SELEXINI – a large and diverse automatically parsed corpus of French

Manon Scholivet[1], Agata Savary[1], Louis Estève[1],
Marie Candito[2], Carlos Ramisch[3],

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France,
[2] Université Paris Cité, CNRS, LLF, Paris, France,
[3] Aix Marseille Univ, CNRS, LIS, Marseille, France

first.last@lisn.fr[1], first.last@u-paris.fr[2], first.last@lis-lab.fr[3]

Abstract

The annotation of large text corpora is essential for many tasks. We present here a large automatically annotated corpus for French. This corpus is divided into two parts: the first from BigScience, and the second from HPLT. The annotated documents from HPLT were selected in order to optimise the lexical diversity of the final corpus SELEXINI. An analysis of the impact of this selection was carried out on syntactic diversity, as well as on the quality of the new words resulting from the HPLT part of SELEXINI. We have shown that despite the introduction of interesting new words, the texts extracted from HPLT are very noisy. Furthermore, increasing lexical diversity did not increase syntactic diversity.

1 Introduction

Morphosyntactic treebanks are cornerstones of grammar induction (Zhu et al., 2020) and of morphosyntactic parsing, whether in monoligual (Dary et al., 2022), multilingual (Straka, 2018) or cross-lingual (Glavaš and Vulić, 2021) contexts. They help to probe language models for linguistic knowledge possibly encoded therein (Shen et al., 2023), and for challenges, e.g. related to syntactic constructions, which these models might fail to appropriately address (Bonial and Tayyar Madabushi, 2024).

Treebanks are also fundamental resources in research on language. They enable studying linguistic properties within or across languages (Levshina et al., 2023), examining the appropriateness of language universals (Brosa-Rodríguez and Kahane, 2024), formalising and searching for complex phenomena such as constructions (Weissweiler et al., 2024a) or documenting low-resourced and endangered languages and dialects (Pugh and Tyers, 2024), inter alia.

For some of such research questions, manually annotated treebanks are not enough to check generalisations and touch upon long-tail phenomena (Sheinfux et al., 2019). In such cases, corpora automatically annotated for morphology (Baroni et al., 2009) and/or syntax (van Noord et al., 2013; Ginter et al., 2013) are used (Schneider, 2011; Bloem et al., 2014).

Our objective is to build such a morphosyntactically parsed corpus for French which would fulfill two conditions. First, it should be large but manageable, i.e. its parsing, storage and maintenance cost should not be prohibitive. Second, it should still have sufficient lexical and syntactic diversity to serve studies in which long-tail phenomena play important roles, such as frame induction (QasemiZadeh et al., 2019), identification of multiword expressions (MWEs) unseen in manually annotated corpora (Ramisch et al., 2020), probing language models for rare but interesting syntactic phenomena (Misra and Mahowald, 2024; Weissweiler et al., 2024b), etc.

To this aim, we use two very large raw corpora: BigScience (Laurençon et al., 2022) and HPLT (**H**igh **P**erformance **L**anguage **T**echnologies) (De Gibert et al., 2024). We select a clean subset of BigScience and we extend it with fragments of HPLT sampled so as to increase the diversity of the whole resulting corpus, henceforth called SELEXINI[1].

Even if both lexical and syntactic diversity are of interest for us, the latter requires pre-existing syntactic annotation, which is prohibitive with a corpus as large as HPLT. Therefore, for data sampling we only use lexical diversity, formally defined as entropy over word types. This sampling strategy likely also has an impact on syntactic diversity, and more generally on the Zipfian distribution of the corpus, as new words and syntactic structures are added and the pre-existing ones change their frequencies. In this context, our research questions

[1] http://hdl.handle.net/11372/LRT-5822

are:

Q1 How does data sampling driven by lexical diversity influence the syntactic diversity of the corpus?

Q2 What are the resulting quantitative and qualitative properties of the corpus in terms of its Zipfian distribution?

Q1 and Q2 are studied in a comparative context. Namely, we compare BigScience and two extracts of HPLT: one sampled by diversity and another random.

The paper is organized as follows. We briefly discuss related work on French syntactic treebanks (Section 2). We define the diversity measures used for data sampling and corpus comparison (Section 3). We describe the guiding principles (Section 4) used in the corpus construction, as well as the source data (Section 5), their sampling (Section 5.3) and parsing (Section 6). We perform a comparative analysis of two parts of the resulting corpus (Section 7). We finally discuss the limitations of our approach (Section 8) and the conclusions (Section 9).

2 Related work

In dependency syntax, two annotation schemas come with large manually annotated treebanks for French. Historically the FTB-dep schema is a French-specific dependency schema, defined as the result of automatic conversion (Candito et al., 2010) of the 18k phrase-structure trees of the French Treebank (Abeillé et al., 2003). An out-of-domain additional corpus of 3k sentences (the Sequoia corpus (Candito and Seddah, 2012)) is also available in this schema. Then, treebanks of various genres were either annotated under or converted to the Universal Dependencies (UD) schema (Nivre et al., 2020), for a total of 29,735 sentences in UD version 2.15. Concerning available large annotated French corpora, the web-based 1.6 billion token corpus frWac[2] was automatically POS-tagged. Available syntactically parsed corpora are either much smaller (a 150 million token regional news corpus (Seddah et al., 2012)) or mono-genre (parsed French Wikipedia distributed for the CoNLL 2017 shared task[3]).

[2] https://wacky.sslmit.unibo.it/doku.php?id=corpora

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989

This shows that no currently existing openly available and morphosyntactically parsed resource is large and diverse enough to serve our needs.

3 Diversity measures

Inspired by formal approaches to diversity (Rényi, 1961; Chao et al., 2014; Morales et al., 2020), we consider it to be a property of populations/systems (here: datasets) whose *elements* can be apportioned into *categories*. For lexical diversity, we define categories as word types and elements as their occurrences in the dataset. For instance, the toy corpus with one sentence from Figure 1(a) contains 8 elements, each one belonging to a different category.

For syntactic diversity, we understand categories as complete syntactic subtrees (where for each node all its children nodes are also included), containing only POS labels and dependency relations. Elements are occurrences of these subtrees in the corpus. Figure 1(b) shows a sample category with two elements in Figure 1(a), highlighted in blue. Figure 1(c) contains another category which does not occur in Figure 1(a), although *y* and *jouent* match the tree fragment in Figure 1(c). This is because the category enrooted in V has to contain all children of V. With this understanding of categories, the example in Figure 1(a) has 5 categories (leaves D, A and PRO, and 2 non-trivial subtrees enrooted in NC and in V) and 8 elements (one per word).

Once elements and categories are defined, diversity can be measures along 3 dimensions: *variety* (which deals with the number of categories), *balance* (which tackles how even the distribution of elements into categories is) and *disparity* (which aggregates pairwise distances between categories). Many diversity measures were proposed in the past, especially in ecology, and most of them are hybrids between at least two of those dimensions. One of them is *richness*, i.e. simply the number of categories n, which is a pure variety. Another one is *entropy* (Shannon and Weaver, 1949), defined by (1), which is a hybrid between variety and balance, where $\Delta_n = \{p_1, ..., p_n\}$ is the distribution of categories. We will use H_{lex} and H_{syn} to refer to entropy over word types and syntactic subtrees, respectively, as defined above.

$$H\left(\Delta_n\right) = -\sum_{i=1}^{n} p_i \log_b\left(p_i\right) \quad (1)$$

In natural language data Zipfian distributions, defined by (2), and their generalisations – Zipf-

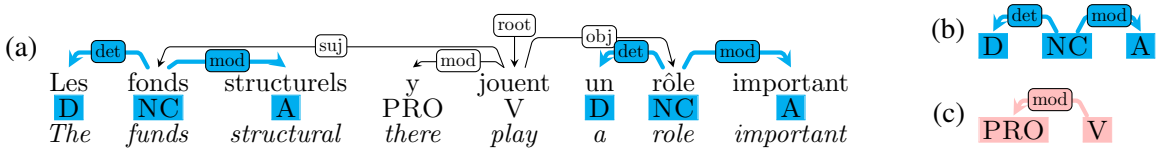


Figure 1: (a) A simplified syntactic tree in FTB-dep schema: *The structural funds play an important role there*; (b) a syntactic category with its two elements highlighted in (a); (c) a syntactic category not occurring in (a), despite the subtree overlap in *y jouent*.

Mandelbrot distributions, defined by (3)[4] – are pervasive. The inverse of their curvature parameter $-s$ can be considered a good balance measure (Lion-Bouton et al., 2022), as it achieves its maximum with $s = 0$, i.e. with a perfectly uniform distribution, and diminishes when the curvature grows (i.e. the data are more and more unbalanced).

$$Z_{s,n}(x) = \left(x^s \sum_{i=1}^{n} i^{-s} \right)^{-1} \quad (2)$$

$$Z^q_{s,n}(x) = (x+q)^{-s} \left(\sum_{i=1}^{n} (i+q)^{-s} \right)^{-1} \quad (3)$$

4 Best practices in corpus construction

There are several best practice recommendations when it comes to creating corpora. The work of (De Pauw, 2006) motivated a number of choices for the construction of this corpus.

Retrieving the data with their **context** makes it possible to analyse the corpus in more detail. It will be easier to understand to whom 'she' or 'he' refers in a text if we have the text that precedes this sentence. For this, the two corpora on which we are relying (BigScience and HPLT) are ideal because they contain complete documents, which we then segment into sentences.

The data collected must match as closely as possible the language studied (Biber, 1993) in order to obtain a certain level of **representativeness**. This question of representativeness is explored when selecting diversified data. However, the **homogeneity** of the data must not be sacrificed for the sake of diversity. This is why we separate data from the two original corpora (HPLT and BigScience), which are very different in their respective genres (web crawls on the one hand, and parliamentary and Wikipedia texts on the other).

In order for the corpus to be reusable by the community, it is important to use **standards** from the community. Data annotation according to the Universal Dependencies schema was also performed for this reason, in addition to the FTB-dep schema which is specialised for French corpora.

5 Source data

The choice of data to annotate was made in two steps: in the first, preference was given to texts with information on their origins, in order to encourage the use of diversified sources. We focused on French data from the BigScience[5] (Laurençon et al., 2022) as the basis for the SELEXINI corpus. The second selection step was done in order to increase the quantity and the diversity of the corpus data. For this, the HPLT[6] (De Gibert et al., 2024) corpus was chosen. Less clean than BigScience, this part of the corpus nevertheless contains the most diverse part of the data.

5.1 BigScience

The **BigScience** initiative aims to make large quantities of data available in many languages, with the intention of facilitating the training of large multilingual language models (LLMs). Created using pseudo crawls (crawls based on certain predefined domain names), this dataset remains fairly clean.

We chose to work on the parts of the dataset from Europarl, the French part of the United Nations Parallel Corpus and Wikipedia, mainly because of their large size (1.5 billion tokens). Henceforth, this subset will be called BASE. Additionally to its large size, BASE fulfills our other criteria: the metadata allow to easily deduce the language and text genres, no or few multilingual texts are included, licenses are clear and compatible with the intended use of our corpus.[7] The Wikisource subset of BigScience was also considered, but presented too many problems (text starting in the middle of

[4] With $q = 0$ we have $Z_{s,n}(x) = Z^q_{s,n}(x)$.

[5] https://huggingface.co/bigscience-data

[6] https://hplt-project.org/datasets/v1.2

[7] The BigScience RAIL license is inspired both from open licenses and fairness principles: https://bigscience.huggingface.co/blog/the-bigscience-rail-license.

a sentence, HTML tags, encoding problems, sentences in Old French, etc.).

5.2 HPLT

BigScience only contains two text genres: Wikipedia articles and parliamentary debates. In order to achieve a better diversity of genres, we benefit from **HPLT** (De Gibert et al., 2024), a massive multilingual dataset of texts provided by Internet Archive and CommonCrawl. These texts were cleaned by the HPLT authors so as to eliminate documents from dubious URLs (possibly pornographic, racist, etc.) and filter out noisy paragraphs. The remaining documents were then sorted according to he majority vote over a number of language predictors. We work with the cleaned version of French HPLT, containing around 99.59M documents and 122.88B words.

This dataset is still not perfect:

- the filter for setting aside problematic documents is based mainly on the document URL, and some undesired texts can still remain
- the language identification is sometimes erroneous, particularly when several languages are present in the same text
- the data cleaning keeps some uninteresting documents (lists of phone numbers, number plates, etc.)

However, this dataset covers a wide variety of fields and should help increase the diversity of the BASE corpus, as discussed in the following section.

5.3 Diversity-driven data sampling

Diversity of datasets is usually strongly dependent on their sizes. Since we are interested in comparative studies, the compared corpora should have similar sizes. Therefore, we sample HPLT for a subset of a size which would be roughly equivalent to BASE (1.5 billion tokens), while keeping entire texts intact. To reduce computation, we only use a subcorpus containing 6B documents randomly selected from HPLT. We sample it by batches and for each batch we select the document which, added to BASE, maximizes its lexical diversity measured by H in (1). We stop when we exceed the intended size of 1.5 billion tokens. If all batches have been processed and the intended size is not reached, we decrease the size of a batch and reiterate.

The final subset of HPLT selected in this way is called HPLT$_{div}$. For comparison, we also randomly select another subset of HPTL of roughly the same size as HPLT$_{div}$ and we call it HPLT$_{rand}$.

Merging BASE with HPLT$_{div}$ on the one hand, and with HPLT$_{rand}$ on the other hand, yields the final SELEXINI corpus and its non-diverse equivalent SELEXINI$_{rand}$. The following section describes the process of automatic parsing of SELEXINI. Section 7 is then dedicated to comparing the quantitative and qualitative properties of BASE, HPLT$_{div}$ and HPLT$_{rand}$, so as to address the research questions Q1 and Q2.

6 Target annotation schemas and model training

From the outset, we opted for dependency syntax. Morphosyntactic annotation of our corpus can only be done automatically, so to choose the target annotation schemas, we were constrained by the availability of large enough training sets. We thus had two candidates: the monolingual FTB-dep schema or the UD schema (cf. Section 2). We aimed at both accurate linguistic description of French, and cross-lingual parallelism, which exactly corresponds to the balance sought for in the UD project. Yet, for specific linguistic traits, it might prove difficult to satisfy both objectives[8]. Indeed, a closer look at the instantiation of UD guidelines in French UD treebanks first shows some diversity in annotation choices (Guillaume et al., 2019). Second, certain specific phenomena were dealt with (i) either by not following the UD guidelines, which breaks the cross-lingual uniformity, or (ii) by following them at the cost of breaking an internal regularity. We provide some examples in Appendix A.

6.1 Models

For all the previously seen reasons, we chose to keep both annotation schemas, FTB-dep and UD, and thus to build two parsed versions of our SELEXINI corpus, thanks to two models.

FTB-dep To train this model we concatenated two treebanks, containing approximately 21k sentences in total:

- the dependency version of the French TreeBank (FTB) (Abeillé et al., 2003), adapted by Seddah et al. (2013);

[8] As put forward in UD's web introduction, which presents UD design as a "subtle compromise" : https://universaldependencies.org/introduction.html.

- the Sequoia[9] treebank (Candito and Seddah, 2012), version 9.2.

While both treebanks have the same main annotation schema (FTB-dep), subtle differences have been introduced over time. In order to get more homogenous training data, we modified the FTB. This harmonisation is described in Appendix B.

UD We use fr_sequoia-ud-2.12 model, one of the models trained on the French treebanks from Universal Dependencies version 2.12, with UDPipe2 (Straka, 2018) and made available by Straka (2023).

6.2 Annotation process

Two different cases were treated to carry out the annotation of the SELEXINI corpus: the annotation with the FTB-dep schema, and the annotation using the Universal Dependencies.

For the annotation using UD, UDPipe 2 was used to carry out all the steps (segmentation, tokenisation, POS tagging, morphological features tagging, lemmas prediction and syntactic analysis).

In the case of the FTB-dep annotation, sentence segmentation and tokenisation were performed using the Bonsai tool[10], designed to specifically handle French. Tagging and parsing were then done with UDPipe 2 as well, but this time using the FTB-dep model described in Section 6.1.

The last step, both for the UD and FTB-dep version, was a lemmas correction phase. While the predicted lemmas on in-domain dev are 99% correct (see Table 1), a qualitative analysis of lemmas for unknown rare word forms on our SELEXINI revealed sometimes absurd predictions[11]. We thus applied lemma correction using the Lefff lexicon (Sagot, 2010)[12].

7 Results

Assessing the quality of annotations is not a trivial task without manually annotated data. We can nevertheless observe the performance of the models on the corresponding dev and test corpora. The results can be seen in Table 1.

Model	Test set	POS	UFeats	LAS
FTB-dep	FTB+Sequoia dev	98.49	94.68	91.11
UD	Sequoia test Gold Tokenisation	99.25	98.01	94.37
	Sequoia test Raw text	98.40	97.19	92.75

Table 1: Scores of the FTB-dep and UD models. The test set for the FTB-dep model is the dev set of the FTB+Sequoia (28 POS tags, 34 dependency labels). For the UD model, the test set of Sequoia is used (17 POS tags, 47 dependency labels)

UDPipe models are frequently used as a baseline thanks to their strong performance. Although the quality of the annotations is better using gold tokenisation than raw text, the results are still good enough to be usable. The model used to annotate in FTB-dep obtains slightly lower scores than the UD model, but as the test corpus and annotation schemes are different, the results are not perfectly comparable and remain acceptable for the annotation task.

We will now compare diversity measures on the different corpora studied. A summary of this information is available in Table 2. The parameters $-s$ et n are computed using equation (3).

7.1 Syntactic Diversity

The algorithm used to select the HPLT$_{div}$ texts aimed to maximize lexical diversity (Section 5.3). We will now evaluate whether this selection also had an impact on syntactic diversity (defined in Section 3) in order to answer our research question Q1.

However, syntactic diversity can only be calculated if we have access to the syntax annotations. The SELEXINI corpus, composed of BASE and HPLT$_{div}$, has been parsed but not HPLT$_{rand}$. Therefore, syntactic diversity is only calculated for the former.

In Table 2, we can observe an increase in the lexical entropy H_{lex} : +0.72 for BASE+HPLT$_{div}$. The opposite trend is visible for syntactic diversity: a decrease of 0.36 point when BASE is augmented with HPLT$_{div}$. Although HPLT$_{div}$ is both more varied (higher n) and more balanced (higher $-s$) than BASE, which leads to a higher entropy from a lexical point of view (8.10 vs. 7.02), HPLT$_{div}$ is less varied and less balanced than BASE from a

[9] https://deep-sequoia.inria.fr/

[10] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

[11] This is the case e.g. for first person verbal form, rare in the FTB+Sequoia training set. Moreover the predicted lemmas sometimes do not match the predicted POS tag. After lemma correction on out-of-domain 2.64 million tokens, 7000 lemmas were modified using the lexicon heuristic. An analysis of the first 100 corrections revealed only one introduced error, and 99 corrected errors.

[12] The heuristic was to replace any predicted lemma unknown in the lexicon by the longest lemma compatible with this word form and POS tag.

Corpus	Size	n_{lex}	$-s_{lex}$	H_{lex}	n_{syn}	$-s_{syn}$	H_{syn}
BASE	1.54	3.5	-1.250	7.02	167	-1.381	7.17
HPLT$_{div}$	1.56	11.7	-1.182	8.10	124	-1.660	6.25
SELEXINI = BASE + HPLT$_{div}$	3.10	13.6	-1.204	7.74	282	-1.457	6.81
HPLT$_{rand}$	1.56	6.4	-1.138	7.42	-	-	-
SELEXINI$_{rand}$ = BASE + HPLT$_{rand}$	3.10	8.7	-1.187	7.41	-	-	-

Table 2: Summary of the sizes of each corpus in billion tokens, the value of their Zipfian parameters, n for the number of categories in millions (higher is better), and $-s$ for the Zipfian curvature (closer to 0 is better). H is the entropy (higher is better). All these measures are computed for the lexical and syntaxtic version.

syntactic perspective.

As a reminder, n_{lex} and n_{syn} correspond respectively to the number of lexical categories (words) and the number of syntactic categories (syntactic subtrees). The number of common lexical categories between BASE and HPLT$_{div}$ is 1.6 million words, i.e. 11.8% of the total final corpus (BASE + HPLT$_{div}$, i.e. SELEXINI). However, in the case of syntax, there are 9 million common trees, which this time represents only 3.2% of the final corpus.

While 74.3% of the lexical categories in SELEXINI originate from HPLT$_{div}$ only, 41.8% of the syntactic categories originate from HPLT$_{div}$. HPLT$_{div}$ therefore has more weight, more impact, on the diversity of SELEXINI than BASE from a lexical point of view. However, this is not true for syntactic diversity.

Now, if we look at the $-s$ parameter of the Zipfian curvature, which is a measure of balance, we can see that in lexical terms, $-s_{lex}$ obtains a better score for HPLT$_{div}$ than for BASE. This is reversed in the case of s_{syn} where HPLT$_{div}$ is clearly less balanced than BASE.

In conclusion, as an answer to Q1, it appears that optimizing lexical diversity with HPLT$_{div}$ did not also improve syntactic diversity. On the contrary, the sampling had the opposite effect, causing syntactic diversity to notably decrease.

7.2 Lexical Zipfian distributions

In this section, we will focus first on the differences between BASE and HPLT$_{div}$. Then, SELEXINI and SELEXINI$_{rand}$ will also be compared. In order to answer the research question Q2, we will first carry out an analysis of the quantitative properties by looking at the different scores in Table 2. Secondly, we will analyse the qualitative properties by exploring the new words added by HPLT$_{div}$ to SELEXINI. This section deals only with lexical diversity.

Quantitative properties Although BASE and HPLT$_{rand}$ have roughly the same size, HPLT$_{rand}$ is more diverse than BASE, whether for entropy H_{lex}, variety n_{lex} or balance s_{lex}. One hypothesis is that the Wikipedia articles and parliamentary debates in BigScience create a certain redundancy in the data, making this dataset a less varied and balanced than those from HPLT.

As seen in the previous subsection, augmenting BASE with HPLT$_{div}$ increased the lexical entropy H_{lex} from 7.02 to 7.74: a gain of 0.72. Augmenting BASE with HPLT$_{rand}$ increased the entropy to 7.41: a gain of only 0.4. HPLT$_{rand}$ has roughly half as many categories as HPLT$_{div}$ (11.7 million and 6.4 million respectively). HPLT$_{rand}$ is therefore much less varied than HPLT$_{div}$ (although it is still more varied than BASE). However, with an $-s_{lex}$ at 1.182 for HPLT$_{div}$ and at 1.138 for HPLT$_{rand}$, the latter is more balanced. So it is likely that the selection algorithm favours variety more than balance.

Qualitative properties For this section, we extracted the vocabularies of BASE and HPLT$_{div}$. We began by identifying new words present in HPLT$_{div}$ that were not present in BASE. A list of around 10 million words was thus extracted. We then got 2 million static embeddings of dimension 300 from Grave et al. (2018). These embeddings were trained on Common Crawl and Wikipedia using fastText, and keeping only the 2 million most frequent words.We can assume that most words without embeddings will be noise. Only 84,526 of the 10 million words have word embeddings. This means that over 99% of the new words in HPLT$_{div}$ are noise.

Nevertheless, we're going to try to identify whether we can find any common points among the non-noisy words in HPLT$_{div}$. We created word embeddings clusters that can be seen in Figure 2. These clusters were obtained by randomly select-

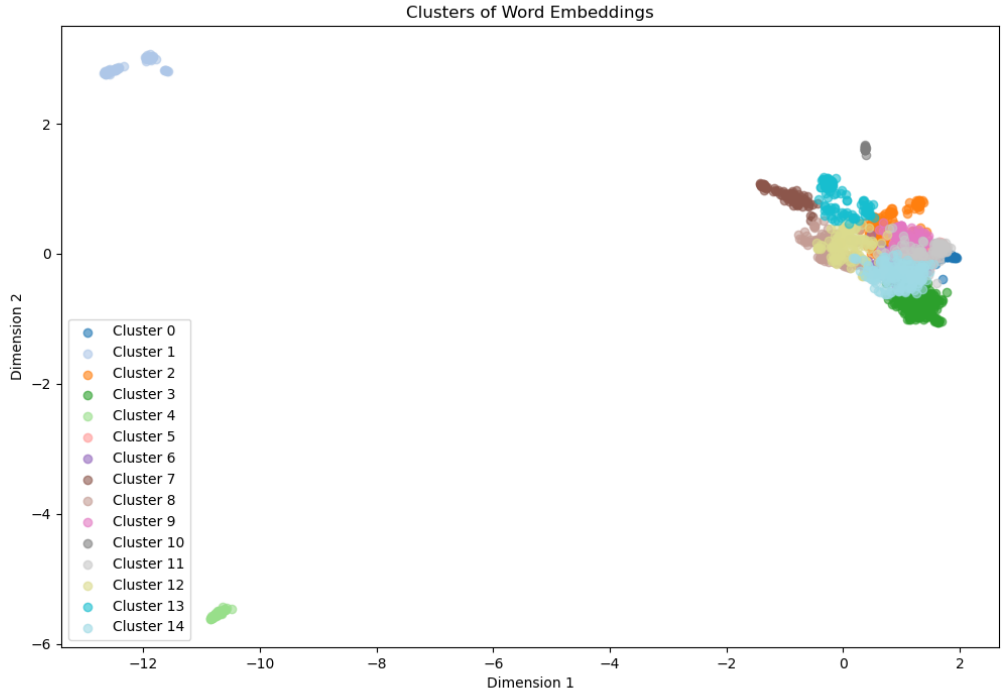


Figure 2: Word embeddings clusters of new words from HPLT$_{div}$.

ing 2,000 words from our list and extracting their embeddings. We then reduce the number of dimensions to 50 using the UMAP algorithm (McInnes et al., 2018). We cluster our embeddings using the K-means algorithm, with K=15[13]. Finally, we reduce our embeddings again to two dimensions using the PCA algorithm, in order to visualise the clusters. Details of the clusters are available in Appendix C. Although the clusters are not perfectly pure, common themes can be identified across most of them.

- Clusters 1 and 4, very isolated from the rest, contain only dates in two different formats (year-month-day for cluster 1 and day-month-year for cluster 4). Cluster 10 also contains numbers only, with a '-'.
- Cluster 8 contains 'sms'-type language: *copinette* (friend), *choupie* (cute) or *merciiii* (thanks).
- Clusters 2, 7 and 13 contain words in other languages : *bedsheets* (English), *polozoni* (Croatian) or *abgerufen* (German).
- Cluster 2 also contains neologisms and concatenations of words: *brocantitude* (flea market attitude), *miseenservice* (commissioning)
- Cluster 7 contains a subcluster with symbols and emojis
- Clusters 0, 5, 11 and 14 contains many spelling mistakes, often due to missing accents : *patrimoin* (heritage), *helices* (propeller), *ludotheque* (toy library), *mesage* (message)
- Clusters 5 and 11 also contain suffixes : *ficiaires*, *ctions*, *geait*, *pondants*
- Cluster 3 contains rare forms of conjugation : *flippent* (they freak out), *débuterez* (you will start) or *chouchoutent* (they pamper)
- Cluster 6 contains words concatenated with a final dot : *normalement.* (normally.), *châteaux.* (castles.), *surf.* (surf.)
- Cluster 12 contains URLs and filenames : *main.php*, *monsite.com*, *top-site*
- Cluster 9 is the only cluster with no specific theme. There are misspells (*pâtissiére* (fe-

[13]The choice of 15 clusters was made empirically.

male baker)), rare words (*non-couvert* (not covered)), foreign word (*cocoon*) and others

In conclusion, although the quantitative study showed that texts that increase variety are more often selected (either because of their greater importance in the entropy, or because they occur more frequently than texts that increase the balance), the qualitative study showed that this variety is almost artificial, because of the very high noise content of the texts from HPLT. Nevertheless, some new "valid" word forms are added, especially rare conjugations, which usefully extend the vocabulary of SELEXINI.

8 Limitations and future work

A first limitation of this work is obviously the presence of a lot of noise in HPLT. Applying the selection algorithm to a corpus without noise could lead to very different results and conclusions. The use of noise reduction techniques could also help to limit the problem (Zhu et al., 2022).

Another limitation is the automatic prediction of labels. These predictions carry the biases of the models used to generate these annotations, which may have only encountered certain rare phenomena on an infrequent basis.

There are many different measures of diversity. Here we focused only on Zipfian parameters and Shannon-Weaver entropies, but some other measures highlight other information. In particular, disparity is another dimension of diversity that we have not explored here, but which would have its rightful place in an analysis of corpus diversity.

9 Conclusions

In this article, we have presented three contributions. The first is the creation of a large automatically parsed French corpus. The second is a study of the impact of lexical diversity-driven data sampling on syntactic diversity. Finally, we also performed a quantitative and qualitative analysis of the lexical diversity resulting from the selection aimed at maximising this same lexical diversity.

The main conclusions are that the selection based on lexical diversity favours variety more than balance, and mainly extracts noise. We also found that there was no positive impact on syntactic diversity, and even that there was a rather negative impact. It would be interesting to understand if this negative impact is due to noisy data or if it is inherent to natural language (e.g. rare and new words might tend to occur in syntactic constructions known for frequent words). More research is still needed to find methods that will maximise lexical diversity while avoiding the problems of noisy texts.

10 Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD010615876 made by GENCI.

References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a French treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*.

Douglas Biber. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Jelke Bloem, Arjen Versloot, and Fred Weerman. 2014. Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1974–1984, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Antoni Brosa-Rodríguez and Sylvain Kahane. 2024. New proposal of greenberg's universal 14 from typometrics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12217–12226, Torino, Italia. ELRA and ICCL.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier,

and Silvio Ricardo Cordeiro. 2020. A french corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2).

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: Treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

Anne Chao, Chun-Huo Chiu, and Lou Jost. 2014. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324. Publisher: Annual Reviews.

Franck Dary, Maxime Petit, and Alexis Nasr. 2022. Dependency parsing with backtracking using deep reinforcement learning. *Transactions of the Association for Computational Linguistics*, 10:888–903.

Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128.

Guy De Pauw. 2006. Developing Linguistic Corpora—A Guide to Good PracticeMartin Wynne (ed.). *Literary and Linguistic Computing*, 22(1):101–102.

Filip Ginter, Jenna Nyblom, Veronika Laippala, Samuel Kohonen, Katri Haverinen, Simo Vihjanen, and Tapio Salakoski. 2013. Building a large automatically parsed corpus of Finnish. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NoDaLiDa'13)*, pages 291–300.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [conversion and improvement of Universal Dependencies French corpora]. *Traitement Automatique des Langues*, 60(2):71–95.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova, and Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *Preprint*, arXiv:2403.19827.

Pedro Ramaciotti Morales, Robin Lamarche-Perrin, Raphael Fournier-S'niehotta, Remy Poulain, Lionel Tabourier, and Fabien Tarissan. 2020. Measuring Diversity in Heterogeneous Information Networks. *arXiv preprint*. Issue: arXiv:2001.01296 arXiv:2001.01296 [cs, math].

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and

Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2024. A Universal Dependencies treebank for Highland Puebla Nahuatl. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1393–1403, Mexico City, Mexico. Association for Computational Linguistics.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

Alfréd Rényi. 1961. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pages 547–562. University of California Press.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Gerold Schneider. 2011. Using automatically parsed corpora to discover lexico-grammatical features of english varieties. In *30th International Conference on Lexis and Grammar, Nicosia, Cyprus*.

Djamé Seddah, Marie Candito, Benoit Crabbé, and Enrique Henestroza Anguiano. 2012. Ubiquitous usage of a broad coverage French corpus: Processing the Est Republicain corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3249–3254, Istanbul, Turkey. European Language Resources Association (ELRA).

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Wolińsk, Alina Wróblewska, and Éric Villemonte de La Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, United States. Association for Computational Linguistics.

Claude Elwood Shannon and Warren Weaver. 1949. *A Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pages 35–68. Language Science Press, Berlin.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to syntax: Probing spoken language models for syntax. In *Proc. INTERSPEECH 2023*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 1259–1263. Publisher Copyright: © 2023 International Speech Communication Association. All rights reserved.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2023. Universal dependencies 2.12 models for UDPipe 2 (2023-07-17). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer Berlin Heidelberg, Berlin, Heidelberg.

Leonie Weissweiler, Nina Böbel, Kirian Guiller, Santiago Herrera, Wesley Scivetti, Arthur Lorenzi, Nurit Melnik, Archna Bhatia, Hinrich Schütze, Lori Levin, Amir Zeldes, Joakim Nivre, William Croft, and Nathan Schneider. 2024a. UCxn: Typologically informed annotation of constructions atop Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16919–16932, Torino, Italia. ELRA and ICCL.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2024b. Hybrid human-llm corpus construction and llm evaluation for rare linguistic phenomena. *Preprint*, arXiv:2403.06965.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Adelani, and Dietrich Klakow. 2022. Is BERT robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67, Dublin, Ireland. Association for Computational Linguistics.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

A Examples of difficulties in the UD annotation of French

In the French UD treebanks, certain specific phenomena were dealt with (i) either by not following the UD guidelines, which breaks the cross-lingual uniformity, or (ii) by following them at the cost of breaking an internal regularity. As examples of (i), (Guillaume et al., 2019) explicitely report not to follow (for now) UD guidelines for copula constructions with clausal predicative complements (which would lead to a verb with two distinct subjects), nor for expletive *il* subjects[14]. An example of (ii) is the use of different dependency labels for dependents of verb, depending on the category of the dependent, differently to what occurs in the FTBdep annotation schema, itself deriving from the FTB annotation (Abeillé and Barrier, 2004). For instance, the verb *souhaiter (to wish)* can take , the direct a direct complement which is either a NP, an infinitival clause, a clause, or a clitic pronoun. All these cases fill the same valency slot (and thus are mutually exclusive) and are pronominalized using the same accusative clitic pronoun *le*. This uniformity is captured by using a single `obj` label in FTBdep, but 3 different labels in UD (`obj`, `xcomp`, `ccomp`). Moreover, the two latter labels are also used for indirect complements, which obfuscates the linking to semantic roles. Another example concerns the use of `iobj`. For instance for *X parle de Y à Z (X talks about Y to Z)*, in UD, the Y argument can be `iobj`, `obl:arg`, `xcomp`, and the Z argument can be `iobj` or `obl:arg`, whereas Y and Z are uniformly annotated as `de_obj` and `a_obj` in FTBdep.

[14]UD guidelines take into account the semantic property of not baring a semantic role, which has clear advantages for downstream semantic analysis, but which causes peculiarities from the stricter syntactic point of view.

B FTB modifications

The FTB has been modified compared to the version described in (Seddah et al., 2013). Corrections were done:

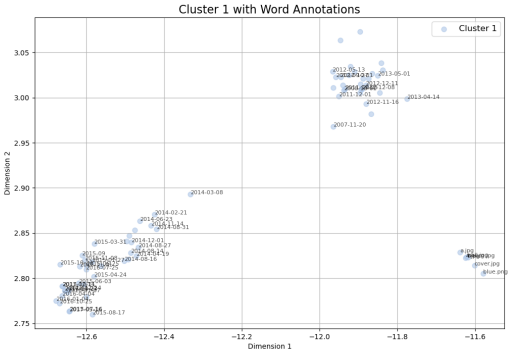
- Automatic corrections to ensure flat representations of MWEs have their linearly first component as head of all other components;
- Manual removal of spurious cycles in surface dependency trees (10 cases).

Some harmonisation with the Sequoia treebank :

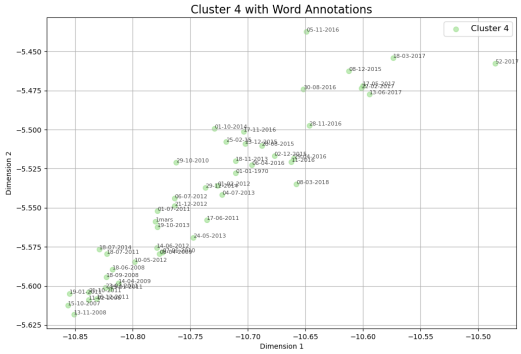
- Representation of MWEs as in Sequoia 9.2, namely as designed in the PARSEME-FR project[15] and described in (Candito et al., 2020);
  - the main change concerns using regular syntax for MWEs whenever possible;
  - for remaining MWEs, final prepositions or complementizers are not included in the MWE (i.e. *que (that)* not included in the MWE *étant donné (given)*.)
- Minor modifications of tokenization:
  - any X - X (- X)* sequence of tokens within a MWE was remerged as one token (i.e. "au - dessus de" → "au-dessus de")
  - numbers: any sequence [0-9]+ (, [0-9]+)+ merged as one token (i.e. "34 , 7" => "34,7")
- Homogenisation of lemmas:
  - reflexive clitics (CLR tag) have lemma *se*;
  - dative and accusative first and second person clitics all receive *le/lui* lemma (ambiguity is to be solved in syntax);
  - distinguish lemma for *madame (madam)* from that of *monsieur (mister)*.

C Word embeddings clusters

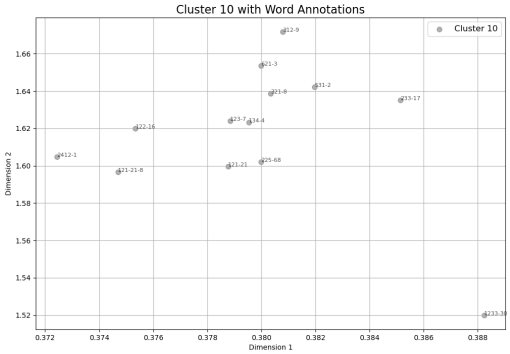
[15]`https://parsemefr.lis-lab.fr`



(a) Cluster 1



(b) Cluster 4



(c) Cluster 10

Figure 3: Word embeddings clusters of new words from HPLT$_{div}$ with dates

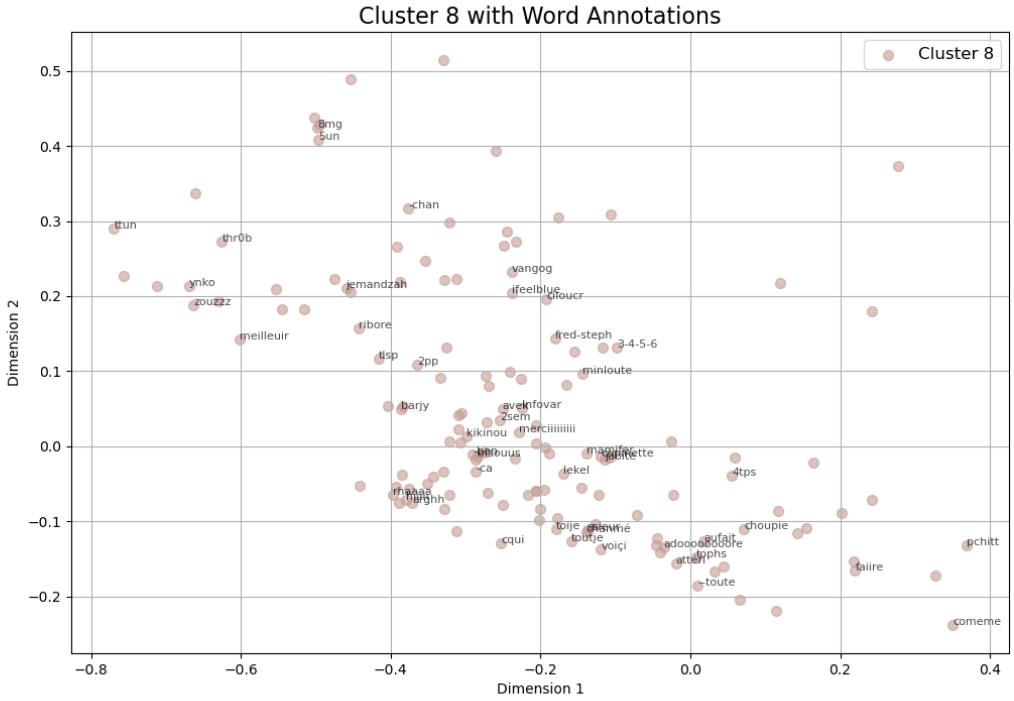
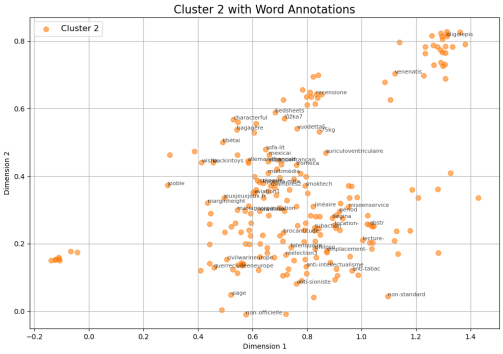
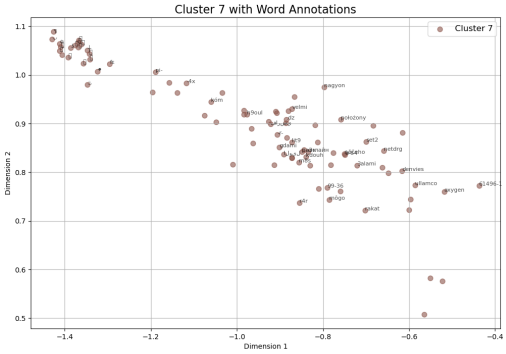


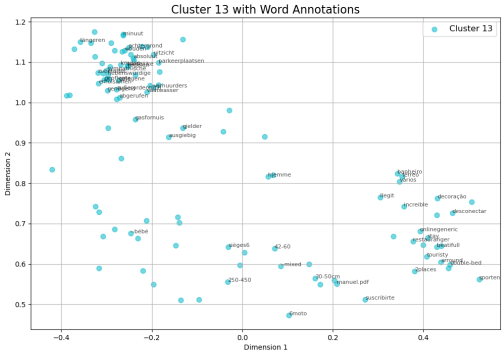
Figure 4: Cluster 8 (SMS language) of Word embeddings clusters of new words from HPLT$_{div}$.



(a) Cluster 2



(b) Cluster 7



(c) Cluster 13

Figure 5: Word embeddings clusters of new words from HPLT$_{div}$ with foreign words

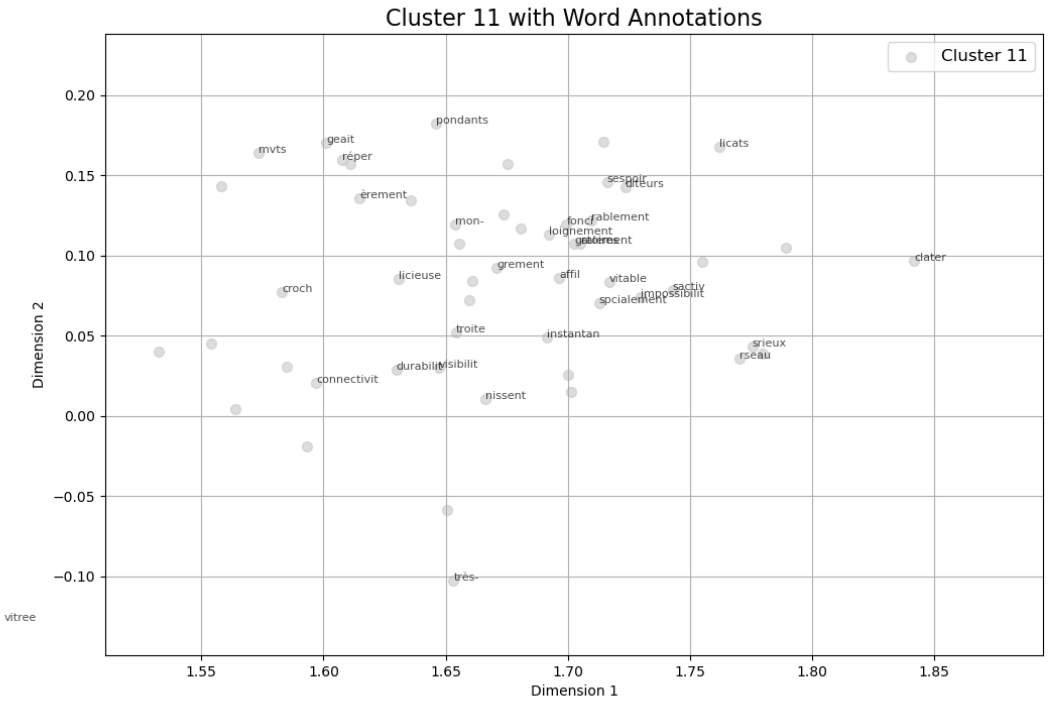
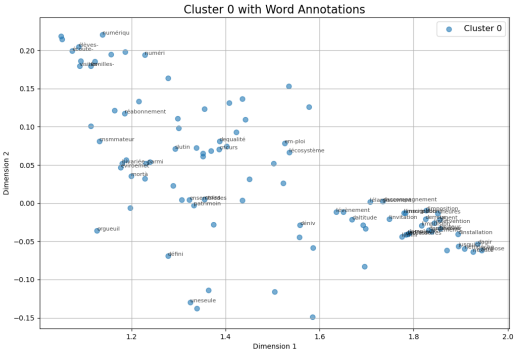
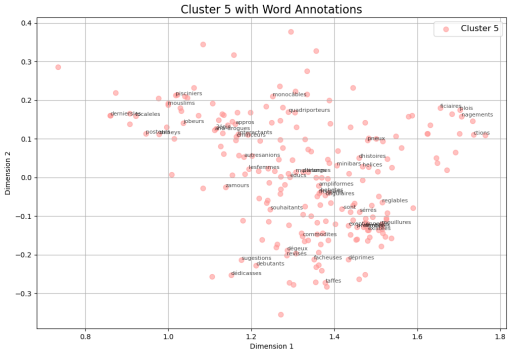


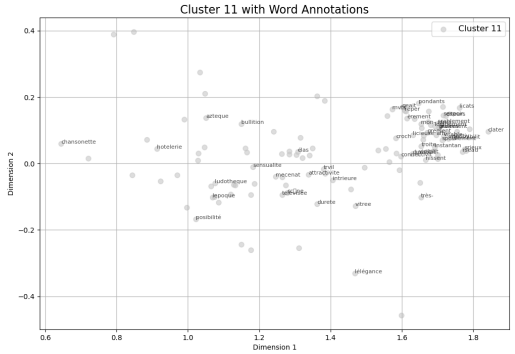
Figure 6: Cluster 11 (suffixes) of Word embeddings clusters of new words from HPLT$_{div}$ (zoom on suffixes part).



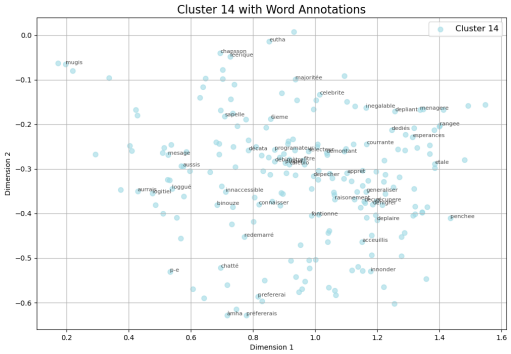
(a) Cluster 0



(b) Cluster 5



(c) Cluster 11



(d) Cluster 14

Figure 7: Word embeddings clusters of new words from HPLT$_{div}$ with spelling mistakes

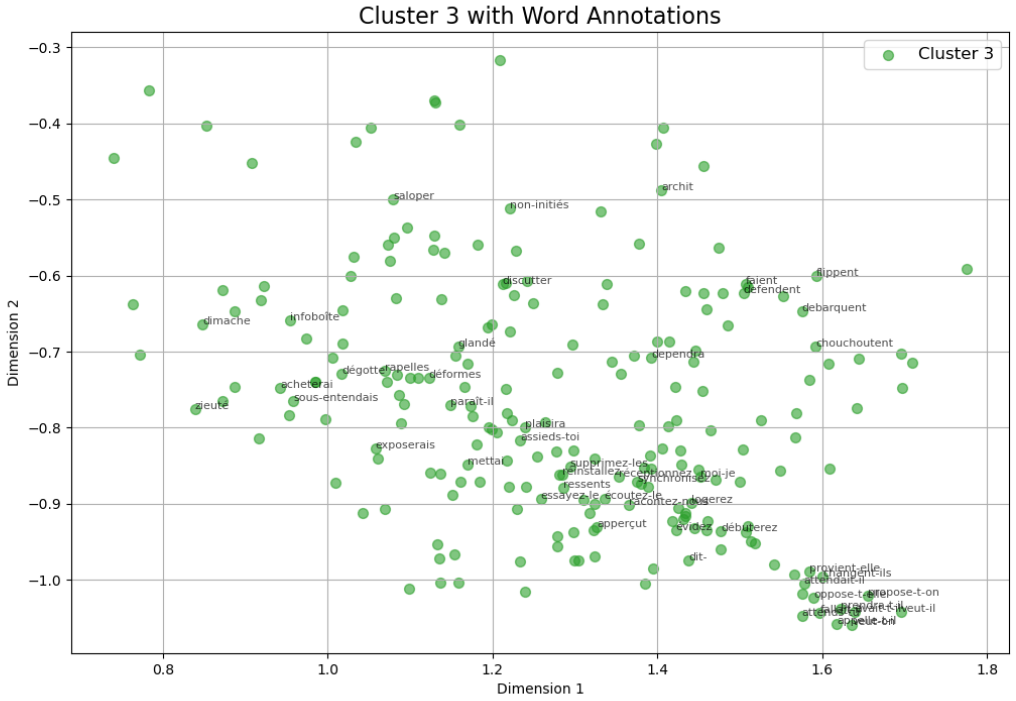


Figure 8: Cluster 3 (rare conjugations) of Word embeddings clusters of new words from HPLT$_{div}$.

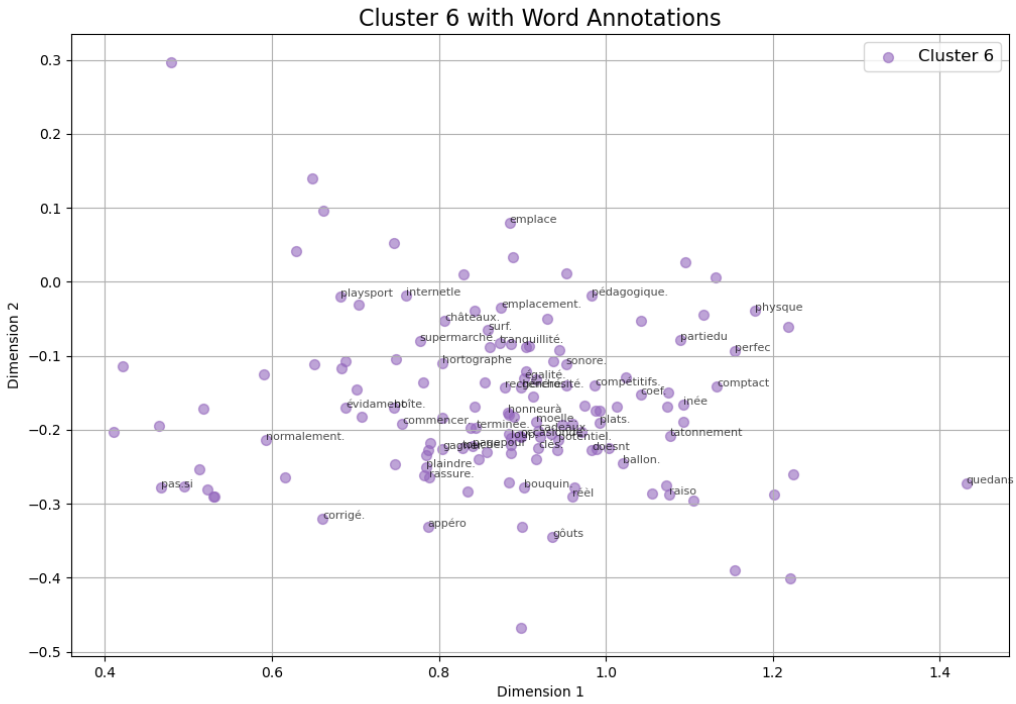


Figure 9: Cluster 6 (final point) of Word embeddings clusters of new words from HPLT$_{div}$.

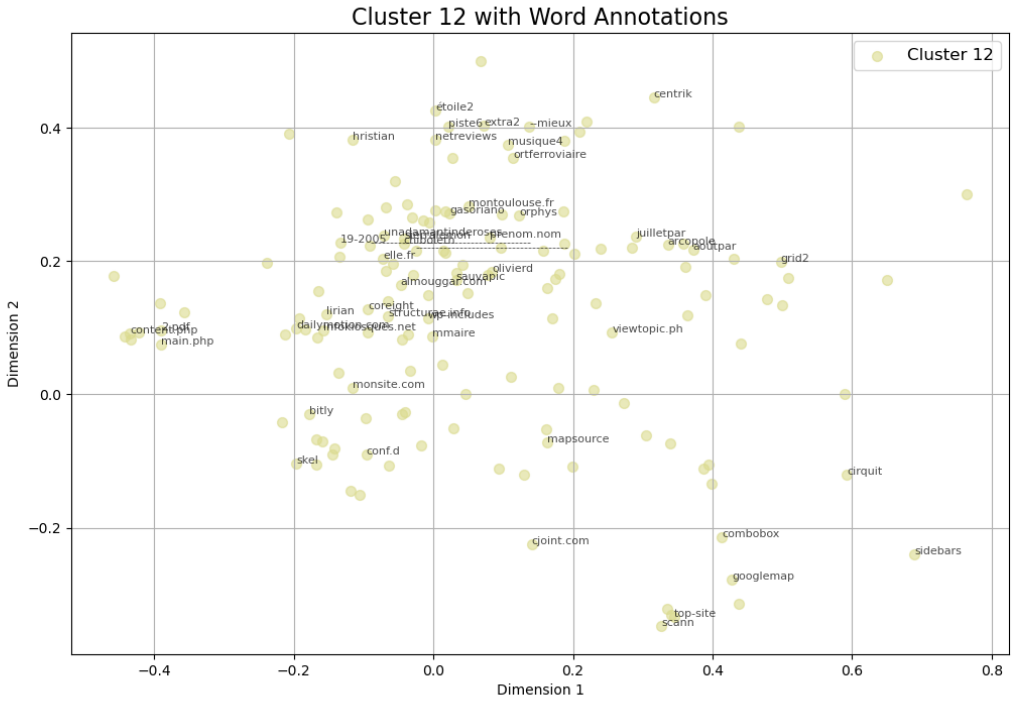


Figure 10: Cluster 12 (url and filenames) of Word embeddings clusters of new words from HPLT$_{div}$.

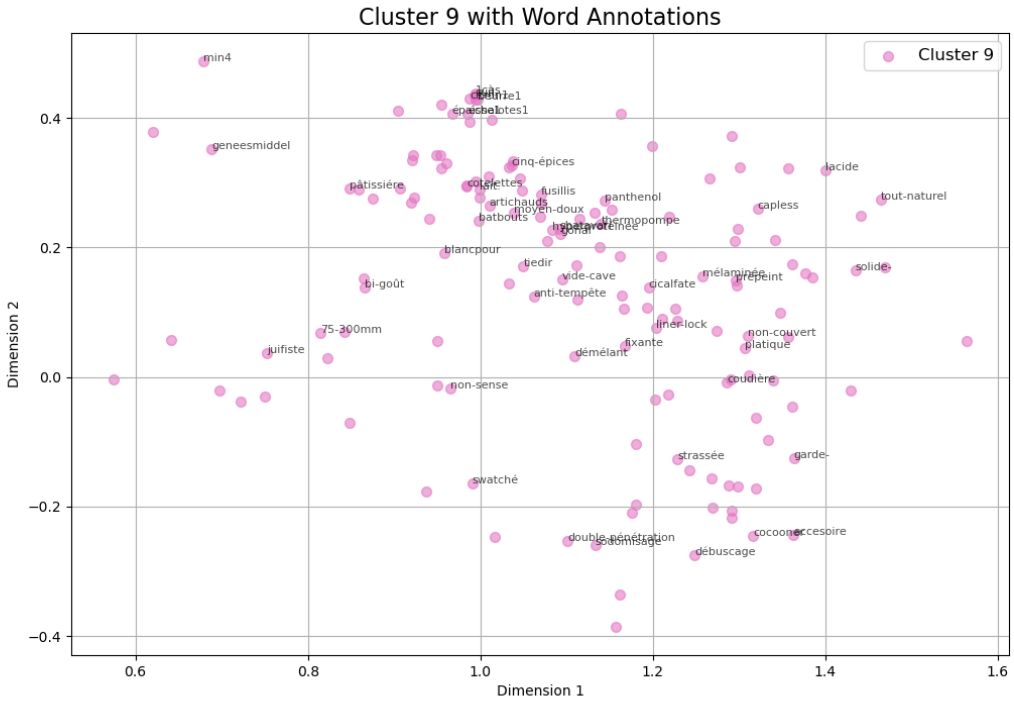


Figure 11: Cluster 9 (diverse) of Word embeddings clusters of new words from HPLT$_{div}$.