A tool for automatic analysis of linguistic clues in dialogue transcripts

Olga Leticevscaia, Ruixue Liu and Yiqing Liang

UNIVERSITÉ DE LORRAINE

Objective

The project aims to create tool kit for automatically performing analysis on **POS** and **disfluency** of the dialogue transcripts with pathologic speech content.

Annotation Tools

Tools for annotation:

MELT: Tool for POS (Part-Of-Speech) and lemma annotation.

```
spk1    voilà/V/voilà
spk1    donc/CC/donc ça/PRO/cela va/V/aller
spk2    oui/NC/oui oui/ADV/oui
spk1    c'/CLS/ce est/V/être c'/CLS/ce est/V
ce est/V/être c'/CLS/ce est/V/être un/DET/u
ce est/V/être exploratoire/ADJ/exploratoire
on/CLS/cln teste/V/tester on/CLS/cln teste/
*calibration c'/CLS/ce est/V/être un/DET/un
V/falloir que/CS/que
spk2    d'/P/de accord/NC/accord
spk1    ceux/PRO/celui qui/PROREL/qui conçoi
les/DET/le caméras/NC/caméra fassent/VS/fai
beaucoup plus/ADV/plus beaucoup/ADV/beaucou
c'/CLS/ce est/V/être pas/ADV/pas gagné/VPP/
```

Figure 1: Fragment of the output file of MELT.

Distagger: It's used for disfluency annotation in speech content. The next picture shows the example of Distagger result file. The example of output is present on the picture below.

```
{S}{#4,.IGN+slot} {spk1,.IGN+speaker} elle parlait tellement doucement que j' ent
{S}{#5,.IGN+slot} {spk2,.IGN+speaker} d' accord c' est parti {S}
{S}{#6,.IGN+slot} {spk1,.IGN+speaker} alors oui donc {euh,.IGN+EUH} j' aimerai qu
{S}{#7,.IGN+slot} {spk2,.IGN+speaker} que je fais {euh,.IGN+EUH} {S}
{S}{#8,.IGN+slot} {spk1,.IGN+speaker} oui vos occupations {S}
{S}{#9,.IGN+slot} {spk2,.IGN+speaker} mes activités {S}
{S}{#10,.IGN+slot} {spk1,.IGN+speaker} oui {S}
{S}{#11,.IGN+slot} {spk2,.IGN+speaker} bon je travaille le soir je travaille donc le s
{S}{#12,.IGN+slot} {spk1,.IGN+speaker} {euh,.IGN+EUH} oui donc {euh,.IGN+EUH}
{S}{#13,.IGN+slot} {spk2,.IGN+speaker} au niveau du boulot {S}
{S}{#14,.IGN+slot} {spk1,.IGN+speaker} oui {c' est,.IGN+REP} c' est quoi c' est la ve
{S}{#15,.IGN+slot} {spk2,.IGN+speaker} je suis responsable d' une salle en fait {S}
```

Figure 2: Fragment of the output file of Distagger.

Corpora

We have access to two corpora consisting **18** transcriptions of pathologic speech, from both **male & female;patients & control group**, . They are organized differently in the file layout and folder structure.

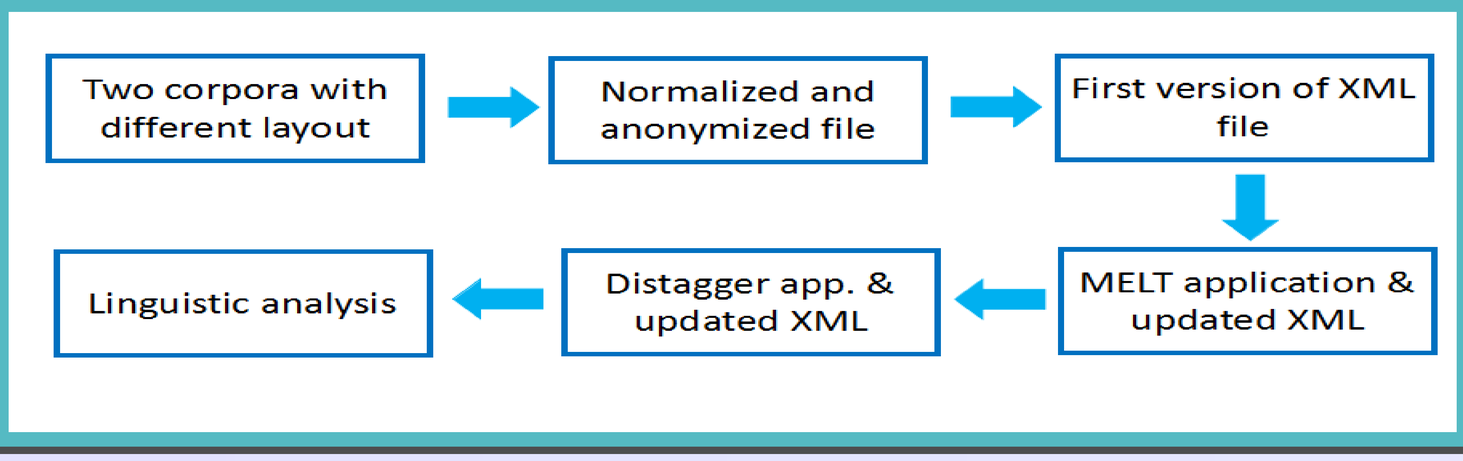
References

[1] Maxime Amblard, Karën Fort, Caroline Demily, Nicolas Franck, and Michel Musiol. Analyse lexicale outill {\'e} e de la parole transcrite de patients schizophr {\e} nes. *arXiv preprint arXiv:1509.01539*, 2015.

Acknowledgements

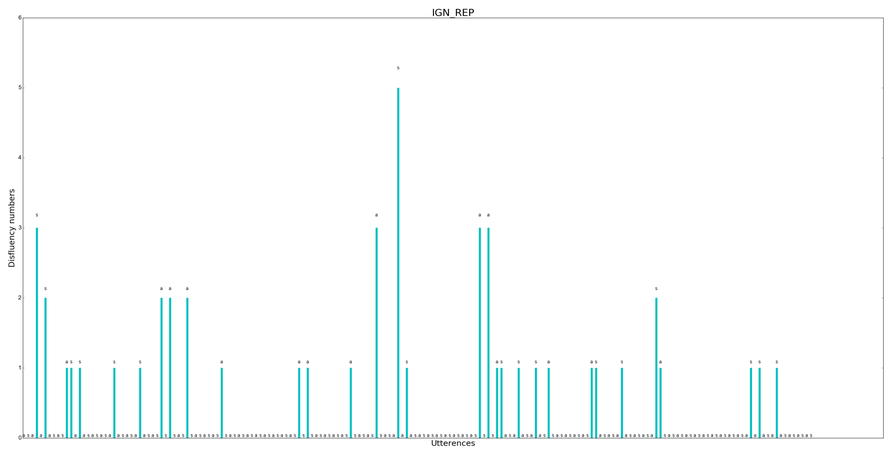
We would like to express our great appreciation to our superviser Maxime Amblard for his guidance and assistance during the project elaboration and for offering us access to the SLAM project corpora.

Tool implementation procedures



Disfluency and POS tagging results

For the disfluency analysis several types of data aggregation and latex/pdf table generation were implemented as well as the figures for every type of disfluencies(disfluency of 'euh',short pause,repetition, fragmentation;self-correction) or their sum.



Male				
Disfluency types	Numbers		nb to words	
	S	T	S	T
EUH	634	2755	0.015718	0.052075
REP	770	1811	0.019090	0.034231
CORR	22	59	0.000545	0.001115
Short-pause	34	183	0.000842	0.003459

Female				
Disfluency types	Numbers		nb to words	
	S	T	S	T
EUH	116	595	0.018110	0.029974
REP	171	789	0.026697	0.039748
CORR	5	19	0.000780	0.000957
Short-pause	24	28	0.003747	0.001410

Figure 3: The figure (left) shows the relation of REP disfluency number per utterance in the speech transcript; The table(right) summarizing disfluency data separately for Male and Female speakers.

The POS analysis focuses on the analysis of POS taggers and lemmas of each word in each corpus. 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.

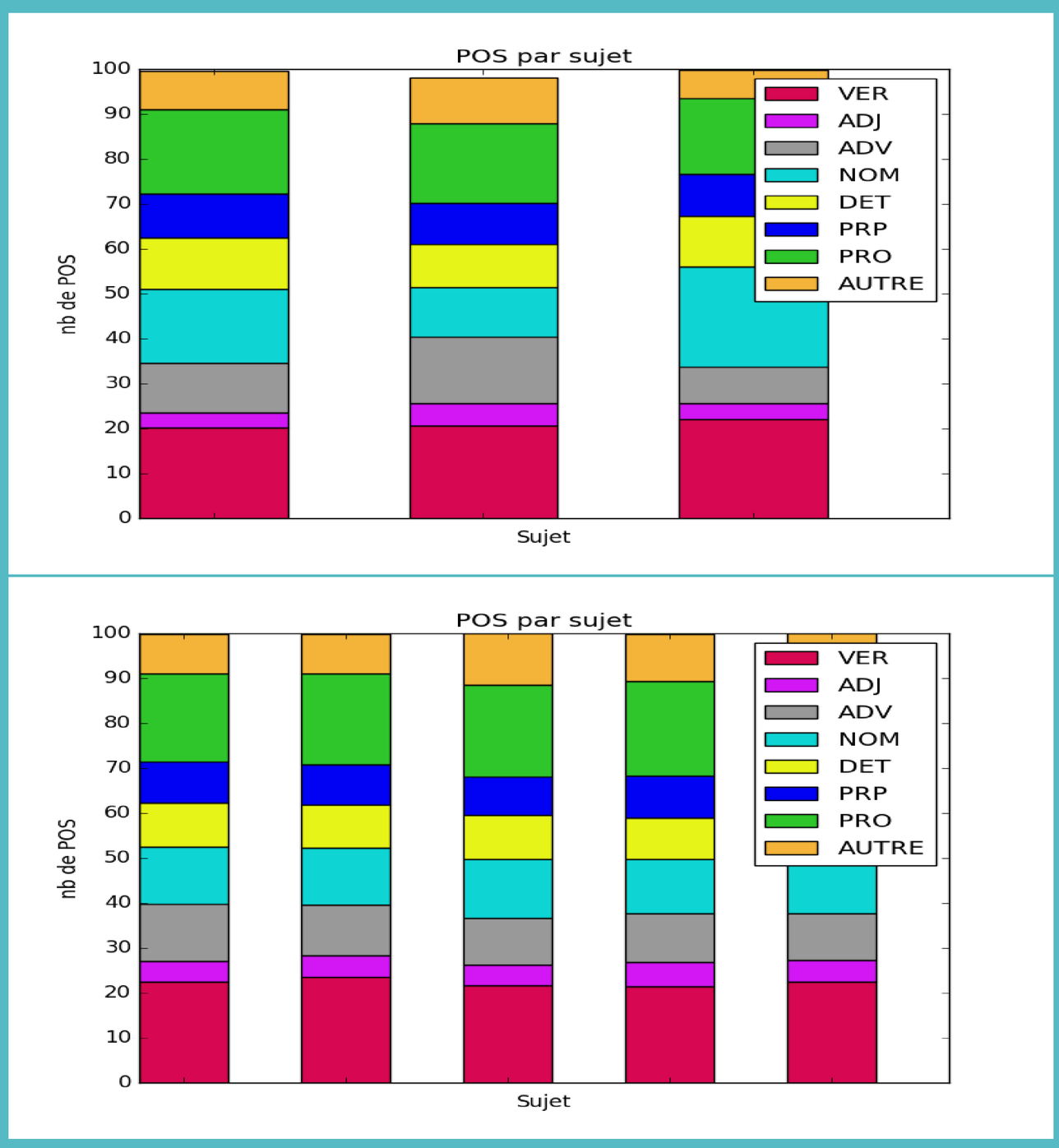


Figure 4: Proportion of each POS for two groups, schizophrenic patients (up) and control group (bottom)