MSc Natural Language Processing - 2020/2021
UE 805 - Supervised Project

Corpus correction

Realization report

Sutdents :
Margot GUETTIER
Kahina MENZOU
Papa Amadou SY

Supervisor :
Bruno GUILLAUME
Reviewer :
Yannick PARMENTIER

June 2021

Contents

1 Introduction

The goal of this project is to detect errors in dependency annotated corpora. In the first part of our project, we analyzed the different methods allowing error detection. The method that we found most often is based on the principle of variation detection. That means that if we can find two similar occurrences but annotated in a different way, then we can assume that one of the two may be a potential error. This method was first used in error detection for part-of-speech annotations and then applied by Boyd et al [1] to dependency annotated corpora. For dependency annotations, each occurrence corresponds to a pair of words (word1, word2), and the relationship between them. For example word1 is the subject of word2. However, in their article it is not only the pairs of words that are taken into account, but also a context. In fact, if we do not take into account a context there would be too many results and the majority would be false positives. In the article they defined three types of context:

- The internal context corresponds to all the elements between word1 and word2

Figure 1: Internal context

- The neighboring or external context corresponds to four elements, it is the immediate context of word1 and the immediate context of word2

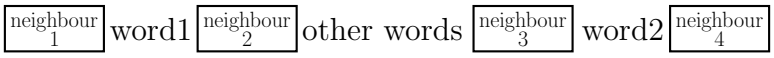


Figure 2: Neighbour context

- The context of dependency corresponds to the type of relationship that the couple's governor has with its own governor.

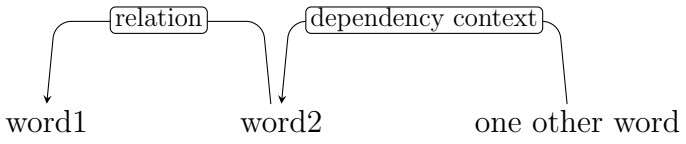


Figure 3: Dependency context

For this second part of our project, we will implement an error detection system based on the methodology of Boyd et al. We will start by collecting data in the form of a couple of words. After that, we will expose the step which concerns the collection of the different types of context followed by the phase of error detection. The results will be analyzed and filtered for each context to allow a better vision to the users. Finally, the last part is dedicated to the display of our results to allow the user to better visualize the potential errors detected.
In the third part of the report which concerns the experimentation, we will start with the presentation of the corpus we are going to work on, followed by the results obtained once the couples are extracted with each context. After that, a manual evaluation of our system will be done in order to see the performance of our system and allow us to plan future solutions to optimize it.

2 Methodology

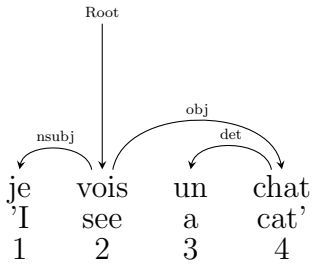
We said in the introduction the methodology we use is based on the variation detection, in this part we will explain the important step of this method. As the extraction step will depends on the format of the corpus, we will not talk about it in this part, however in the experimentation part we will explain it for the conllu format, which is the format of the corpus we use for the experimentation.

2.1 Collecting the couples

2.1.1 Couples with relation

After extracting the data, the next step is to establish a list of all the couples of words for which there exist a dependency relation. In order to do that for each word we have to look at its governor and the relation name.

(1) a.



b. (je, vois) ⇒ (1, 2, nsubj-R)
(vois, chat) ⇒ (2, 4, obj-L)
(un, chat) ⇒ (3, 4, det-R)

c. (je, un)⇒ (1, 3, NIL)
(je, chat) ⇒ (1, 4, NIL)
(vois, un) ⇒ (2, 3, NIL)

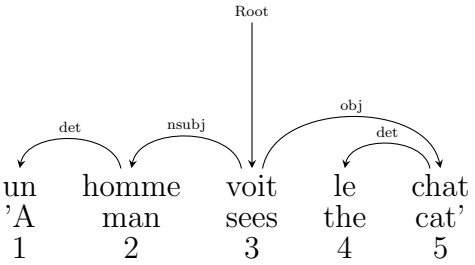
For the creation of the couple of pairs we must always keep the identifier of the sentence in order to be able to find it easily afterwards, but also indicate the position

of the words which interest us. Then for the name of the relation we followed the method of Boyd which consists in extending the label in order to specify which of the two words is the governor of the couple. For this we add -R and -L depending on whether the governor is on the right or on the left. For instance, in sentence 1 we havea couple (je, vois) , "je" is in position 1, "vois" is in position 2, there exist a relation nsubj between the two element, and -R indicates that the governor is on the right that means it is "vois"

2.1.2 NIL relation

Once we have listed those couples who have a dependency relation, we must then list the couples who have not. If two words are not in a relation we will say that they have a NIL relation. This type of relation is useful for our project because suppose that a relation between two words is missing for one occurrence but present everywhere else in the corpus, we want to be able to detect this lack. And conversely, if two words are never linked in the corpus except for one occurrence, we want to be able to detect this irregularity, in order to know if it is an error or simply a particular formulation. Note that a couple with a NIL relation is only relevant if there is another occurrence of this couple in the corpus but with a real dependency relation. If we go back to our example 1, there are three NIL couples 1c but none of them is relevant, so we ignore them. However, if we take into account a larger corpus the NIL relations could have an importance. For example, in the sentence 2, there is a NIL relation such as *(un, chat)⇒ (1, 5, NIL)*. If we only take this sentence, then this NIL relation is not relevant to analyze, but in a corpus that would include the sentence 1 and the sentence 2, the relation *(un, chat)* appears once with a *det-R* relation and once with a *NIL* relation, so it is a matter of checking if this is a real error.

(2)



2.2 Collecting the context

As explained in the introduction, this methodology implies to take into account a context. In our system, it is chosen by the person who want to check the annotations when launching the program.

The internal context is in the form of a tuple containing each element between the two words. For the neighboring context, it is a tuple of four elements, if the two words are contiguous then we replace the second and the third context element by 0, similarly if the word 1 is the first word of the sentence or the word 2 the last word of the sentence, then respectively, we will assign the value 0 to the first or last context element. For the dependency context it is a string with the name of the relation and the label *-R* or *-L* that we added when we created the couples.

(3) a. (je, vois) () ⇒ (1, 2, nsubj-R)
(vois, chat) (un) ⇒ (2, 4, obj-L)
(un, chat) () ⇒ (3, 4, det-R)

b. (je, vois) (0, 0, 0, un) ⇒ (1, 2, nsubj-R)
(vois, chat) (je, un, 0, 0) ⇒ (2, 4, obj-L)
(un, chat) (vois, 0, 0, 0) ⇒ (3, 4, det-R)

c. (je, vois) root ⇒ (1, 2, nsubj-R)
(vois, chat) root ⇒ (2, 4, obj-L)
(un, chat) obj-L ⇒ (3, 4, det-R)

Depending on the choice of the user, for the three couples of the sentence (1a), we have (3a) for the internal context, (3b) for the neighboring context and (3c) for the dependency context.

The list of couples with relation and the list of couples with NIL relation are organized in the same way. It is a dictionary with for key the two words of the relation as value a list of tuples such as (sentence identifier, position word1, position word2, name relation, context)
In this part, we used a simple example to present the different concepts that are useful for the rest of the project. However, variation detection only makes sense when a word pair appears more than once in the corpus and with at least two different annotations. If we only consider the sentence "I see the cat" this method is useless.

2.3 Detecting potential errors

Once we created all the couples with relation, the NIL couples and retrieve the context of interest, we have to determine which couples can have potential errors.
Here the user can choose if we want to take into account all the couples or only the one with actual dependency relation. However, when choosing the dependency context it is not relevant to take into account the NIL relation for the detection of potential errors. Indeed, in order to determine the context of dependency it is necessary to know the governor of the relation, but in a NIL relation the relation is non-existent so there is no defined governor.

For the detection of errors, it is necessary to verify that the couple appears at least twice in the corpus in the same context, and that among the occurrences there are at least two different dependency annotations.

The couple (sahara, occidental) (figure 4) with an neighboring context ('le', 0, 0, 'a') appears three times in the train part of the GSD (section 3.1), twice with *amod* - *L* annotation and one time with *flat:name* - *L* annotation. The sentences (4a) and (4b) correspond to two of the three sentences of figure(4) and the couple (sahara, occidental) is annotated differently. If we look the sentences there is no element

that can explain that difference. Therefore here we can assume there is a wrong annotation. We have two choices, either we consider "Sahara Occidental" as a single entity, a noun, and we use the label *flat:name* - *L*, or we consider that 'occidental' modify the noun 'Sahara' and in that case we use the label *amod* - *L*. Whatever we choose, the two sentences (and the third one in figure (4)) must have the same label for this relation.

```
(('sahara', 'occidental'), ['le', 0, 0, 'a']) {'amod - L': [('fr-ud-train_10680', 6, 7),
('fr-ud-train_12551', 5, 6)],'flat:name - L': [('fr-ud-train_04795', 31, 32)]}
```

Figure 4: example couple with potential error

(4) a. sent_id = 'fr-ud-train_12551'

Pour rappel, Le **Sahara Occidental** a été annexé par le Maroc en 1975 après le retrait de l'Espagne de ce territoire.

'As a reminder, Western Sahara was annexed by Morocco in 1975 after the withdrawal of Spain from the territory.'

b. sent_id = 'fr-ud-train_04795'

Quand le prix du phosphate a culminé à près de 500 dollars par tonne en 2008, la production de la mine de phosphate de Bou Craa au **Sahara Occidental** a été de près de 4 millions de tonnes.

'When the price of phosphate peaked at nearly $500 per ton in 2008, production at the Bou Craa phosphate mine in Western Sahara was nearly 4 million tons.'

2.4 Filter the results

By observing the couples of potential error and the sentences in which these couples appear we realized that in the same sentence a couple can appear several times. If for each occurrence in the sentence it is a couple with a dependency relationship,

we keep both occurrences. However when one of the couples is a NIL couple we have to check the overlapping of these couples. Because in the same sentence each word form can only have one governor. In this case, we will suppose that if the NIL relation is an actual error it would take place in the same direction as the relation of the occurrence with which we compare it. For example in a relation -R the governor is on the right so we have to check that the element on the left is different for the couple having a dependency relation and the couple NIL. Otherwise we will delete this NIL occurrence because it would be a false positive.

(5) a. Le **chat** marron clair **voit** la vache et le chien beige clair **voit** la poule
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
'The light brown cat sees the cow and the light beige dog sees the chicken'

b. (chat, voit) [le,marron,clair,la] ⇒ (mot2, mot5,nsubj-R)
(chat, voit) [le,marron,clair,la] ⇒ (mot2, mot13,NIL)

In the sentence 5a if the relation NIL were really an error it would mean that word2 would have 2 governors, word5 and word15. In the case of this sentence we can delete the occurrence with the NIL relation.

We created a filter to delete all the false positive, after applying this filter on each result we obtain, we notice that the filter have an effect only with the neighboring context (you can see the results in table 2 or when we don't take into account any context. For the dependency context it can be explain by the fact that NIL relation are not relevant, therefore there is no change between the results and the filtered results. For the internal relation it is actually impossible that for a same internal context, we obtain two couples with one identical element.

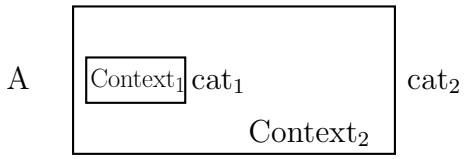


Figure 5: Internal context

For instance (Figure 5) if we imagine a sentence where there is twice the noun *cat*, and once the determiner *a*, and additional other words to make a real sentence. It is not possible to obtain the exact same internal context. There will be at least one additional word. Here context_2 contains the word cat_1, therefore context_2 is different from context_1. Here is the reason for which the false positive filter we use has no impact on couples with internal context.

2.5 Display

For the display, we used two graphical interfaces. A first one (Figure 6) with python which allows to select the parameters of the error detection. To make this interface, we used the gooey API of python. It is a very practical API that allows to make simple graphical interfaces that meet the needs and do not require large resources. The input field takes the conllu file returned by the first part of the algorithm. In the four remaining fields we have drop-down lists that allow to choose respectively the context (internal, neighbor, dependency, none), the NIL relation (NIL, Not NIL), the punctuation (punct, not punct), and finally the representation of the words (lemma, wordform). Figure 6 is the rendering of the graphical interface.

Just after this step, the algorithm generates, according to the chosen parameters, a list of word (Figure 7) pairs which appear in two or more sentences and which are annotated differently with the sentences in question. To ensure a more pleasant rendering, we opted for a display in the form of an HTML page. To do this, we used Kirian Guiller's API [1] available on his github account. This tool allows to draw dependency trees based on a conllu file. To avoid any ambiguity, we highlighted the annotations that could be wrong by coloring them in red. The purpose of this maneuver is to make it easier for the proofreader. To go back to the page where the word pairs are, we have put a "back" button. In Figure 7 we have a preview.

[1] https://github.com/kirianguiller/reactive-dep-tree

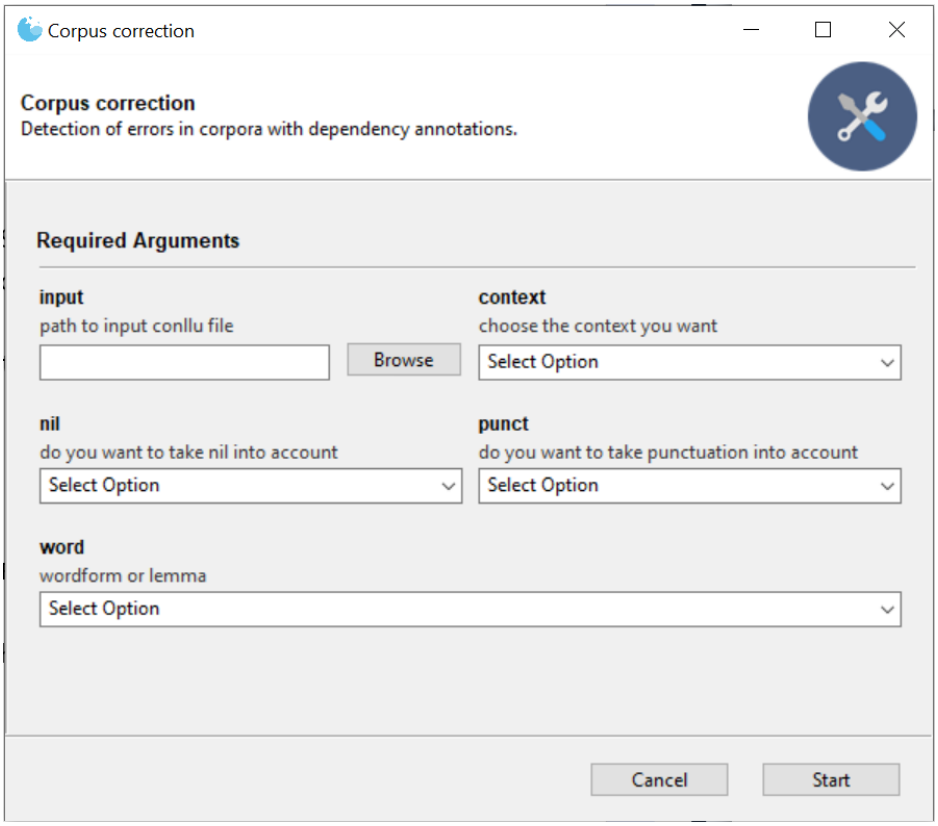


Figure 6: Interface for selection of error detection parameters

In cases where the wrong annotation concerns several sentences, we have provided a link "Others sentences with this relation" to see the other sentences that contain exactly the word pair that is likely to be wrongly annotated. Figure 8 is an overview of this page.

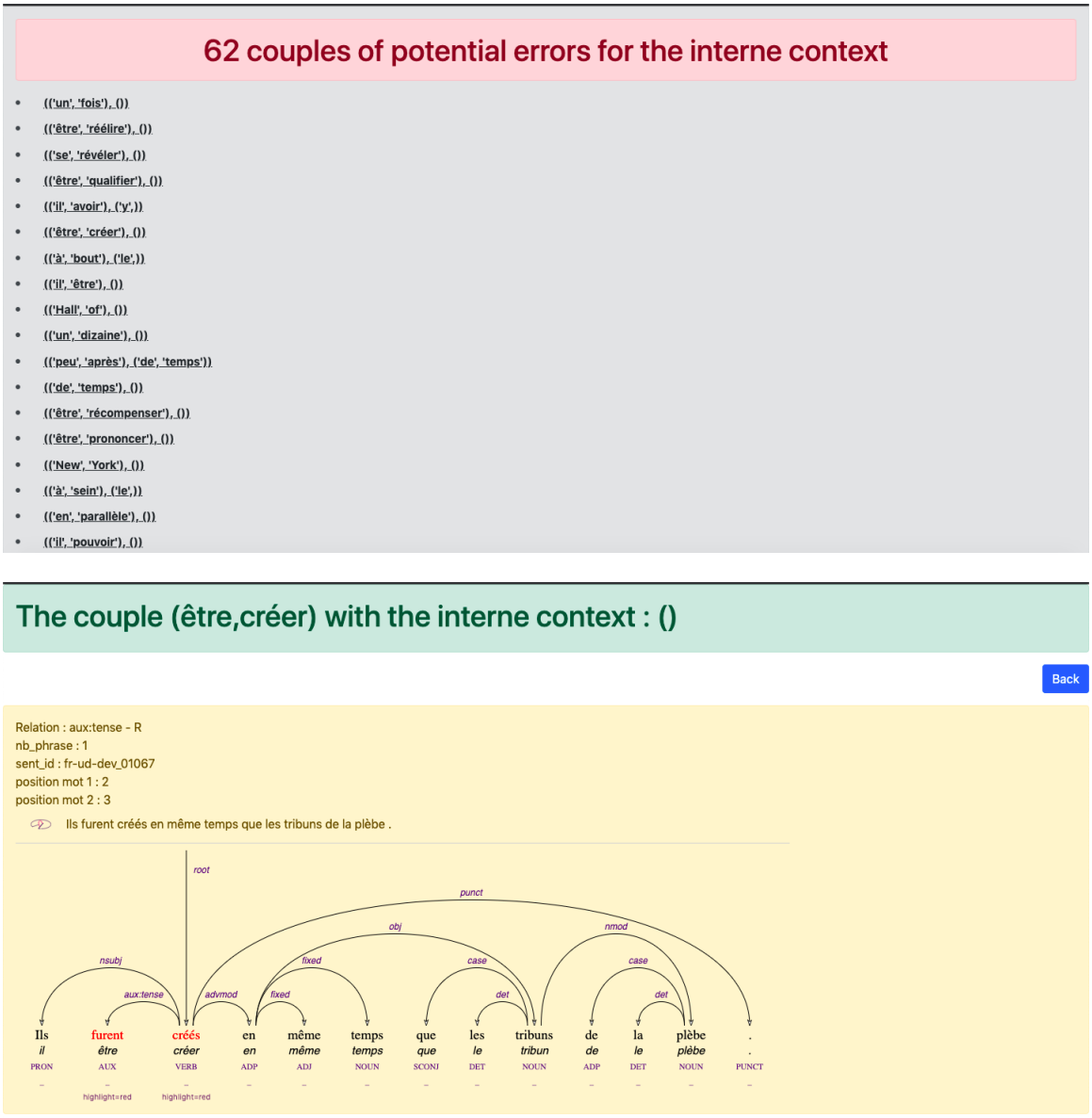


Figure 7: Home page and a sample of couple

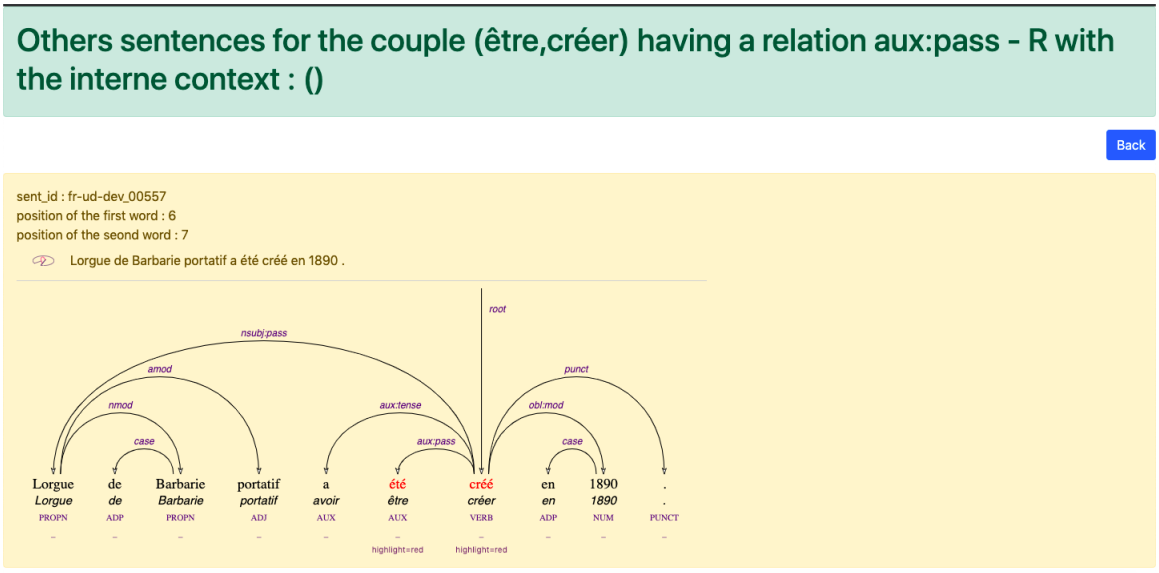


Figure 8: In cases where we have several sentences

3 Experimentation

3.1 Corpus presentation

For this second part of the project we decide to use the version 2.7 of the GSD corpus of French. We download it on Universal dependencies website[2]. This corpus is in conllu format (figure 9) and divide in three files (development, train, test). For the experimentation part we use the train part of the corpus, it contains 14449 sentences and 354662 tokens. In conllu format, for each sentences there is two metadata lines, the identifier of the sentence and the sentence itself. After, for each word of the sentence there is a line with 10 columns that give us morphological and grammatical information about the word but also the dependency relationship that this word has with the other words of the sentence.

[2]https://universaldependencies.org/treebanks/fr_gsd/

```
# sent_id = fr-ud-dev_00001
# text = Aviator, un film sur la vie de Hughes.
1   Aviator Aviator PROPN   _   _   0   root    _   SpaceAfter=No
2   ,   ,   PUNCT   _   _   4   punct   _   _
3   un  un  DET _   Definite=Ind|Gender=Masc|Number=Sing|PronType=Art   4   det _   _
4   film    film    NOUN    _   Gender=Masc|Number=Sing 1   appos   _   _
5   sur sur ADP _   _   7   case    _   _
6   la  le  DET _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art    7   det _   _
7   vie vie NOUN    _   Gender=Fem|Number=Sing  4   nmod    _   _
8   de  de  ADP _   _   9   case    _   _
9   Hughes  Hughes  PROPN   _   _   7   nmod    _   SpaceAfter=No
10  .   .   PUNCT   _   _   1   punct   _   _
```

Figure 9: format conllu

For our project, we will only be interested in six of these columns: the first three, which give information about the position, the wordform and the lemma. The fourth one to have the part of speech. Finally the seventh and the eighth give us the position of the governor and the relation between this governor and the word. In order to use these data, we have organized them in a dictionary. The key is the identifier of the sentence and the value is a list of list containing the elements of the six columns for each word of the sentence (Figure 10).

```
fr-ud-dev_00001 : [['1', 'aviator', 'Aviator', 'PROPN', '0', 'root'],
['2', ',', ',', 'PUNCT', '4', 'punct'], ['3', 'un', 'un', 'DET', '4', 'det'],
['4', 'film', 'film', 'NOUN', '1', 'appos'], ['5', 'sur', 'sur', 'ADP', '7', 'case'],
['6', 'la', 'le', 'DET', '7', 'det'], ['7', 'vie', 'vie', 'NOUN', '4', 'nmod'],
['8', 'de', 'de', 'ADP', '9', 'case'], ['9', 'hughes', 'Hughes', 'PROPN', '7', 'nmod'],
['10', '.', '.', 'PUNCT', '1', 'punct']]
```

Figure 10: Dictionary entry corresponding to the figure 1 sentence

3.2 Result

From the dictionary containing all the sentences of the train part of the GSD, we can apply the methodology. First we retrieve all the couples(section 2.1) and the context (section2.2) and then we detect the potential errors.

	no context		Internal		Neighboring		Dependency	
	Wordform	Lemma	Wordform	Lemma	Wordform	Lemma	Wordform	Lemma
NIL/Punct	70549	78267	4076	4553	1070	1392	1050	2070
NIL/ not_Punct	55894	65508	2594	2836	910	1140	1050	2070
not_NIL/not_Punct	2704	4430	803	902	105	122	1050	2070

Table 1: Number of potential errors for each context

	neighbour	
	Wordform	Lemma
NIL/Punct	735	992
NIL/ not_Punct	628	807
not_NIL/not_Punct	105	122

Table 2: Number of potential errors for neighboring context after applying the filter

The summary table (Table 1) shows the results obtained with or without context, and according to different criteria. With respect to these results, we can see the importance of NILs in both internal and neighboring context. We decided finally to remove the couples with punctuation and keep those with NILs. We also choose to do the error detection on wordform and lemma in order to determine which method gives us better results For the evaluation of our system for the neighboring context we take into account the results obtained after the application of the filter (Table 2).

Those results were compared with those obtained with the tool Errator[3], created by Guillaume Wisniewski, and also based on the principle of variation detection. It is

[3]https://perso.limsi.fr/wisniews/errator/

a tool which allows the detection of errors in dependency annotated corpora. This tool uses the concept of pair of words but also the concept of surrounding context. The comparison was not obvious at first because on the one hand Errator detects 4919 errors however it takes punctuation into account which we do not, and on the other hand our system works with pairs of 2 words whereas Errator compares the largest common character chain. We looked only for our pairs in the results proposed by Errator, those we found correspond either to the internal context or to the dependency context. To find out if we have more or less the same number of results, we have started the error detection by taking into account the NILs and the punctuation on the internal context and we obtained 4076 couples.

3.2.1 Analysis

For the internal context with NIL and without punctuation, for 70% of the detected pairs the internal context turns out to be of zero length which means that in most cases for this context we are interested in contiguous words.Then, for 22% of the couples the internal context contains only 1 element As said before, in order to be considered as a potential error, a couple must have at least 2 occurrences with two different annotations. For the internal context, 95.5% of the couples oppose only 2 relations, 4.16% oppose 3 relations and 0.35% of the couples oppose 4 relations.

In the same situation but observing the lemmas instead of the wordforms, 65.8% of the pairs have an internal context of zero length, and 25.45% have a length of one. Concerning the number of relations put in opposition, for 95.35% there are 2 relations, for 4.05% there are 3 relations and 0.42% oppose 4 relations.

As we can see, whether it is with the lemmas or with the word forms, for the internal context, in most cases we are interested in contiguous words, or very close words. As the neighboring context takes into account a part of the internal context we could expect that the two middle elements of the neighboring context are both *0* such that [neighbour 1, 0,0, neighbour 2] or at least one middle element is *0* such that [neighbour 1, 0, neighbour 2, neighbour 3]. However in the situation with NIL

and without punctuation, in 78.9% of the couples there are four true neighbours in the neighboring context, in 17.14% of the couples the two middle element are *0* (that means null internal context), and 3.95% of the couples have only one true middle neighbour(that means one neighbour and one*0*) between the two words which corresponds to an internal context of length 1.

This shows us that the internal context will be more useful for error detection for relations between two close or contiguous words, while the neighboring context can handle relations with a longer distance. Indeed, the neighboring context is both more restrictive because it imposes more context elements, which is why the number of results is lower, but it allows a certain freedom concerning the spacing of the words.

3.3 Evaluation

The first phase of our experimentation concerns the manual evaluation of the system. The purpose of this manual evaluation is to see the performance of our system on detecting dependency errors and to analyze the types of errors that occur the most. This analysis will allow us to make changes to better optimize the program and have better results thereafter. To perform the evaluation, first we extracted the potential errors from the *fr_gsd-ud-train.conllu* file. We obtained 5 files:

- 2 files for the neighboring context, one with NIL and the other without NIL.
- 2 files for the internal context, one with NIL and the other without NIL.
- 1 file for the dependency context

For each of these files we have chosen randomly 20 couples, for each couple we have compared the different annotations and assigned a label to this couple. The labels are defined as follows:

- (+) If we consider that the annotations of the dependency tree are wrong.
- (-) If we consider that the annotations of the dependency tree are correct.

- (?) If we have doubts about the annotations of the dependency tree.
  we obtained those results (Tables (3) , (4) and (5))

The evaluation method is presented in the form of pairs of words of all errors obtained by our system and that for each context.
Each pair of words will be accompanied by two sentences which show why we decided to annotate the error like that.

3.3.1 Evaluation of neighboring context

	+	-	?
Neighboring context without NIL	75%	10%	15%
Neighboring context with NIL	50%	50%	0%

Table 3: Neighboring context

In this first example of neighboring context Without NIL (Table (3)) the couple: (('il', 'a'), [0, 'y', 0, 'deux']) (Figure 11) appears twice with different annotations but those annotations are justified this is why we attribute (-) to it. For the first sentence 6a the dependency is annotated as *fixed - L* and for the second 6b it is annotated as *expl:subj - R*.

(6) a. sent_id = 'fr-ud-train_12908'

Il y **a** deux ou trois siècles, le Jaunay était un petit fleuve qui se jetait directement en mer, au droit de son parcours terrestre, vers l'endroit de la côte connu sous le nom de Roche-Biron, au nord de Brétignolles-sur-Mer.

'Two or three centuries ago, the Jaunay was a small river that flowed directly into the sea, at the right of its land course, towards the place on the coast known as Roche-Biron, north of Brétignolles-sur-Mer'.

b. sent_id = 'fr-ud-train_01426'

Il y **a** deux types de névroglie qui contiennent chacune des types cellulaires différents.

'There are two types of neuroglia, each of which contains different cell types.'

The first dependency is *fixed -L* (Figure 11) because the relation between the couple and the context (meaning of the sentence) is about a notion of time *two or three centuries ago*. The second dependency is annotated as *expl:subj - R* (Figure 11) because parts of speech are connected with expl:subj VERB-PRON ("il" , "a") in the sense that *it exists two types of neuroglia*.

```
1   Il  il  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Prs   7   case    _   ExtPos=ADP|PhraseType=Idiom|wordform=il
2   y   y   PRON    _   Person=3|PronType=Prs   1   fixed   _   _
3   a   avoir   VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   1   fixed   _   _
4   deux    deux    NUM _   Number=Plur 7   nummod  _   _
```

```
1   Il  il  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Prs   3   expl:subj   _   wordform=il
2   y   y   PRON    _   Person=3|PronType=Prs   3   dep:comp    _   _
3   a   avoir   VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
4   deux    deux    NUM _   Number=Plur 5   nummod  _   _
```

Figure 11: Dependencies of the 2 sentences without NIL with annotation (-)

For the second example of neighboring context Without NIL the couple (("s'", 'étend'), ['elle', 0, 0, 'sur']) (Figure 12) we have two sentences where the couple appeared with two different annotations and we consider that one of the two annotations is wrong (+):

(7) a. sent_id = 'fr-ud-train_00025'

Située dans la partie nord de la côte est de cette île de l'océan Indien, elle **s'étend** sur soixante kilomètres de long et s'étire sur trente de large jusqu'à la péninsule de Masoala, qui délimite ses rivages orientaux.

'Located in the northern part of the east coast of this island in the Indian Ocean, it stretches sixty kilometers long and stretches thirty wide to the Masoala peninsula, which defines its eastern shores.'

b. sent_id = 'fr-ud-train_01426'

Elle **s'étend** sur 105,3 km^2 et comptait 5633 habitants en 2010.

'It covers 105.3 km² and had 5,633 inhabitants in 2010.'

For the first sentence the dependency is annotated as *expl:pass - R* and for the second it is annotated as *dep:comp - R*,

```
18  elle    il  PRON    _   Gender=Fem|Number=Sing|Person=3|PronType=Prs    20  nsubj:pass  _   _
19  s'  se  PRON    _   Person=3|PronType=Prs   20  expl:pass    _   SpaceAfter=No
20  étend   étendre VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
21  sur sur ADP _   _   23  case    _   _
```

```
1   Elle    il  PRON    _   Gender=Fem|Number=Sing|Person=3|PronType=Prs    3   nsubj   _   wordform=elle
2   s'  se  PRON    _   Person=3|PronType=Prs   3   dep:comp    _   SpaceAfter=No
3   étend   étendre VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
4   sur sur ADP _   _   6   case    _   _
```

Figure 12: Dependencies of the 2 sentences without NIL with annotation (+)

Finally, for the third case of neighboring context whithout NIL (?) we have two sentences with the annotation: (('une', 'dizaine'), ['d', 0, 0, 'de']) (Figure 13) where one time is annotated as *nummod - R* and another time as *det - R*:

(8) a. sent_id = 'fr-ud-train_10848'

La préparation du projet commence dans le courant des années 1980, mais sa concrétisation prend plus d'**une dizaine** d'années.

'The preparation of the project began in the course of the 1980s, but its realization took more than dozen of years.'

b. sent_id = 'fr-ud-train_07365'

L'archéologie a, en effet, démontré le maintien d'une présence grecque à Alalia jusqu'à la prise de possession par Rome en 259, et une courte occupation punique d'**une dizaine** d' années à l'extrême fin de la période.

'Archeology has, in fact, demonstrated the maintenance of a Greek presence in Alalia until the taking of possession by Rome in 259, and a short Punic occupation of dozen of years at the end of the period.'

```
20  d'  de  ADP _   _   22  case    _   SpaceAfter=No
21  une un  NUM _   Gender=Fem|Number=Sing  22  nummod  _   _
22  dizaine dizaine NOUN    _   Gender=Fem|Number=Sing  19  dep:comp    _   _
23  d'  de  ADP _   _   24  case    _   SpaceAfter=No
```

```
33  d'  de  ADP _   _   35  case    _   SpaceAfter=No
34  une un  DET _   Definite=Ind|Gender=Fem|Number=Sing|PronType=Art    35  det _   _
35  dizaine dizaine NOUN    _   Gender=Fem|Number=Sing  31  nmod    _   _
36  d'  de  ADP _   _   37  case    _   SpaceAfter=No
```

Figure 13: Dependencies of the 2 sentences without NIL with annotation (?)

Concerning the neighboring context with NIL, the results are presented as follows (Table 3):

In this first example, for the couple (("n'", 'est'), ['ce', 0, 0, 'pas']) (Figure 14) appears 23 times in the file, 22 times as *NIL* and one time as *advmod*. We choose one example for each relation as follow:

(9) a. sent_id = 'fr-ud-train_00908'

Ce **n'est** pas du tout un bar, mais bel et bien un restaurant haut de gamme.

'It is not a bar at all, but indeed an upscale restaurant.'

b. sent_id = 'fr-ud-train_11770'

Ce **n'est** pas à la France et aux États-Unis d'imposer un président à la Côte d'Ivoire.

'It is not for France and the United States to impose a president on Côte d'Ivoire.'

We consider in this case that the annotations are correct (-) because the verb to be is not equal the auxiliary to be.

```
1   Ce  ce  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Dem   9   nsubj   _   wordform=ce
2   n'  ne  ADV _   Polarity=Neg    9   advmod  _   SpaceAfter=No
3   est être    AUX _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   9   cop _   _
4   pas pas ADV _   Polarity=Neg    9   advmod  _   _
```

```
2   Ce  ce  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Dem   4   expl:subj   _   wordform=ce
3   n'  ne  ADV _   Polarity=Neg    4   advmod  _   SpaceAfter=No
4   est être    VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
5   pas pas ADV _   Polarity=Neg    4   advmod  _   _
```

Figure 14: Dependencies of the 2 sentences with NIL with annotation (-)

For the case of neighboring context with NIL and where the annotation is wrong (+) we have for the couple (('est', 'plus'), ["n'", 0, 0, 'en']) (Figure 15) which appears twice in the file two different annotations. One sentences where we have the annotation *NIL* and a second sentence with the annotation *Advmod* in as follow:

(10) a. sent_id = 'fr-ud-train_00619'

Il n'**est plus** en activité.

'He is no longer active'

b. sent_id = 'fr-ud-train_00003'

Le comportement de la Turquie vis-à-vis du problème palestinien a fait qu'elle n'**est plus** en odeur de sainteté auprès de la communauté juive en générale, et américaine en particulier.

'Turkey's behavior vis-à-vis the Palestinian problem has meant that it no longer smells of holiness with the Jewish community in general, and the American community in particular.'

```
1   Il  il  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Prs   6   nsubj   _   wordform=il
2   n'  ne  ADV _   Polarity=Neg    6   advmod  _   SpaceAfter=No
3   est être    AUX _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   6   cop _   _
4   plus    plus    ADV _   Polarity=Neg    6   advmod  _   _
```

```
14  elle    il  PRON    _   Gender=Fem|Number=Sing|Person=3|PronType=Prs    16  nsubj   _   _
15  n'  ne  ADV _   Polarity=Neg    16  advmod  _   SpaceAfter=No
16  est être    VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   12  ccomp:obj   _   _
17  plus    plus    ADV _   _   16  advmod  _   _
```

Figure 15: Dependencies of the 2 sentences with NIL with annotation (+)

3.3.2 Evaluation of internal context

The second part of the evaluation concerns the internal context. The 20 couples that we chose randomly from the *fr_gsd-ud-train.conllu* file gived this result in (Table 4):

	+	-	?
Internal context without NIL	65%	25%	15%
Internal context with NIL	30%	70%	0%

Table 4: Internal context

For the Internal context without NIL, the couple (('en', 'commun'), []) (Figure 16) appears eight times with two different annotations, one time as *case* - *R* and seven times as *fixed* but those annotations are justified (-) because

(11) a. sent_id = 'fr-ud-train_06152'

Elle a, **en commun** avec d'autres Églises, un Service protestant de mission (le Défap), qui entretient le lien de solidarité et de mission commune avec d'autres Églises, notamment l'Église réformée de France et des Églises réformées africaines et océaniennes, principalement issues de l'ancienne Société des missions évangéliques de Paris.

'It has in common with other Churches, a Protestant Mission Service (Défap)...'

b. sent_id = 'fr-ud-train_01710'

Réseau Mistral est le service de transport **en commun** desservant les villes de Carqueiranne, Hyères, La Crau, La Garde, La Seyne-sur-Mer, La Valette-du-Var, Le Pradet, Le Revest-les-Eaux, Ollioules, Saint-Mandrier-sur-Mer, Six-Fours-les-Plages et Toulon, le chef-lieu départemental.

'Mistral network is the public transport service serving the towns of Carqueiranne...'

```
4   en   en   ADP  _   _   5   case      _   _
5   commun   commun   NOUN      _   Gender=Masc|Number=Sing 2   obl:mod _   _
```

```
8   en   en   ADP  _   _   7   advmod   _   ExtPos=ADV|PhraseType=Idiom
9   commun   commun   ADJ  _   Gender=Masc|Number=Sing 8   fixed   _   _
```

Figure 16: Dependencies of the 2 sentences without NIL with annotation (-)

The annotations are correct because the meaning of *elle a en commun avec les autres* and *transport en commun* is different.

Concerning the case where there is a real error (+) in the file, we have the example of the couple (('né', 'famille'), ['dans', 'une']) (Figure 17) where the annotation *obl:arg - L* appears one time and the annotation *obl:mod* appears four times we consider that one of the annotations is wrong because they are inconsistent annotations.

(12) a. sent_id = 'fr-ud-train_09430'

Né dans une **famille** de la bourgeoisie Foyalaise, sa mère, Renée Daniel, est bijoutière et son père, Louis Appoline-Darsières, est inspecteur des contributions à Fort-de-France, on lui a souvent reproché ses origines, précisément parce qu'il était l'avocat des plus humbles.

'Born in a family of the Foyalaise bourgeoisie...'

b. sent_id = 'fr-ud-train_00362'

Il est **né** dans une **famille** d'immigrants portugais qui retourne dans son pays d'origine en 1923.

'He was born into a Portuguese immigrant family...'

```
1  Né  naître  VERB   _  Gender=Masc|Number=Sing|Tense=Past|Typo=Yes|VerbForm=Part  17  advcl  _  CorrectGender=Fem|wordform=née
2  dans  dans  ADP _  _  4  case  _  _
3  une un  DET _  Definite=Ind|Gender=Fem|Number=Sing|PronType=Art  4  det _  |
4  famille famille NOUN  _  Gender=Fem|Number=Sing  1  obl:arg _  _
```

```
3  né  naître  VERB  _  Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part  0  root  _  _
4  dans  dans  ADP _  _  6  case  _  _
5  une un  DET _  Definite=Ind|Gender=Fem|Number=Sing|PronType=Art  6  det _  _
6  famille famille NOUN  _  Gender=Fem|Number=Sing  3  obl:mod _  _
```

Figure 17: Dependencies of the 2 sentences without NIL with annotation (+)

Finally, for cases where we have doubts, we have for example the couple: (('les', 'anciens'), ['plus']) (Figure 18) where the annotation *det - R* appears twice and the annotation *advmod - R* appears six times.

(13) a. sent_id = 'fr-ud-train_00854'

La Chronique de Peterborough passe du vieil anglais livresque classique au moyen anglais primitif après 1131, fournissant des textes comptant parmi **les** plus **anciens** connus en moyen anglais.

'....providing some of the oldest texts known in Middle English.'

b. sent_id = 'fr-ud-train_00356'

Certains hôpitaux ont été invités à faire de même (série H-dépôt) pour leurs documents **les** plus **anciens**.

'Some hospitals have been asked to do the same (H-depot series) for their oldest documents.'

```
27  les le  DET _   Definite=Def|Number=Plur|PronType=Art   29  det _   _
28  plus    plus    ADV _   _   29  advmod  _   _
29  anciens ancien  ADJ _   Gender=Masc|Number=Plur 25  obl:arg _   _
```

```
17  les le  DET _   Definite=Def|Number=Plur|PronType=Art   19  advmod  _   ExtPos=ADV|PhraseType=Idiom
18  plus    plus    ADV _   _   17  fixed   _   _
19  anciens ancien  ADJ _   Gender=Masc|Number=Plur 16  amod    _   SpaceAfter=No
```

Figure 18: Dependencies of the 2 sentences without NIL with doubts

For the Internal context with NIL the couple (Table 4) the couple ('les', 'premiers'), ['trois']) (Figure 14) appears four times with two different annotations, twice as *NIL* and twice as *det - R* but those annotations are justified (-) because we have for the one case *les trois premiers* + NOUN and in the other case *les trois premiers* + VERB:

(14) a. sent_id = 'fr-ud-train_07609'

À Arles, des objets provenant de différents sondages, en particulier du site de l'hôpital Van-Gogh attestent l'existence d'une occupation sur cet îlot rocheux dès la fin du VIIe siècle et durant **les** trois **premiers** quarts du VIe siècle av. J.-C.

'...during the first three quarters of the 6th century BC. J.-C.'

b. sent_id = 'fr-ud-train_13659'

Les trois **premiers** sont des tissus parenchymateux, en opposition au tissu conjonctif (ou stroma).

'The first three are parenchymal tissue...'

```
39  les le  DET _   Definite=Def|Number=Plur|PronType=Art   42  det _   _
40  trois   trois   NUM _   Number=Plur 42  nummod  _   _
41  premiers    premier ADJ _   Gender=Masc|Number=Plur 42  amod    _   _
```

```
1   Les le  DET _   Definite=Def|Number=Plur|PronType=Art   3   det _   wordform=les
2   trois   trois   NUM _   Number=Plur 3   nummod  _   _
3   premiers    premier ADJ _   Gender=Masc|Number=Plur 6   nsubj   _   _
```

Figure 19: Dependencies of the 2 sentences with NIL with annotation (-)

Concerning the case where we have an errors in annotations we have for example the couple:(('compositeur', 'groupe'), ['de', 'le']) where the annotation *NIL* appears one time and the annotation *nmod - L* also appears one time.

(15) a. sent_id = 'fr-ud-train_04380'

Finn Andrews (né à Brixton, Londres, le 24 août 1983) est le chanteur/ **compositeur** du **groupe** The Veils basé à Londres.

b. sent_id = 'fr-ud-train_00365'

DeStijl est un groupe de rock français, chantant en anglais, à formation variable s'articulant autour du créateur et principal **compositeur** du **groupe** P. DeStijl.

```
19  compositeur compositeur NOUN    _   Gender=Masc|Number=Sing 17  conj    _   _
20-21   du  _   _   _   _   _   _   _   _
20  de  de  ADP _   _   22  case    _   _
21  le  le  DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   22  det _   _
22  groupe  groupe  NOUN    _   Gender=Masc|Number=Sing 17  nmod    _   _
```

```
24  compositeur compositeur NOUN    _   Gender=Masc|Number=Sing 21  conj    _   _
25-26   du  _   _   _   _   _   _   _   _
25  de  de  ADP _   _   27  case    _   _
26  le  le  DET _   Definite=Def|Gender=Masc|Number=Sing|PronType=Art   27  det _   _
27  groupe  groupe  NOUN    _   Gender=Masc|Number=Sing 24  nmod    _   _
```

Figure 20: Dependencies of the 2 sentences with NIL with annotation (-)

3.3.3 Evaluation of dependency context

	+	-	?
Dependency context	35%	55%	10%

Table 5: Dependency context

In this context, there is no NIL relation because we must have a word pair with an existing relation (Table 5).
For the case (('il', 'doit'), 'root') (Figure 21), we have judged that there are no errors in the annotations detected because the meaning of the sentences is totally different.

(16) a. sent_id = fr-ud-train_12724

Il doit y exister un trésor mystérieux que l'on cache aux Européens.

'There must be some mysterious treasure hidden from Europeans.'

b. sent_id = fr-ud-train_00489

Il doit son nom au mathématicien, physicien, naturaliste, politologue et navigateur français Jean-Charles de Borda (1733-1799)

'It owes its name to the mathematician...'

```
1   Il  il  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Prs   2   expl:subj   _   wordform=il
2   doit    devoir  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
```

```
1   Il  il  PRON    _   Gender=Masc|Number=Sing|Person=3|PronType=Prs   2   nsubj   _   wordform=il
2   doit    devoir  VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root    _   _
```

Figure 21: Dependency context with annotation (-)

For the case where we have a real dependency annotation error we have the example (('est', 'prévu'), 'root') (Figure 22). The meaning and the tense are the same for the two sentences but they are annotated differently.

(17) a. sent_id = fr-ud-train_00336

Le prochain passage au périhélie de 22P/Kopff **est prévu** le 25 octobre 2015.

'The next switch to 22P / Kopff perihelion is scheduled for October 25, 2015.'

b. sent_id = fr-ud-train_06153

Il **est prévu** 8 sections régionales qui travaillent avec les conseils d'aires coutumières.

'8 regional sections are planned to work...'

```
9   est être    AUX _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   10  aux:pass    _   _
10  prévu   prévoir VERB    _   Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part   0   root    _   _
```

```
2   est être    AUX _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   3   aux:tense   _   _
3   prévu   prévoir VERB    _   Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part   0   root    _   _
```

Figure 22: Dependency context with annotation (+)

For the last case, we have doubts about the annotation of the couple (('sont', 'produits'), 'root'). In the example 18a and 18c we have doubts about the tenses if it's *aux:tense* or *aux:pass*.

(18) a. sent_id = fr-ud-train_05472

Lorsque la porcine zona pellucida est injectée à d'autres mammifères, des anticorps **sont produits** et se rattachent à la zone pellucide de cet animal, empêchant les spermatozoïdes de se fixer à l'ovule, et empêchant ainsi la fécondation.

'...antibodies are produced...'

b. sent_id = fr-ud-train_01172

80 % des importations **sont** des **produits** manufacturés à partir des matières préalablement exportées.

'80% of imports are manufactured products...'

c. sent_id = fr-ud-train_12060

80% de cette drogue **sont produits** dans les laboratoires européens.

'80% of this drug is produced in European laboratories.'

```
15  sont    être    AUX _   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   16  aux:pass    _   _
16  produits    produire    VERB    _   Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part    0   root    _   _
```

```
6   sont    être    AUX _   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   8   cop _   _
7   des un  DET _   Definite=Ind|Number=Plur|PronType=Art   8   det _   _
8   produits    produit NOUN    _   Gender=Masc|Number=Plur 0   root    _   _
```

```
6   sont    être    AUX _   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   7   aux:tense   _   _
7   produits    produire    VERB    _   Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part    0   root    _   _
```

Figure 23: Dependency context with annotation (?)

As shown in the summary tables, NIL relations can be only applied on the internal and neighboring contexts and the results show that the neighboring context is more interesting if we apply the NIL. However, ignoring the NIL relations, we can clearly see that the neighboring context detects more errors followed by the internal context and finally we have the dependency context. In view of the results, we notice that each chosen context detects errors that are not seen by the other contexts. Given this information, to have a more exhaustive list of errors, we must cross the results of the different contexts and eliminate the duplicates.

4 Conclusion and future work

In this second part of the project which was devoted to the practical part which follows the bibliographic part, first we implemented the algorithm proposed by Boyd et al [1] which aims to detect the dependency errors of annotated corpora. Our system currently manages to detect errors in three different contexts of dependency (Neighboring, internal, dependency) and with or without NIL relation. The system is also available and usable on any corpus coded in Conllu. There are other methods of detecting dependency errors such as parser but these are dedicated to the corpus annotated automatically and the corpus we used is annotated manually. The objective is not to correct automatically with a parser but create a tool in order to facilitate the future correction.

This tool detect some dependency errors, but each one must be verify manually. regarding the future work it would be interesting to be able to correct directly from the tool's display. Kirian's tool allows us to correct the annotations directly on the HTML page by orienting the arrows or modifying the labels. However, despite all the possibilities offered by this tool, the modifications are not applied on the conllu file. Our perspective on this point is to continue the work so that any modification operated on the HTML page is automatically reflected on the conllu file. The second objective is to ensure that each pair of corrected words is no longer detected by the system. Another goal is to optimize our system so that it can take less time when we have a large corpus. This project allowed us first of all to live the experience of working in a group during the whole academic year. It also allowed us to acquire more knowledge for computer scientists of how to annotate a text and the methods used in the bibliographic part, but also to practice more the Python programming language and HTML markup language.

References

[1] Adriane Boyd, Markus Dickinson, and Detmar Meurers. "On Detecting Errors in Dependency Treebanks". In: *Research on Language and Computation* 6 (Oct. 2008), pp. 113–137. DOI: 10.1007/s11168-008-9051-9.

[2] Marie-Catherine de Marneffe et al. "Assessing the Annotation Consistency of the Universal Dependencies Corpora". In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Pisa,Italy: Linköping University Electronic Press, Sept. 2017, pp. 108–115. URL: https://www.aclweb.org/anthology/W17-6514.

[3] Guillaume Wisniewski. "Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. URL: https://www.aclweb.org/anthology/L18-1711.