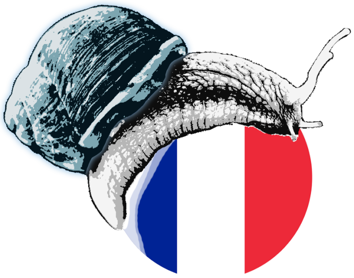
Traducteur Épicène



NLP and Written Corpora Project

by

Group C*hampion*: Eduardo Calò, Olivier Ratenon, Ludivine Robert, Thibo Rosemplatt

under the guidance of

Maxime Amblard, Bruno Guillaume

December, 2019

Chapter 1

Introduction

The current report will present the results obtained in our NLP and Written Corpora's courses joint project, in the framework of MSc first year. The aim of this project was to develop an application to become familiar with the resources, techniques and tools used in the field of Natural Language Processing (NLP).

As the choice of the subject was free, we decided to work on the epicene side of French language, building a translator. In particular, in our project, an intralingual translator, which turns a standard French text into its epicene form, was created.

Intralingual translation, or rewording, was defined by Jakobson as "an interpretation of verbal signs by means of other signs of the same language" [5].

An epicene[1] form aims at representing both genders at the same time, when the context does not need it to be specified. The purpose of this form is to avoid gender discrimination. The utilization of the epicene French is a hot topic, due to the feminist philosophy and the LGBTQIA+ community.

Epicene language (also known as gender-inclusive language) is a sub-part of inclusive writing. The latter's purpose is to avoid any kind of discrimination, not only narrowed to gender-discrimination. Some languages are by nature more epicene than others. For example, English does not take grammatical gender, as opposed to French. Hence, as previously mentioned, our tool focuses only on the French language.

Moreover, it is also important to mention that there are no official rules for writing epicene language. This is the reason why the tool relies on a set of rules chosen by the creators, following the common utilization of the French epicene language.

[1] from Ancient Greek $\grave{\varepsilon}\pi\acute{\iota}\kappa o\iota\nu o\zeta$ / épikoinos, "owned in common", meaning "that represents, under one single form, one and the other sex".

Chapter 2

Linguistic description

In the realm of linguistic categorization, grammatical gender is regarded by many linguists as a specific form of noun class system. Categorizing means labelling each element of a determined group sharing some features within mutually exclusive categories. That being said, noun classes are categories of nouns. A criterion of belonging to a given class may be the characteristic features of the referent to which the noun is related, such as biological sex, animacy, shape, etc. If a language presents noun classes, they can be even up to 20 [3]. Grammatical gender can be regarded as a specific type of noun class system when, in a language, these classes are few (2 or 3).

In grammatical gender, common divisions are masculine, feminine and neuter. The assignment of these grammatical gender might be determined by the biological sex of the referent (but not always) or can be apparently arbitrary (i.e., gender assignment of inanimate entities in languages with only masculine/feminine subdivision criteria) [4].

An important feature of grammatical gender phenomenon is agreement (when words related to a noun inflect according to the gender of noun) [2]. This aspect has been crucial for the project, since we had to catch all the lexical elements (adjectives, determiners, verbs) related to an *epicenizable* noun and *epicenize* them, as well. Epicene language, given its aim, focuses only on nouns related to human referents.

For the purpose of our project, we divided French nouns which have a human referent in 4 different groups. We used grammatical gender of the word form, biological gender of the referents, and agreement as division criteria. Examples of each group can be seen in Figure 2.2.

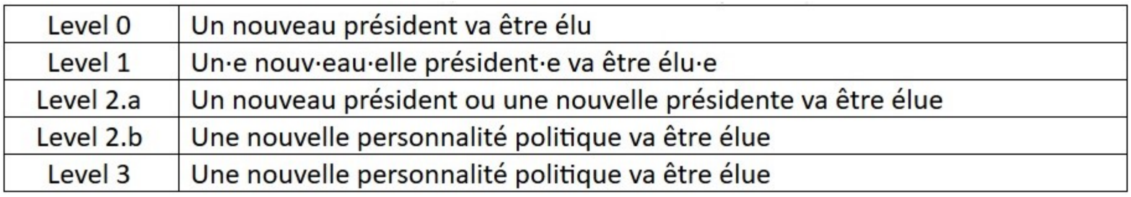
- Group 1: These nouns are already epicene by nature. Both grammatical genders refer to the same word form, which, in turn, refers both to male and female referents. Agreement of related words change accordingly to the referent. In this case no change is applied to the noun, but all the related words are *epicenized*.

- Group 2: two different word forms sharing the same stem, one for each grammatical gender. One word form refers to a male referent, the other to a female referent. In this case both forms and all syntactically related words are *epicenized*.

- Group 3: nouns which have a fixed grammatical gender but refer to both male and female referents. In this case, agreement follows grammatical gender. No *epicenization* is performed here.

- Group 4: semantically related doublets which do not share the same stem. One element of the doublet refers to male referent, the other to female referent. These doublets are not addressed by our tool.

On the basis of readability, the subdivision of the epicene language we suggest is as Figure 2.1 shows. Our tool focuses only on Level 1. Level 0 is a text written in standard French (input of our tool). Level 1 is the epicene level achieved with the use of mid-dot notation (output of our tool). Mid-points positions have been decided using a set of rules of our creation, with the general idea of keeping the common part in masculine and feminine forms before the first mid-point, having the masculine part after it and then the feminine part after another mid-point. Eventually, in case of plural forms, another mid-point and plural indicator were added. We tried to follow French morphemes, but it was not always possible. Level 2.a is the epicene level achieved using conjunctions and both forms for masculine and feminine. Level 2.b is achieved using genderless periphrastics. Level 3 is the final aim of epicene language, a fluent readable written text (in this case corresponds to Level 2.b, but it is not always the case).



Level 0	Un nouveau président va être élu
Level 1	Un·e nouv·eau·elle président·e va être élu·e
Level 2.a	Un nouveau président ou une nouvelle présidente va être élue
Level 2.b	Une nouvelle personnalité politique va être élue
Level 3	Une nouvelle personnalité politique va être élue

Figure 2.1: Levels of *epicenity*.

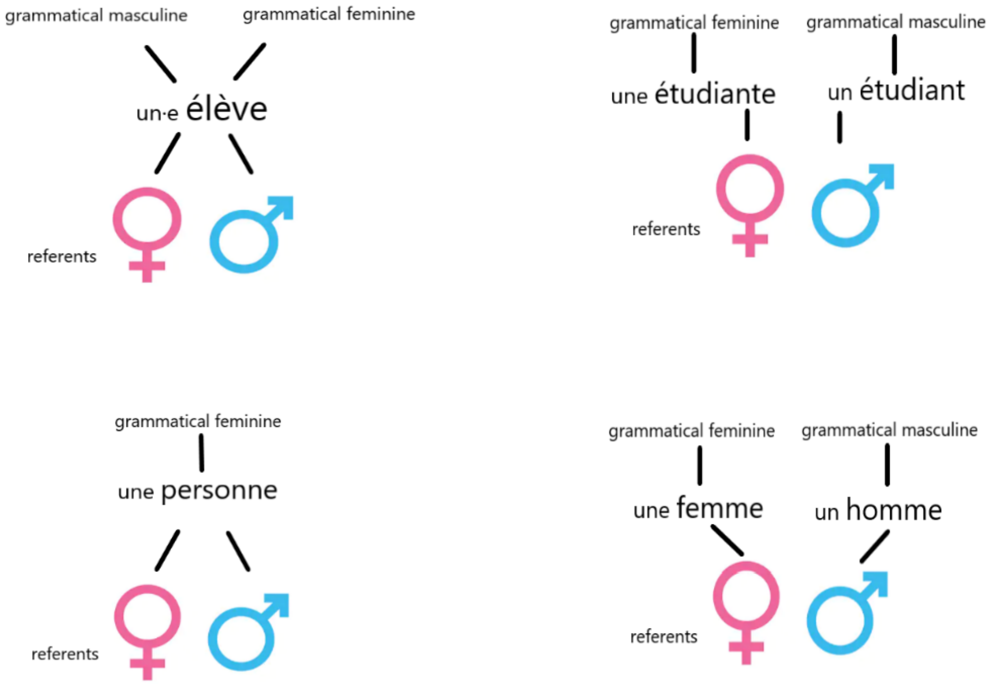


Figure 2.2: Groups of nouns. Top left: Group 1, top right: Group 2, bottom left: Group 3, bottom right: Group 4.

Chapter 3

Approach

3.1 Pipeline

3.1.1 Spot the nouns

As mentioned in Section 2, in the French language, only nouns, adjectives, pronouns, verbs in the past participle, and determinants can take a grammatical gender, and all of these entities are inflected according to the noun. Thus, by spotting the nouns, the dependent tokens can be retrieved using the language model, and then, they can be epicenized the right way, if needed.

3.1.2 Find the referent of a noun

Not all the nouns and their dependencies need to be epicenized. Actually, a necessary (but not sufficient) condition for a token to be epicenized is that the noun to which it is attached refers to a human being (see Section 2). So, each noun of an input which is not related to a human being will be discarded.

This condition of referring to a human being is not sufficient because sometimes the purpose of an utterance is to actually refer to a specific gender (*"Les footballeuses françaises sont les meilleures du monde"*). Our program will not handle such cases, and will epicenize the noun and its dependencies every time it is referring to a human referent.

3.1.3 Retrieve the dependencies

Once all the nouns which refer to a human are retrieved, the next step is to retrieve also their dependent tokens that can take a grammatical gender (see Section 3.1.1).

3.1.4 Epicenize token by token

Since now we know exactly which tokens need to be epicenized, the final step is to apply the transformation to each token by concatenating masculine and feminine morphemes using mid-dots (·).

Epicenization of nouns and adjectives

This process is as follows:
In all the cases, take the feminine and the masculine form of the word (for further details, check Section 3.2.3). If they are the same, this means that the word is epicene and it does not need to be processed. In any other case, they will be compared such that we know which part is common to both forms (noted *common*), which part is the masculine suffix (noted *masc*), and which one is the feminine suffix (noted *fem*).

If the head noun[1] related to the token is singular, then write the token as follows:

$$common \cdot masc \cdot fem$$

Note that sometimes the masculine form is included in the feminine form in its whole (*e.g : Joli, Jolie*). Thus, *masc* is the empty string, and the preceding formula becomes *common · fem*.

However, if the noun is plural, then the process is a little trickier.
If the masculine and the feminine have the same plural suffix (noted *masc_plur*, *fem_plur* or *plur* if common), then the token will be written:

$$common \cdot masc \cdot fem \cdot plur$$

This case is the most common, and the plural suffix is usually a "s". (*e.g : gentils (s), gentilles (s)*)
However, if the masculine and feminine forms do not share the same plural suffix (see how this is checked in Section 3.2.3), the token will be written as:

$$common \cdot masc\ masc\_plur \cdot fem\ fem\_plur$$

(*e.g: "Beau" becomes "beaux", "belle" becomes "belles"*)

[1] For example: in *"Les talentueux footballeurs sont sortis victorieux de cette compétition"*, *footballeurs* is the head noun because *talentueux*, *victorieux*, and *sortis* are inflected in accordance with *footballeurs*.

Epicenization of determinants

There are really few determinants, compared to the amount of nouns or adjectives. Thus, it was not useful to write some general inflexion rules, and we decided to *hard-code* the rules. It is only a matter of pre-written rules, working on the basis of string-replacement.

Epicenization of verbs in the past participle

As far as the verbs are concerned, the only thing to do is to add at the end of the token *"e"*, if the head noun is singular, *"e·s"* if it is plural.

3.1.5 Output

Finally, the words that have been processed will replace the old ones in the document. Now the linguistic aspect of the project has been covered, the functioning is going to be presented.

3.2 Code

3.2.1 Libraries used

In this project, the libraries spaCy, os, ElementTree have been used.

spaCy has been used to load the language model "fr_core_news_sm". Many of them were available, but this one was the most suitable because it includes the parts-of-speech tagger, and also a syntactic dependencies annotation.

ElementTree is the library used to browse the Morphalou and Wolf XML files (see Section 4.2 to know more about them, see section 3.2.3 to know how they have been used).

Os is used to read the content of the working directories, so that the program knows what file needs to be processed.

3.2.2 Code pipeline

As shown in Figure 3.1, we created 7 files (adj_inflector.py, det_rules.py, dottize.py, get_relations.py, main.py, nouns_inflector.py, primitives.py), for a total of more than eight hundred lines of code. The pipeline starts in the **main.py** file.

1. The input files from the *./inputs* folder are listed. The ones which have already been processed are discarded with the help of the os library (see Section 3.2.1). The files will then be processed, one by one, as follows.

2. The content of the currently processed file will be read, and loaded by the spaCy language model; a new object called *doc* will be created. This doc contains each token originally in the input text, now converted into objects containing the text, their part-of-speech, their syntactic dependencies, and their position within the document (or index).

3. A function based on a filter, and a lambda function will create a list containing the indexes of all the nouns of the doc.

4. For each of these nouns, a check for their belonging to the set of the human-beings will be made by passing them as input in the is_human_from_noun function from the **primitives.py** file. Only the nouns referring to a human will be kept. The functioning of this script is explained in Section 3.2.3.

5. Using the functions of the **get_relations.py** file, indexes of the dependencies of the nouns will be retrieved. The functioning of this script is explained section 3.2.3.

6. We now have a list of the indexes of the nouns related to human, and the indexes of their dependencies that might be inflected. The next step is to pass them through the *epicenize* function of **main.py**, which takes as input the token that need to be epicenized and its head noun. According to the part of speech of the token, they will be sent to the right function of epicenization of the **dottize.py** file. They work as covered in Section 3.1.4.

7. At this point, each token has been processed and some of them have been transformed into their epicene form. The final epicene text will now be printed inside a new .txt file, created inside the ./outputs folder. The whole process will reiterate for each non processed file of the ./inputs folder.

3.2.3 Miscellaneous details about the code

Inflectors

In order to epicenize the nouns and adjective, we must be able to move instantaneously from the singular to the plural and from the feminine to the masculine.

Indeed, this will allow to compare the morphological differences within the same lexeme.
This is done using the Morphalou XML files (see Section 4.2). One is dedicated to the nouns, the other to the adjectives. These files contain the grammatically inflected (including the gendered and numbered) forms of a good amount of the French words. Thus, it was possible to create a set of functions allowing switch from one form to another, and to obtain grammatical information about any noun or adjective.

Getting syntactic relationships

The aim of this script is to get all the lexical items that have to be epicenized syntactically related to the epicenizable noun. The functioning is quite straightforward. It exploits the syntactic parser present in spaCy. To study the relations, Universal Dependencies[2] and Grew-match[3] have been used. Each function of the file is specific for a particular type of lexical item (adjective, determinant, etc.). Each function take a processed spaCy doc and the index of a noun eligible for epicenization as inputs, retrieve the items related on the base of the relations and output lists of indexes which correspond to the indexes of the epicenizable items related to that noun by that syntactic relation.

Retrieving the referent of a noun

The main problem with epicenization is choosing which nouns are eligible for being epicenized. Since only the ones which refer to a human being have the right properties, we had to find a way to check if a noun's referent was a human being or not. To do that, we have used WOLF (WOrdnet Libre du Français) (see Section 4.2 for a detailed description). When a noun had to be handled, the XML file was opened and explored with ElementTree (see Section 3.2.1) in order to find the line containing our lexical entry; its ID was stored and then we moved up from hyperonym to hyperonym until finding a primitive that allowed us to understand if the noun referred to a human being or not. To do that, we used a recursive function that goes to the next hyperonym, until one of the general meaning ”things”, ”animal”, ”idea” or ”human” is found. If ”human” is found, the function stops and the word is considered to refer to a human being.

[2] https://universaldependencies.org/

[3] http://match.grew.fr/

3.3 User guide

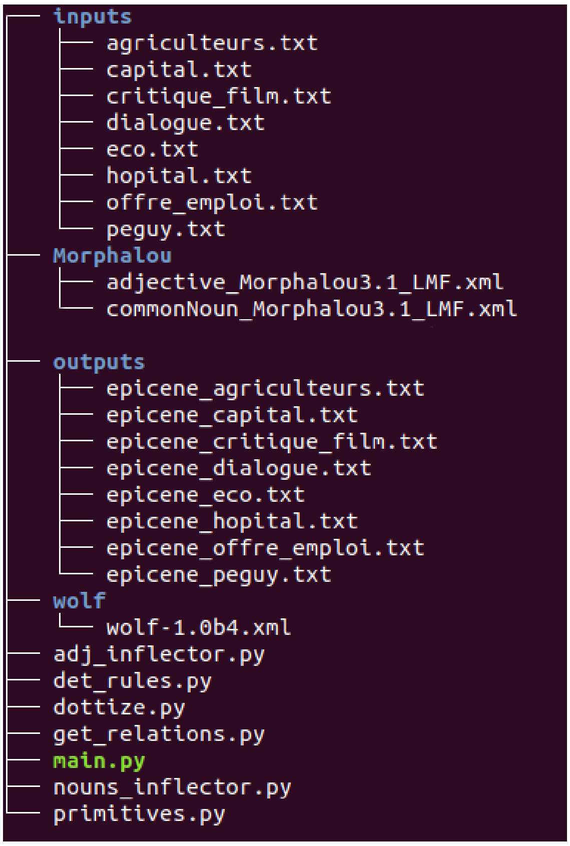
The first thing to do is to download spaCy (see Section 3.2.1) via pip.

```
pip install spacy
```

Then, the French language model "fr_core_news_sm" must be downloaded by typing in the terminal:

```
python -m spacy download fr_core_news_sm
```

Now that the prerequisites are in place, it is time to consider the practical work. The directory tree is presented in Figure 3.1.



```
├── inputs
│   ├── agriculteurs.txt
│   ├── capital.txt
│   ├── critique_film.txt
│   ├── dialogue.txt
│   ├── eco.txt
│   ├── hopital.txt
│   ├── offre_emploi.txt
│   └── peguy.txt
├── Morphalou
│   ├── adjective_Morphalou3.1_LMF.xml
│   └── commonNoun_Morphalou3.1_LMF.xml

├── outputs
│   ├── epicene_agriculteurs.txt
│   ├── epicene_capital.txt
│   ├── epicene_critique_film.txt
│   ├── epicene_dialogue.txt
│   ├── epicene_eco.txt
│   ├── epicene_hopital.txt
│   ├── epicene_offre_emploi.txt
│   └── epicene_peguy.txt
├── wolf
│   └── wolf-1.0b4.xml
├── adj_inflector.py
├── det_rules.py
├── dottize.py
├── get_relations.py
├── main.py
├── nouns_inflector.py
└── primitives.py
```

Figure 3.1: Directory tree of the project.

The *X.txt* files that one wants to epicenize must be put inside the *inputs* folder. Now, the only thing left is to run the main function, and the output files will be created as *output_X.txt* in the outputs folder.

Chapter 4

Datasets

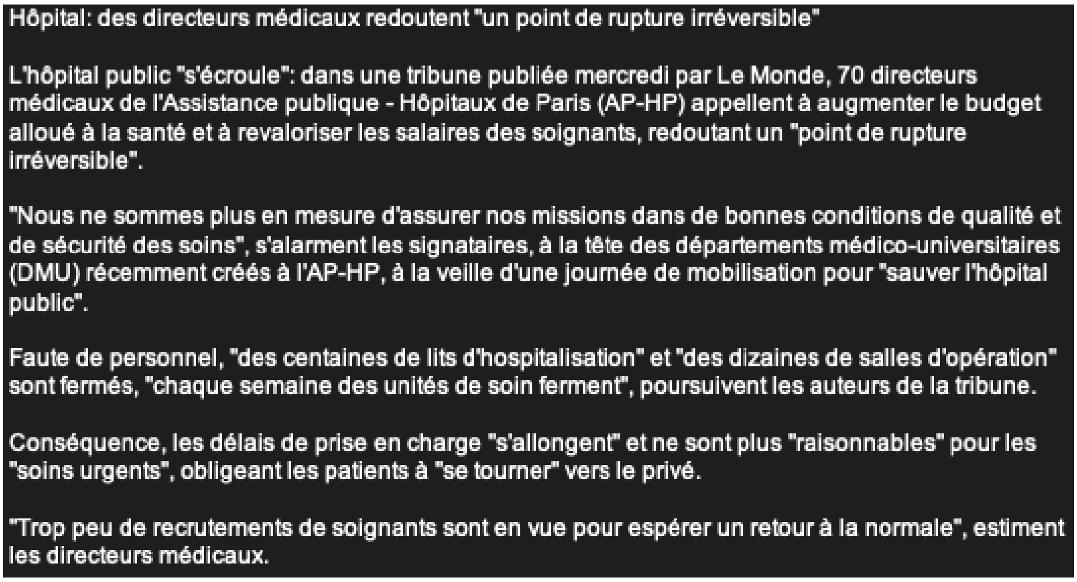
Within the scope of our project, 2 types of datasets have been used: written corpora and linguistic resources. In the following paragraphs an overview will be given.

4.1 Written corpora

Written corpora have been used for mainly two tasks in our project: creation of epicenization rules (see Section 3.1.4) and evaluation (see Section 5). These corpora include texts of various types (press articles, job offers, film reviews, short news, etc.) selected online on public and free websites (see Figure 4.1 for an example). All these texts have been manually translated and aligned, helping us establishing and adjusting the rules. These manually translated epicene texts (see Figure 4.2) have become our gold standard for the first evaluation (the one in which we compare the output of our tool with these texts to see if our tool over-epicenized or missed some epicenizations, see Section 5).

For the second evaluation (see Section 5 for details), the output texts of our program have become the corpus. In this evaluation, we give our output texts as input to Antidote software[1] to check if our tool missed some epicenizations and, above all, to have a third-party method of evaluation, different from our own texts.

[1] https://www.antidote.info/fr/



Hôpital: des directeurs médicaux redoutent "un point de rupture irréversible"

L'hôpital public "s'écroule": dans une tribune publiée mercredi par Le Monde, 70 directeurs médicaux de l'Assistance publique - Hôpitaux de Paris (AP-HP) appellent à augmenter le budget alloué à la santé et à revaloriser les salaires des soignants, redoutant un "point de rupture irréversible".

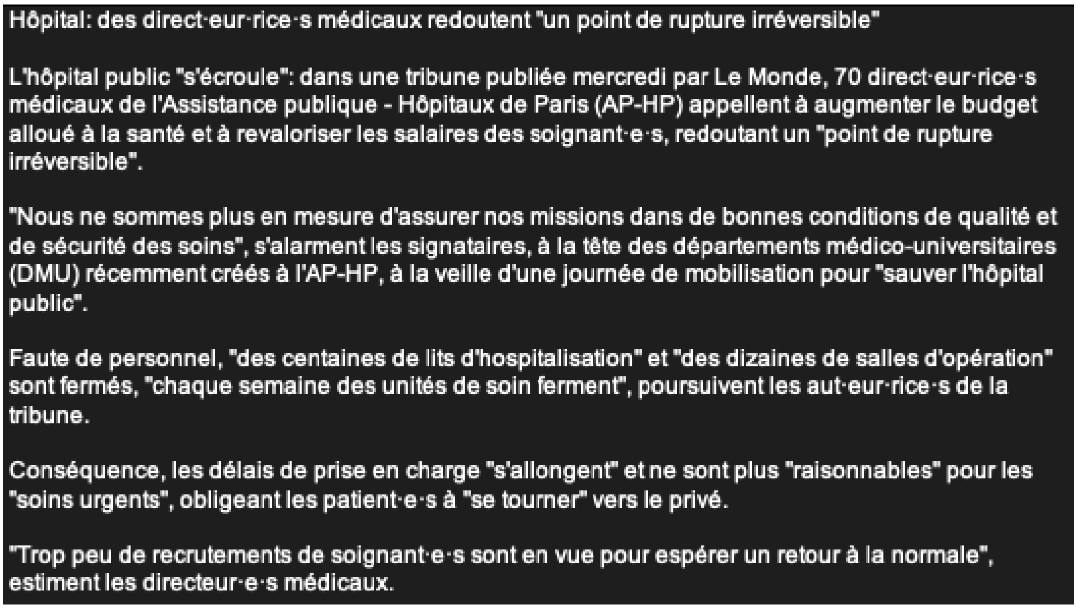
"Nous ne sommes plus en mesure d'assurer nos missions dans de bonnes conditions de qualité et de sécurité des soins", s'alarment les signataires, à la tête des départements médico-universitaires (DMU) récemment créés à l'AP-HP, à la veille d'une journée de mobilisation pour "sauver l'hôpital public".

Faute de personnel, "des centaines de lits d'hospitalisation" et "des dizaines de salles d'opération" sont fermés, "chaque semaine des unités de soin ferment", poursuivent les auteurs de la tribune.

Conséquence, les délais de prise en charge "s'allongent" et ne sont plus "raisonnables" pour les "soins urgents", obligeant les patients à "se tourner" vers le privé.

"Trop peu de recrutements de soignants sont en vue pour espérer un retour à la normale", estiment les directeurs médicaux.

Figure 4.1: A standard French text.



Hôpital: des direct·eur·rice·s médicaux redoutent "un point de rupture irréversible"

L'hôpital public "s'écroule": dans une tribune publiée mercredi par Le Monde, 70 direct·eur·rice·s médicaux de l'Assistance publique - Hôpitaux de Paris (AP-HP) appellent à augmenter le budget alloué à la santé et à revaloriser les salaires des soignant·e·s, redoutant un "point de rupture irréversible".

"Nous ne sommes plus en mesure d'assurer nos missions dans de bonnes conditions de qualité et de sécurité des soins", s'alarment les signataires, à la tête des départements médico-universitaires (DMU) récemment créés à l'AP-HP, à la veille d'une journée de mobilisation pour "sauver l'hôpital public".

Faute de personnel, "des centaines de lits d'hospitalisation" et "des dizaines de salles d'opération" sont fermés, "chaque semaine des unités de soin ferment", poursuivent les aut·eur·rice·s de la tribune.

Conséquence, les délais de prise en charge "s'allongent" et ne sont plus "raisonnables" pour les "soins urgents", obligeant les patient·e·s à "se tourner" vers le privé.

"Trop peu de recrutements de soignant·e·s sont en vue pour espérer un retour à la normale", estiment les directeur·e·s médicaux.

Figure 4.2: Epicene form of the standard French text in Figure 4.1.

4.2 Linguistic resources

In addition to the corpora, other linguistic resources have been used for the realization of this project.

In order to perform epicenizations and compare all morphological differences within the same lexeme (see Section 3.2.2), we used the latest version (3.1) of *Morphalou* [1], a freely available lexicon, created by ATILF[2]. It groups 159.271 lemmas and 976.570 inflected forms of French. This lexicon was obtained by grouping together other 5 lexicons: Morphalou 2, DELA, Dicollecte, LGLex et LGLexLefff and Lefff.

Morphalou 3.1 is available in different formats, but the most suitable one for our needs was the eXtensible Markup Language (XML) format, since the the XML tree structure is relatively easy to explore with the use of appropriate tools. We exploited, in particular, data which groups adjectives and common nouns. For each lexical entry, there are orthographic and morphological information, as well as all the inflected forms related. An example of an entry is shown in Figure 4.3.

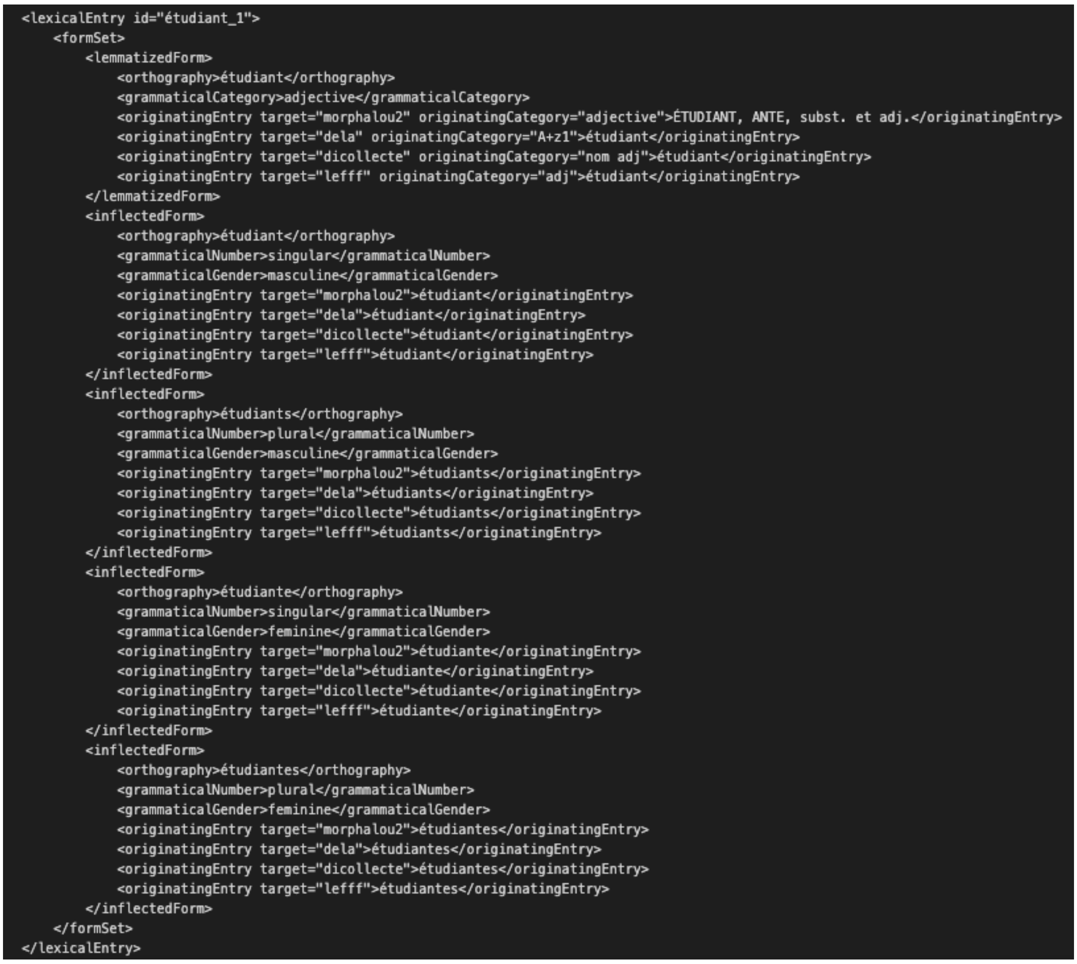
In order to retrieve which type of entity was referred to each noun (namely, which referent was associated to each noun) (see Section 3.2.3), another useful resource has been used: *WOrdnet Libre du Français* (WOLF)[3], a free semantic lexical resource for French, created by Sagot and Fišer. It is mainly based on Princeton WordNet (PWN)[4] and various multilingual resources [6]. WordNet is a lexical database for English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WOLF is essentially the French version of WordNet.

Thanks to the semantic relations between synonyms, hypernyms and hyponyms present inside the document, we were able to retrieve if a nouns was indeed referring to a human referent of not. WOLF was relatively easy to explore, thanks to its XML format.

[2] `https://hdl.handle.net/11403/morphalou/v3.1`

[3] `http://pauillac.inria.fr/~sagot/index.html#wolf`

[4] `https://wordnet.princeton.edu/`



```
<lexicalEntry id="étudiant_1">
    <formSet>
        <lemmatizedForm>
            <orthography>étudiant</orthography>
            <grammaticalCategory>adjective</grammaticalCategory>
            <originatingEntry target="morphalou2" originatingCategory="adjective">ÉTUDIANT, ANTE, subst. et adj.</originatingEntry>
            <originatingEntry target="dela" originatingCategory="A+z1">étudiant</originatingEntry>
            <originatingEntry target="dicollecte" originatingCategory="nom adj">étudiant</originatingEntry>
            <originatingEntry target="lefff" originatingCategory="adj">étudiant</originatingEntry>
        </lemmatizedForm>
        <inflectedForm>
            <orthography>étudiant</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
            <originatingEntry target="morphalou2">étudiant</originatingEntry>
            <originatingEntry target="dela">étudiant</originatingEntry>
            <originatingEntry target="dicollecte">étudiant</originatingEntry>
            <originatingEntry target="lefff">étudiant</originatingEntry>
        </inflectedForm>
        <inflectedForm>
            <orthography>étudiants</orthography>
            <grammaticalNumber>plural</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
            <originatingEntry target="morphalou2">étudiants</originatingEntry>
            <originatingEntry target="dela">étudiants</originatingEntry>
            <originatingEntry target="dicollecte">étudiants</originatingEntry>
            <originatingEntry target="lefff">étudiants</originatingEntry>
        </inflectedForm>
        <inflectedForm>
            <orthography>étudiante</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>feminine</grammaticalGender>
            <originatingEntry target="morphalou2">étudiante</originatingEntry>
            <originatingEntry target="dela">étudiante</originatingEntry>
            <originatingEntry target="dicollecte">étudiante</originatingEntry>
            <originatingEntry target="lefff">étudiante</originatingEntry>
        </inflectedForm>
        <inflectedForm>
            <orthography>étudiantes</orthography>
            <grammaticalNumber>plural</grammaticalNumber>
            <grammaticalGender>feminine</grammaticalGender>
            <originatingEntry target="morphalou2">étudiantes</originatingEntry>
            <originatingEntry target="dela">étudiantes</originatingEntry>
            <originatingEntry target="dicollecte">étudiantes</originatingEntry>
            <originatingEntry target="lefff">étudiantes</originatingEntry>
        </inflectedForm>
    </formSet>
</lexicalEntry>
```

Figure 4.3: Morphalou entry of the French word *”étudiant”*.

Chapter 5

Evaluation

For the evaluation we used corpora coming from different origins and sources, in order to have a wider vision over the language (see Section 4.1 for further details). A dual evaluation has been performed, one using f-measure and one using the external software Antidote. This Chapter will present the evaluation methods; the results will be shown in Chapter 6.

5.1 Manual evaluation of our corpora

For this first evaluation, we compared the output of our tool with the manually translated gold standard texts (expected output) to see if our tool over-epicenized or missed some epicenizations.

Common indexes have been used:

- Precision: % of selected items that are correct.
- Recall: % of correct items that are selected.

$$precision = \frac{pertinent\ epicenizations}{all\ epicenizations\ occured}$$

$$recall = \frac{pertinent\ epicenizations}{all\ pertinent\ epicenizations}$$

For the evaluation we combined both to calculate the f-measure, which is the harmonic mean of precision and recall:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

To calculate the f-measure, we focused on the words changed in the expected output and in our tool's output. A word which is changed in the expected output and in our tool's output is considered as true positive. A word which is changed in the expected output but not changed in our tool's output is considered as false negative. A word changed in our tool's output but not in the expected output is considered as false positive and, finally, a word which is not changed in our tool's output and not changed in the expected output is considered as true negative. True negatives are not taken into account by f-measure, rightfully within the scope of our project. In fact, the longer the text, the higher the precision would be, since the words to epicenize would be few with respect to the total amount.

For this purpose, the texts have been cleaned automatically using the script present in *clean.py*. Then, the files have been compared manually using the software BBedit[1], which provides an helpful tool to visually spot the differences between two files. The results were finally noted in an Excel file (see Figure 6.1).

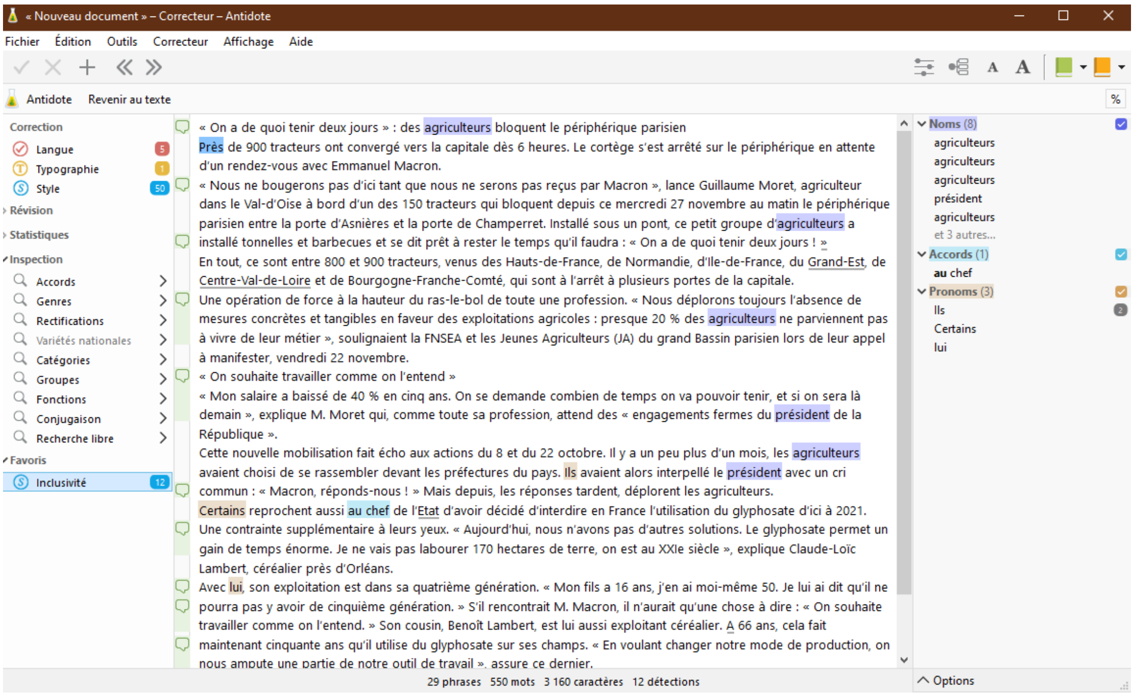
5.2 Evaluation using Antidote

The second evaluation has been performed thanks to Antidote software (see Figure 5.1 for a sample screen of this software). Antidote has a function that allows the user to see if the text is epicene or not. We used Antidote as reference to check how many words should have been changed, if our tool changed the non epicene words correctly, and to see which part of speech are affected.

This method allows us to have an idea in which specific part of speech our tool lacks of performance and have an automatic tool to help us. Results are shown in Table 6.1.

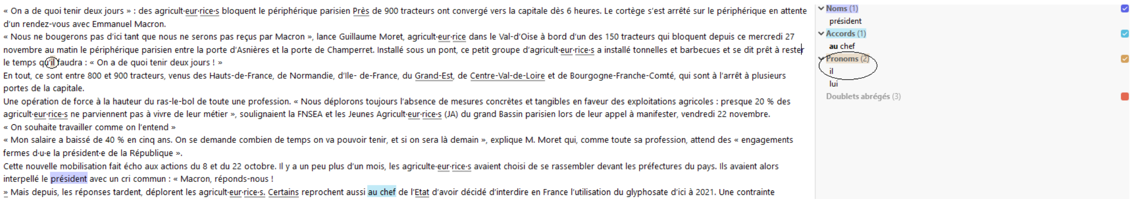
Eventually, we noticed that Antidote has some flaws, two in particular. It notes as non epicene some words that should not be epicenized, like the pronoun ”il” (see Figure 5.2) and sometimes counts epicenized words as mistakes (see Figure 5.3).

[1] `https://www.barebones.com/products/bbedit/`



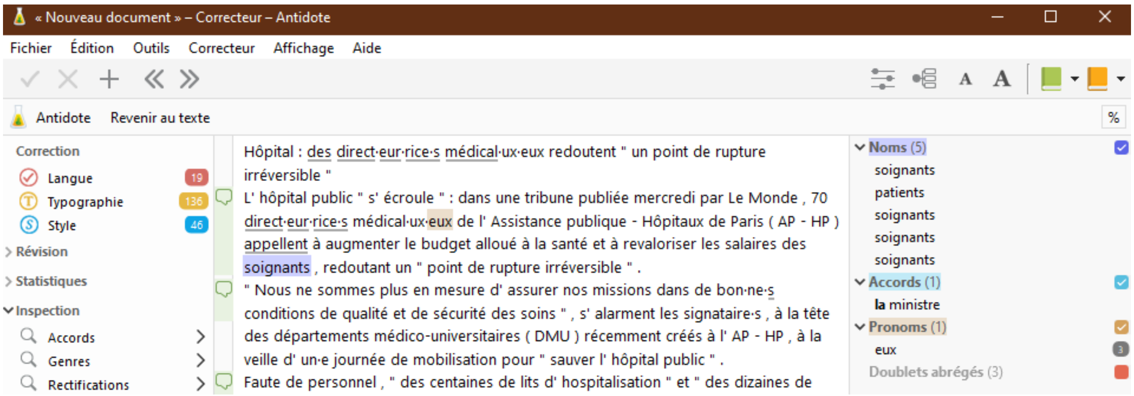
« On a de quoi tenir deux jours » : des agriculteurs bloquent le périphérique parisien
Près de 900 tracteurs ont convergé vers la capitale dès 6 heures. Le cortège s'est arrêté sur le périphérique en attente d'un rendez-vous avec Emmanuel Macron.
« Nous ne bougerons pas d'ici tant que nous ne serons pas reçus par Macron », lance Guillaume Moret, agriculteur dans le Val-d'Oise à bord d'un des 150 tracteurs qui bloquent depuis ce mercredi 27 novembre au matin le périphérique parisien entre la porte d'Asnières et la porte de Champerret. Installé sous un pont, ce petit groupe d'agriculteurs a installé tonnelles et barbecues et se dit prêt à rester le temps qu'il faudra : « On a de quoi tenir deux jours ! »
En tout, ce sont entre 800 et 900 tracteurs, venus des Hauts-de-France, de Normandie, d'Ile-de-France, du Grand-Est, de Centre-Val-de-Loire et de Bourgogne-Franche-Comté, qui sont à l'arrêt à plusieurs portes de la capitale.
Une opération de force à la hauteur du ras-le-bol de toute une profession. « Nous déplorons toujours l'absence de mesures concrètes et tangibles en faveur des exploitations agricoles : presque 20 % des agriculteurs ne parviennent pas à vivre de leur métier », soulignaient la FNSEA et les Jeunes Agriculteurs (JA) du grand Bassin parisien lors de leur appel à manifester, vendredi 22 novembre.
« On souhaite travailler comme on l'entend »
« Mon salaire a baissé de 40 % en cinq ans. On se demande combien de temps on va pouvoir tenir, et si on sera là demain », explique M. Moret qui, comme toute sa profession, attend des « engagements fermes du président de la République ».
Cette nouvelle mobilisation fait écho aux actions du 8 et du 22 octobre. Il y a un peu plus d'un mois, les agriculteurs avaient choisi de se rassembler devant les préfectures du pays. Ils avaient alors interpellé le président avec un cri commun : « Macron, réponds-nous ! » Mais depuis, les réponses tardent, déplorent les agriculteurs.
Certains reprochent aussi au chef de l'Etat d'avoir décidé d'interdire en France l'utilisation du glyphosate d'ici à 2021.
Une contrainte supplémentaire à leurs yeux. « Aujourd'hui, nous n'avons pas d'autres solutions. Le glyphosate permet un gain de temps énorme. Je ne vais pas labourer 170 hectares de terre, on est au XXIe siècle », explique Claude-Loïc Lambert, céréalier près d'Orléans.
Avec lui, son exploitation est dans sa quatrième génération. « Mon fils a 16 ans, j'en ai moi-même 50. Je lui ai dit qu'il ne pourra pas y avoir de cinquième génération. » S'il rencontrait M. Macron, il n'aurait qu'une chose à dire : « On souhaite travailler comme on l'entend. » Son cousin, Benoît Lambert, est lui aussi exploitant céréalier. A 66 ans, cela fait maintenant cinquante ans qu'il utilise du glyphosate sur ses champs. « En voulant changer notre mode de production, on nous ampute une partie de notre outil de travail », assure ce dernier.

Figure 5.1: Antidote shows the non epicene words sorted by part of speech.



« On a de quoi tenir deux jours » : des agricult·eur·rice·s bloquent le périphérique parisien Près de 900 tracteurs ont convergé vers la capitale dès 6 heures. Le cortège s'est arrêté sur le périphérique en attente d'un rendez-vous avec Emmanuel Macron.
« Nous ne bougerons pas d'ici tant que nous ne serons pas reçus par Macron », lance Guillaume Moret, agricult·eur·rice dans le Val-d'Oise à bord d'un des 150 tracteurs qui bloquent depuis ce mercredi 27 novembre au matin le périphérique parisien entre la porte d'Asnières et la porte de Champerret. Installé sous un pont, ce petit groupe d'agricult·eur·rice·s a installé tonnelles et barbecues et se dit prêt à rester le temps qu'il faudra : « On a de quoi tenir deux jours ! »
En tout, ce sont entre 800 et 900 tracteurs, venus des Hauts-de-France, de Normandie, d'Ile- de-France, du Grand-Est, de Centre-Val-de-Loire et de Bourgogne-Franche-Comté, qui sont à l'arrêt à plusieurs portes de la capitale.
Une opération de force à la hauteur du ras-le-bol de toute une profession. « Nous déplorons toujours l'absence de mesures concrètes et tangibles en faveur des exploitations agricoles : presque 20 % des agricult·eur·rice·s ne parviennent pas à vivre de leur métier », soulignaient la FNSEA et les Jeunes Agricult·eur·rice·s (JA) du grand Bassin parisien lors de leur appel à manifester, vendredi 22 novembre.
« On souhaite travailler comme on l'entend »
« Mon salaire a baissé de 40 % en cinq ans. On se demande combien de temps on va pouvoir tenir, et si on sera là demain », explique M. Moret qui, comme toute sa profession, attend des « engagements fermes d·u·e la président·e de la République ».
Cette nouvelle mobilisation fait écho aux actions du 8 et du 22 octobre. Il y a un peu plus d'un mois, les agriculte·eur·rice·s avaient choisi de se rassembler devant les préfectures du pays. Ils avaient alors interpellé le président avec un cri commun : « Macron, réponds-nous !
» Mais depuis, les réponses tardent, déplorent les agricult·eur·rice·s. Certains reprochent aussi au chef de l'Etat d'avoir décidé d'interdire en France l'utilisation du glyphosate d'ici à 2021. Une contrainte

Figure 5.2: Antidote shows the pronoun "il" as non epicene, but "il" does not refer to a human begin in this context.



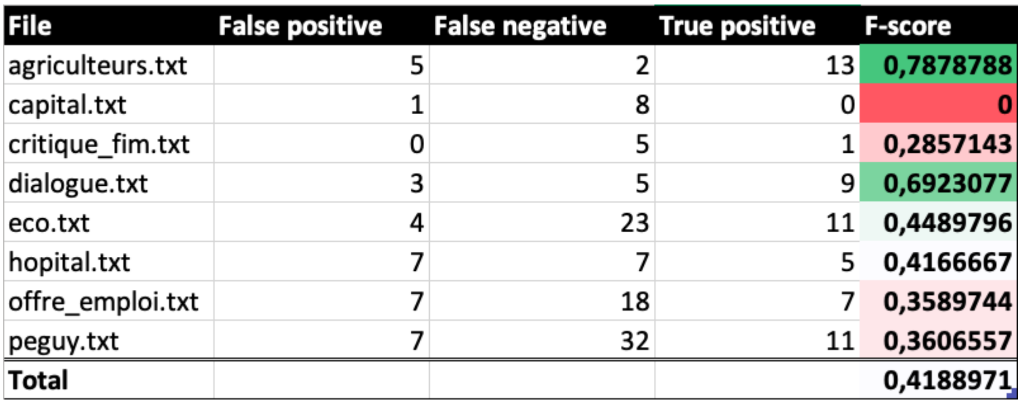
Hôpital : des direct·eur·rice·s médical·ux·eux redoutent " un point de rupture irréversible "
L' hôpital public " s' écroule " : dans une tribune publiée mercredi par Le Monde , 70 direct·eur·rice·s médical·ux·eux de l' Assistance publique - Hôpitaux de Paris (AP - HP) appellent à augmenter le budget alloué à la santé et à revaloriser les salaires des soignants , redoutant un " point de rupture irréversible " .
" Nous ne sommes plus en mesure d' assurer nos missions dans de bon·ne·s conditions de qualité et de sécurité des soins " , s' alarment les signataire·s , à la tête des départements médico-universitaires (DMU) récemment créés à l' AP - HP , à la veille d' un·e journée de mobilisation pour " sauver l' hôpital public " .
Faute de personnel , " des centaines de lits d' hospitalisation " et " des dizaines de

Figure 5.3: Antidote shows "eux" as a mistakes, but "eux" is part of the word "médical·ux·eux".

Chapter 6

Results

6.1 F-measure results



File	False positive	False negative	True positive	F-score
agriculteurs.txt	5	2	13	0,7878788
capital.txt	1	8	0	0
critique_fim.txt	0	5	1	0,2857143
dialogue.txt	3	5	9	0,6923077
eco.txt	4	23	11	0,4489796
hopital.txt	7	7	5	0,4166667
offre_emploi.txt	7	18	7	0,3589744
peguy.txt	7	32	11	0,3606557
Total				0,4188971

Figure 6.1: Type 1 and 2 errors, f-measure for each text.

As shown in Figure 6.1, the result is maximal in the file agriculteurs.txt, and minimal in capital.txt. The error that occurs the most is the false-negative, and there are usually few false-positives. It means that the main weakness of the program is that it is omitting some changes. The detection of non-epicene forms is not accurate. In short, the program is globally not really efficient, as the average f-measure is around 40%. However, the possibilities for improvement are really encouraging. Section 7.1 is an attempt of explaining these drawbacks.

6.2 Antidote results

The tables shown below are the results got from Antidote. We can see how many mistakes Antidote detected on each type of corpora and on which part of speech the mistake is detected. Standard French text is referred as *input*, manually epicenized text is referred as *expected output* and result from our tool is referred as *output*. These data are useful to improve the performance of our tool and have a global view of the actual performance by comparing the number of mistakes detected on expected output and our tool's output.

With regards to the Antidote-based evaluation, we can expect good result from the files agriculteurs.txt and dialogue.txt, because the number of mistakes shown in *expected output* and in *output* are similar. Conversely, we can expect bad result from capital.txt, hopital.txt and peguy.txt, where the number of mistakes in *expected output* and the number of mistakes in *output* are far from each other.

These hypothesis are confirmed by the f-measures (see Table 6.1), which show good results on agriculteurs.txt and dialogue.txt, with an f-measure > 0.69. The f-measure on capital.txt, hopital.txt and peguy.txt are bad as expected with f-measure < 0.36. The files critique de film.txt and eco.txt seem irrelevant in the Antidote evaluation because the number of mistakes in *input*, *expected output* and *output* are the same and we cannot predict the f-measure of these files; eco.txt has an average f-measure with 0.45, and critique de film.txt a low f-measure with 0.28.

Table 6.1: Antidote evaluation: number of non epicene words detected in each file.

agriculteurs	Noun	Agreement	Pronoun	Total
input	8	1	3	12
expected output	0	1	3	4
output of our tool	0	1	3	4

critique de film	Noun	Agreement	Pronoun	Total
input	3	0	0	3
expected output	3	0	0	3
output of our tool	3	0	0	3

eco	Noun	Agreement	Pronoun	Total
input	1	2	2	5
expected output	1	1	3	5
output of our tool	1	2	2	5

offre emploi	Noun	Agreement	Pronoun	Total
input	6	1	2	9
expected output	1	0	2	3
output of our tool	4	0	1	5

capital	Noun	Agreement	Pronoun	Total
input	6	0	2	8
expected output	1	0	1	2
output of our tool	6	0	1	7

dialogue	Noun	Agreement	Pronoun	Total
input	5	2	1	8
expected output	3	0	1	4
output of our tool	0	2	1	3

hôpital	Noun	Agreement	Pronoun	Total
input	7	1	1	9
expected output	1	0	1	2
output of our tool	5	1	1	7

peguy	Noun	Agreement	Pronoun	Total
input	18	7	1	26
expected output	0	0	1	1
output of our tool	11	3	1	15

Chapter 7

Discussion

7.1 Functional issues

7.1.1 Language model

In some cases, the language model ”fr_core_news_sm” is not accurate and makes some mistakes in the part-of-speech and syntactic relations tagging. The POS-tagging is theoretically 94.62% accurate, and the dependencies tagger 84.48% accurate[1].

A mistake in the POS-tagging can result in missing a noun needing for epicenization, or also in epicenizing it improperly. Similarly, a mistake in the dependencies tagging can result in omitting some dependencies of a head noun. In a nutshell, it leads to lower the coverage performance of the program.

7.1.2 Anaphora resolution

The search for links between anaphor and its antecedent is one of the most complicated topics in NLP. Anaphora resolution deals with both syntax, semantics and pragmatics, with the latter known as being very problematic in NLP. As far as this topic is concerned, our program is not sufficiently efficient, since relations between anaphors (pronouns in the first place, but not only) and antecedents are not handled.

[1] `https://spacy.io/models/fr`

7.1.3 Encoding error

Because of an encoding error in the Morphalou XML file (see Section 4.3), an error was originally raised when trying to load it trough the XML parser. We had to manually delete a specific lexical entry (*R&D*).

7.2 Performance issues

7.2.1 XML browsing

The program is really long. It lasted approximately five minutes to process all the files present in the *inputs* folder (see Figure 3.1). This is mainly due to the XML files browsing. The XML files (Morphalou in particular) are huge, and the program presents difficulties in finding a given lexical entry in a good amount of time.

Bibliography

[1] ATILF. Morphalou, 2019. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[2] G. G. Corbett. Number of genders. *World Atlas of Language Structures (WALS)*, 2005.

[3] C. G. Craig. *Noun classes and categorization: Proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983*, volume 7. John Benjamins Publishing Company, 1986.

[4] J. H. Greenberg. How does a language acquire gender markers. *Universals of human language*, 3:47–82, 1978.

[5] R. Jakobson. On linguistic aspects of translation. *On translation*, 3:30–39, 1959.

[6] B. Sagot and D. Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco, May 2008.