University of Lorraine

Report for Tutorial Project

A tool for automatic analysis of linguistic clues in dialogue transcripts

Author:

Yiqing Liang

Ruixue LiU

Olga Leticevscaia

Supervisor:

Maxime Amblard

May 31, 2016

Abstract

Our project is involved in the research project named SLAM, which aims at systematizing the study of pathological dialogues: dialogue with patients of schizophrenia. The main topic of it is the study of dialogic interaction, which includes several levels of linguistic annotation such as disfluence tagging and part of speeching (POS) tagging. The current research object is to develope a tool that integrates several other tools for annotation, to automatically identify pathological linguistic clues in transcriptions.

Two stages are involved in the development of the tool. The first stage is to re-implement the SLAMtk tools in a generic manner. Developed with the Python programming language, the tool works on different corpora with specific characteristics. In order to build an XML structure to represent the generic version of different corpora, we perform a series of pre-treatments to normalize them. With the XML structure, all the necessary information needed for annotation and data extration are stored, including annotation from Distagger for disfluencies ('euh ...', repetition of words, etc.) and morphosyntactic segmentation (part-of- speech) from Melt. The second stage refers to the analysis of linguistic features and identificantion of clues. Several treatments are performed to create representations of these interactions and mathematics tests are applied to identify specific clues.

Contents

1 Introduction

As part of the Master One Cognitive Science program and Language and Communication Technology program, this tutorial project is finished during a period of 5 month. This project is involved in the research project SLAM (Schizophrenia and Language: Analysis and Modeling).

The research objective of SLAM project is to analyze the speech in schizophrenia patients. It aims to systematize the study of pathological conversations as share of an interdisciplinary approach-combining psychology, linguistics, computational linguistics and philosophy. The analysis are performed from a text corpus. These transcripts of actual conversations between schizophrenic patients and psychologists from two different cities in France. However, transcripts of these two different cities are organized in different format. For the linguistic analysis, all the transcripts should be normalized into similar format (an XML file) and several tools should be involved in annotation.

The research objective of the current project is to develop a tool that can work on different transcripts with specific characters and integrate several annotation tools together, to identify automatically the linguistic clues in various corpora. In that case, there are two main stages involved in the project: firstly, with programming language of python, we will normalize the different transcript corpus and store information into XML format. Then, with the implementation of of annotation tools, the annotated information is updated into the XML format files. The second steps involves the analysis of linguistic features including POS(Part Of Speech) tagger and disfluency clues used by interviewees. Most of the analysis or working procedures is followed

by the previous work from the SLAM project [Amblard et al., 2015].

In the following part, we will firstly introduce you the context of the project, mainly from the perspective of the organization as scientific context. Then detailed information of tools that we need to use for annotation will also be presented. Two tools are used during the data processing, namely Distagger for disfluence annotation and MELT for POS annotation. In order to organize the two different types of transcripts from two corpora into a similar format, we choose an XML structured file to store the necessary information.Thus, in the practical part, we focus on how the data are formalized and constructed into XML format, how the annotated information from MELT and Distagger are updated into the XML files. Before conclusion, we will demonstrate briefly the result of linguistic analysis.

2 Theoretical Framework

In this section, we will present you information on the corpora we are dealing with for linguistic analysis. We will also introduce you the detailed information on two annotation tools we need to use, and how the annotation is going to be analyzed later.

2.1 Research data

Like many other linguistic research, the interview transcripts are conducted between schizophrenic patients and control groups. Two genders of them, male and female are also taken into consideration among the two group of interviewees. The body of data is divided into two parts, corresponding to the cities of medical units specialized collections. Out of respect for the patients, these cities' names are anonymised as City1 and City 2. The collection of interviews are firstly recorded as MP3 files and then being transcribed into text files.They are transformed by different annotators, following the recommendation for fine transcriptions. These transcriptions were also treated according to the recommendation of [Blanche-Benveniste and Jeanjean, 1987]. However, as the transcripts in two corpora were made quiet a long time apart (more than 10 years apart), the transcriptions in two groups were formed differently.

In general, SLAM project from two corpora includes 80 subjects for interviews, 49 of whom are schizophrenic patients and 31 are from control group. However, concerning the authorization of the context, and as only part of the SLAM project, we can only access to limited number of data for investigation

of the automatic tool construction. Thus , we can have access to 18 files of interviews, 10 of which come from the City1 group and 8 from the City2 group. Among the 10 files in City1, half of the data relate to the interaction with patients and the rest relate with the control group. While in the files of City2, there are 3 files for patients and 5 for the control group.

Even through, there are many researches focus on the language production of schizophrenia, it's hard to draw some conclusion on their linguistic features. One reason dues to the various fields involved in the study (such as psychology, medicine, linguistics, etc.). Besides, the conditions of the experiments described are so variable that it is difficult to put the results consistent with. However, meta-study from [Maher, 1972] interpret the data from a different view, which focuses on the repetition, based on the TTR (Type-Token Ratio). His research shows that, schizophrenia patients have lower TTR, which means they will repeat more. And some other studies ([Feldstein, 1962] and [Kremen et al., 2003]) also shows the high repetition among the schizophrenia patients. Thus, in this research, the disfluence of speakers is one of our focus. Further more, the POS is taken into consideration during the linguistic analysis.

As introduced in the previous part, though the transcriptions for interviews in the two corpora follow the same recommendation during the annotation, their organization are quite different from each other. In order to create a tool for analyzing linguistics units in different types of transcripts, we choose to use an XML format to store the needed information from two corpora with the same XML structures.Besides, the adaptation of XML file also make it easily to update or extract the information for annotation details.

To realize the linguistic analysis mentioned above, in the practical part, we will anonymize and normalize the data for each corpus and construct XML file format for storing and updating all the necessary information for annotation tool for disfluence and POS. Information required for linguistic analysis is also updated in the XML file.

2.2 MELT for POS tagger

MELT is developed by [Denis and Sagot, 2012] and allows for using multiclass Maximum-Entropy Markov models (MEMMs) or multiclass perceptrons (multitrons) as underlying statistical devices. Its output is in the Brown format (one sentence per line, each sentence being a space-separated sequence of annotated words in the word/tag format). The performance of this tool with this model reach 97.61% accuracy in French.

The format of MELT is listed below, and the character of $*$ is used for annotating the lemma of unknown characters.

spk2 non/ADV/non ça/PRO/cela me/CLO/cld plait/V/*pdre pas/ADV/pas
spk1 d'/P/de accord/NC/accord
spk1 y/CLO/cll a/V/avoir rien/PRO/rien du/P+D/*du tout/PRO/tout qui/PROREL/qui euh/V/*euh
spk2 si/CS/si le/DET/le premier/ADJ/premier semestre/NC/semestre je/CLS/cln j'/CLS/cln allais/V/aller à/P/à tous/ADJ/tout les/DET/le cours/NC/cours j'/CLS/cln aime/V/aimer bien/ADV/bien

There are 29 types of POS annotation in the French text (such as 'V', 'VIMP', 'Nc', 'NPP'), based on the types and function of these word, we

have then categorized them into eight main categories (including verb, adjective, adverb,noun,determiner,preposition,pronoun,and other types). The categorization is listed in Figure 1. Based on this, we will analyze the number of each category used by each group of speakers and then compare the proportion of each type of POS among them.

```
POS_dic={
'VER':['V','VIMP','VINF','VPP','VPR','VS'],
'ADJ':['ADJ','ADJWH'],
'ADV':['ADV','ADVWH'],
'NOM':['NC','NPP'],
'DET':['DET','DETWH','P+D'],
'PRP':['P','P+D','P+PRO'],
'PRO':['CLO','CLR','CLS','P+PRO','PRO','PROREL','PROWH'],
'AUTRE':['CC','CS','ET','I','PONCT','PREF']
}
```

Figure 1: The category of POS for linguistic analysis

MELT can also provide the annotation for lemma of each word. We will also analyze the number of lemma for each group of speakers. Then the information is used for TTR (Type/Token Ratio) analysis. Here 'type' refers to the number of lemma used in each file, while 'token' refers to the word number. TTR is an effective way to illustrate the word diversity used by speakers.

2.3 Distagger for disfluence

The annotation of disfluence is applied to study the practice of schizophrenia patients. Developed by [Blanc et al., 2010], distagger is a annotation tool for disfluences and its performance is evaluated on a corpus with 22 476 words and 1 280 disfluences, with 95.5% F-score (Precision 95.3%, 95.8% recall).

Speech disfluency refers to any types of various breaks, irregularities or

some non-lexical vocables that occurs within the flow of a speech. In our project, Distagger will identify four types of disfluency in the corpus, namely 'euh', repetition, short pause, and self-correction. Below, are examples of disfluency annotated in our corpus.

(1) Disfluency of 'euh'

Vous avez travaillé comme {euh,.IGN+EUH} en mécanique ajusteur P {euh,.IGN+EUH} mécanicien ajusteur .

In this example, if there is an 'euh' shown during the conversation, the distagger will annotate it as '{euh,.IGN+EUH}'.

(2)Disfluency of repetition

Non {dans,.IGN+REP} dans les domaines j'étais {j j,.IGN+REP} j j'étais pas {très /,.IGN+REP} très bon.

Non même si il pleut si le temps {si,.IGN+REP} si il faut faire des courses je vais faire des courses.

In the speech of patients, repetition usually occurs as one word, such as 'dans, dans', 'vous, vous ' or as some phrase with 'j'étais, je, je, j'étais'. All these types of repetition will be marked as {.IGN+REP}.

(3)Disfluency of self-correction

Self-correction refers to the phenomenon that the speaker try to correct and repeat his word after realizing the error he has made during the speech. The self-correction will be marked as {.IGN+CORR}.

(4)Disfluency of pause

Ah d'accord donc on peut {/,.IGN+short-pause } ca vous donne le droit de {/,.IGN+short-pause}.

During the transcript of records for interviews, all the information of pause is also recorded and is rewritten as "/", which will be annotated as {/,.IGN+short-pause} by Distagger. With the four types of annotation, all the disfluency will be analyzed in the fourth section of our report.

2.4 Method for data analysis

To provide a valid way of interpreting result of linguistic analysis, we have applied the measure of significance [Amblard et al., 2015]. This measurement is used to calculate a distribution index as a function of number of words between two categories of speakers. Three possible combination are calculated, namely psychologists and schizophrenic patients, psychologists and control group , control group and schizophrenic patients. The formula to be used for significance is list below and is taking [Amblard et al., 2015] as reference:

$$s = \frac{(p_1 - p_2))}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{1}$$

where :
- $p = \frac{(n_1 p_1 + n_2 p_2)}{(n_1 + n_2)}$
- n_1 is the number of words of the first category of speakers
- n_2 is the number of words of the second category of speakers
- p_1 is the proportion of the phenomenon attributed to the first category of speakers,
- p_2 is the proportion of the phenomenon attributed to the second category of speakers

This measure has the advantage of comparing work on the disfluencies French. According to [Amblard et al., 2015], we will adapt the Hypothesis H_0 ,and will reject the result if s is lower that 2.5% or higher than 97.5% of a normal distribution, that, is to say, 1.96. The result that is higher than 1.96 can be considered as significant, with an error risk of 5%. The detailed result is shown in the following section 4.

3 Implementation of SLAMtk

In this part, we will present you the practical work we have done for the tutorial project, namely how this automatic tool will work on two different corpora and how the text file will be annotated and how the information will be stored and updated for future analysis.

3.1 Corpus Preprocessing

As mentioned in previous part, there are two subcorpus included in the research and the format for transcripts included in each corpus differs from each other.In order to develop an automatic tool to analyze linguistic units in each of the corpus, we need to normalize the format of each corpus. Then , we choose to use an XML structure to store all the information of utterance. This enables us to present information in two corpora in a similar way and also enables us to update or extract the annotated information later effectively.

In the corpus of City1, each transcript is saved as '.txt' file, and the file name present an unique code, such as '101BAK, 002LOA, 105CAI'. In each file, interviewee's personal information, including gender, interviewee' group, age, education level and so on are included insides. Besides, during the conversation, there are characters simplified as 'P, H' or 'A,B' to present the speaker of interviewer (psychologist) and interviewee (patient or control group). What's more, different types of arrows are also used to present intonation, pause,repetition.

The following are examples of layout of City1 corpus.

```
P26 :   Bof euh (→) quand j'ai envie de (→) de faire un tour je fais
    un tour si je vois que j'ai pas envie quand il c- quand il pleut
    quand il fait mauvais eh ben je reste à la maison (↓)
H27 :   D'accord (↓) et où est-ce que vous faites vos courses par
    exemple (↑)
P28 :   Bô (↑) dans les magasins (↓)
H29 :   N'importe où (↑)
```

Figure 2: The layout of City1 corpus

In the examples showed above, information of pause and repetition are presented as right arrow. Besides, in order to identify the gender and identity of speaker, at the end of each file, there is one line information indicating the identity of psychologist, such as :
Ps = A

While in corpus of City2, each transcript is saved as '.cha' file and the file name is presented with an unique code for each interviewee. And mainly the content of utterance is stored in the file. Each interviewee (patient or control group member) is marked as it file name, such as '005AUB, 001PEP'. While for the content, time duration for several utterances sharing the same topic is also recorded. During the conversation, the repetition, speakers' laughter or pause are also marked with different singles.

Some examples are shown in below.

In the examples shown above, the '[//]'is used to mark the repetition of speakers and '[/]'markers the pause. Time duration for one topic (such as '47112_58422'to '58422_66718') is also demonstrated in the transcripts.

From the above examples, it's demonstrated clearly that the layout format

```
*PSY:   Bon. Je vous remercie d'accepter de
    reparticiper. De(.) voilà, de [/] de refaire cet
    entretien parce que c'est vrai qu'on entendait rien
    du tout sur le premier +... NAK47112_58422NAK
*005AUB:    hum hum
*PSY:   +, et euh c'était un peu compliqué (.) Donc
    voilà (.) alors! <qu'est-ce qui> +..? [/] qu'est-ce
    qui c'est passé depuis la dernière fois?
    NAK58422_66718NAK
*005AUB:    bah j(e) vais toujours à l'hôpital de jour
    de la M. +...
```

Figure 3: The layout of City2 corpus

for the two corpora are totally different. To process the data into the different annotation tools and to store the information in the same xml format, both corpora need to be normalized. The work for normalization of two corpus is already finished during the former work, here they are reused for our tool.

To normalize the corpus of City1, the first step is to delete all the personal information in the file. However, as later the linguistic analysis is based on the interviewees' gender and group (patient or control group), thus all the files' structure will be manually reorganized into folders of 'Schizophrenes'and 'Temoins(control group)', and also 'femmes(female)'and 'hommes(male)'. The structure of the corpus is showed in Figure 4 below:

In order to easily get the information of their folder's group, the path information such as '../corpus/schizophrènes/hommes/001pep/001pep.cha'is stored in the first line of the normalized file.

After that, the content of utterance are normalized by several way. One of the important step is to get the anonymous name of speakers. For example, according to the information for speaker of psychologist's character,

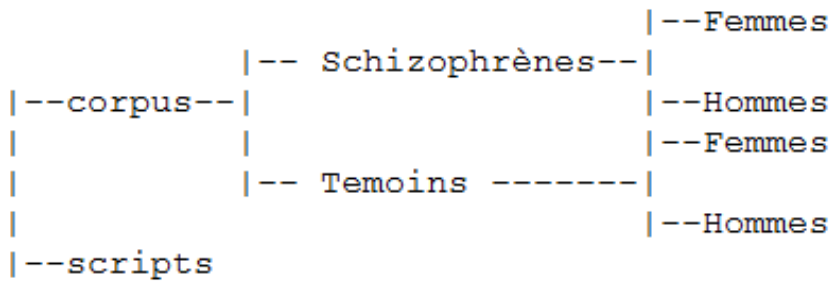

Figure 4: The required structure for files in corpus

such as 'Ps = A'showed as example above, all the speaker's name start as A (like A1, A3, A5) are replaced with 'spk1', while the characters for interviewee are replaced with 'spk2'. After that, all the arrows or figures such as '(→),(+)'showed in the file need to be normalized into blank space. All the emotional information such as laugh will not be taken into consideration for linguistics analysis and will thus be replaced as blank space. If only one marks such as in appears one speaker's utterance turn, then the utterance will be deleted.

Apart from that, to fill the requirement for Distagger input file, all the pause or repetition marked as '(→)'will be replaced by '/'. While in order to get the input file of MELT for POS annotation, all the phrase such as 'j'ai, qu'est-ce'will be divided with one more space: 'j' ai, qu' est- ce'. The all the normalized file are stored in the folder name 'normalise'and are saved as'.nor'type of file.
The final normalized file for City1 corpus is listed below:

spk2 oh je inaud je ne sais pas quoi dire
spk1 oh ben je ne sais pas vous vous levez à quel heure par exemple

spk2 vers les 6 heures du matin

spk1 6 heures c' est tôt

spk2 oui ben je peux me rendormir après

spk1 ah d' accord

spk2 mais le soir je me couche vers les 8 heures aussi

For the data in City2 corpus, the first steps for data processing is also to anonymize any specific names mentioned in the context. To maintain the readability of the file, 7 categories will be hidden and replaced with unspecific names instead, based on the previous work of the SLAM project. The 7 category include person, city, country, hospital, mountain, capital, department, institution. To maintain the consistency of the names mentioned during the conversation, each name of entity is replaced by an unique identifier. Thus if the first reference entity is substituted by a certain name, then the same name will be used throughout the whole conversation. Apart from that, names for people are substituted with indication of their gender, such as 'pernomF1'or 'pernomM2'. To realize the anonymisation, several python codes are implemented with the regular expression. In the following part, some examples are used for demonstration of anonymised file.

*PSY: Et euh (2 sec.) vous faites euh, vous ça, ça fait longtemps que vous êtes en, vous êtes au **Institution1?**

*PSY: **Ville2**. D'accord alors moi qui suis pas du tout d'ici, c'est pas très loin.

*PSY: Si **PrenomM9** se r(e)trouve ...

The information of anonymized file is stored in the folder named 'normalise'and then is used for normalization. During the normalization part,

all the information referring to time duration, emotion are deleted.The same with corpus of City1, all the pause and repetition are replaced by '/'for Distagger and all phrased like 'j'ai, qu'est-ce'will be divided with one more space: 'j' ai, qu' est- ce'for MELT.
The following are examples of final normalized version.

spk1 bon je vous remercie d' accepter de reparticiper de voilà de / de refaire cet entretien parce que c' est vrai qu' on entendait rien du tout sur le premier
spk2 hum hum
spk1 et euh c' était un peu compliqué donc voilà alors qu' est- ce qui / qu' est- ce qui c' est passé depuis la dernière fois
spk2 bah je vais toujours à l' hôpital de jour de la m

As demonstrated from the example of normalization of two corpora, each utterance of one speaker is listed in one line of the file. The interviewer (psychologist) is marked as 'spk1'in all corpora and the interviewees (patients or control group) are marked as 'spk2'. All the punctuation and empty multiple lines are deleted and the capital letters are replaced by the lower case one. All the normalized file is stored in the folder named 'normalise' and as '.nor' file with the unique code for each patient as its file name.

3.2 XML construction

With the normalized file as input, we need to produce a xml file for storing information of the content of utterances of each speakers. Based on the normalized file, the number of utterances and number of words for each

speaker should be calculated and stored in the XML file.

Based on the requirement above, the XML file will include mainly two parts, namely the utterance content and the analysis. The detailed information on the XML structure is listed below:

- the general information of the interviewees, including the gender and group of interviewee

- the analysis of total utterance, including the number of words and turns of utterance in the context.

- the analysis of utterance of each speaker, including the number of words and turns of utterance.

- the information of two speakers conversation content1.

- the utterance turn for each speaker (including number of words in the utterance, the number of turns of utterance and identity of speaker for each utterance).

- each word included in the utterance.

As there are several types of information needed in the analysis part, we have applied various code with python language to calculate the utterance number, word number for each speaker and in total.

Then based on the structure of the XML file, we need to add information such as total word numbers, word number for each speaker, utterance number for each speaker, word number in each utterance. To get the information needed above, we will use the following process:

(1) The process for getting total word number, and word number for each speaker is shown as followings:

1. Reading lines of normalized file as a list;

2. For each line in the list, splitting the words into a new list;

3. Building dictionary, with the identity of speaker:'spk1'and 'spk2' as the dictionary key and their utterance as the value for each key.

4. Checking each line of list, if the first element is already exist as key in the dictionary, then append the rest of the list as value of this key.Other wise, adding the first element as new key, and the rest of the element as its value.

5. For each key in the dictionary, counting on the number of elements included as its value. The result we get refers to the total number of utterance for each speaker.Adding the utterance number for each speaker together, we will get the total word number. Then summing up the length of each value element and this number refers to the total word number for each speaker. Then, adding up word number for each speaker together, we can get the total number of utterance in the file.

6. Adding information of total word number and total utterance as the attribute into the node named 'participants'. Then for each speaker, creating sub-element named 'speaker' and adding information on word and utterance number as attribute inside.

(2) Process for creating each utterance and its words inside as elements and subelements:

1. Creating node named ,'topic' which will include all the utterance content inside.

2. Reading lines of the normalized file and split each line as a list.

3. Setting the sequence of utterance as '1' at the beginning.Then, for each line of the file, creating an XML sub element, belong to the 'topic' node. Taking the first element as the attribute which will indicate the identity of speaker. Then, counting the length of the rest of the list, setting it as the word number in the attribute. Taking the sequence as its utterance number in the attribute, too. For the lines blow, repeating the same procedure, with adding value of sequence '+1' each time.

4. For each elements in the list, creating the subsubelement, named 'word', and setting each element as the text content for the 'word' elements.

The following figure 5, figure 6 will show the detailed information for the xml file.

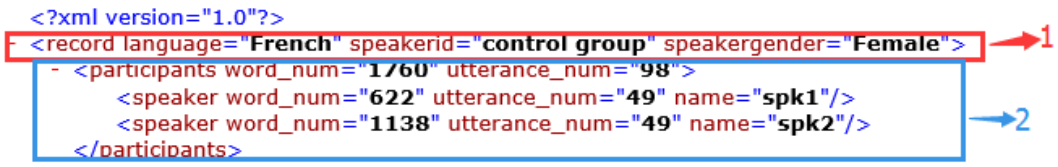

Figure 5: The first part of the XML file

The figure 5 shown above is the first part of the XML file. While the first subsubsection shown in red refers to the general information of the whole transcript, and is marked as 'record'. The attributes of this tree shows in detail the gender (male or female)of the interviewee and also which group (patient or control group) it belongs to. The second subsubsection shown in blue in this figure refers to the first subelement of 'record' in the xml file, which is named as 'participants'. The attributes of 'participants' indicates

the total number of word and of utterance in this file. While the second part of the XML file is showed in the figure below.

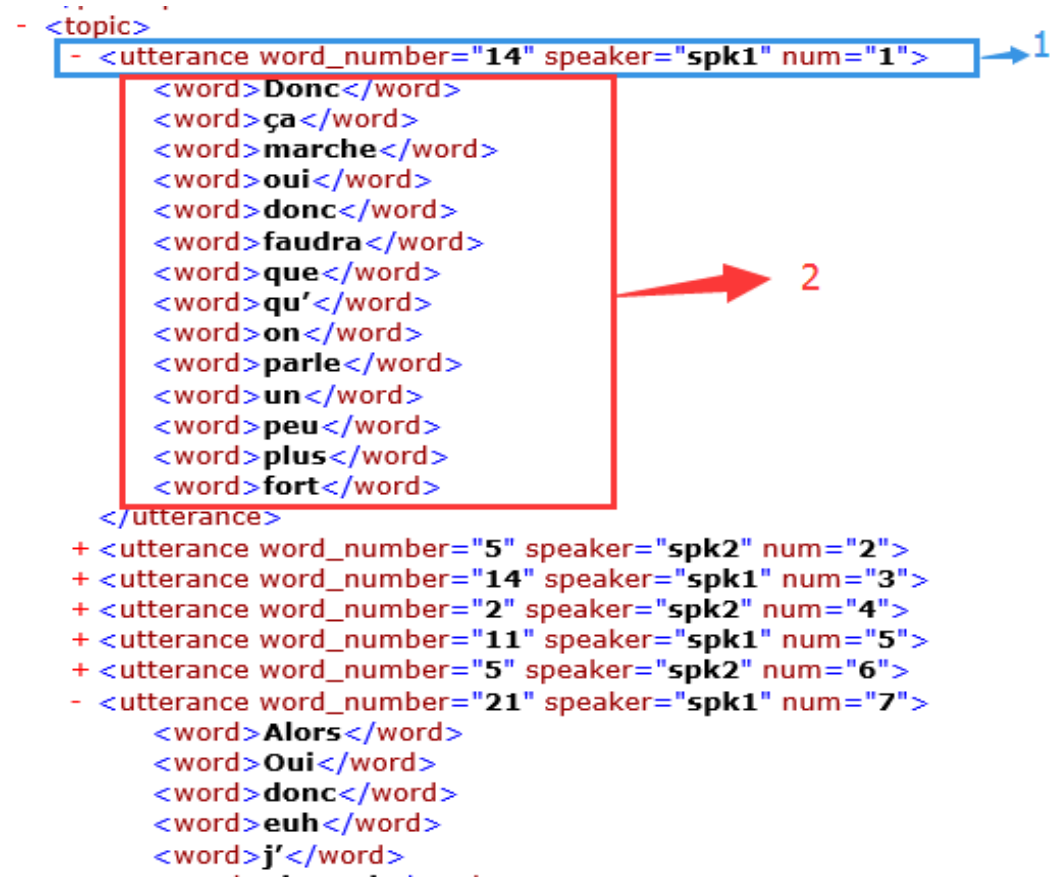

Figure 6: The second part of the XML file

The second part of the XML file refers to the whole utterances of two speakers in the interview and they are all included in the element named 'topic'. As shown in the figure 6, the first subsubsection marked in blue is the subelement of 'topic', which refers to each turn of utterance of each speaker during the interview. The attribute of the 'utterance' element can show us the number of word included in and also the speaker of it. The attribute

'num' here indicates the turn of utterance among the whole interview. The second subsubsection in red shows the subelement of the 'utterance'.Each word appears in the utterance is included in its subelement name 'word'.

With the XML file shown above, we will extract all the information of utterance with the name of its speaker, and stored all the needed information into a '.txt' file. After that, we will apply the annotation tool of Distagger and MELT for disfluence taggers and POS taggers with the '.txt' file as input, and then update the information and also their analysis into this XML file. The detailed information is illustrated in the following parts.

3.3 Distagger application and xml update

The disfluency analysis was performed using Distagger - automatic disfluency detection tool for speech transcripts. In present work the Distagger 0.2 was applied. For that purpose the input files were modified in order to correspond to the input format for Distagger and the application of the tool was performed. Then the output .snt files are used for update implementation of the initial xml files. The matching is done by file names.

Every .snt file is parsed in the way to extract the overall numbers of every disfluency made by each speaker, the disfluencies by every utterance and the exact location of disfluencies is marked with tags named by disfluency pronounced including the word or phrase producing this disfluency type. The updated version of xml is used then for analysis extraction (see section 3.5.1).

The figure 7 shows the format of initial file which is modified (tags ¡deb¿ and ¡fin¿ are added in the beginning and end of file respectively). Then

Figure 7: Fragment of original input .txt file used for Distagger application

Figure 8: Fragment of the Distagger result .snt file.The tags mark the disfluency location and its type

Distagger is applied to the modified txt file. The resulting file format has extension snt and has the following format - see figure 8. The extraction of the statistical data about each disfluency type occurrence per file, per utterance and the exact location of the disfluency made was performed by parsing snt files and adding new tags(or updating existing) disfluency analysis tags. The resulting updated xml file example is presented on the figure 9. The summarized data for a speaker is placed under every speaker tag(¡speaker¿), the number of occurrences and disfluency types made during an utterance are added as attributes to each utterance tag(¡utterance¿). The words being disfluencies are included in tags corresponding to disfluency type it reflects. Some of the IGN-REP type disfluencies are not relevant to analysis, thus the prefiltring of tags was applied to remove unimportant data.

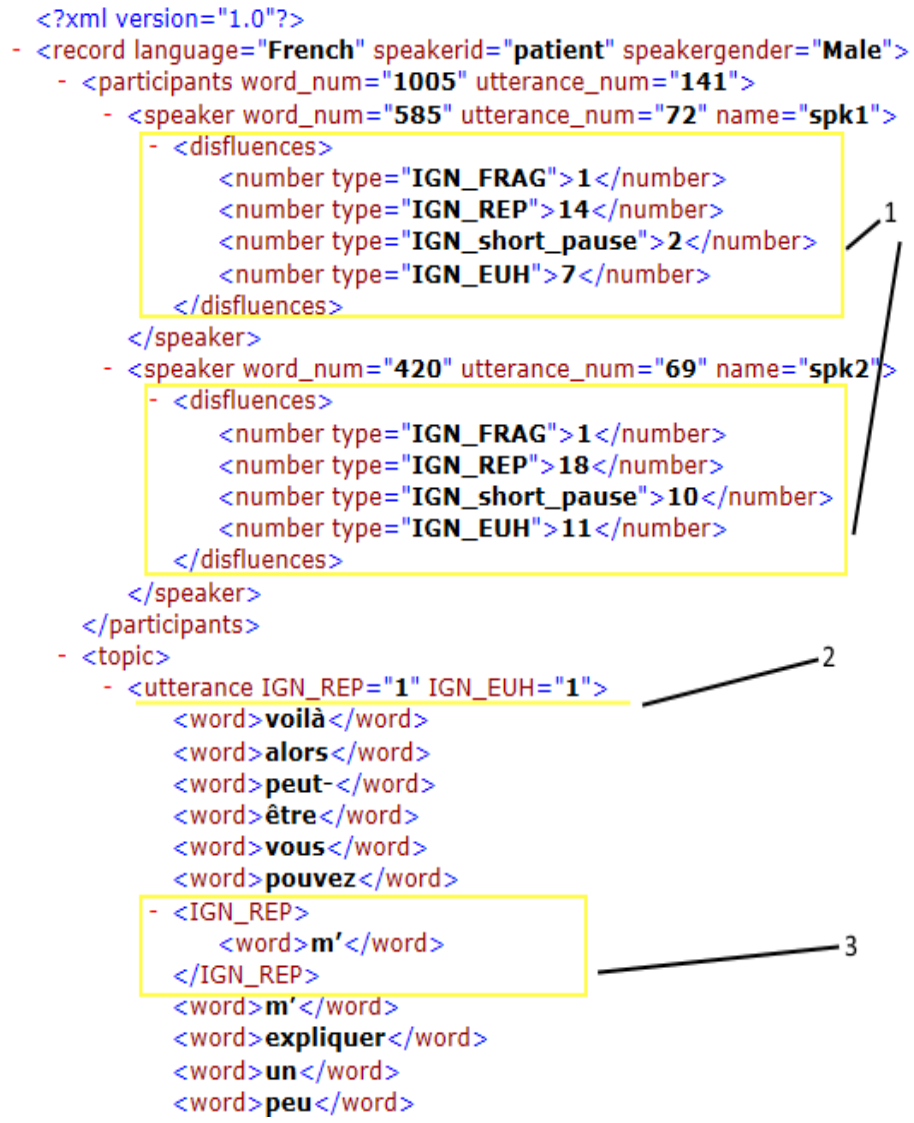

Figure 9: Fragment of xml file updated with the Distagger data from .snt file. The summarized data is added(1) for each speakers tag, also the data for every utterance is calculated(2) and the exact disfluency location is marked with according tag

3.4 Melt application and xml update

Melt is a tool for POS tagging in French. It takes transcripts of conversation in txt format as input, and returns the tagged file also as txt file. It gives not only the POS of every word in the text, but also the their lemma. An example of tagged file is already given in previous chapters.

In this project, the input files for POS tagging are the normalized version of transcripts extracted from the basic xml, which are produced before this step.

The main tasks in this part includes:

- Taking original xml and transcripts extracted from original xml as input to produce POS tagged files
- Adding POS and lemma information for each word
- Adding numbers of each POS for each speaker in xml
- Adding numbers of POS categories, word forms and lemmas for each speaker in xml

The algorithm to use MELT for POS tagging is already finished in former work. Here, the code for this purpose is reused. However, the process of the MELT application is still worth mentioning, because it helps giving a complete view of the whole POS updating part. The process for POS tagging is as followed:

1. Dividing each input file into two separate files, the left part and the

right part. The left part contains only speaker information, and the right part contains all the information left.

2. Taking the right part files as input for MELT, and then getting MELTed right part files as output

3. Merging the MELTed files and the left part files, producing the final results for POS tagging and then putting the results in another folder named *POSresults*

An example of the original transcript, the left part, the right part, the MELTed right part and the final merged file are shown in Figure 10

The POS result files are what we need for updating the existing XML file.

The process for adding POS and lemma information to XML files is as followed:

1. Taking two directories as input for *add_pos* function, corresponding to paths of original XML files and POS results

2. For every XML file, find the corresponding POS result file based on the first six character of file names of two files

3. Parsing the XML file as element tree

4. Using *iter* to find every utterance node in XML, at the same time, using *readline* to read the corresponding line in POS result file

(a) Original file

(b) left part

(c) right part

(d) MELTed right part

(e) merged file

Figure 10: Different versions of file for MELT

5. Using *iter* again to find every word node based on the utterance node found in the former step. Splitting the line in POS result file to get all the separate words

6. Splitting every word based on symbol "/" in POS result in order to get the original word, the POS tag and the lemma

7. Setting two new attributes, POS and lemma, to the corresponding word node

8. Removing all the intermediate files generated in this process

Now we have new XML files containing all the POS and lemma information for all words. However, this is still not convenient enough for further analysis.

Later, we'll need the numbers of POS categories, number of lemmas, number of word forms, number of word for every POS category spoken by each speaker. Thus, such information also needs to be added in the XML.

Instead of counting the numbers in txt file every time when needed, as done before, here, we count all the numbers from the XML, and write them back in XML again, once and for all. So, every time when such numbers are needed, they don't need to be counted again and again, they just need to be read directly from the XML. This method not only makes the analysis process more convenient, but also increases the efficiency and saves time and space, making the program more extensible.

The process for counting needed numbers and updating XML is as followed:

1. Parsing XML file as element tree

2. Setting up lists to record every POS, lemma and word form ever occurred for each two speakers in the file

3. Scanning all the word nodes in the XML. If the POS, lemma or word form of this node is not recorded for the corresponding speaker, appending them to the list

4. After the scanning, the length of the list is the number of the element we need. For example, the length of lemma list for one of the speakers is 250, it means that 250 different lemmas are spoken by this speaker in this transcript

5. Adding the number of different POS, lemmas and word forms for the two speakers

The steps above get the number of different POS, lemma and word form for each speaker, but there's still one kind of information we need to put in XML, the frequency of occurrences for each POS for each speaker.

This information is also obtained from XML. Process as followed:

1. Parsing XML as element tree

2. Setting two empty dictionaries, one for each speaker

3. Scanning every word node, getting its POS and speaker information

4. If the POS is not in the dictionary of the corresponding speaker, adding it as a key and setting its value as 1; if it's already in the dictionary, increasing its value by 1

5. After the scanning, writing the POS and their frequency under each speaker node

The result of POS number counting is shown in Figure 11

Figure 11: Updated XML file

After all these steps, the updating part from MELT is finished.

3.5 Data extraction for analysis

3.5.1 Parsing xml for disfluence analysis

The updated xml after applying Distagger was used for analysis extraction. For that purpose were implemented methods for graph construction, method for producing summing table for every dialog presenting the average

data for every type of speaker in correspondance with disfluency numbers in pdf/Latex formats was adopted in accordance with new data input format, the significativity numbers were calculated for each speaker group for the whole corpus using the methods implemented later.

Also the results for male/female division were produced for further analysis.

3.5.2 Parsing xml for POS analysis

For most of the POS analysis for now, only the data under the *participants* node is needed. Thus, when extracting information for POS analysis, steps are usually as followed:

1. Parsing xml as element tree
2. Checking root node's attribute *speakerid*, deciding if it's a file for patient or control group
3. Checking speaker node's attribute *name*, deciding if it's psychologist (spk1) or patient/control group (spk2)
4. Getting information from speaker node, like word numbers or number of each POS

4 Results of the corpus

In this part, we will present you the quantitative analysis of different types of linguistic units in corpus, namely the POS taggers and disfluency taggers used by each group of speakers. We will present the numbers of them used in each file and also the significance of the linguistic units compared between two pair of groups.

4.1 Disfluence analysis

The disfluency analysis implemented, as for the previous version of the SLAMtk, by constructing graphs for every dialog representing several kinds of data: sum of all disfluencies made by an utterance(see figure 13), all types of disfluencies made by an utterance(see figure 12), each type of disfluencies presented in this dialog per utterance(see example in figure 14). The bar for every disfluency number is marked by the letter "a" if it was produced by a psychologist and by letter "s" if it was produced by a patient.

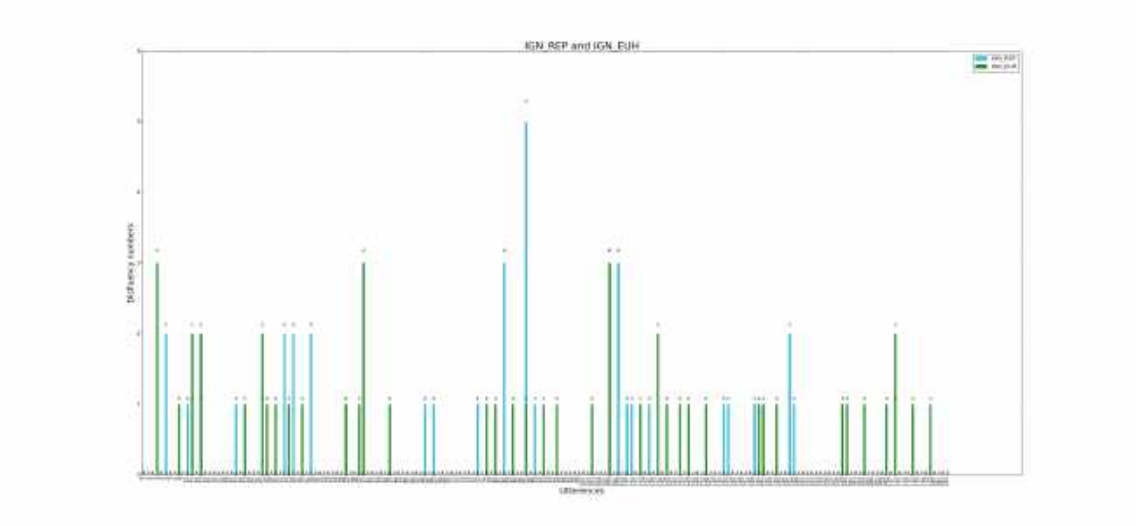

Figure 12: The chart for three types of disfluencies- the occurrences of it per utterance

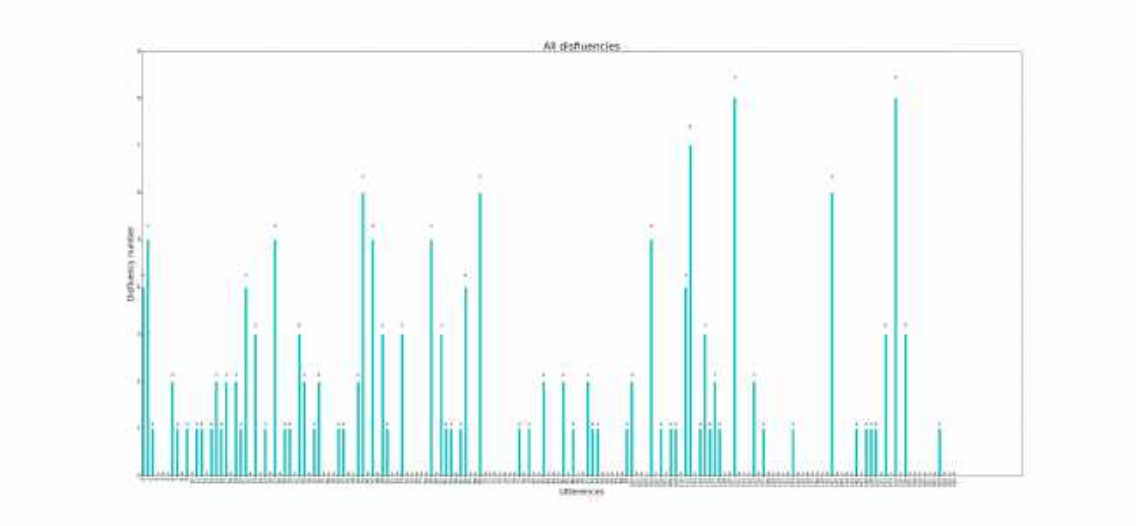

Figure 13: The chart for all the disfluencies made per utterance

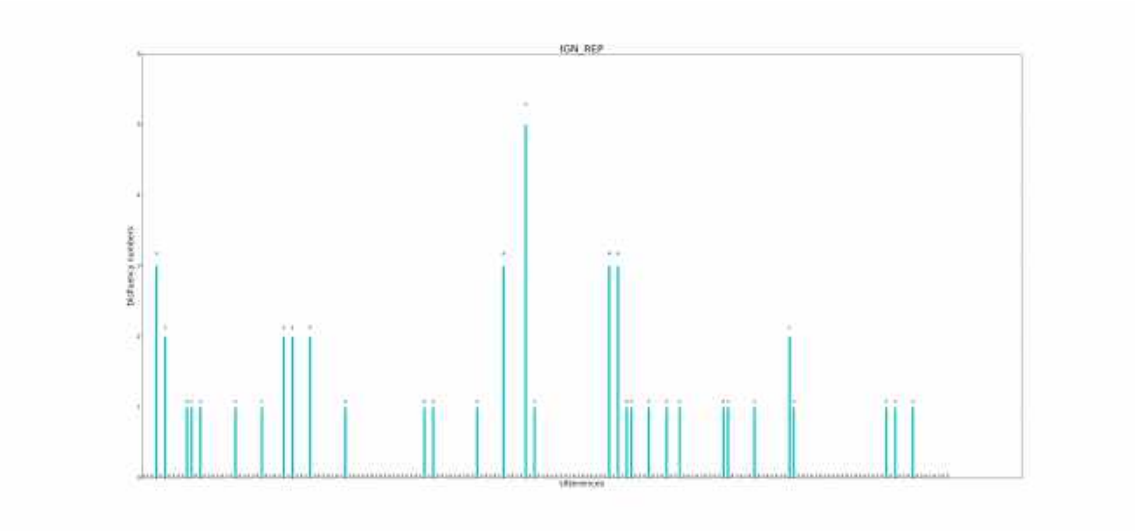

Figure 14: The chart for the disfluency type IGN-REP, the occurrences of it per utterance

Then the table containing the disfluency data for six speaker categories (”Schizo” - schizophrenic, ”temoins” - control group,”nonPsy” - all utterance made by control group and schizophrenic,”Psy+S” - all utterances of psychologist and schizophrenic,”Psy+T” - all utterances of psychologist and control group, ”Psy”- all utterances of psychologist) per each dialog was generated. The summing this data can be represented by the following table(see the table 1) The data for this table was obtained from the subcorpus for City1.

Also the analysis for the male/female data obtained was implemented and presented on the table 2. The calculation was performed on the subcorpus of City1. Although this result suggest that schizophrenic male patients produce larger number of disfluencies than female speakers, it should be checked on the larger corpus. The present result is mostly influenced by the fact there

Disfluency types	S	T	P
REP	0.014070	0.007790	0.013601
CORR	0.0	0.0	0.0
EUH	0.004885	0.003169	0.005087
FRAG	0.000977	0.0	0.000519
short-pause	0.019542	0.001848	0.011836
total	0.0394762	0.012808	0.031045

Table 1: The relation of numbers of every disfluency with the number of words per speaker category, where S is patient, T is control group speaker and P is psychologist.

are more female speakers in control group.

4.2 POS analysis

The study aim in this part, focuses on the analysis of POS taggers and lemmas of each word in each corpus. As introduced in the sections above, with the utilization of MELT, we will annotate the POS and lemma of each word, then update the information into the xml file. In this part, we will present the numbers calculated for POS and lemma.

First of all, with the updated XML version, we can easily get the numbers of each type of POS used in each file. Based on the category defined in section 2before, mainly eight types of POS taggers are analyzed : verb, adjective, noun, adverb, preposition,determiner, pronoun and other types. In order to interpret their values easily, we will use some figures to demonstration the distribution of each type of POS used in the files. The figures shown

Male				
Disfluency types	Numbers		nb to words	
	S	T	S	T
EUH	25	6	0.004885	0.002926
REP	72	9	0.014070	0.004390
CORR	0	0	0	0
FRAG	5	0	0.000977	0
Short-pause	100	5	0.019542	0.002439

Female				
Disfluency types	Numbers		nb to words	
	S	T	S	T
EUH	4	18	0.001080	0.003259
REP	31	50	0.057010	0.009053
CORR	16	0	0.000121	0
FRAG	0	0	0	0
Short-pause	0	9	0	0.001629

Table 2: The numbers of each disfluency type with the number of occurances per speaker category, where S is patient, T is control group speaker and relation of its occurances to the number of words produced for the male or female patient with the category division to control group and real patient.

below present the POS used by three groups of speakers in City2 corpus. Among these figures, figure 15 presents the POS appears in schizophrenic patients' conversation; figure 16 shows the ratio of POS used by control group interviewees while figure 17 shows the proportion of POS used by all the psychologists.

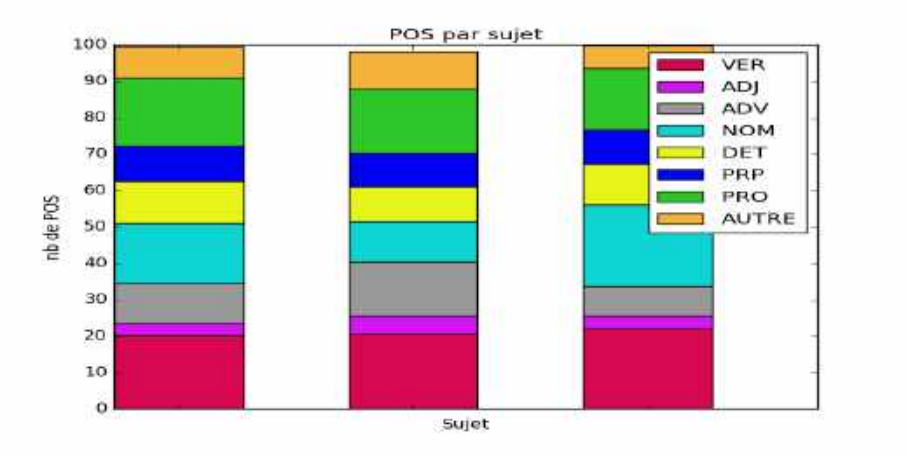

Figure 15: The proportion of POS used by schizophrenic patients

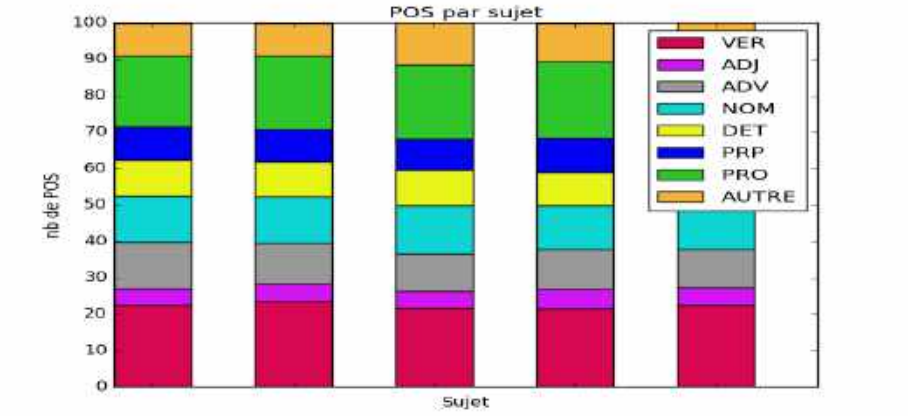

Figure 16: The proportion of POS used by control group

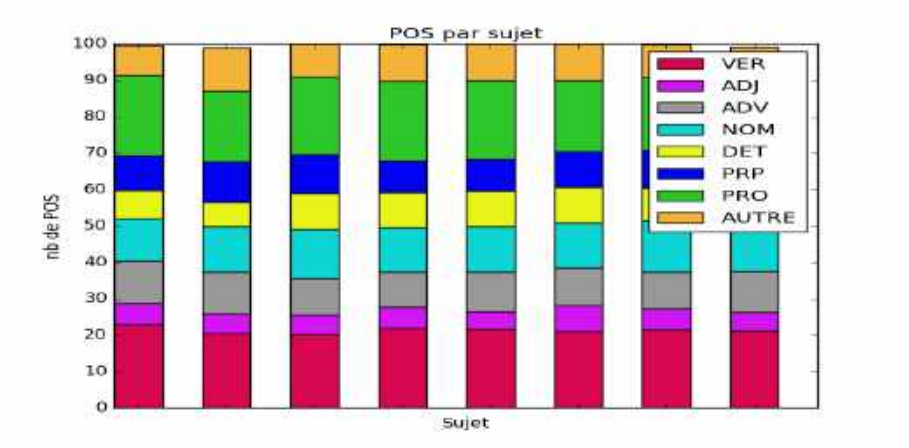

Figure 17: The proportion of POS used by psychologists

As demonstrated in the three figures, each category of POS tagger is presented by an unique color, while each bar presents the information of POS used in each file. From these figures, we can easily indicate that, the proportion for each category of POS is homogeneous in three group of speakers. Among the eight categories of POS taggers, **verb** and **pronoun** are the most frequently used categories, while **adjective** appears the least in the whole corpus.

As we have access to only a limited number of files in the two corpus, the proportion of POS can not demonstrate a valid comparison between three groups of speakers, thus in the next section, we will present the significance of each speakers and also the TTR (Type/Token Ratio) for each file.

Table 3 presents us the detailed information for significance between three groups, and each time we compare two of the three: patients and control group; patients and psychologists; control group and psychologists. As mentioned in section 2.4, the result will be considered as significant only when

Comapared groups		cat/word	cat/utter	form/word	lemma/word
City 1	Pa-Con	2.54	1.42	3.92	4.16
	Pa-Psy	2.42	1.16	4.19	4.36
	Con-Psy	0.13	0.28	0.33	2.41
City 2	Pa-Con	1.26	3.52	1.68	1.90
	Pa-Psy	0.26	1.32	1.02	1.08
	Con-Psy	1.53	2.17	0.53	0.68

Table 3: Significance between goups. Pa = Patient, Con = Control group, Psy = Psychologist. cat = POS category, word = number of words, utter = number of utterances, form = number of word forms, lemma = number of lemmas.

the s value is higher than 1.96. In the City 1 corpus, there is significant difference between patients-control group , and patient-psychologist group. For the patients-control group, there are significant difference in proportion of POS with word number, in the total word number and also the TTR (here the lemma refers to type, word refers to token). That is the same case in the patient-psychologist group. While in City 2 corpus, there is not so much difference neither between patients and control group , nor patients and psychologist group or psychologist and control group. The only difference lies in the proportion of POS category with each utterance between patients and control group and also between psychologist and control group.

Table 4 illustrates the gender difference in each group of speakers in City 1 corpus. The data shows us that, for the whole interviewee group (patients and control group), there is much difference between male and female speakers in the words number and also TTR. While, for the patient group, the gender difference lies in proportion of POS category and word number, the total

Comapared groups	cat/word	cat/utter	form/word	lemma/word
Pa+Con	1.23	0.98	2.93	2.80
Pa	1.22	3.75	2.78	2.97
Con	2.62	0.43	5.33	5.29

Table 4: Significance between male and female in different groups. This table is arranged similar to Table 3

	City 1	City 2
Pa	0.186	0.216
Con	0.140	0.185
Pa+Con	0.157	0.200
Psy with Pa	0.151	0.190
Psy with Con	0.134	0.205
Psy	0.141	0.198

Table 5: Type/Token Ratio for patients, control groups and psychologists

word number and also TTR. While for the control group, there is much significant gender difference in category and word ratio; total words number and TTR.

Here Table 5 shown above illustrates the TTR for each group of speakers in two corpora. As mentioned in section 2.2, the TTR presents the lexical variety used by speakers,and the higher the Value is , the more lexical diversity there is . In City1 corpus, there is more lexical diversity in patients (0.186) than in control group(0.140). Besides, compared with the psychologists in control group (0.134), the psychologists in patient group (0.151) has a little bit higher lexical variety. Whereas the same in City2 corpus, the patients group also enjoys the highest value (0.216) than the other speaker groups.

The reason for why there are a higher lexical variety in patients group needs to be explored in the further study.

5 Conclusion

During the present work the re-implementation of tool kit was performed. The input specification was produced. The pre-processing part made it possible to use the tool for both corpora(of City1 and City2). Xml structure for every dialog was constructed making it easier to extract statistical data about numbers of utterances, words etc. Application of MELT and Distagger tools was performed and the corresponding xml structure update was implemented.

The analysis part is able to produce the data produced by the later version of SLAMtk and also the new type of analysis based on sex of the patient speaker division was introduced. For further analysis, it is crucial to use not only small subcorpora, but more data to obtain more accurate and representative results.

References

[Amblard et al., 2015] Amblard, M., Fort, K., Demily, C., Franck, N., and Musiol, M. (2015). Analyse lexicale outill {\'e} e de la parole transcrite de patients schizophr {\e} nes. *arXiv preprint arXiv:1509.01539*.

[Blanc et al., 2010] Blanc, O., Constant, M., Dister, A., and Watrin, P. (2010). Partial parsing of spontaneous spoken french. In *LREC*.

[Blanche-Benveniste and Jeanjean, 1987] Blanche-Benveniste, C. and Jeanjean, C. (1987). *Le français parlé: transcription et édition*. Éditions Interco.

[Denis and Sagot, 2012] Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language resources and evaluation*, 46(4):721–736.

[Feldstein, 1962] Feldstein, S. (1962). The relationship of interpersonal involvement and affectiveness of content to the verbal communication of schizophrenic patients. *The Journal of Abnormal and Social Psychology*, 64(1):39.

[Kremen et al., 2003] Kremen, W. S., Seidman, L. J., Faraone, S. V., and Tsuang, M. T. (2003). Is there disproportionate impairment in semantic or phonemic fluency in schizophrenia? *Journal of the International Neuropsychological Society*, 9(01):79–88.

[Maher, 1972] Maher, B. (1972). The language of schizophrenia: A review and interpretation. *The British Journal of Psychiatry*, 120(554):3–17.