2023; Sun and Hewitt, 2023), the multilingual Backpack LM demonstrates the ability to learn multiple senses of words. Each sense vector encoded by the model captures different meanings that appear to align closely with semantic and grammatical structures, such as nouns and verbs, in both French and English. This suggests that the model effectively learns similar linguistic functions across languages and for each sense simultaneously. As a result, the multilingual Backpack senses are largely language-independent; no specific sense was clearly associated with a particular language, as observed in the sense visualization experiment (see Section 6.3). Moreover, the variations found in the categories mentioned above further support this finding for sense 4 (see Section 6.4).

# 7. Conclusion and Perspective

To summarize, in this report, we explored the Backpack LMs in multilingual contexts. We showed that these models efficiently encode polysemous words across French and English languages and present similar performance baseline transformers in both language modeling and understanding tasks in most instances. The ability to capture various meanings of words allows the Backpack LMs to address meaning conflation issues. This is important in multilingual settings where words or subwords may carry different meanings depending on the language. Furthermore, the experiments achieved indicate that Backpack LMs not only excel in predictive performance compared to vanilla transformers but also give insights into word sense distribution across contexts. The sense vectors represent different aspects of word meaning but are not language-specific. Also, the multilingual Backpack senses capture similar linguistic function across languages.

Next to these findings, future work could expand our experiments to further understand the multilingual Backpack LMs. This includes scaling the experiments to more languages, which would allow us to assess how well the models generalise across more than two linguistic settings. In addition, testing Backpack LMs on a wide range of downstream tasks could provide precise insights into their practical usage in multilingual applications. Analysing performance in scenarios where languages are imbalanced in the training data–such as low-resource languages–would also be valuable, as it could reveal how effectively the models encode senses for underrepresented languages and help improve their adaptability in multilingual environments.

# Acknowledgements

I want to acknowledge AIMS and it's funders for supporting this work, Google DeepMind, as well as my supervisor Francois Meyer, Assistant Professor at University of Cape Town.

# References

Bandrivska, N. English and french vocabulary as a product of historical development. *ScienceRise: Pedagogical Education*, pages 12–16, 07 2022. doi: 10.15587/2519-4984.2022.261748.

Bender, E. *An Introduction to Mathematical Modeling*. Krieger, 1991. ISBN 9780894645822. URL https://books.google.co.za/books?id=AVM_AQAAIAAJ.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. In *Journal of machine learning research*, 2003. URL https://api.semanticscholar.org/CorpusID:221275765.

Camacho-Collados, J. and Pilevar, M. T. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, 12 2018. doi: 10.1613/jair.1.11259.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Eisele, A. and Chen, Y. MultiUN: A multilingual corpus from united nation documents. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/686_Paper.pdf.

Elman, J. L. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. ISSN 0364-0213. doi: https://doi.org/10.1016/0364-0213(90)90002-E. URL https://www.sciencedirect.com/science/article/pii/036402139090002E.

Hewitt, J., Thickstun, J., Manning, C. D., and Liang, P. Backpack language models, 2023. URL https://arxiv.org/abs/2305.16765.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

Hu, Z., Chanumolu, R., Lin, X., Ayaz, N., and Chi, V. Evaluating nlp systems on a novel cloze task: Judging the plausibility of possible fillers in instructional texts, 2021. URL https://arxiv.org/abs/2112.01867.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 2024. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released August 20, 2024.

Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks, 2015. URL https://arxiv.org/abs/1506.02078.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

Lipton, Z. C. The mythos of model interpretability, 2017. URL https://arxiv.org/abs/1606.03490.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, 12 1990. ISSN 0950-3846. doi: 10.1093/ijl/3.4.235. URL https://doi.org/10.1093/ijl/3.4.235.

Mohebbi, H., Jumelet, J., Hanna, M., Alishahi, A., and Zuidema, W. Transformer-specific interpretability. In Mesgar, M. and Loáiciga, S., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 21–26, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-tutorials.4.

Nair, S., Srinivasan, M., and Meylan, S. Contextualized word embeddings encode aspects of human-like word sense knowledge, 2020. URL https://arxiv.org/abs/2010.13057.

Navigli, R. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), Feb. 2009. ISSN 0360-0300. doi: 10.1145/1459352.1459355. URL https://doi.org/10.1145/1459352.1459355.

Navigli, R. and Ponzetto, S. P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.07.001. URL https://www.sciencedirect.com/science/article/pii/S0004370212000793.

Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. volume 14, pages 1532–1543, 01 2014. doi: 10.3115/v1/D14-1162.

Radford, A. and Narasimhan, K. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Rodrigues da Silva, J. and Caseli, H. d. M. Sense representations for portuguese: experiments with sense embeddings and deep neural language models. *Language Resources and Evaluation*, 55(4):901–924, Feb. 2021. ISSN 1574-0218. doi: 10.1007/s10579-020-09525-1. URL http://dx.doi.org/10.1007/s10579-020-09525-1.

Roh, J., Park, S., Kim, B.-K., Oh, S.-H., and Lee, S.-Y. Unsupervised multi-sense language models for natural language processing tasks. *Neural Networks*, 142:397–409, 2021. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.05.023. URL https://www.sciencedirect.com/science/article/pii/S0893608021002197.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. URL https://api.semanticscholar.org/CorpusID:205001834.

Salton, G., Wong, A., and Yang, C.-S. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, 1975. URL https://api.semanticscholar.org/CorpusID:6473756.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Shi, Y., Zheng, Y., Guo, K., Zhu, L., and Qu, Y. Intrinsic or extrinsic evaluation: An overview of word embedding evaluation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1255–1262, 2018. doi: 10.1109/ICDMW.2018.00179.

Song, L., Wang, Z., Mi, H., and Gildea, D. Sense embedding learning for word sense induction, 2016. URL https://arxiv.org/abs/1606.05409.

Sun, H. and Hewitt, J. Character-level chinese backpack language models, 2023. URL https://arxiv.org/abs/2310.12751.

Tan, Z., Chen, T., Zhang, Z., and Liu, H. Sparsity-guided holistic explanation for llms with interpretable inference-time intervention, 2023. URL https://arxiv.org/abs/2312.15033.

Tan, Z., Peng, J., Chen, T., and Liu, H. Tuning-free accountable intervention for llm deployment – a metacognitive approach, 2024. URL https://arxiv.org/abs/2403.05636.

Tiedemann, J. Parallel data, tools and interfaces in OPUS. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Turian, J., Ratinov, L.-A., and Bengio, Y. Word representations: A simple and general method for semi-supervised learning. volume 2010, pages 384–394, 01 2010.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Voita, E., Sennrich, R., and Titov, I. Analyzing the source and target contributions to predictions in neural machine translation, 2021. URL https://arxiv.org/abs/2010.10907.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., and Korhonen, A. Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity, 2020. URL https://arxiv.org/abs/2003.04866.

Weaver, W. Translation. In Locke, W. N. and Booth, D. A., editors, *Machine Translation of Languages*. MIT Press, Cambridge, MA, 1955. ISBN 0-8371-8434-7.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. The United Nations parallel corpus v1.0. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1561.

# A Basic concepts

## A.1 Probabilistic Models

In mathematical modeling, a **model** is a system of mathematical equations describing a given phenomenon. It provides a mathematical way to understand the phenomena in the nature around us. Models are characterized by the input variables (also called parameters), output variables and assumptions (Bender, 1991). There are two main types of models: deterministic and probabilistic. **Deterministic models** produced the same output while **Probabilistic models** include uncertainty, i.e their outputs change based on probabilistic factors. This report is focused on probabilistic models.

## A.2 Autoregressive Language Models

Probabilistic models are present in Natural Language Processing (NLP) and constitute the foundation for many language modeling tasks, including text classification, and sentiment analysis. In NLP, probabilistic models appear in the terms **Language Models** or **Autoregressive Language Models**. Language Models predict the next word or subword given a sequence of previous words. Autoregressive Language Models are Language Models that include their own previous output as input to predict the next word or subword.

**Example:** Given an non-negative integer $n$, a $n$-gram is Language Model that predicts the next word or subword based on the $n-1$ words in the context (Jurafsky and Martin, 2024, chap. 3). The unigram ($1$-gram) is not an autoregressive language model because its assignes probability to next word independently.

## A.3 Log-linear Model

A probabilistic model is **log-linear** if the logarithm of its output is a linear combination of its parameters.

## A.4 Transformers

The rapid advancement of NLP has set the transformers as benchmark in Natural Language Models. Their popularity and efficiency lie in the fact that they learn a contextual representation of each word at each position. The main concept beyond this ability of transformers is the attention mechanism.

### FeedForward Layer

A **feedforward network** is a type of artificial neural network where information moves in only one direction: from the input layer, through a given number of hidden layers, to the output layer (see Figure 1).

### Self-attention mechanism

The core idea is, given a sequence of words $w_1$, $w_2$, $\dots$, $w_n$, to compare each word $w_j$ within it to a given word $w_i$ in order to understand and capture their relevance in the current context[1]. In reality, the comparison is achieved with the embedding vectors of each word using the dot product.

The attention mechanism in transformers uses three representations of each word in the input sequence named **key**, **query** and **value** and learnt via the weight matrices $K$, $Q$ and $V$ (Vaswani et al., 2017). We describe the causal attention mechanism in four steps.

[1] Here, *context* refers to both prior and succeeding words, depending on whether the attention mechanism is causal or bidirectional.

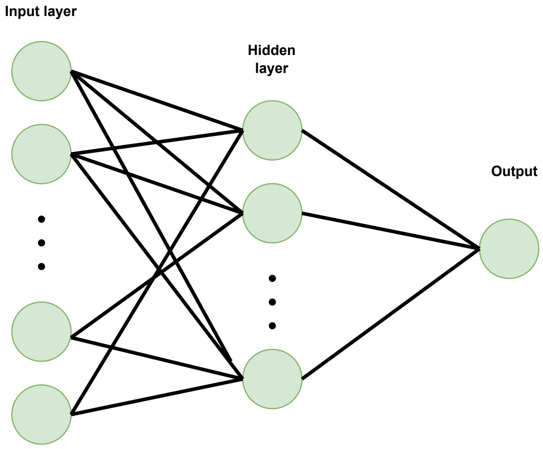


Figure 1: Feedforward neural network: Illustration of a neural network with an input layer, one hidden layer, and an output layer.

First, considering that words are embedded in $\mathbb{R}^d$ vector space, the key, query and value representations of the word $w_j$ are respectively defined by linear projection:

$$\begin{aligned} k_j &= Kx_j \\ q_j &= Qx_j \,, \\ v_j &= Vx_j \end{aligned} \tag{A.1}$$

where the weight matrices are $K \in \mathbb{R}^{d_k \times d}$, $Q \in \mathbb{R}^{d_q \times d}$; $d_k$, $d_q$ and $d_v$ are respectively the dimensionalities of the key, query and value vectors.

Second, given a current word $w_i$, its relevance with the preceding words in the input is measured by the dot product their query and key representations:

$$score(x_i, x_j) = \frac{q_i \cdot k_j}{\sqrt{d}}, \ \forall j \leq i. \tag{A.2}$$

which is down scaled to avoid to avoid large values due to high-dimensional vectors (Vaswani et al., 2017; Jurafsky and Martin, 2024).

Third, these scores are then normalised using the softmax function and define proportional relationship between words through the weights:

$$\alpha_{ij} = \frac{\exp\left(score(q_i, k_j)\right)}{\sum\limits_{j=1}^{i} \exp\left(score(q_i, k_j)\right)}, \ \forall j \leq i. \tag{A.3}$$

Fourth, for the current word $i$, the output of the self-attention is computed by the weighted sum:

$$a_i = \sum_{j=1}^{i} \alpha_{ij} v_j, \tag{A.4}$$

where the weights $\alpha_{ij}$ define the contribution of each word $w_j$ to the representation of $w_i$.

The equations (A.1), (A.2), (A.3) and (A.4) describe the causal attention mechanism focusing on one single word at the same. However, the whole process can be parallelise and aslo can be extent to bidirectional case as well (Jurafsky and Martin, 2024).

In general, transformers use multiple attention mechanisms (named heads) to capture different aspects of the relationships between words in the input, and their outputs are concatenated to improve the model's ability to focus on diverse information (Vaswani et al., 2017; Jurafsky and Martin, 2024). In the literature, this kind of attention mechanisms are commonly called **multi-head attention**.

### Transformers

Introduced by Vaswani et al. (2017), a transformer architecture has, in general, two principal components: encoder and decoder (see fig. 2). Each of these components are made of transformer blocks which consist in multi-head self-attention following by feed-forward layer (Jurafsky and Martin, 2024). Also the self-attention and feed-forward layers are followed by residual connections and layer normalization to ensure stability during training.

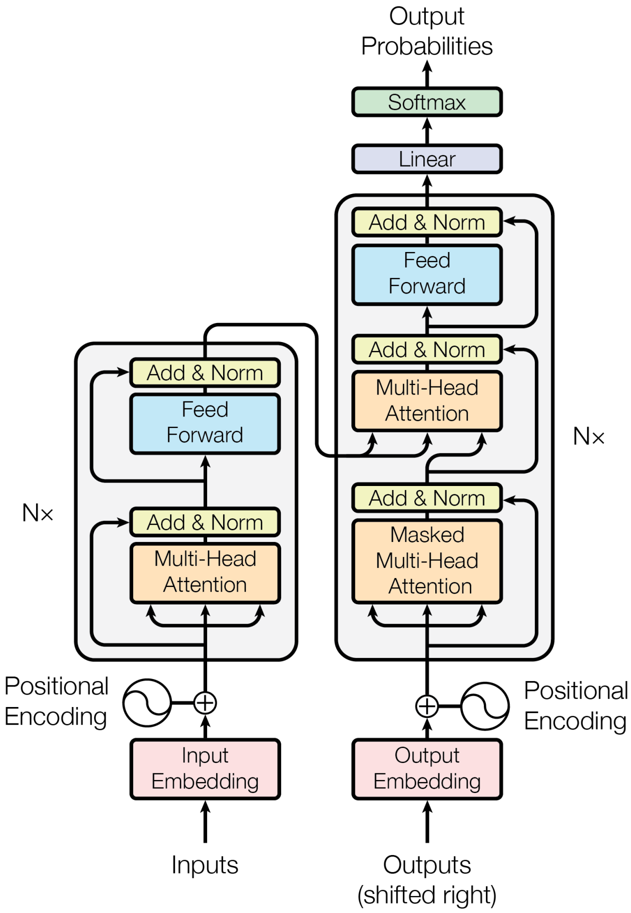


Figure 2: Transformer Architecture. Illustration of the architecture of the Transformer model, showing both the encoder (left) and decoder (right) layers. The encoder consists of a multi-head attention mechanism followed by a feed-forward network with residual connections and layer normalization. The decoder has an additional masked multi-head attention layer to handle autoregressive predictions, feeding into a softmax output. Source: (Vaswani et al., 2017).

# B Validation loss

Figure 3 presents the validation loss over a series of training steps for Small Backpack-eu-8 and Small GPT2-eu trained on Europarl. This loss decreases as the number of training steps increases. For small Backpack-eu-8, validation loss consistently remains lower than the one of small GPT2-eu model across training steps, indicating better generalization performance.

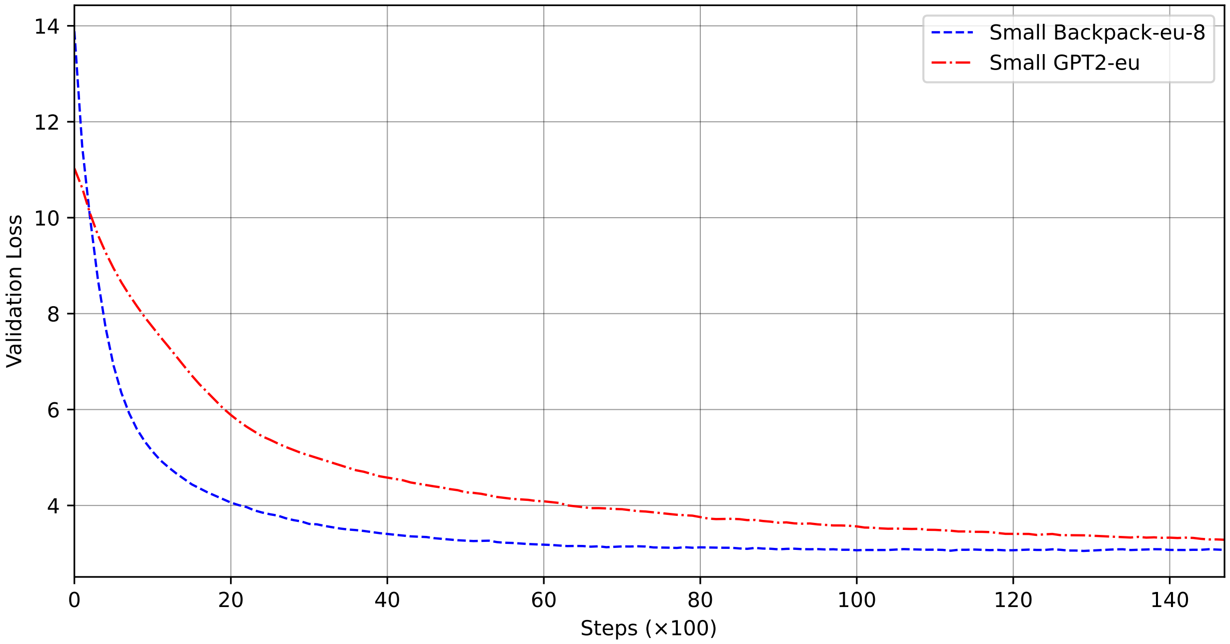


Figure 3: Validation Loss vs. Epochs for Small Backpack-eu-8 and Small GPT-2-eu Models.

# C Sense Visualization

These following tables present the senses of some words. Each line corresponds to a sense and lists the top-5 meanings of the target words. The concerned words are: democracy, droits (rights), démocratie, emploi (employment), first, good, how, justice, law, moins (less), nouveau (new), nécessaire (necessary), où (where), politique (politics), quick, rights, and égalité (equality).

| Sense | democracy | | | | |
|---|---|---|---|---|---|
| 1 | fragile | troubled | caricature | oppressed | opp |
| 2 | gouvern | rule | human | govern | régner |
| 3 | movement | functioning | autour | stables | . |
| 4 | democracy | democracies | Belarussian | ballot | democratic |
| 5 | Rule | approuvera | market | everyone | chacun |
| 6 | Nepal | anywhere | poses | remains | throughout |
| 7 | the | ’ | ’s | ’ | deserves |
| 8 | activist | campaigners | activists | versus | religieuse |

| Sense | droits | | | | |
|---|---|---|---|---|---|
| 1 | restreints | bafoués | foul | violés | fondamentaux |
| 2 | devoirs | obligations | inviolable | obligations | libertés |
| 3 | procéduraux | antidumping | civils | syndicaux | sociaux |
| 4 | constitutions | indivisible | droits | universels | CEDH |
| 5 | polluer | insured | homme | douane | débiteur |
| 6 | enfant | ami | ère | tibétaine | I |
| 7 | de | suffrages | libertés | peuples | voix |
| 8 | douanes | douane | prélevés | émission | émission |

| Sense | démocratie | | | | |
|---|---|---|---|---|---|
| 1 | façade | malmen | VD | caricature | inexistantes |
| 2 | gouvern | droits | séparation | liberté | pluralité |
| 3 | parlementarisme | naissante | humanité | multi | parlementaire |
| 4 | démocratie | urnes | ocratie | ballot | Belarussian |
| 5 | Rule | approuvera | puisse | puissent | peuvent |
| 6 | Birmanie | date | Ouzbékistan | Indonésie | Afghanistan |
| 7 | à | et | de | du | avec |
| 8 | représentative | participative | directe | interne | chil |

| Sense | emploi | | | | |
|---|---|---|---|---|---|
| 1 | perdus | noir | précaires | instables | menacés |
| 2 | allocations | formations | réinsertion | qualifications | horaires |
| 3 | décent | décents | sûr | “ | appropriés |
| 4 | employabilité | emploi | Employment | chômeurs | emploi |
| 5 | entrepreneur | accompl | regroupant | maritime | upgrading |
| 6 | suites | dépend | 1997 | 1999 | attendus |
| 7 | à | exactement | avec | concernant | de |
| 8 | salarié | rémunéré | abusif | vacant | atypiques |

| Sense | first | | | | |
|---|---|---|---|---|---|
| 1 | foremost | all | corrects | alerted | became |
| 2 | step | reading | échéance | milestone | urgency |
| 3 | glance | reading | resort | steps | sight |
| 4 | second | Secondly | foremost | secondly | troisième |
| 5 | ever | hand | ever | rate | filed |
| 6 | half | quarter | few | tranche | pillar |
| 7 | time | times | these | consecutive | contact |
| 8 | token | recommendation | reason | instance | sentence |

| Sense | good | | | | |
|---|---|---|---|---|---|
| 1 | practice | engaging | selectively | sense | notices |
| 2 | faith | evil | conscience | sensible | than |
| 3 | intentions | fortune | neighbourly | neighbourliness | quality |
| 4 | bad | evil | good | enemy | imperfect |
| 5 | men | alike | of | throughout | besides |
| 6 | sir | ay | fer | solving | job |
| 7 | at | our | the | how | its |
| 8 | stead | thing | deal | illustration | idea |

| Sense | how | | | | |
|---|---|---|---|---|---|
| 1 | describe | interp | appreciates | schedule | views |
| 2 | best | ? | employability | devraient | resource |
| 3 | oni | else | itt | matters | valuable |
| 4 | how | what | sink | aspire | How |
| 5 | progresses | evolve | weighting | square | when |
| 6 | unwilling | fores | icious | ents | ignorant |
| 7 | much | worked | things | many | the |
| 8 | circumstance | reference | exception | catalyst | amendment |

| Sense | justice | | | | |
|---|---|---|---|---|---|
| 1 | unfairly | accusées | Communautés | generations | denied |
| 2 | home | administered | predictability | dispensed | redress |
| 3 | sociale | system | mondiale | . | indépendante |
| 4 | Tampere | magistrats | juges | justice | Hague |
| 5 | puisse | puissent | assurés | peut | chacun |
| 6 | 11 | date | contre | tempête | 02 |
| 7 | to | ' | à | ' | défis |
| 8 | civile | pénale | sommaire | delayed | chil |

| Sense | law | | | | |
|---|---|---|---|---|---|
| 1 | discriminate | jungle | violated | contro | circum |
| 2 | jurisprudence | order | repealed | amnesty | abrog |
| 3 | su | enforcement | making | garantissant | passed |
| 4 | Constitutions | law | jur | Arrest | juges |
| 5 | makers | informing | expel | location | acquisitions |
| 6 | centuries | hitherto | decades | behind | . |
| 7 | hopes | préoccupations | expectations | optimism | requêtes |
| 8 | marti | su | takeover | incitement | Helms |

| Sense | | | moins | | |
|---|---|---|---|---|---|
| 1 | mieux | légiférer | vite | parle | nerait |
| 2 | que | dollars | qu | dollar | évid |
| 3 | nombreux | amples | valu | enclin | chers |
| 4 | possible | possibles | moins | davantage | obtient |
| 5 | jamais | disant | monde | nécessite | coûteuse |
| 6 | répandues | répandue | favorisées | spectaculaires | bien |
| 7 | insurmont | équivalente | équivalent | partielle | paradoxal |
| 8 | quarante | 1 | dépendre | dix | trois |

| Sense | | | nouveau | | |
|---|---|---|---|---|---|
| 1 | prometteur | autocritique | rass | enj | instructive |
| 2 | 2016 | entrera | CFP | succéder | 2020 |
| 3 | entrants | ièmement | oque | arrivants | remerciant |
| 4 | nouvelles | nouveaux | nouveau | nouvelle | ancienne |
| 5 | comptera | vingtième | jamais | allong | franchi |
| 6 | z | CN | millénaire | mi | rez |
| 7 | départ | cycle | paradigme | vaccin | raisonnement |
| 8 | millénaire | round | Traité | nationalisme | mode |

| Sense | | | nécessaire | | |
|---|---|---|---|---|---|
| 1 | autocritique | urgent | répéter | persuasion | évoquer |
| 2 | amélioreront | spécifier | harmonisée | résorber | disposer |
| 3 | . | lic | pressant | and | tés |
| 4 | efficacement | indispensables | nécessaire | indispensable | primordiales |
| 5 | importé | millions | lumière | textiles | tissu |
| 6 | défendue | vu | avancée | résid | . |
| 7 | annexe | faille | Questions | les | à |
| 8 | Premièrement | urgemment | neutraliser | parameter | territori |

| Sense | | | où | | |
|---|---|---|---|---|---|
| 1 | suffira | remplit | réagit | suffit | commune |
| 2 | “. | où | devraient | Jérusalem | planète |
| 3 | vir | ues | rir | préparatifs | ir |
| 4 | assistons | tolèrent | où | procurent | abritent |
| 5 | bât | où | pêché | excédentaire | doigt |
| 6 | imaginables | inexistant | mérit | inapproprié | justifi |
| 7 | règne | sévit | prévaut | siègent | dominent |
| 8 | upon | attendre | parents | ils | gens |

| Sense | politique | | | | |
|---|---|---|---|---|---|
| 1 | érales | inadéquates | unes | libérales | divergentes |
| 2 | consisté | échoué | renational | cooperation | échec |
| 3 | étrangère | énergétique | spatiale | extérieure | industrielle |
| 4 | politique | COS | politiques | indissociable | Politique |
| 5 | mène | adoptera | spectre | substances | advertis |
| 6 | annuelle | jusqu | ère | inchangée | vis |
| 7 | emotions | suffrages | setbacks | oppositions | interrogation |
| 8 | voisinage | agricole | structurelle | asile | concurrence |

| Sense | quick | | | | |
|---|---|---|---|---|---|
| 1 | fixes | notices | appeals | reins | pil |
| 2 | than | plau | simple | satisfying | tier |
| 3 | ach | haleine | fixes | fix | accessions |
| 4 | quickly | quick | faster | fast | quickest |
| 5 | sides | besides | emerge | on | of |
| 6 | vation | deaths | oning | candidatures | resolving |
| 7 | to | enough | this | the | its |
| 8 | fix | ly | ment | word | fix |

| Sense | rights | | | | |
|---|---|---|---|---|---|
| 1 | unmarried | suspects | indigenous | violated | derive |
| 2 | duties | obligations | wrongs | freedoms | exercised |
| 3 | records | PEV | multilatéraux | record | zéro |
| 4 | rights | Universal | constitutions | right | Covenant |
| 5 | holders | insured | child | regardless | defence |
| 6 | raging | around | decades | over | there |
| 7 | life | freedoms | dignity | everything | expectations |
| 8 | campaigners | activist | lawyer | abuses | defenders |

| Sense | égalité | | | | |
|---|---|---|---|---|---|
| 1 | existed | exercés | nul | gay | obtenues |
| 2 | chances | égalité | égales | mutual | fraternité |
| 3 | APP | les | genres | chances | interculture |
| 4 | égalité | masculins | discriminations | inégalité | féminin |
| 5 | recevront | tous | utilisateur | Dire | strata |
| 6 | traitement | 2010 | 2002 | 2006 | 2009 |
| 7 | de | à | avec | suivantes | 67 |
| 8 | entre | salariale | nel | raciale | nellement |

# D Sense Distribution

We present additional examples of the distribution of senses across different contexts, including case studies of cognates and shared words (see Figures 4, 5, 6, 7, 8 and 9).

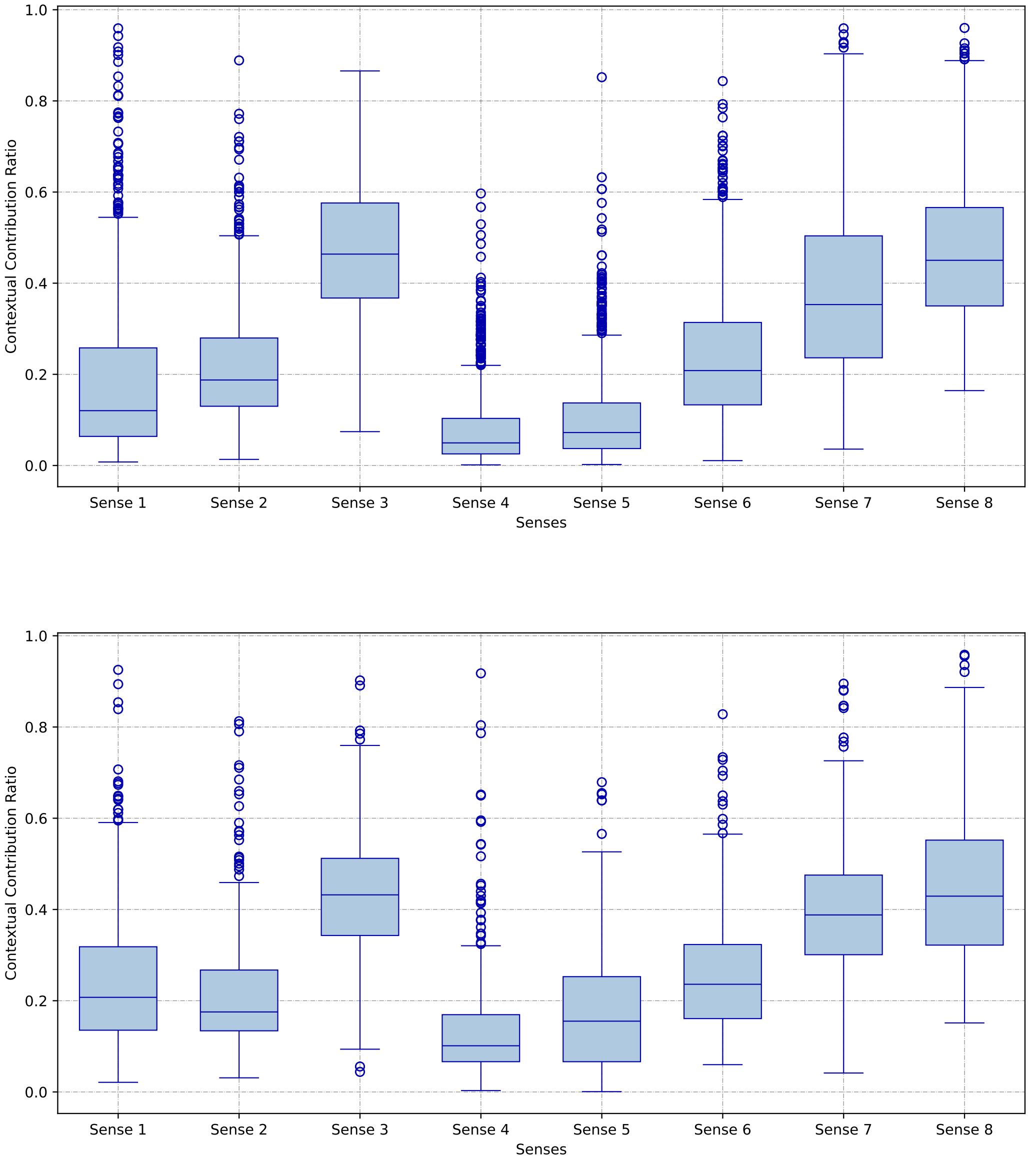


Figure 4: Box plot of the contextual ratios of words minister (top) and ministre (below) per senses.

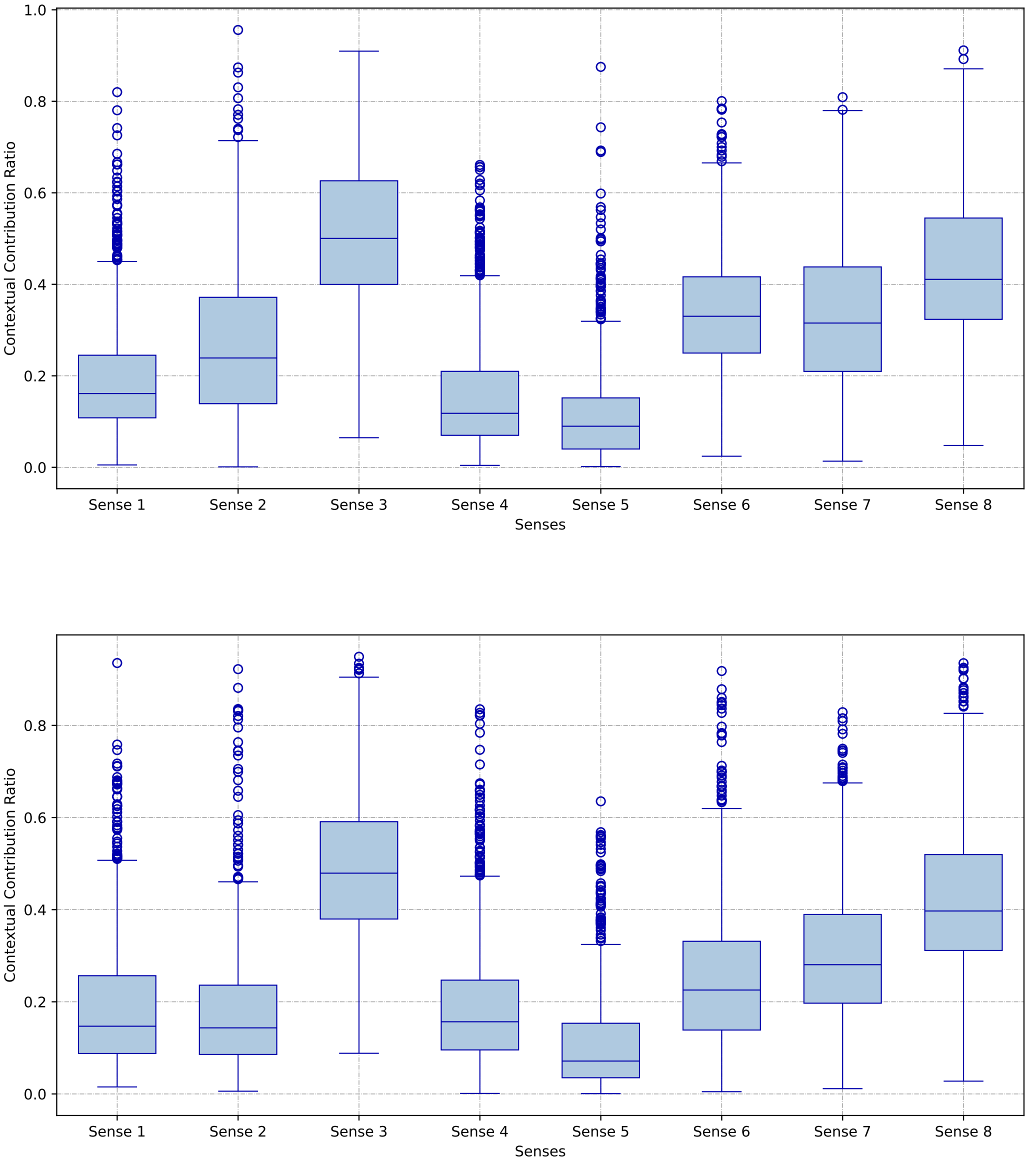
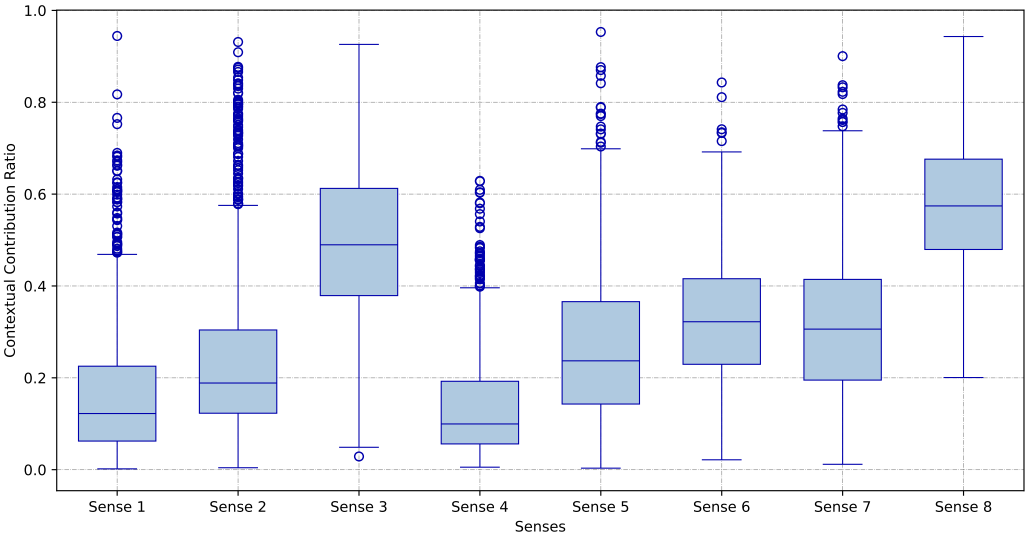


Figure 5: Box plot of the contextual ratios of words problem (top) and problème (below) per senses.



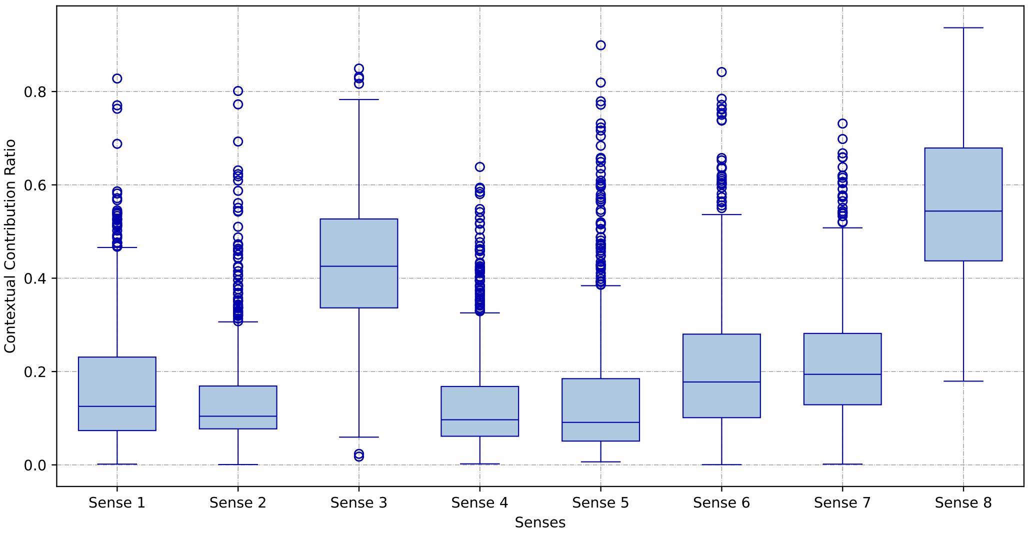


Figure 6: Box plot of the contextual ratios of words education (top) and éducation (below) per senses.

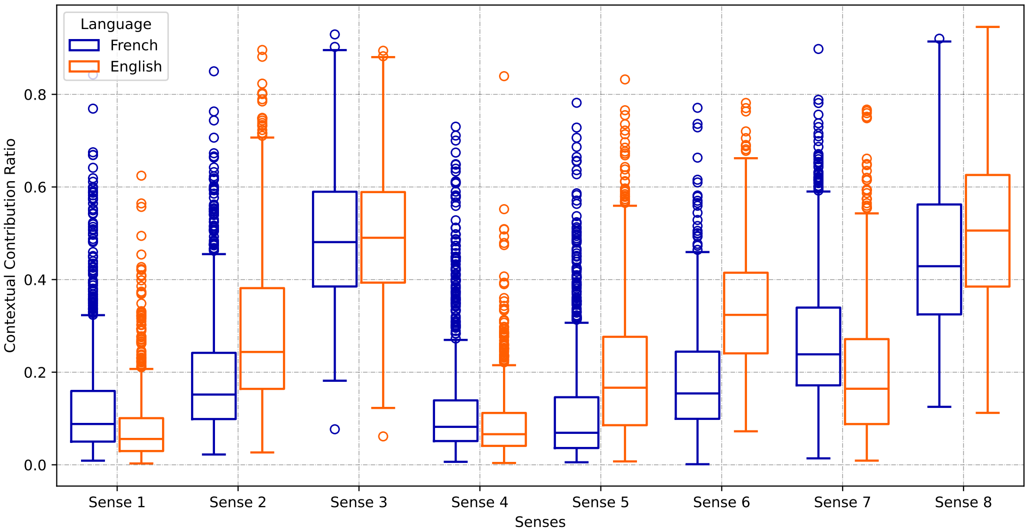


Figure 7: Box plots of the contextual ratios of word dialogue in French and English contexts. Blue is associated with the French contexts and Orange to English contexts.

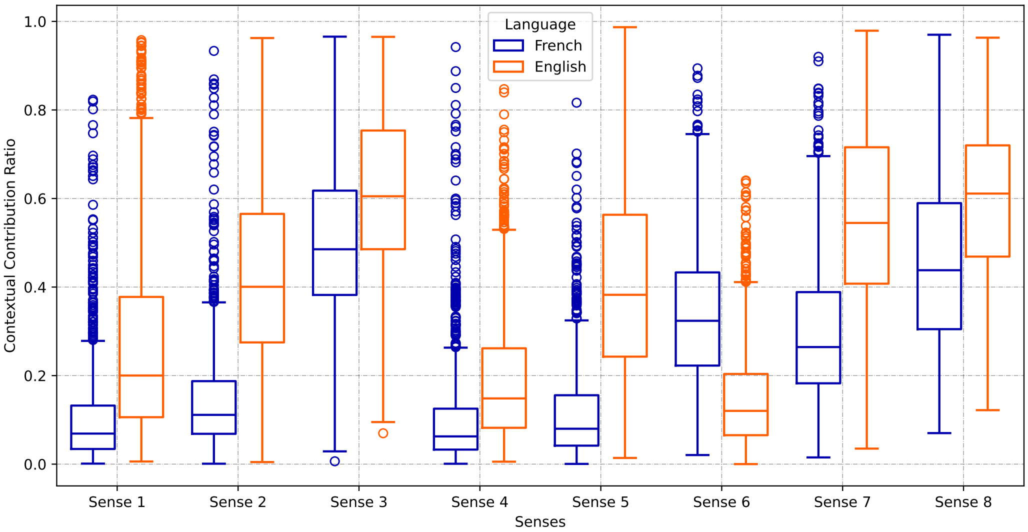


Figure 8: Box plots of the contextual ratios of word international in French and English contexts. Blue is associated with the French contexts and Orange to English contexts.

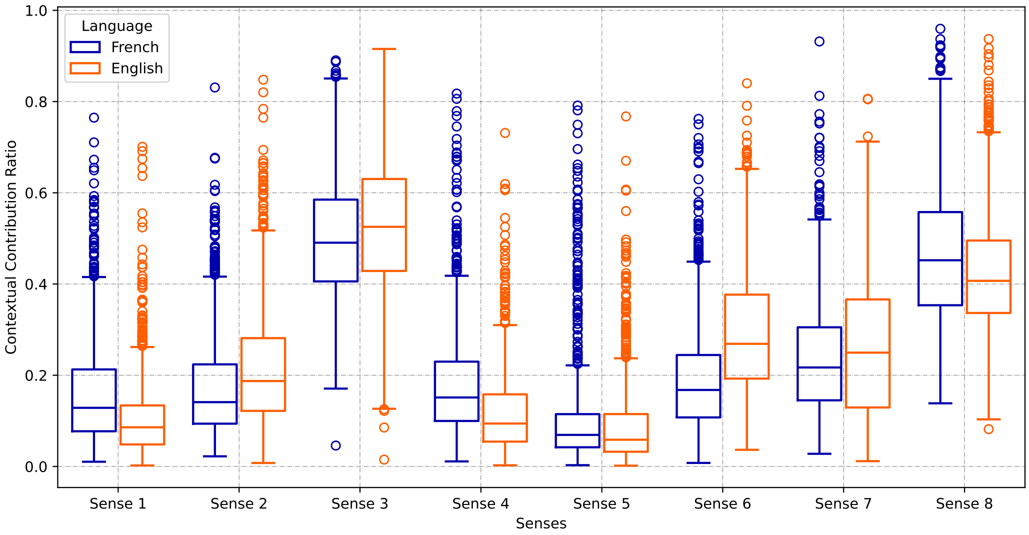


Figure 9: Box plots of the contextual ratios of word structure in French and English contexts. Blue is associated with the French contexts and Orange to English contexts.