Institut des sciences du Digital
Management & Cognition

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Msc Natural Language Processing 2022-2023

UE 805 - Supervised Project

Intrinsic Evaluation of Word Embeddings

Students:
Clémentine BLEUZE
Ekaterina GOLIAKOVA
Chun YANG

Supervisor:
David LANGLOIS

Contents

Introduction

In the bibliographic part of the report, we got familiar with the notions of *Word Embeddings* (WE) or *Word Representations*, and *Word Embeddings Evaluation* that is used to measure their performance in a range of NLP tasks (extrinsic evaluation) or independently of any concrete application, using linguistic resources (intrinsic evaluation). We also took note of strategies that can be used to incorporate linguistic knowledge in pre-trained WE: the produced embeddings can be guided towards a finer encoding of morphological features (Cotterell and Schütze, 2015), or relational information such as synonymy or paraphrase (Faruqui et al., 2015).

In this paper, we propose a framework to extract morphological information for binary encoded features and by doing this, evaluate the encoding quality of morphological information in the WE. In addition to this, we will provide our own evaluation results for WE of the French language produced by the FlauBERT model for encoding Part of Speech (PoS) for nouns, verbs, and adjectives; grammatical number (GN) for nouns and adjectives; and grammatical gender (GG) for nouns and adjectives.

1 Project Presentation

As we have seen in the bibliography part, extracting morphological or semantic information can be a difficult task since the information can be encoded in multiple dimensions (Gladkova and Drozd, 2016). In this paper, we will make an assumption that it is possible to retrieve grammatical information from a small subset of dimensions in the vector representation of a given word. The methodology for extracting morphological information in our research consists of performing multiple feature importance tests, including the analysis of individual vector component values, weights of linear and Neural Network classifiers, and clustering algorithms performed using single and multiple dimensions. The frequency with which candidate dimension subsets are found within the multiple experiments is used as a way to evaluate the encoding quality of grammatical information.

If there are consistently found dimensions that are connected to certain morphological information, we will assess the quality of encoding of the given information by using these dimensions to cluster WE into 2 clusters (e.g. masculine and feminine nouns). Should the clustering be of high quality, we can assume that given WE have clearly encoded grammatical information. As this concerns the evaluation of WE in the way they allow retrieving linguistic information, we consider this to be an intrinsic evaluation task.

To test our framework for extracting morphological information, we chose to work with WE of FlauBERT, a pre-trained Transformer model for the French language, and Morphalou3, a corpus of French language that contains the morphological information which will be described in more detail in the subsections below.

1.1 Morphalou3

As a main lexical reference to conduct our experiments, we used data from the Morphalou3[1] corpus. It is an open morphological lexicon of French, developed by the ATILF laboratory[2], which is the result of merging of 5 lexicons:

- Morphalou2 (december 2013)[3]
- DELA (december 2011)[4]
- Dicollecte 4.3[5]
- LGLex and LGLexLefff 3.4 [6]
- Lefff 2.1[7]

Version 3.1 of Morphalou includes 159,271 lemmas and 976,570 inflected forms, which makes it a valuable resource to cover a wide range of the French language. Detailed counts of lemmas and inflected forms by grammatical category are presented in Table 1. It also contains grammatical information about its lexical units: PoS, GG, GN, verb tense, etc. The data of Morphalou3 is grouped into two sets of lemmas:

- *Lexical words:* adjectives (ADJ), adverbs (ADV), common nouns (NOUN), verbs (VERB), interjections (INTJ). For our work, we focused exclusively on NOUN, VERB, and ADJ.
- *Grammatical words:* numbers, prepositions, pronouns, conjunctions, determiners, and uncategorized lemmas. These categories were not studied in our work.

1.1.1 Data extraction

Morphalou3 files are encoded in UTF-8 and available in Lexical Markup Framework (LMF), Text Encoding Initiative (XML), Comma-Separated-Values (CSV) and MySQL format. An extract of LMF encoding (which is the format we selected for our research) is shown in Figure 3 starting at the `<lexicalEntry>` element, which is the root of the tree for this lexical unit. Below the root, a `<FormSet>` tag contains a `<lemmatizedForm>` node, and a list of `<inflectedForm>` nodes. They correspond to the lemma of the unit, and to all the possible inflected forms of it. Each of them contain useful grammatical information in nodes `<orthography>`, `<grammaticalCategory>`, `<grammaticalGender>` and `<originatingEntry>`.

[1] `https://repository.ortolang.fr/api/content/morphalou/2/LISEZ_MOI.html`
[2] ATILF: Analyse et Traitement Informatique de la Langue Française. `https://www.atilf.fr`
[3] `http://www.cnrtl.fr/lexiques/morphalou/`
[4] `http://infolingu.univ-mlv.fr/`
[5] `http://dicollecte.org/home.php?prj=fr`
[6] `http://infolingu.univ-mlv.fr/`
[7] `http://www.labri.fr/perso/clement/lefff/`

	Lemmas	Inflected Forms
Common nouns	102,238	188,912
Adjectives	36,523	98,970
Verbs	14,762	660,537
Adverbs	4,157	4,167
Interjections	422	422
Prepositions	258	412
Numbers	198	202
Conjunctions	179	297
Pronouns	121	234
Determiners	57	181
Uncategorized lemmas	356	356

Table 1: Lemmas and inflected forms counts in Morphalou3

```
<lexicalEntry id="abaissable_1">
        <formSet>
            <lemmatizedForm>
                <orthography>abaissable</orthography>
                <grammaticalCategory>adjective</grammaticalCategory>
                <originatingEntry target="morphalou2" originatingCategory="adjective">ABAISSABLE, adj.
</originatingEntry>
                <originatingEntry target="dela" originatingCategory="A+z2">abaissable</originatingEntry>
                <originatingEntry target="dicollecte" originatingCategory="adj">abaissable</originatingEntry>
                <originatingEntry target="lefff" originatingCategory="adj">abaissable</originatingEntry>
            </lemmatizedForm>
            <inflectedForm>
                <orthography>abaissable</orthography>
                <grammaticalNumber>singular</grammaticalNumber>
                <grammaticalGender>invariable</grammaticalGender>
                <originatingEntry target="morphalou2">abaissable</originatingEntry>
                <originatingEntry target="dela">abaissable</originatingEntry>
                <originatingEntry target="dicollecte">abaissable</originatingEntry>
                <originatingEntry target="lefff">abaissable</originatingEntry>
            </inflectedForm>
            <inflectedForm>
                <orthography>abaissables</orthography>
                <grammaticalNumber>plural</grammaticalNumber>
                <grammaticalGender>invariable</grammaticalGender>
                <originatingEntry target="morphalou2">abaissables</originatingEntry>
                <originatingEntry target="dela">abaissables</originatingEntry>
                <originatingEntry target="dicollecte">abaissables</originatingEntry>
                <originatingEntry target="lefff">abaissables</originatingEntry>
            </inflectedForm>
        </formSet>
    </lexicalEntry>
```

Figure 3: Extract from the original Morphalou3 content (LMF): the adjectival lemma *abaissable* and its forms *abaissable* and *abaissables*

The tree-like structure of the files allowed us to use *xml.etree.ElementTree* module[8] to parse them and extract the data as structured CSV files (see an example of a parsed file in Figure 4). In addition to this, we wrote a function to allow queries on Morphalou3 content, in case we need information about a given form, which can be very useful in case of a possible ambiguity. Take for instance the form *mais*. As

[8] `https://docs.python.org/3/library/xml.etree.elementtree.html#xpath-support`

can be seen in Figure 5, it can appear in multiple grammatical categories: NOUN, ADV, INTJ or CCONJ.

	lemma	gender	category	invariable	singular	plural
0	100-mètres	masculine	commonNoun	100-mètres	NaN	NaN
1	2D	feminine	commonNoun	2D	NaN	NaN
2	3D	feminine	commonNoun	3D	NaN	NaN
3	A	masculine	commonNoun	µA	NaN	NaN
4	a	masculine	commonNoun	a	NaN	NaN

Figure 4: Conversion of LMF content into .csv files: this file contains grammatical information about NOUNS

```
queryf("mais")

-'mais' is the inflected plural form of common Noun 'mai'.
-'mais' is an adverb.
-'mais' is an interjection.
-'mais' is a grammatical word with following attributes:
    -grammatical category: conjunction
    -grammatical subcategory: coordination
```

Figure 5: A query on Morphalou3 for the form *mais*

1.2 FlauBERT

Since training a language model for WE can be a time- and resource-consuming task, for this stage of our project we decided to work with WE from a pre-trained model. Among French-language models, we have chosen FlauBERT due to its reported high results on the PoS tagging task (Le et al., 2020) which could be potentially linked to rich morphological information encoded in their WE. In this section, we will briefly discuss how and on what data the FlauBERT model was trained.

1.2.1 Training data

FlauBERT was trained using data from 24 different subcorpora, among which are CommonCrawl (Buck et al., 2014), NewsCrawl (Li et al., 2019), Wikipedia and Wikimedia data dumps[9] and others. Overall, the training corpus of FlauBERT

[9] `https://meta.wikimedia.org/wiki/Data_dumps`

consisted of 71GB of data which was composed out of 12.79B tokens and 478.78M sentences, all of which are in French language. The collected data was pre-processed: short sentences and content such as telephone/fax numbers, emails, addresses, etc. were removed; the data was Unicode-normalized.

1.2.2 Model architecture

The FlauBERT project provided several versions of their model with different sizes of corresponding embeddings, out of which we have chosen to work with the 512-dimensional FlauBERT-small to reduce the number of investigated dimensions for our project. The model follows BERT architecture which consists of a multi-level bidirectional Transformer (Devlin et al., 2019). The chosen FlauBERT-small model has 6 layers, a hidden size of 512, and 8 attention heads. The model was trained using pre-norm attention and stochastic depths.

Unlike the original BERT models, FlauBERT was trained using only a masked language model that learns to predict randomly masked tokens in a sentence as its supervised task. The next sentence prediction model was not used during the training of the model.

1.2.3 Building the dataset

The FlauBERT model is publically available[10]. For simplicity of usage, we created an instance of the model using the HuggingFace library[11]. Using the model and Morphalou3 data sets of nouns, verbs, and adjectives we found the values of the last layer for all the words in the sets. During the process of matching a Morphalou3 wordform with a corresponding embedding from FlauBERT, we found that some of the wordforms do not have their own unique embedding in FlauBERT-small but instead, their word parts are tokenized separately, which could be due to the size of the model. The wordforms that were tokenized into two or more separate parts by FlauBERT, were excluded from our train and test datasets. Such filtering left us with a dataset of 13,239 nouns, 10,089 verbs, and 6,335 adjectives of the French language. For each wordform the following information was stored: its 512 dimensions from FlauBERT, its PoS, gender, and plurality (where it's applicable) from Morphalou3.

1.3 A first visualisation

Before proceeding further, let's have a look at a small sample of 15 randomly selected embeddings: 5 verbs (presenting various tenses), 5 nouns, and 5 adjectives (presenting various grammatical numbers and genders). You will be able to find the full plots in Appendix Figures 11 and 12. We plotted the heatmaps of these embeddings, with dimensions on the x-axis and word labels on the y-axis. The 512 dimensions are cut

[10] `https://github.com/getalp/Flaubert`

[11] https://huggingface.co/flaubert/flaubert_small_cased

into dimension subsets for better readability: dark cells correspond to higher values for a given embedding and dimension, whereas pale cells correspond to lower values. It can be noticed that:

- The different dimensions groups represent unequal amplitude and extremum values. For instance, the maximum value for dimensions 0-102 is around 10, however, it exceeds 50 for dimensions 309-411. We observe that this skew is due to dimension 371, which seems to have very high values (relative to the other dimensions) for almost all the embeddings of our sample, regardless of their grammatical characteristics. Due to this outlier, the heatmap is actually quite hard to read for this dimension grouping.
- We note other tendencies that apply to all the embeddings: relatively low values for dimensions 52, 274, 508, and relatively high values for dimensions 44, 195, and 250.

From this quick overview, we expect to find similar phenomena in our complete corpus. This highlights the importance of *data normalization* before making any conclusion in terms of value distribution. However, besides these constatations, the distribution of remaining values seems fuzzy, and uninterpretable at first glance. This last point is what we are going to tackle with upcoming experiments.

2 Single dimensions encoding GG, GN and PoS

In this first part, we will present the strategies we implemented in order to retrieve single dimensions (among the 512 ones of our WE dataset) that are most likely to encode some grammatical characteristics in nouns, verbs, and adjectives. Although it may seem quite unreasonable to expect that grammatical information is encoded in a single dimension of a WE (Gladkova and Drozd, 2016), this must be considered as a first step towards candidate subsets identification. We may hypothesize that a single dimension appearing on its own as a potential candidate for grammatical feature encoding will also appear in the experiment results of dimension combinations along with other dimensions. This hypothesis will be tested in Section 3. In order to find these singletons, we narrowed our task to three grammatical features:

- *Grammatical Gender (GG)* (masculine vs. feminine) in nouns and adjectives
- *Grammatical Number (GN)* (singular vs. plural) in nouns and adjectives
- *Part of Speech (PoS)* in nouns, adjectives and verbs: we split this task into three tasks NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ. Because of potential PoS ambiguity for a given form (e.g *lit* can be a NOUN or a VERB, from *lire*), we filtered data to produce datasets of "pure" NOUN, ADJ and VERB, according to Morphalou3 tags.

As can be seen from the setup, all the experiments are binary: a tested feature (e.g. plurality, NOUN/not-NOUN) is encoded as 0 for one of the options and 1 for the other option. Following this logic we have:

- *GG* encoded as 0 for feminine nouns and adjectives and 1 for masculine ones;
- *GN* encoded as 0 for singular nouns and adjectives and 1 for plural forms;
- *PoS* encoding is one-hot encoding, for each of the NOUN, VERB, ADJ the words are encoded as 1 if they belong to the PoS and as 0 if they don't. Since we excluded ambiguous words, there is no ambiguity in the encodings as well.

The strategies we implemented are of various types. First, we investigate the distribution of values in a dimension among the dataset, attempting to find dimensions with distribution interesting for us which could be potentially linked to a morphological feature. Then, we perform a correlation study with the targeted grammatical features. Finally, we train classifiers and a clustering algorithm to separate WE according to these grammatical features and study the dimensions that have the highest contribution to the results. In the final subsection, we provide a comparative analysis of all these strategies, assessing the consistency of their results.

2.1 Study of the distribution of dimensions values

As for our initial investigation, we took to investigating several distributional metrics of WE of a category: masculine/feminine nouns and adjectives, singular/plural nouns and adjectives, and words belonging and not belonging to a certain PoS. We started by investigating the standard deviation of the categories, followed by finding the dimensions with the highest average difference between binary subcategories (e.g. masculine nouns vs feminine nouns). Finally, we introduce our own WE evaluation metric: AABCC-score for assessing how well a dimension groups the values of the experiment groups.

2.1.1 Minimum intra-class standard deviation

Principle:

- For each WE dimension, compute its standard deviation within each of the sub-datasets for the considered grammatical feature (e.g singular vs. plural).
- Retrieve dimensions having the lowest intra-class standard deviation.

The intuition behind this first strategy is that if a certain dimension has values close to one another, this dimension is potentially linked to a grammatical feature with the expected values in the small range. However, Figure 6 illustrates the encountered problem for this method. It shows distributions of values for dimensions

314, 287 and 365 in NOUN, which were found to have among the lowest intra-class standard deviation both for masculine and feminine nouns. This phenomenon was replicated for all dimensions found in the GG, GN and PoS experiments. Since the values completely overlap, we can not use the standard deviation as a good predictor of grammatical information, therefore, we will not be using it in the main part of our framework. Nevertheless, we believe it is an important first step in the investigation.

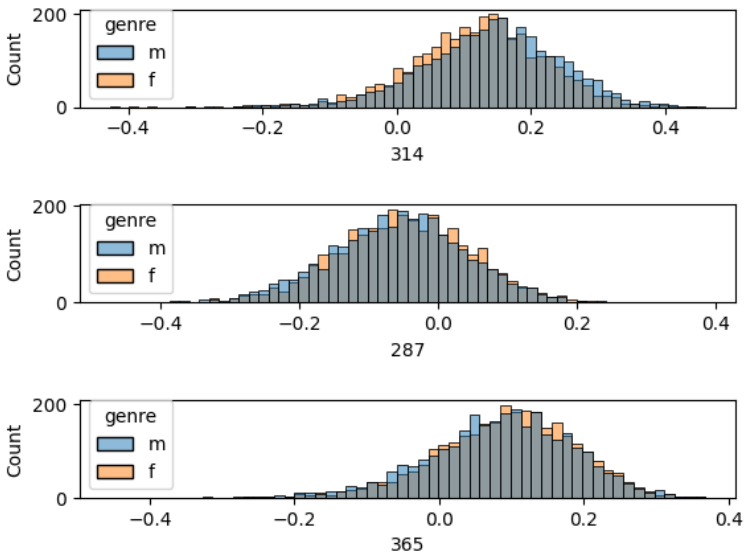


Figure 6: Distributions of values for dimensions 314, 287 and 365 in masculine and feminine nouns embeddings (data has been normalised). We see that the distributions almost perfectly overlap, suggesting that the highlighted dimensions do not separate classes masculine and feminine.

In order to tackle this issue, we tried a new strategy that would focus more on the difference between classes of our classification tasks.

2.1.2 Binary SIG-PROPS

Principle:

- Create two subsets (Group 1 and Group 2, e.g. singular adjectives and plural adjectives)
- Compute averages for each dimension for each group
- Calculate the difference between dimension averages for Group 1 and Group 2
- Find the dimensions with the highest differences

This second strategy is inspired by the work about the elucidation of conceptual properties from WE (Jang and Myaeng, 2017). By analyzing the distribution of values for word2vec WE, Jang and Myaeng achieved to isolate specific dimensions encoding for semantic properties. More precisely, they found high correlation values between the typicality score of words for a given *semantic category* (e.g for the category *fruit*, *apple* has a high score, *guitar* has a low one - but the contrary holds for the category *music instrument*) and dimensions having the highest average value within groups of WE of considered categories. Such dimensions (which researchers named SIG-PROPS) were found by finding dimensions with the highest value for each category and subcategory.

Since our experiment is set up in a binary way, we adapted this idea for our task as binary SIG-PROPS with the process as described above. This approach allowed us to find dimensions that are better at separating grammatical information. You will be able to find the top 10 dimensions for each experiment type in Table 2. You can find the dimensions 100 and others investigated in Section A.1 in Appendix.

Note that, for GG and GN, we ran the experiment first on the dataset of nouns only, then on the dataset of adjectives only, and finally on the dataset of nouns and adjectives. This is to test the consistency of the dimensions retrieved: ideally, we would like to isolate very similar sets of dimensions, meaning that the GG/GN-encoding is PoS-independent. In order to check this, for each subtable in Table 2, we color in blue dimensions found in 2 columns, and in red dimensions found in 3 columns. For PoS also, we use the same code. This will be replicated in the following results presentations as well.

NOUN	ADJ	NOUN + ADJ
100	466	245
316	250	192
245	439	250
195	503	121
192	245	5
202	181	276
117	5	438
121	192	181
507	234	507
499	88	195

(a) Results for GG

NOUN	ADJ	NOUN+ADJ
310	310	310
288	54	54
81	192	288
54	274	81
250	84	285
285	56	384
172	384	278
278	474	172
507	5	495
359	285	311

(b) Results for GN

- For GG, we find 7 dimensions (245, 195, 192, 121, 507, 250, 181) appearing in the top 10 of more than one dataset, 2 of them appearing in the top 10 of all datasets (245, 192). This consistency despite PoS variation makes them interesting candidates for encoding GG.

NOUN vs. not-NOUN	VERB vs. not-VERB	ADJ vs. not-ADJ
159	310	276
409	159	2
458	480	478
504	401	158
346	89	370
480	158	24
401	282	220
212	192	409
128	504	464
305	458	139

(c) Results for PoS

Table 2: Top 10 dimensions with the highest difference between average values of masculine and feminine (GG) / singular and plural class (GN) : NOUN, ADJ and both. For PoS, we show NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ. Dimensions found in 2 experiments of the same set in blue, and dimensions found in 3 experiments are in red.

- GN experiments show even better consistency, with 8 dimensions (310, 288, 81, 54, 285, 172, 278, 384) common to 2 datasets, 3 of them to all datasets (310, 54, 285). Note also that dimension 310 appears first each time. We notice almost no overlap at all with the dimensions found in GG experiments: so both these grammatical informations must be encoded in different parts of WE.

- As for PoS information, we can notice that there is no dimension that would be highlighted by all three experiments, meaning that no dimension encodes all PoS at once. In addition to that, we can notice that one of the top dimensions highlighted for differentiating VERB and not-VERB is 310, also highlighted during GN experiments, which is interesting since, indeed, we can't define a grammatical number for VERB. We can also notice that there is quite a lot of overlap for NOUN/not-NOUN and VERB/not-VERB (dimensions 159, 480, 458, 504, 401) in this experiment. This coincides with what was previously shown in (Musil, 2019) for Czech language where values for NOUN and VERB were found on the opposite sides of a Principal Component. As for ADJ, we can see that there is very little overlap with other PoS: only dimension 158 is shared with VERB/not-VERB and only dimension 409 is shared with NOUN/not-NOUN.

2.1.3 AABCC-score

Principle:

- Join WE of Group 1 and Group 2 together for each experiment (e.g. NOUN and not-NOUN)
- Take every dimension and sort by it
- Apply AABCC calculation
- Find the dimensions with the highest AABCC scores

To continue the investigation, we introduced *AABCC-score*, a metric to measure the meaningfulness of a dimension in encoding a given grammatical characteristic. The computation method is explained in Figure 7, and favors dimensions presenting large sequences of identical tags when ordered.

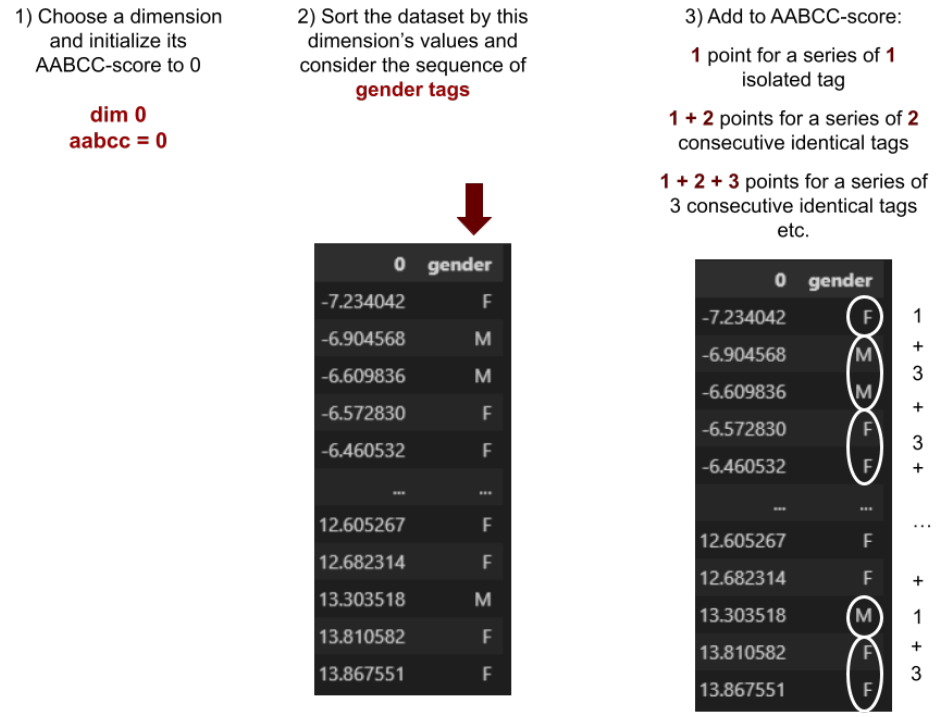


Figure 7: Method for computing the AABCC-score of a dimension for GG. The principle remains the same for GN and PoS, with the sequence of appropriate tags.

More precisely, using a dataset of WE of size N:

- The worst AABCC score corresponds to an alternate sequence of tags (for instance, M F M F ... (GG) or S P S P ... (GN)). Each tag increases the score by only 1 point, so we get a final AABCC score of N.

- The best AABCC score corresponds to a sequence of tags that perfectly splits the two categories (for instance, M M M ... F F F (GG) or P P P ... S S S (GN)). In case there is the exact same number of WE for the two categories, N is even, and we get a final AABCC score of:

$$2 \times \sum_{i=1}^{N/2} i = 2 \times \frac{N/2 \times (N/2+1)}{2} = \frac{N^2+2N}{4}$$

  This is greater than N for $N>2$, therefore, the higher the AABCC-score, the better candidate for encoding GG (GN or PoS) the dimension.

You will be able to find the top 10 dimensions with the highest AABCC-score for each experiment group in Table 3 , where we can observe that:

- For GG, we observe some dimensions already seen in the binary SIG-PROPS experiment in NOUN and ADJ (100, 316, 195, 192, 121, 507, 466, 250, 439, 245). However, none of the dimensions identified in NOUN + ADJ match, suggesting that this common subset is not PoS-independent for encoding GG.
- The matching is more consistent in GN experiments, both between the three datasets NOUN, ADJ and NOUN + ADJ, and between the AABCC and binary SIG-PROPS experiments. Dimension 310 in particular appeared first every time, and dimensions (54, 285, 81, 278, 384) also keep high ranks.
- Finally, for PoS, the matching of AABCC results with binary SIG-PROPS experiment is very poor (dimensions 159, 346 for NOUN vs. not-NOUN and 89, 504 for VERB vs. not-VERB). Interestingly, we once again observe that important dimensions for separating NOUN and not-NOUN overlap with the ones for separating VERB and not-VERB quite significantly.

NOUN	ADJ	NOUN + ADJ
100	466	432
192	250	213
403	245	249
245	256	57
377	439	364
316	133	325
195	432	331
121	192	137
507	503	82
202	121	272

(a) Results for GG

NOUN	ADJ	NOUN+ADJ
310	310	310
54	25	54
288	54	57
285	278	278
278	84	25
384	274	384
81	384	420
25	285	213
250	81	325
172	311	81

(b) Results for GN

NOUN vs. not-NOUN	VERB vs. not-VERB	ADJ vs. not-ADJ
185	341	318
127	162	310
341	89	64
346	185	487
209	346	47
299	299	69
96	127	291
159	182	365
89	4	425
4	504	182

(c) Results for PoS

Table 3: Top 10 dimensions with the highest AABCC-score. For GG and GN we show NOUN, ADJ and both. For PoS, we show NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ. Dimensions found in 2 experiments of the same set in blue, and dimensions found in 3 experiments are in red.

2.2 Correlation study

Principle: Retrieve dimensions having the highest Point-Biserial Correlation with the considered grammatical feature (encoded as 0-1 labels).

Previously, the researchers attempted to evaluate the extracted dimensions from the experiment of maximum difference between average values (see Section 2.1.2) by correlating them with the values of typically (Jang and Myaeng, 2017). We extended this idea and performed a correlation test for each dimension with a grammatical feature using Point-Biserial Correlation. Its values are always between -1 to +1: +1 means a perfect positive correlation between two variables, -1 means a perfect negative relationship correlation between two variables, and 0 represents no meaningful relationship between studied variables. Our hypothesis is that dimensions with high correlation coefficients may encode GG, GN of PoS information. You will be able to find the top 10 dimensions for each experiment group in 4.

- For *GG*, the best correlation is found with dimension 466 in adjectives. Yet we can see that the correlation remains very low between single dimensions and gender, with maximum coefficients of 0.20 (NOUN), 0.35 (ADJ), and 0.21 (NOUN + ADJ). Dimensions 121 and 245 stand for all three experiments. But we find consistent results with the previous experiment of binary SIG-PROPS: the dimensions appearing first (100, 466, 245) are exactly the same; and with AABCC score (100, 466). Dimensions 245 and 192 (common to 3 datasets in the previous experiment) are present in the correlation top 10, and dimension 121 (common to 2 datasets in correlation) is present in the top 10 of binary SIG-PROPS.

- The results for *GN* are better than *GG*, with the largest correlation coefficients ranging between 0.28 and 0.48, depending on the considered dataset. There are 4 common dimensions: 310, 285, 54, 285, which appear in all three experiments. Same as *GG*, we also find consistent results with the previous experiment of binary SIG-PROPS: the dimension (310) still appears first for both NOUN, ADJ and NOUN + ADJ. Dimensions 310, 285, 54 (common to 3 datasets in maximum difference) are present in the correlation top 10, and dimensions 288, 81, 278, 172, 384 (common to 2 datasets in correlation) are present in the top 10 of binary SIG-PROPS.

- As for *PoS* results, we can observe that VERB vs not-VERB experiment produces the highest correlation of a dimension with the *PoS* information at 0.51. The correlation coefficient of NOUN vs. not-NOUN is a little worse than VERB vs not-VERB, with the highest result reaching 0.5. However, the correlation for ADJ vs. not-ADJ shows a significant difference in results: a highest correlation of only 0.25. We can't find a dimension overlap for all *PoS*, nonetheless, we can still find 6 common dimensions for NOUN vs. not-NOUN and VERB vs. not-VERB : 159, 480, 401, 310, 458, 29. The consistent dimensions found with binary SIG-PROPS are 159, 480, 401, 458 (common to 2 datasets in correlation).

- We observed that dimension 310 not only has the largest correlation coefficient in *GN*, but also has the highest correlation coefficient with *PoS* in VERB vs. not-VERB, and it is also among the top 10 dimensions in NOUN vs.not-NOUN. It seems that this dimension plays an important role in encoding linguistic information.

NOUN	Corr	ADJ	Corr	NOUN + ADJ	Corr
100	0.20	466	0.35	245	0.22
195	0.20	439	0.31	192	0.20
316	0.19	250	0.31	507	0.19
245	0.18	503	0.31	121	0.19
507	0.18	133	0.30	250	0.18
192	0.17	245	0.29	5	0.17
377	0.16	234	0.28	181	0.17
121	0.16	432	0.28	377	0.17
117	0.16	181	0.26	195	0.17
403	0.15	121	0.25	439	0.16

(a) Results for GG

NOUN	Corr	ADJ	Corr	NOUN+ADJ	Corr
310	0.48	310	0.41	310	0.45
81	0.35	54	0.40	54	0.34
288	0.35	192	0.36	81	0.33
250	0.32	384	0.35	288	0.33
507	0.31	274	0.35	285	0.31
278	0.31	84	0.34	278	0.30
285	0.31	56	0.34	384	0.29
54	0.31	318	0.32	25	0.29
172	0.29	285	0.32	172	0.26
25	0.28	25	0.31	311	0.26

(b) Results for GN

NOUN vs.not-NOUN	Corr	VERB vs.not-VERB	Corr	ADJ vs.not-ADJ	Corr
159	0.5	310	0.51	158	0.25
480	0.45	159	0.51	220	0.21
401	0.44	480	0.50	478	0.19
310	0.42	401	0.49	439	0.18
346	0.41	192	0.44	464	0.17
458	0.41	89	0.43	250	0.17
29	0.40	29	0.42	50	0.16
504	0.38	458	0.42	47	0.16
128	0.37	198	0.41	222	0.16
341	0.37	504	0.41	119	0.16

(c) Results for PoS

Table 4: Top 10 dimensions having highest correlation (Pearson coefficient) in GG (0: feminine, 1: masculine) and GN (0: plural, 1: singular): NOUN, ADJ and both. For PoS, we show NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ. Dimensions found in 2 experiments of the same set are in blue, and dimensions found in 3 experiments are in red.

Looking at the results above, most correlations of a dimension with *number, gender and PoS* information are less than 0.5, which indicates that the relationship between two variables is not strong. Therefore, based on the analysis, we can say that our hypothesis is not valid and a single dimension cannot encode grammatical information alone. Even though as we see, for some dimensions, the correlation coefficient between some single dimension and grammatical features information can reach still 0.50, this is not common. We can further assume that the combination of multiple dimensions may be more suitable to encode grammatical feature information. This will be tested in Section 3.

2.3 Classification and Clustering experiments

In this part, we no longer study the dimensions based on their value distribution, but rather on their contribution to classification tasks for gender, number, and PoS. By performing these tasks, we assume that certain dimensions would either contribute as a higher weight for a classifier or generate a higher quality clustering of grammatical feature and therefore these dimensions can be potential candidates for encoding certain grammatical information.

2.3.1 Logistic Regression

Principle:

- Combine Group 1 and Group 2 of an experiment together
- Train a Logistic Regression classification model to predict a grammatical feature (e.g. GN)
- Retrieve the weights of the model and select top 10 with the highest absolute weights

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes, which is exactly what we are interested in. If *Linear* Regression makes the assumption of a linear relation between data features X and a target Y (taking values in set $\{0, 1\}$), that is, β_0 being the bias and β_1 the weights :

$$Y = \beta_0 + \beta_1 \cdot X$$

Logistic regression wraps this hypothesis with a *sigmoid function* σ:

$$Y = \sigma(\beta_0 + \beta_1 \cdot X) = \frac{1}{1 + e^{\beta_0 + \beta_1 \cdot X}}$$

This allows predictions to be strictly distributed between 0 and 1, with a threshold value (0.5) used to discriminate between the two classes. From this formula, we see that the higher a weight's absolute value, the greater its contribution to the classification result. High positive weights will make the prediction come closer to 0 (class 0) and high negative weights will make it closer to 1 (class 1). Therefore, the study of the weights of a Logistic Regression model can allow us to retrieve informative components for the classification task, that is, dimensions encoding singular vs. plural, etc.

Using Logistic Regression to predict GG, GN and PoS, we find good to excellent classification results (best f1 scores: 0.9894 (GG), 0.9960 (GN) and 0.8740 (PoS)), which gives credit to the retrieved important dimensions. Top 10 dimensions with

the highest weights in GG, GN and PoS classification tasks are presented in Table 5, where we can see that:

For GG, we find a totally new dimension, that was encountered in none of the previous top 10: 250 comes second in the dataset of nouns, and first in the two other ones (please note that we remarked this dimension in 1.3 as having relatively high values for all WE types). Dimensions 162 and 390 also are new, and common to all three datasets' top 10: we see that Logistic Regression relies on different dimensions than the previous approaches. Given the excellent F1 scores for GG classification that were obtained (0.9544 on nouns, 0.9894 on adjectives, 0.9556 on nouns and adjectives, we might consider these as interesting candidates.

NOUN	ADJ	NOUN + ADJ
100	250	250
250	121	162
195	162	100
265	390	195
162	88	269
269	89	390
377	432	214
390	177	265
468	175	377
434	245	468

(a) Results for GG

NOUN	ADJ	NOUN + ADJ
310	310	310
54	54	54
208	384	384
384	359	208
359	285	359
158	81	81
81	200	360
285	360	182
182	455	158
172	192	285

(b) Results for GN

NOUN vs. not-NOUN	VERB vs. not-VERB	ADJ vs. not-ADJ
52	12	133
92	261	328
261	216	81
275	192	21
229	291	310
427	310	292
132	341	260
37	92	26
345	275	369
223	56	110

(c) Results for PoS

Table 5: Top 10 dimensions with highest weights for Logistic Regression in GG and GN classification in 3 experiments: NOUN, ADJ and both. For PoS, we show NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ. Dimensions found in 2 experiments of the same set are in blue, and dimensions found in 3 experiments are in red.

GN's top 10 come with a common subset of size 6: (310, 54, 384, 359, 81, 285), all of them being encountered at least once in the previous experiments, and 310 still being first. The classification F1 scores are also excellent (0.9936 on nouns, 0.9956 on adjectives, 0.9960 on nouns and adjectives.

As for PoS, the classification results are less good (F1 scores of 0.7972 on NOUN vs. not-NOUN, 0.8420 on VERB vs. not-VERB and 0.8740 on ADJ vs. not-ADJ), and the overlap between the three columns is less consequent. None of the dimension pairs appearing first ((52, 92), (12, 261), and (133, 328)) appeared previously. This further confirms the observation that Logistic Regression may rely on other dimensions than our other strategies.

2.3.2 Perceptron

Principle:

- Combine Group 1 and Group 2 of an experiment together
- Train a Perceptron classification model to predict a grammatical feature (e.g. GN)
- Retrieve the weights of the model and rank the dimensions by their weights (dimension with rank 1 would have the highest absolute weight)
- Repeat the training 9 more times and repeat the dimension ranking
- Calculate the average dimension rank after 10 runs and retrieve the top 10 dimensions with the highest average rank

As the next step of our investigation framework, we implemented Perceptron classification models using Python library *torch*. Perceptron is one of the simplest Neural Network models, with only 1 input layer and 1 output layer. We chose the following parameters:

- 1 input layer of size (512, 2)
- 1 output layer of size (2, 1)
- CrossEntropyLoss as loss function
- Adam optimizer, with learning rate 0.001
- Batch size = 32, 20 epochs per run, 10 runs in total

We performed multiple runs per experiment since the random initialization of weights can make the results differ a lot from one run to another. Given that we are interested in studying the values of the weights, we took their average values over the 10 runs and considered dimensions with the highest weights. As we can see in Table 6, for GG, as with most of the experimental results above, the dimensions 100 and 466 are in the first place of experiments NOUN and ADJ. Dimension 162 first

appeared in the top 1 place for NOUN+ADJ. We can find only two dimensions (245, 250) presenting in the top 10 of all experiments.

For GN, in all experiments until now, dimension 310 has always ranked first in all datasets. 5 common dimensions (310, 54, 285, 384,495) are in the top 10, among which (310, 54) are the common dimensions of all other experiments for GN.

As for PoS, we cannot highlight vectors for all datasets. There are four common dimensions (159, 378, 310) between the two datasets. We observe that, as with the Logistic Regression results above, dimension 310 occurs in both VERB vs. not-VERB and ADJ vs. not-ADJ experiments. Besides that we also see dimension 310 in the top 10 dimensions in both NOUN vs. not-NOUN and VERB vs. not-VERB of the correlation study. This dimension not only appears at the top of the GN experiment every time but also frequently appears in the PoS experiment, which can make an assumption that a single dimension may not only encode one but also several grammatical information at the same time.

NOUN	ADJ	NOUN + ADJ
100	466	162
434	250	377
162	245	507
316	439	245
377	5	250
245	181	100
507	133	316
250	177	434
186	88	28
117	503	499

(a) Results for GG

NOUN	ADJ	NOUN + ADJ
310	310	310
54	285	54
158	54	158
285	384	285
359	455	359
172	495	384
384	200	172
495	360	495
250	192	250
182	25	200

(b) Results for GN

NOUN vs. not-NOUN	VERB vs. not-VERB	ADJ vs. not-ADJ
159	192	256
409	310	133
305	378	426
465	508	98
275	480	310
378	158	381
260	159	12
387	175	188
462	89	1
37	282	412

(c) Results for PoS

Table 6: Top 10 dimensions with highest weights for Multi-Layer Perceptron in GG and GN classification in 3 experiments: NOUN, ADJ and both. For PoS, we show NOUN vs. not-NOUN, VERB vs. not-VERB and ADJ vs. not-ADJ.

2.3.3 KMeans

The KMeans algorithm is a *clustering algorithm* that separates data into distinct clusters of equal variance, minimizing their inertia (or intra-cluster sum of squares)[12] which is commonly used in evaluating WE (Baroni et al., 2014; Musil, 2019). After using classification, we thought that clustering might bring new light to our problem. Contrary to classification which consists of optimizing weights in order to fit a training set to labels, clustering aims at *discovering structures and natural groupings inherent to the dataset*, without knowing what the real labels are (in some cases, they aren't known at all).

It is therefore an unsupervised Machine Learning algorithm, however, it is possible to evaluate its performance on true labels (Upadhyay et al., 2017) using *Adjusted Rand Index (ARI)*[13], an index that measures similarity between two clusterings (here, the one produced by KMeans, and the one produced by the true labels). As the name suggests, it is adjusted for chance, meaning that the ARI should be close to 0 in case we compare random clusterings. A negative score suggests that the compared clusterings show more difference than random ones, and a positive score suggests that they show less difference than random ones. Thus: the closer the ARI score is to 1, the better the clusterings match.

In order to retrieve single dimensions using the KMeans algorithm as a tool, we established the following strategy: from our original dataset of 512 dimensions WE,

[12]See `https://fr.wikipedia.org/wiki/K-moyennes` and `https://scikit-learn.org/stable/modules/clustering.html#k-means`

[13]See `https://en.wikipedia.org/wiki/Rand_index` and `https://scikit-learn.org/stable/modules/clustering.html#adjusted-rand-score`

we created 512 datasets of 1-dimension WE, by truncating the vectors to only 1 dimension. We then applied the KMeans algorithm on these truncated vectors and computed their ARI score. If the clustering quality is high for a given dimension, it means that it is, in itself, a good container of information about GG, GN or PoS. If the clustering quality is low, then the considered dimension may not be selected as an interesting candidate.

NOUN	ARI	ADJ	ARI	NOUN + ADJ	ARI
100	0.0285	466	0.0779	245	0.0280
195	0.0238	503	0.0703	192	0.0253
316	0.0213	250	0.0651	121	0.0239
245	0.0209	133	0.0611	507	0.0231
507	0.0194	439	0.0585	5	0.0203
192	0.0178	38	0.0536	250	0.0195
403	0.0172	39	0.0497	470	0.0190
121	0.0167	234	0.0490	439	0.0190
377	0.0156	432	0.0468	276	0.0187
202	0.0151	181	0.0454	1133	0.0187

(a) Results for GG

NOUN	ARI	ADJ	ARI	NOUN + ADJ	ARI
310	0.1135	310	0.1126	310	0.1078
54	0.0713	54	0.0745	54	0.0703
285	0.0569	285	0.0614	285	0.0575
288	0.0527	278	0.0568	288	0.0500
278	0.0510	81	0.0553	278	0.0497
81	0.0490	288	0.0520	81	0.0496
243	0.0427	25	0.0482	25	0.0405
359	0.0421	455	0.0453	495	0.0359
172	0.0390	56	0.0449	250	0.0354
182	0.0363	495	0.0448	243	0.0341

(b) Results for GN

NOUN vs. not-NOUN	ARI	ADJ vs. not-ADJ	ARI	VERB vs. not-VERB	ARI
159	0.1861	158	0.0312	310	0.1825
480	0.1614	220	0.0226	480	0.1804
310	0.1345	182	0.0216	159	0.1804
458	0.1337	119	0.0183	89	0.1477
401	0.1231	39	0.0180	458	0.1440
346	0.1272	318	0.0172	401	0.1346
29	0.1204	250	0.0172	192	0.1213
89	0.1009	464	0.0141	341	0.1131
128	0.0995	222	0.0132	29	0.1093
341	0.0964	208	0.0129	282	0.1076

(c) Results for PoS

Table 7: Top 10 subsets of size 1 to produce the best ARI score in KMeans clustering (n_clusters = 2) for GG, GN and PoS.

We can see the results of this approach in Table 7: it is worth noting that for all the experiments we see dimensions observed in multiple previous experiments: 100 and 466 for GG, 310 and 54 for GN, 159 and 310 for PoS and so on. However, for all GN, GG and PoS running KMeans on a single dimension creates a clustering of not very high quality judging by ARI metric. The highest results are observed for PoS: NOUN vs not-NOUN and VERB vs not-VERB. In order to repeat the experiment using several dimensions at a time, let's first summarize the results of all conducted experiments and the dimensions that appear in the results of multiple experiments.

2.4 Comparative analysis of different experiments

As we could see above, we obtained different dimensions that correspond to grammatical information using different experiments. In this section, we will compile all the highlighted dimensions and investigate the patterns among them. In our framework we suggest selecting only dimensions that appeared at least in 50% of conducted experiments (in our case 3 out of 6: (binary SIG-PROPS, AABCC-score, correlation, Logistic Regression and Perceptron weights, best clustering by one dimension).

Experiment count	NOUN	ADJ	NOUN + ADJ
6 experiments	100	250	-
5 experiments	195, 245, 316, 377, 507	245, 439, 466, 503	250
4 experiments	121, 192	181, 432, 133	245, 507
3 experiments	117, 403	234, 88, 121	181, 5, 192

(a) Results for GG

Experiment count	NOUN	ADJ	NOUN + ADJ
6 experiments	310, 54, 285, 172	310, 54, 285	310, 54
5 experiments	81	384	285, 81, 384
4 experiments	278, 288	25	278, 172
3 experiments	182, 250, 384	192, 81	359, 495, 288, 25

(b) Results for GN

Experiment count	NOUN vs. not-NOUN	VERB vs. not-VERB	ADJ vs. not-ADJ
6 experiments	-	-	-
5 experiments	159	89	-
4 experiments	346	159, 310 , 192	-
3 experiments	341, 401, 458, 480	341, 480, 401, 504	158, 220, 310

(c) Results for PoS

Table 8: Number of times a dimension was observed during the experiments for GG, GN and PoS among the top 10 of best dimensions.

In the table 8 above, we can notice the following:

- For GG the dimensions highlighted in most experiments for NOUN and ADJ are mostly not overlapping (the overlap is only in dimensions 245 and 121). This can lead to a hypothesis that there is a distinct way of encoding gender information for NOUN and ADJ. This can be further confirmed by the experiment combining both parts of speech, where we can see much fewer reproducible dimensions throughout different experiments (6 reproducible dimensions for NOUN+ADJ against 10 dimensions for NOUN and 11 for ADJ).

- On the contrary, for GN we can notice a significant overlap in the observed dimensions for NOUN and ADJ. This is further confirmed by the experiment combining NOUN and ADJ: 11 dimensions are reproduced in multiple experiments. Therefore, we can hypothesize that the way the number information is encoded in the French language is shared for NOUN and ADJ.

- As for PoS, we can notice that there are much fewer dimensions reproducible in several experiments in comparison to GN and GG. For ADJ/not-ADJ we previously saw very low correlation results, clustering precision, *etc.* We can now also see that there are only 3 dimensions that were repeated in 3 experiments. This can lead to an assumption, that information about ADJ is not directly encoded in the studied word embeddings and can be a sign of potentially poor WE. As for NOUN and VERB, we can notice that the dimensions appearing in different experiments are shared between the two categories in some cases (dimensions 159, 341, 401, 480).

- As for all experiments, we can notice that certain dimensions appear in the results of different experiment categories. For example, dimension 192 is highlighted for GG, GN and PoS, making this dimension ambiguous despite the results being reproducible across different experiments. Dimension 250 is also observed for both GG and GN. However, as we previously noticed, dimension 310 being seen for both GN and VERB/not-VERB classification can be explained by the absence of the GN parameter for VERB category which can allow to both separate VERB from NOUN and ADJ and store the plurality information.

You can find examples of words with lowest and highest values of some of the highlighted dimensions in Appendix A.1.

3 Subsets of dimensions encoding GG, GN and PoS

Up to now, we only tried to identify candidate *singletons* of dimensions to encode GG, GN or PoS. However, we know it is possible that some dimensions are very important for these features not on their own, but only when combined with other dimensions (for instance, in a relation of linear combination) Gladkova and Drozd (2016); Musil (2019). As we saw above, KMeans clustering of WE using just one dimension did not generate high-quality clusters, especially for GN and GG. In section 2.4, it was observed that a lot more dimensions were highlighted in multiple experiments for GG and GN, compared to PoS. This could be a sign that GG and GN information is encoded in more dimensions than PoS. In order to test this hypothesis, we decided to extend our previous KMeans strategy to dimension subsets of size ≥ 2.

3.1 Dimension couples (KMeans)

We first decided to test *dimension couples*: for every possible dimension pair we obtained 2 vectors and on these 2 vectors we performed KMeans clustering, and compared the obtained clusters with true labels (masculine vs. feminine, singular vs. plural, etc.) using ARI score. This is illustrated in Figure 8 (the protocol is an extension of the one used in Section 2.3.3).

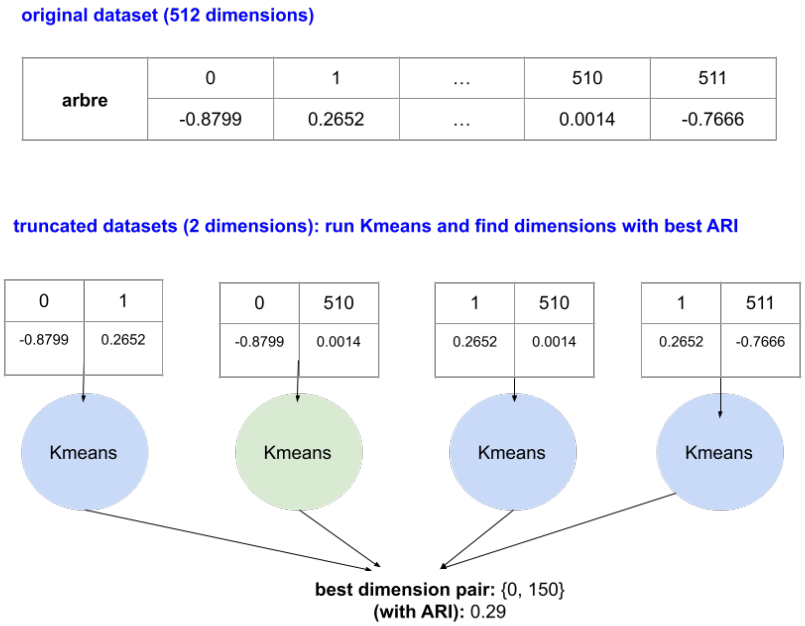


Figure 8: Strategy for discovering best dimension pairs in encoding GG, GN and PoS, using KMeans algorithm

Table 9 shows the lists of 10 best dimension pairs for all the experiments already conducted in Section 2.3.3. It can be seen that, in every experiment, the best ARI score when using 2 dimensions is at least 50% better than using only one (see results from Table 7). We also see that the majority of pairs is composed from dimensions found either in the previous experiment (KMeans for 1 dimension), or in at least 3 other experiments (see Table 2.4). In Table 9, these dimensions are bolded.

These results credit our hypothesis that combinations of dimensions are better in encoding GG, GN and PoS than single dimensions are. It also shows that the single dimensions retrieved from all the experiments of Section 2 are of first importance in constituting these combinations.

NOUN	ARI	ADJ	ARI	NOUN + ADJ	ARI
100, **245**	0.0460	**245**, **466**	0.1316	**121**, **245**	0.0480
195, **316**	0.0400	**439**, **466**	0.1227	**121**, **250**	0.0448
100, **316**	0.0394	**121**, **466**	0.1163	**245**, 316	0.0427
195, **507**	0.0390	**250**, **439**	0.1152	**192**, 195	0.0414
100, **195**	0.0380	**250**, **466**	0.1150	**5**, **245**	0.0412
245, **316**	0.0368	**466**, **503**	0.1144	195, **245**	0.0407
192, **195**	0.0365	**234**, **466**	0.1117	**245**, **507**	0.0403
100, **121**	0.0365	206, **466**	0.1114	**192**, **439**	0.0402
100, **377**	0.0364	**250**, **503**	0.1089	**245**, **439**	0.0401
121, **316**	0.0363	260, **466**	0.1086	**245**, 377	0.0400

(a) Results for GG

NOUN	ARI	ADJ	ARI	NOUN + ADJ	ARI
278, **310**	0.1633	**285**, **310**	0.1797	**285**, **310**	0.1741
285, **310**	0.1605	278, **310**	0.1791	**278**, **310**	0.1661
310, 359	0.1448	**54**, **310**	0.1783	**54**, **310**	0.1660
136, **310**	0.1400	**310**, 455	0.1663	136, **310**	0.1424
246, **310**	0.1397	**25**, **310**	0.1650	**310**, **359**	0.1424
54, **310**	0.1385	306, 310	0.1599	257, **310**	0.1408
81, **310**	0.1384	136, **310**	0.1586	**81**, **310**	0.1400
205, **310**	0.1347	**310**, 470	0.1551	200, **310**	0.1389
25, **310**	0.1304	191, **310**	0.1541	**25**, **310**	0.1388
259, **310**	0.1299	175, **310**	0.1463	246, **310**	0.1353

(b) Results for GN

NOUN vs. not-NOUN	ARI	ADJ vs. not-ADJ	ARI	VERB vs. not-VERB	ARI
159, 462	0.2899	124, 371	0.0413	**159**, 310	0.3126
159, **480**	0.2731	**220**, 485	0.0410	89, 480	0.3076
159, **310**	0.2674	**158**, 313	0.0390	**310**, **458**	0.2982
159, 434	0.2645	**158**, 429	0.0389	**89**, **310**	0.2965
159, **401**	0.2596	**182**, **220**	0.0373	**89**, **159**	0.2954
159, 198	0.2567	82, **158**	0.0373	**159**, **480**	0.2885
31, **159**	0.2530	138, **158**	0.0370	51, **310**	0.2807
159, **458**	0.2529	**220**, 439	0.0369	**159**, 434	0.2801
89, 159	0.2503	**158**, 297	0.0367	**159**, **401**	0.2774
401, **458**	0.2462	**158**, 224	0.0365	**401**, **458**	0.2773

(c) Results for PoS

Table 9: Top 10 subsets of size 2 to produce the best ARI score in KMeans clustering (n_clusters = 2) for GG, GN and PoS. Dimensions that were already found in KMeans for 1 dimension, or in 3+ previous experiments are bolded.

3.2 Larger subsets of dimensions

This very last result is crucial to guide our investigations on larger dimension subsets, given that computational costs don't allow to reproduce the present experiment on all possible combinations of dimensions.

Indeed, note that, for 512-dimension vectors, we find[14]:

- C^1_{512} = 512 dimension subsets of size 1 (singletons)
- C^2_{512} = 130,816 dimension subsets of size 2
- C^3_{512} = 22,238,720 dimension subsets of size 3

The number of combinations exponentially increases with the size of the considered subsets. For the present experiment, the computation on all dimension pairs (130,816) already took an average of 2.5 hours per experiment. This is the reason why we decided to restrict the investigation of larger subsets to those that are exclusively composed of dimensions from Table 2.4. Table 10 shows, for each experiment, the best subset composed of dimensions from Table 8 along with its ARI score. As a reference, we also show ARI scores when using all the dimensions.

Regarding the meaningfulness of the highlighted subsets in encoding grammatical information, when compared to the whole set of dimensions, we see that:

- In every experiment, the best subset obtains a higher ARI score than the whole set of dimensions.
- The largest difference is observed in GN (on ADJ dataset): the ARI increases by 0.3986 points when keeping only dimensions of the best subset. The quality of clustering improves from quasi-random to quite good (remember that 1 is the perfect ARI score). A very similar improvement is observed in GG, also on the dataset of adjectives. For the other datasets, GG clustering fails to obtain interesting ARI scores.
- In NOUN vs. not-NOUN and VERB vs. not-VERB, we observe that the ARI score on all dimensions is already significantly high, whereas it is close to 0 for all the other experiments. This suggests that the useful information for encoding NOUN-ness and VERB-ness is actually distributed in the whole set of dimensions. Keeping dimensions found in the best subsets still increases the score, but this is relatively less impressive than in the GN experiment.

As for the conclusions we draw about the encoding of GG, GN and PoS:

[14] C^i_{512} represents the number of combinations of size i from a set of 512 elements (we look at combinations, because the order of dimensions within the subsets doesn't matter). It is computed using the formula $C^i_{512} = \frac{512 \times 511 \times ... \times (512-i)}{1 \times 2 \times ... \times i}$.

Category	Best dimension combination	ARI on best subset	ARI on all 512 dimensions
GG (NOUN)	100, 117, 192, 195, 245, 316, 377, 403, 507	**0.1058**	0.0066
GG (ADJ)	88, 121, 133, 181, 234, 245, 250, 432, 439, 466, 503	**0.3836**	-0.0008
GG (NOUN + ADJ)	5, 192, 245, 250, 377, 439, 507	**0.1052**	0.0025
GN (NOUN)	54, 81, 172, 250, 285, 310, 359, 384	**0.3848**	-0.0003
GN (ADJ)	25, 54, 81, 84, 274, 285, 310, 384, 455	**0.4013**	0.0027
GN (NOUN + ADJ)	25, 54, 81, 172, 278, 285, 310, 359, 384, 495	**0.3489**	- 0.0056
PoS (NOUN vs. not-NOUN)	128, 159, 341, 401, 458, 480	**0.4258**	0.4112
PoS (ADJ vs. not-ADJ)	158, 220, 464	**0.0457**	- 0.0080
PoS (VERB vs. not-VERB)	89, 159, 192, 282, 310, 341, 401, 458, 480, 504	**0.5932**	0.4832

Table 10: ARI score of the best subset for each experiment, compared with the ARI score on the whole WE (512 dimensions).

- We can see a further confirmation that GG is encoded differently for ADJ and NOUN, as the best subsets for these two categories do only overlap on dimension 245. Moreover, the highlighted dimensions for NOUN generate clusters of much lower accuracy than for ADJ, which can lead us to the conclusion that GG information for nouns is not encoded directly in the WE.

- As for GN, we can see that combinations of highlighted dimensions generate much more accurate clusters than KMeans using just one dimension. We can also notice that even though there are a lot of shared dimensions for NOUN+ADJ and ARI is relatively high for it, it is still slightly less accurate than clustering for just NOUN and ADJ, which can be a sign that treating NOUN and ADJ as separate categories is beneficial for GN as well.

- For PoS, we can see that despite having fewer highlighted dimensions in comparison to GG and GN, VERB clusters are of the highest quality we observed during the experiments. This can suggest that fewer dimensions contain more

information about "verb-ness" than, for example, about GG and GN. On the contrary, clustering for ADJ/not-ADJ is still very poor which confirms the theory that there is no clear encoding of ADJ/not-ADJ information in the studied word embeddings.

We propose using the ARI scores on the best subset from 10 as the way to evaluate the quality of grammatical information encoding in the given WE. As we can see, encoding of GG for NOUN and PoS information for ADJ/not-ADJ is quite poor in the studied WE, while GN and PoS encoding appears to be of a much higher quality. We suggest using the same framework to compare different WE in order to evaluate their grammatical richness.

4 Conclusion and further work

In this paper, we have investigated the WE produced by the FlauBERT model. In our work, we proposed a framework for extracting morphological information from WE using the following steps:

1) Create a dataset with a morphological feature encoded in a binary way.

2) Perform a series of experiments including training a classifier (or several) on the dataset and extracting the dimensions with the highest weights, and retrieving dimensions with the highest correlation to the morphological feature.

3) We as well propose a binary SIG-PROPS and AABCC-score as extra metrics for the extraction of morphologically meaningful dimensions.

4) After performing several experiments, highlight the dimensions observed at least in 50% of the experiments.

5) Obtain all possible combinations of dimensions from step 4 and perform KMeans clustering on each of them. Find the combination with the highest ARI score.

6) Use the ARI score from step 5 to compare WE obtained from different models in terms of how well they pertain grammatical information.

Using this guideline, we isolated dimension subsets that are the most likely to encode GG, GN and PoS in FlauBERT WE (see Table 10). We observed several facts about these Word Embeddings:

- Generally speaking, grammatical information is likely to be encoded in multiple dimensions at the same time: this can be both observed in the clustering experiment results and the dimension investigation results;
- *Grammatical Gender* information is encoded in different dimensions for NOUN and ADJ;

- *Grammatical Number* information mostly encoded in the same dimensions for NOUN and ADJ;
- *Part of Speech* information about adjectives was not conclusive in any of the performed experiments. However, Part of Speech information about nouns and verbs is frequently encoded by the same dimensions.

This work put a first step in elucidating the problem of grammatical information retrieval in WE. But it raises new questions and perspectives that could be interesting to explore in the future:

- Validate the framework with WE produced by other models than FlauBERT and by other FlauBERT models with a higher number of dimensions;
- Extend to other grammatical characteristics: verb tenses, other PoS (adverbs, prepositions, etc.) and to other languages;
- Deepen the study in order to go towards a linguistic information predictor: from the value of certain dimensions of a WE, predict its GG, GN, PoS, etc;
- Deepen the study of ambiguous forms (that have for the moment been left out in our work) such as *lit* or *est* (that can be a VERB or a NOUN): this will lead to an incorporation of semantic information.
- Take into consideration the fact that an ambiguity of grammatical gender or grammatical number is also possible for a given form of a french word. For instance, forms like *livre* and *litre* present masculine and feminine noun homonyms[15]. As for grammatical number, forms ending in "-s", "-z", or "-x" in the singular form are unchanged in the plural form, such as *gaz* [16]. In the present study, we focused on the Part-of-speech ambiguity, which seemed way more common to us, but we could also imagine removing these ambiguous forms from the dataset to reproduce the experiments about gender and number.

[15] *livre* and *litre* can be masculine or feminine nouns, see `https://www.cnrtl.fr/definition/livre` and `https://www.cnrtl.fr/definition/litre`

[16] *gaz* is invariable in number, see: `https://www.cnrtl.fr/definition/gaz`

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. Association for Computational Linguistics.

Kyoung-Rok Jang and Sung-Hyon Myaeng. 2017. Elucidating conceptual properties from word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 91–95, Valencia, Spain. Association for Computational Linguistics.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Tomá Musil. 2019. Examining structure of word embeddings with pca. In *International Conference on Text, Speech and Dialogue*.

Shyam Upadhyay, Kai-Wei Chang, Matt Taddy, Adam Kalai, and James Zou. 2017. Beyond bilingual: Multi-sense word embeddings using multilingual context. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 101–110, Vancouver, Canada. Association for Computational Linguistics.

Appendices

Link to GitHub repository of the project (publicly available): `https://github.com/ClementineBleuze/WordEmbeddings`

A Examination of a few dimensions

In this part, we sketch an examination of a few dimensions that revealed interesting in our study:

- For GG: 100 (NOUN), 250 (ADJ, NOUN + ADJ)
- For GN: 310

A.1 Dimensions 100 and 250 in GG

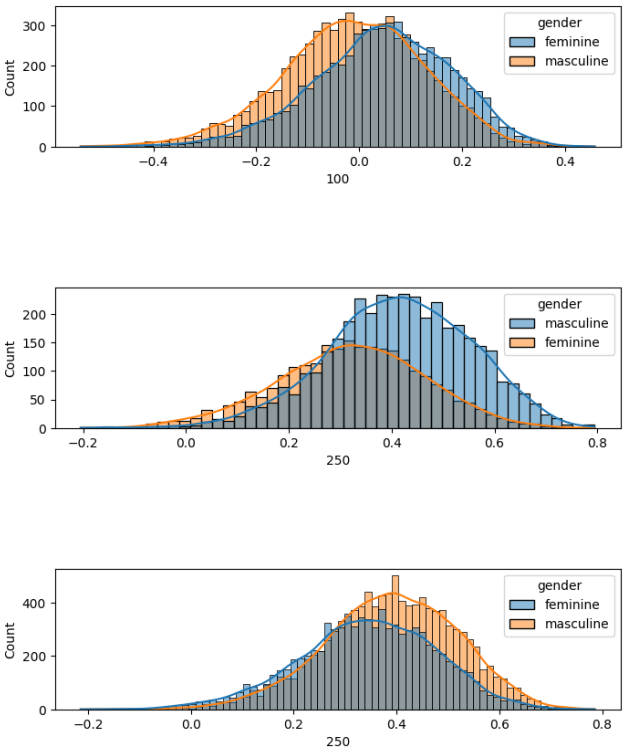


Figure 9: Distributions of values for dimensions 100 (NOUN), 250 (ADJ) and 250 (NOUN + ADJ) in masculine and feminine nouns embeddings (data has been normalised).

NOUN			
Top 10 values		Bottom 10 value	
Word	Gender	Word	Gender
voyant	m	incomplète	f
attitude	f	bambou	m
visibilité	f	vice	m
organisatrice	f	découvert	m
affectant	m	liège	m
gêne	f	fossile	m
items	m	beurre	m
cout	m	yoga	m
prestation	f	tibétain	m
foutre	m	caisson	m

(a) Results for NOUN

From the distribution of dimensions 100 and 250 (see Figure 9), we see that ,generally, masculine nouns have higher values for dimension 100 than feminine nouns. Masculine adjectives have higher values than feminine ones for dimension 250, but the contrary holds for the dataset of nouns and adjectives. This suggests that nouns show the opposite trend. We observe in Figure 11 the Word Embeddings from our datasets presenting the 10 highest and 10 lowest values for these two dimensions. It can be seen that outliers are still found in these lists.

ADJ			
Top 10 values		Bottom 10 value	
Word	Gender	Word	Gender
entrevues	f	peint	m
syndicaux	m	présente	f
éloignés	m	décroissante	f
médicateurs	m	ordonnées	f
étonnés	m	postal	m
survaillants	m	immédiats	m
associatifs	m	démaquillants	m
croyants	m	intercommunale	f
buvant	m	recherchée	f
confidentiel	m	biennal	m

(b) Results for ADJ

NOUN + ADJ			
Top 10 values		Bottom 10 value	
Word	Gender	Word	Gender
forums	m	fonctionnellle	f
cris	m	capitale	f
tribunes	f	sensualité	f
couloirs	m	négociations	f
pleurs	m	cycle	m
versets	m	famille	f
caresses	f	réconciliation	f
verset	m	signature	f
entrevues	f	composée	f
groupes	m	trêve	f

(c) Results for NOUN+ADJ

Table 11: Examples of words sorted by value of top 10 and bottom 10 for dimensions 100 (NOUN) , 250(ADJ) and 250(NOUN+ADJ). Note: m for masculine, f for feminine.

A.2 Dimension 310 in GN

Performing the same overview for dimension 310 in GN on all the datasets (given that this single dimension was found first to encode GN in every experiment), we observe that singular nouns and adjectives present high values for this dimension, whereas plural nouns and ajectives present low values (see Figure 10). This is neatly seen in the top 10 words having highest and lowest values for this dimension in Table 12 where very few outliers are present.

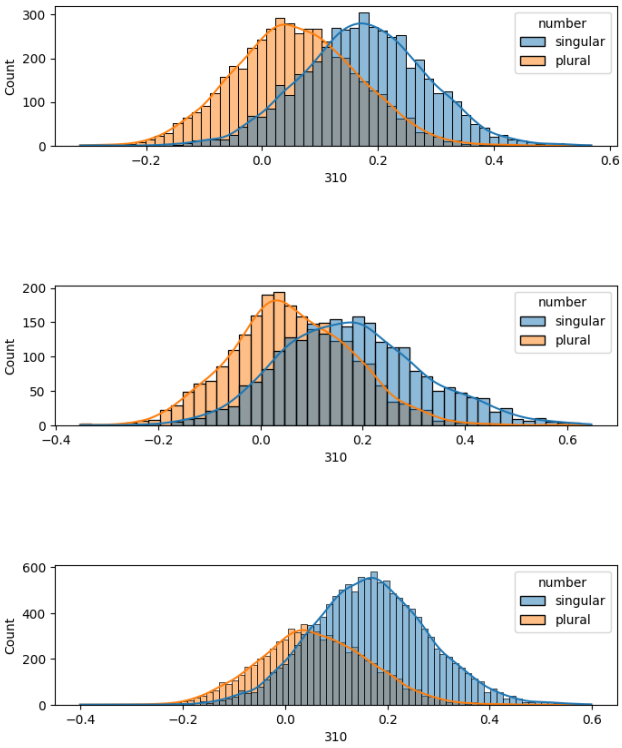


Figure 10: Distributions of values for dimensions 310 (NOUN, ADJ and NOUN + ADJ), in singular and plural nouns embeddings (data has been normalised).

NOUN			
Top 10 values		Bottom 10 value	
Word	Number	Word	Numer
affirmé	s	préférés	p
commiddaire	s	phones	p
affirmant	s	internautes	p
aff	s	pédophiles	p
adjointe	s	partages	p
représentante	s	blogueurs	p
nécessité	s	smartphones	p
RDA	s	urne	s
unanimité	s	bagues	p
dérogation	s	votes	p

(a) Results for NOUN

ADJ			
Top 10 values		Bottom 10 value	
Word	Number	Word	Numer
considérant	s	manufacturier	s
insisté	s	oraux	p
rappelé	s	transfrontaliers	p
affirmé	s	verticaux	p
exhorté	s	entrants	p
plaidé	s	préférés	p
réunis	p	lesdits	p
délégué	s	connectés	p
insistant	s	sécurisés	p
pre	s	verbaux	p

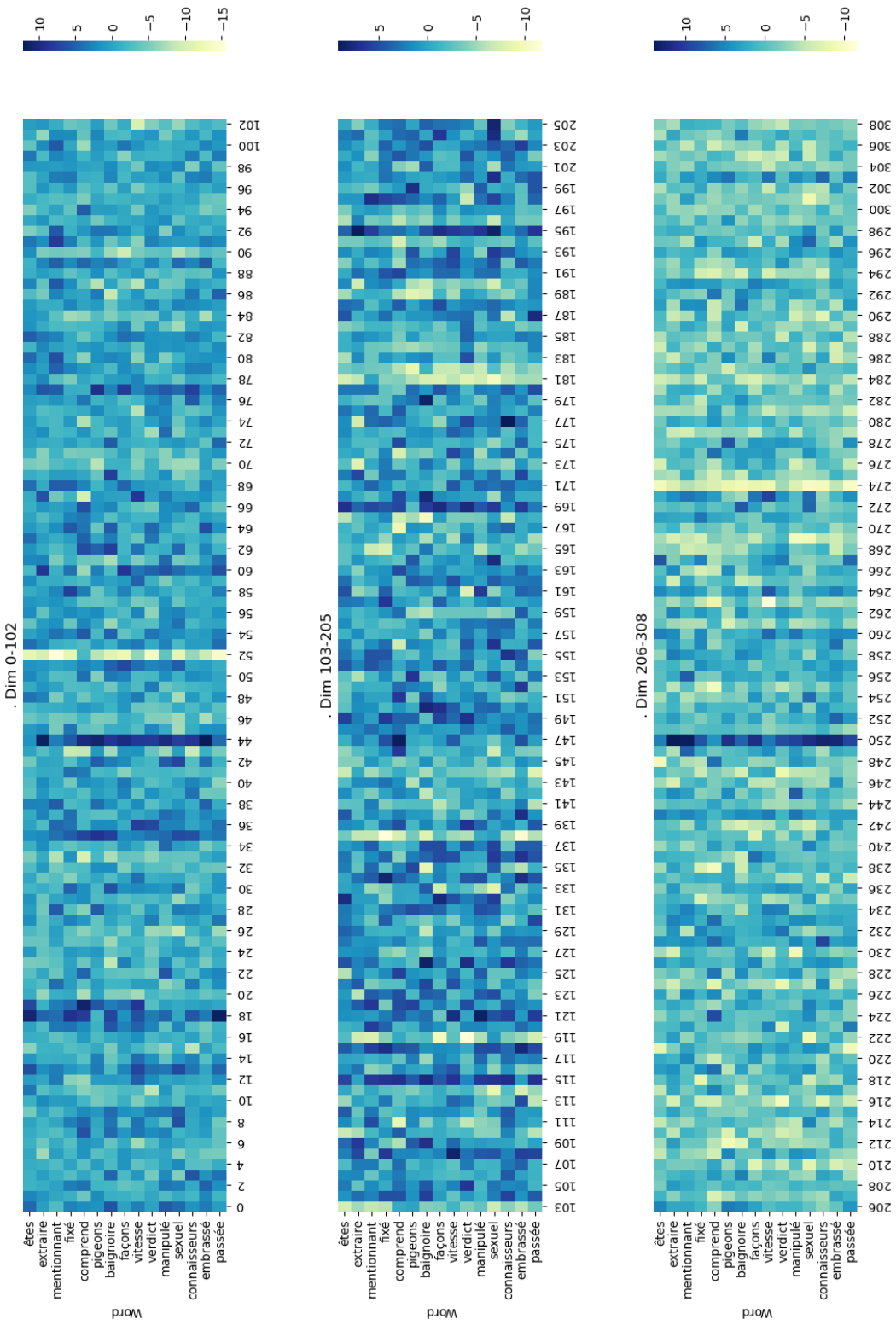
(b) Results for ADJ

NOUN + ADJ			
Top 10 values		Bottom 10 value	
Word	Number	Word	Numer
considérant	s	connectés	p
constant	s	partages	p
insisté	s	blogueurs	p
ré	s	smartphones	p
rappelé	s	urne	s
affirmé	s	bagues	p
exhorté	s	votes	p
réunie	s	sécurisés	p
plaidé	s	verbaux	p
réunis	p	requin	s

(c) Results for NOUN+ADJ

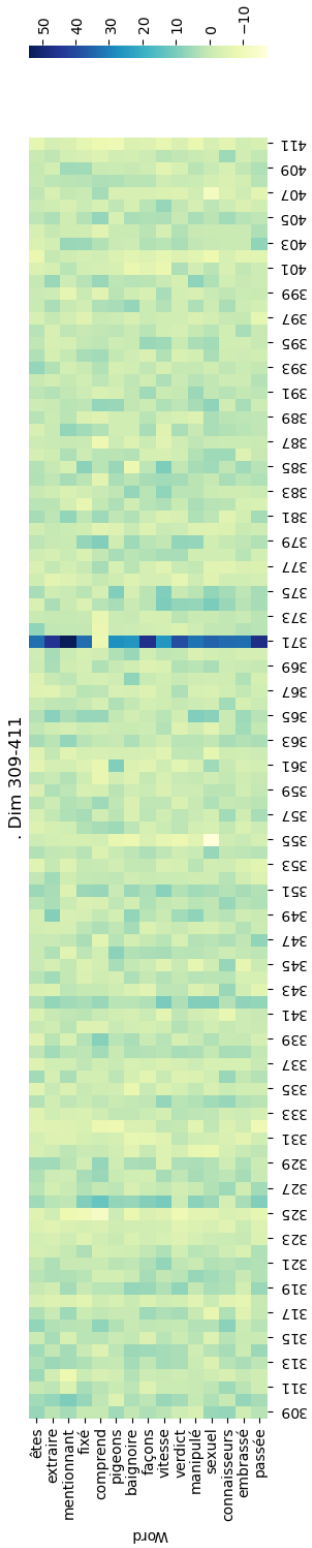
Table 12: Examples of words sorted by value of top 10 and bottom 10 for dimensions 310 (NOUN, ADJ and NOUN+ADJ). Note: m for masculine, f for feminine.

B Visualization of Word Embeddings

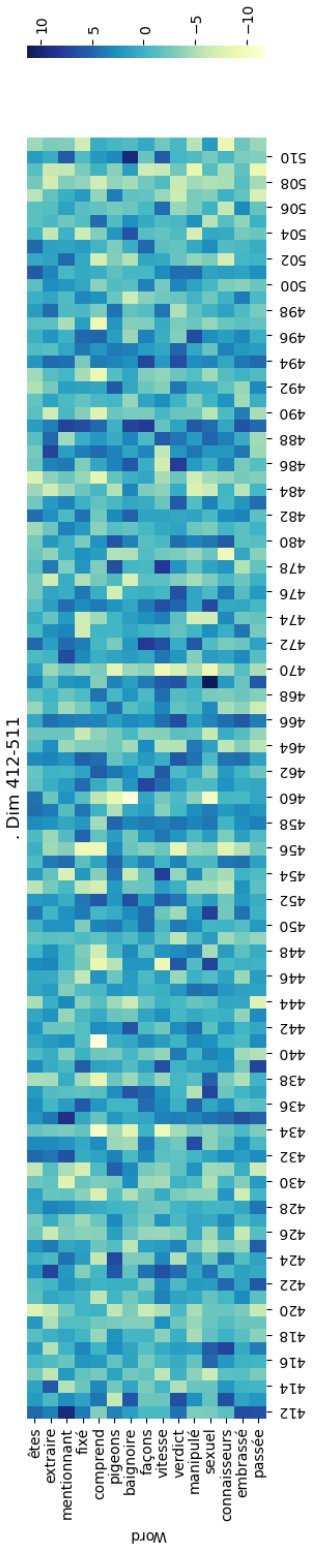


(a) Dimensions 0-102 (b) Dimensions 103-205 (c) Dimensions 206-308

Figure 11: Plot of dimension values of WE of a random sample of words (dimensions 0-308).



(a) Dimensions 309-411



(b) Dimensions 412-511

Figure 12: Plot of dimension values of WE of a random sample of words (dimensions 308-512).