NATURAL LANGUAGE PROCESSING

1. REQUIREMENTS

1.1 Introduction

1.2 Software

1.3 File

1.4 The Project Directory Structure

2. START

2.1 SPPAS

2.2 PRAAT

3. WORKING ON THE FILE

4. JAVA

5. STATISTICS

6. REFERENCES

1. REQUIREMENTS

1.1 Introduction

Natural language processing is a word and a sentence tokenization, text classification and sentiment analysis, spelling correction, information extraction, analysis, i.e the extraction and response to questions.

In this work, we will introduce the underlying theory of probability, statistics, and machine learning that are crucial for the field, and cover fundamental algorithms like n-gram language modeling naïve Bayes and MaxEnt classifiers, sequence models like hidden models Markov, probabilistic dependency and constituent analysis and vector space models of meaning.

1.2 Software

For this project, these software packages are needed:

SPPAS
http://aune.lpl.univ-aix.fr/~bigi/sppas/download.php

AUDACITY
http://audacity.sourceforge.net/

PRAAT
http://www.fon.hum.uva.nl/praat/

The other important software packages are:

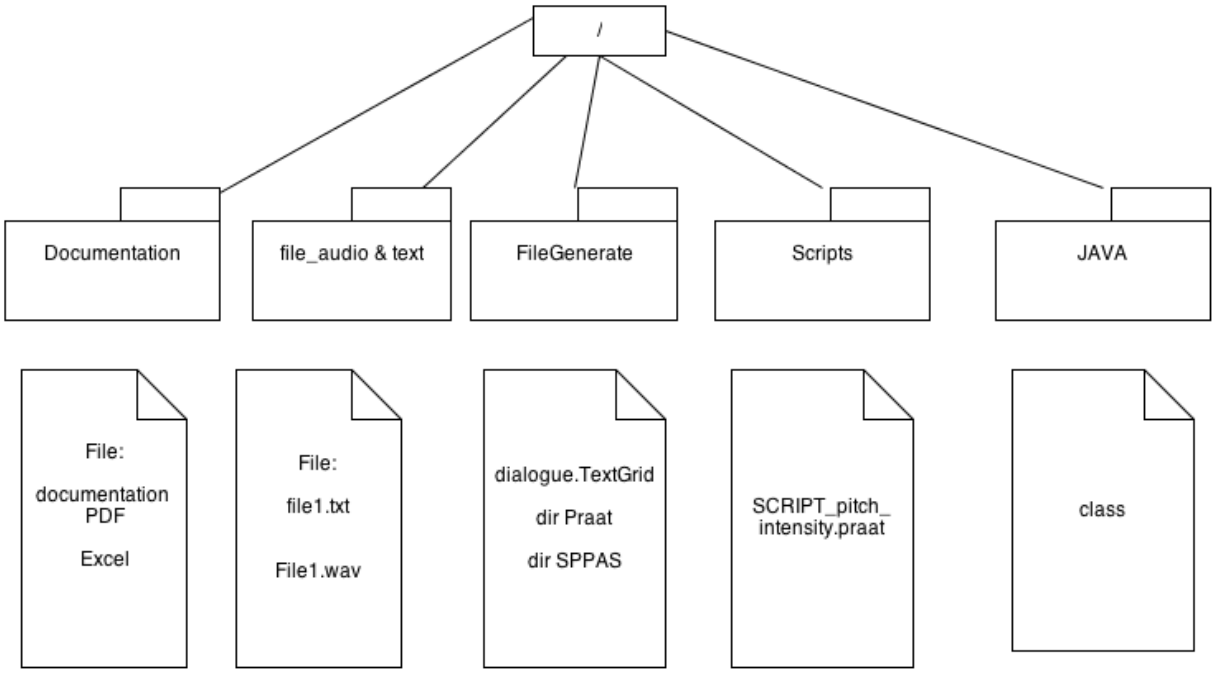
EXCEL, ECLIPSE, JDK

1.3 File

For the project, we needed one file with format .wav or with extension .mp3. In the case of using the .mp3 file, it was necessary to convert with audacity.

There is a dialogue between two people in the audio file and for this dialogue we also created the file with extension .txt, which is the transcript of this dialogue.

1.4 The Project Directory Structure:

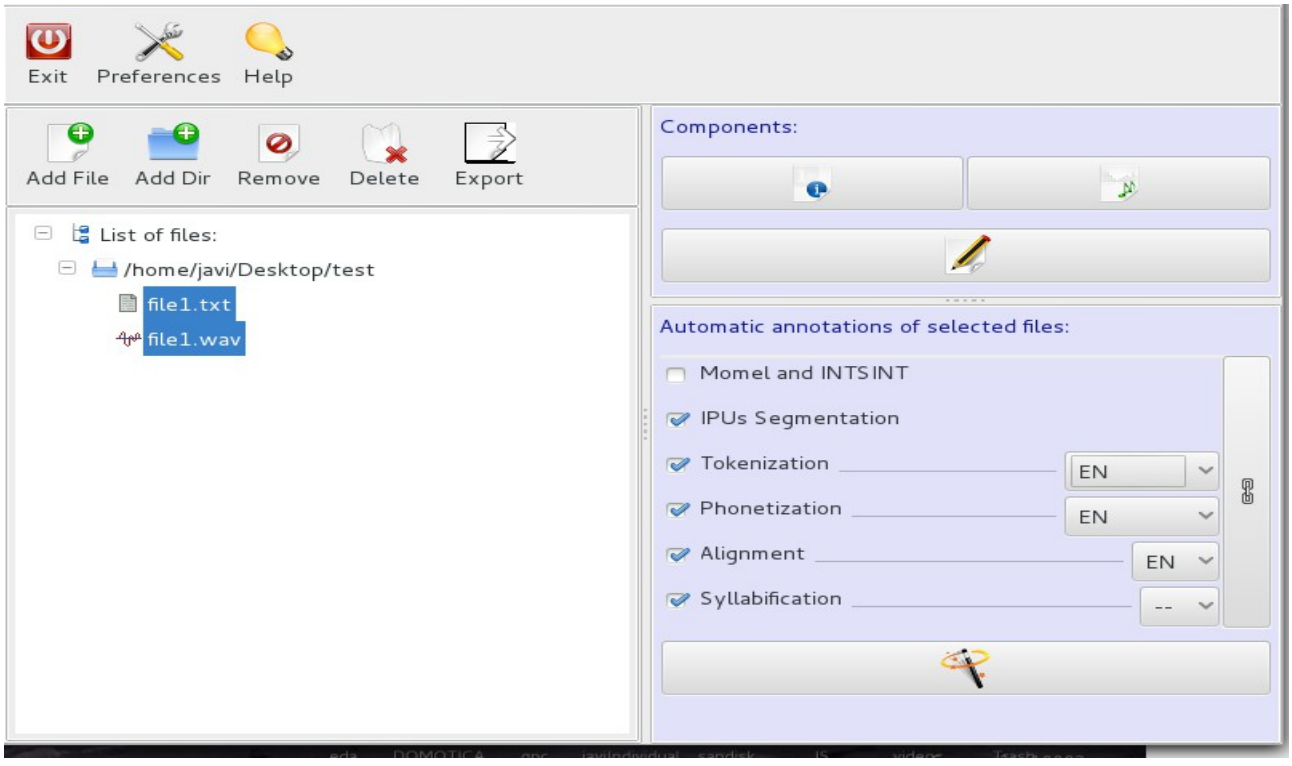


2. START

2.1 SPPAS

The audio files (. Wav) and text file (. Txt) should have the same name, since the software may have some problems recognize them.

We have to select the two files and mark all the options as we can see it on the picture below:



This program generates series of files which are listed below. For our project, we just need the ones which are highlighted.

file1-merge.TextGrid

file1-phon.palign.TextGrid

file1-phon.TextGrid

file1.TextGrid

file1-tokens.TextGrid

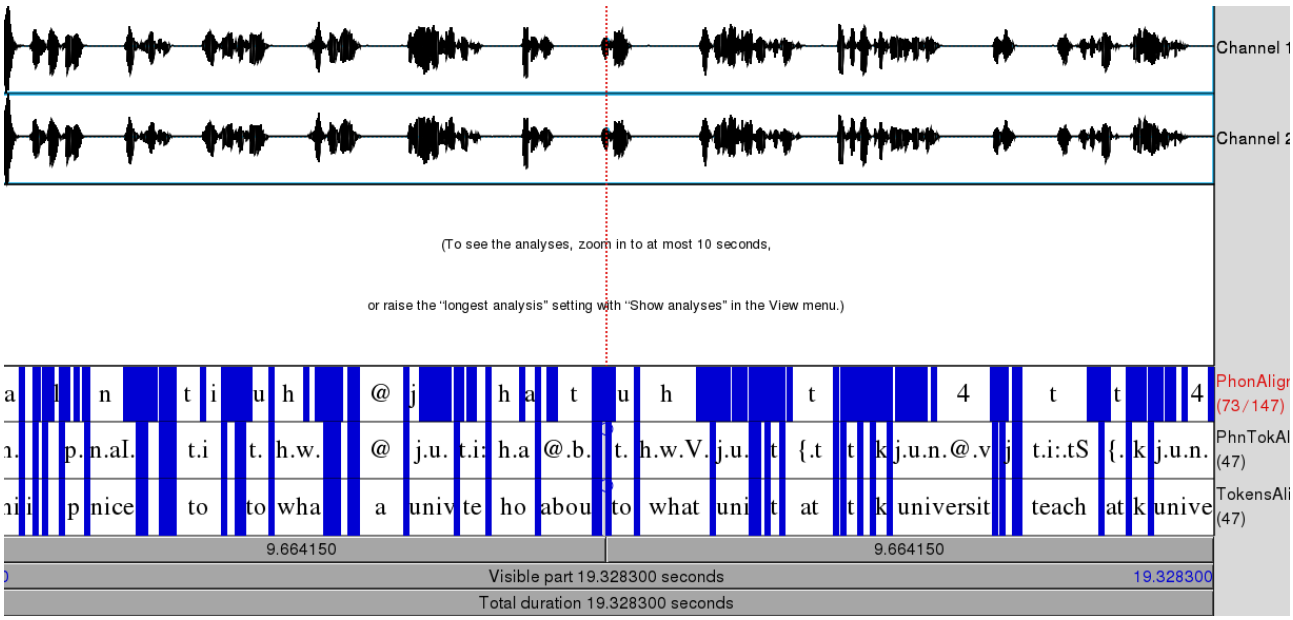
file1.txt

file1.wav

2 .2 PRAAT
We have to open the Praat program and introduce the following files:
Open/Read From File

file1-phon.palign.TextGrid

file1.wav
We select the two files together and go to the View & Edit button.



In the picture on the previous page, we can see the audio waves through the uploaded file SPPAS,

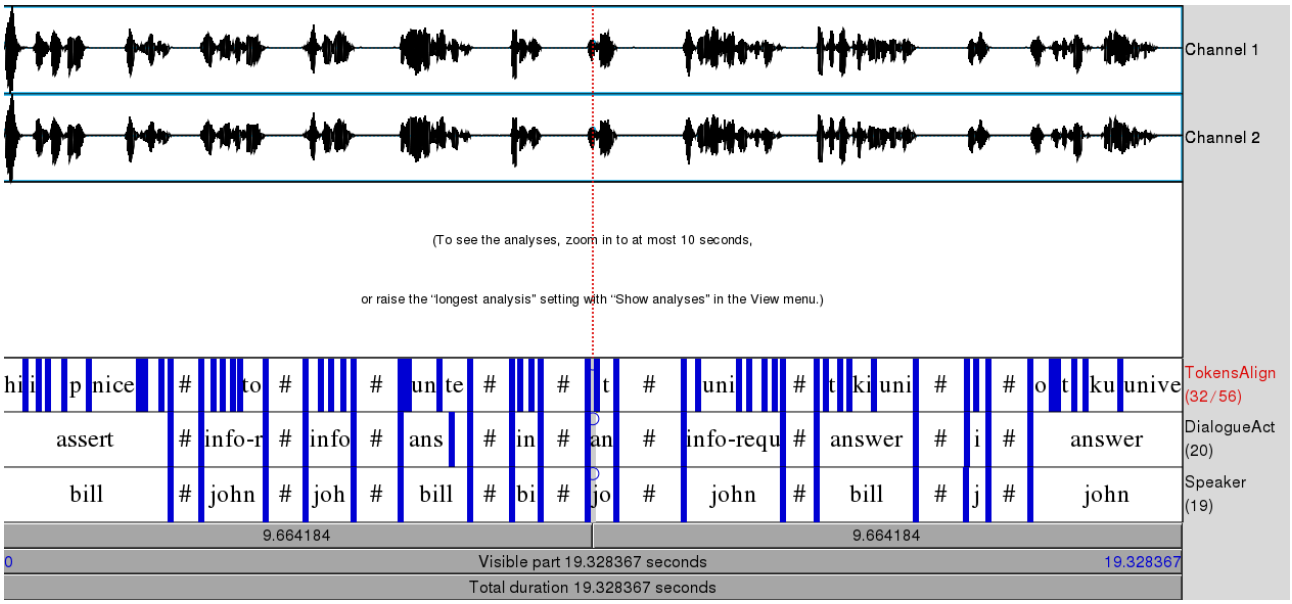
which will stay with TokensAlign, the other above must be all removed from the menu:

Tier/Remove Tier

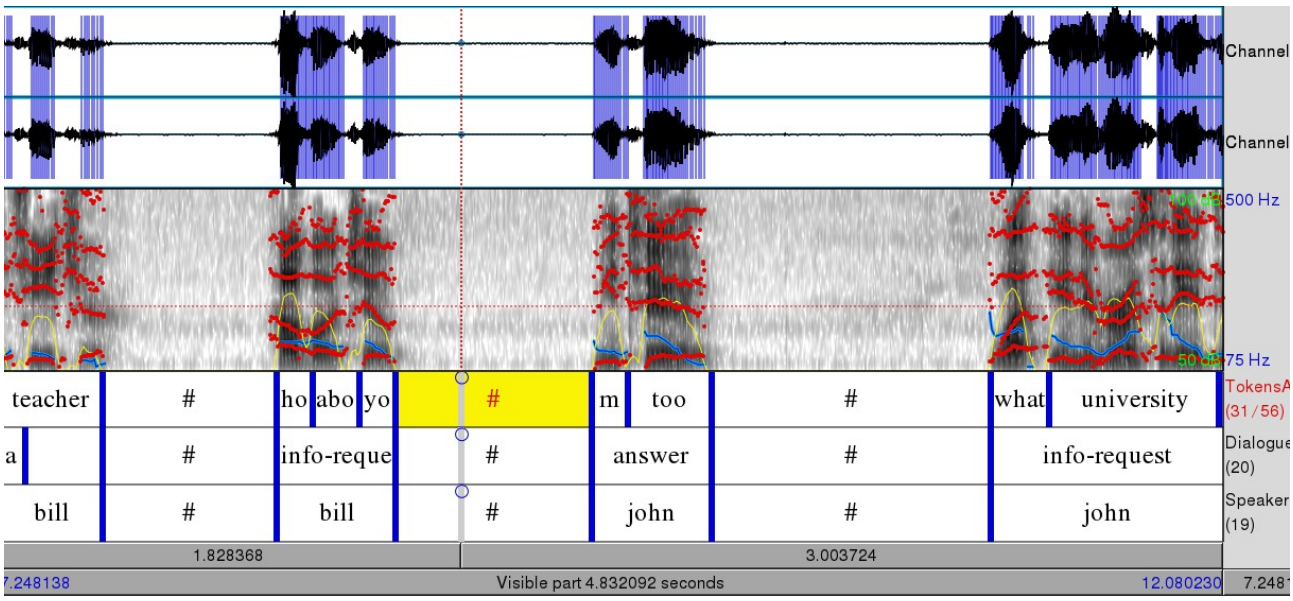
Next, we will insert two Tiers, called:

Dialogue Act
Speakers

Next we need to adjust the timing of each TokenAlign with their respective DialogueAct and speaker.
(If there is the audio which is minimal or empty, we have to introduce the "#")



We can see as approximate separation times as possible on the picture by ZOOM.



Once we managed to align everything with the audio, the tiers must be saved in a file TextGrid
We continue to work with it.
Once we have generated the files, we will work on them.

3. WORKING ON THE FILE

Open the Praat and load only the files the audio.wav and .TextGrid.
We will run the script we created. (script.praat)
We will get an output as we can see in the photo below.

Where we can see these tags there:

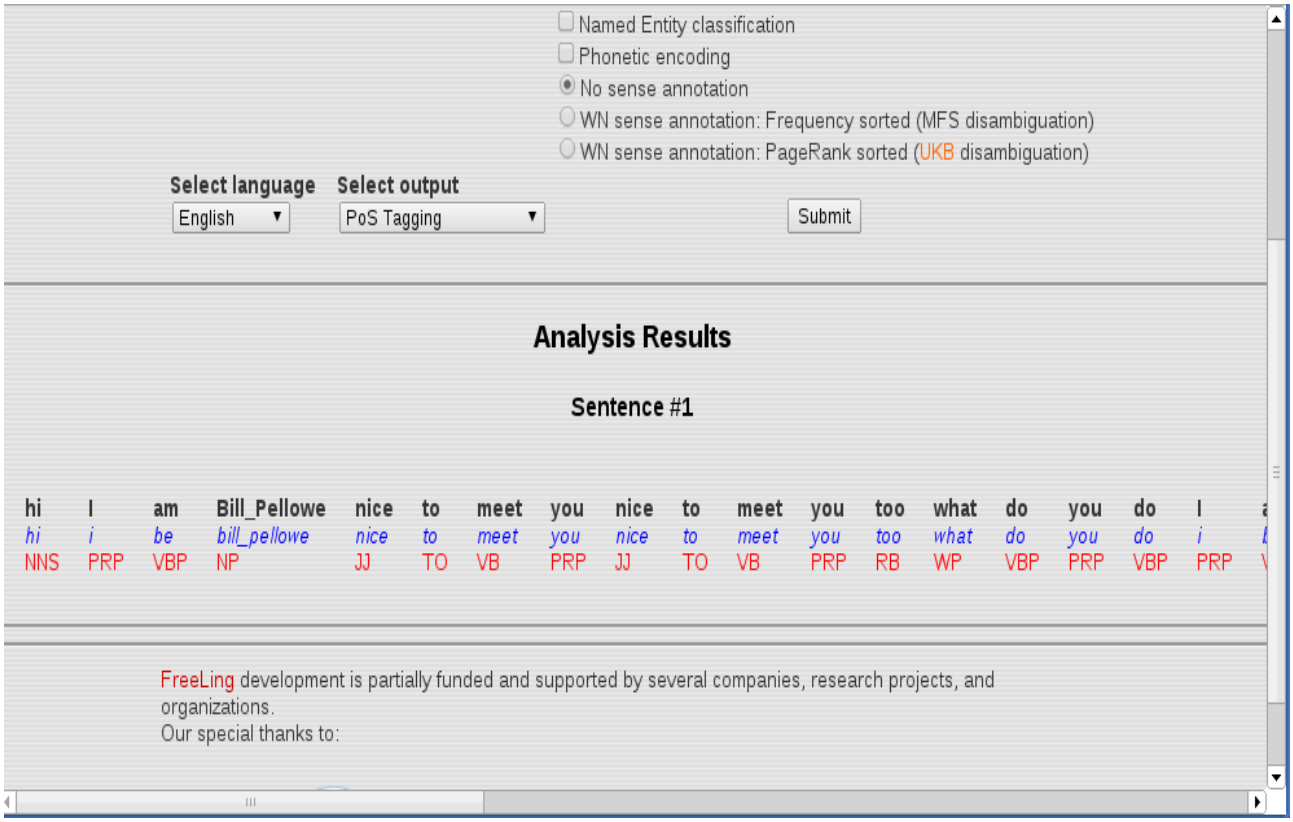
- Token	- Speaker	- Intensity
- Pitch	- DialogueAct	-

We can save this output to a file for later use in the graphs.
Generated POS TAGGER

We select our text file with the dialogue and devote ourselves to the following link:

http://nlp.lsi.upc.edu/freeling/

There is a demo:
We select the language: English and option-TAGGING POS



This will be the result, which we have to add (the POS TAGGING) to our output file generated by the script.

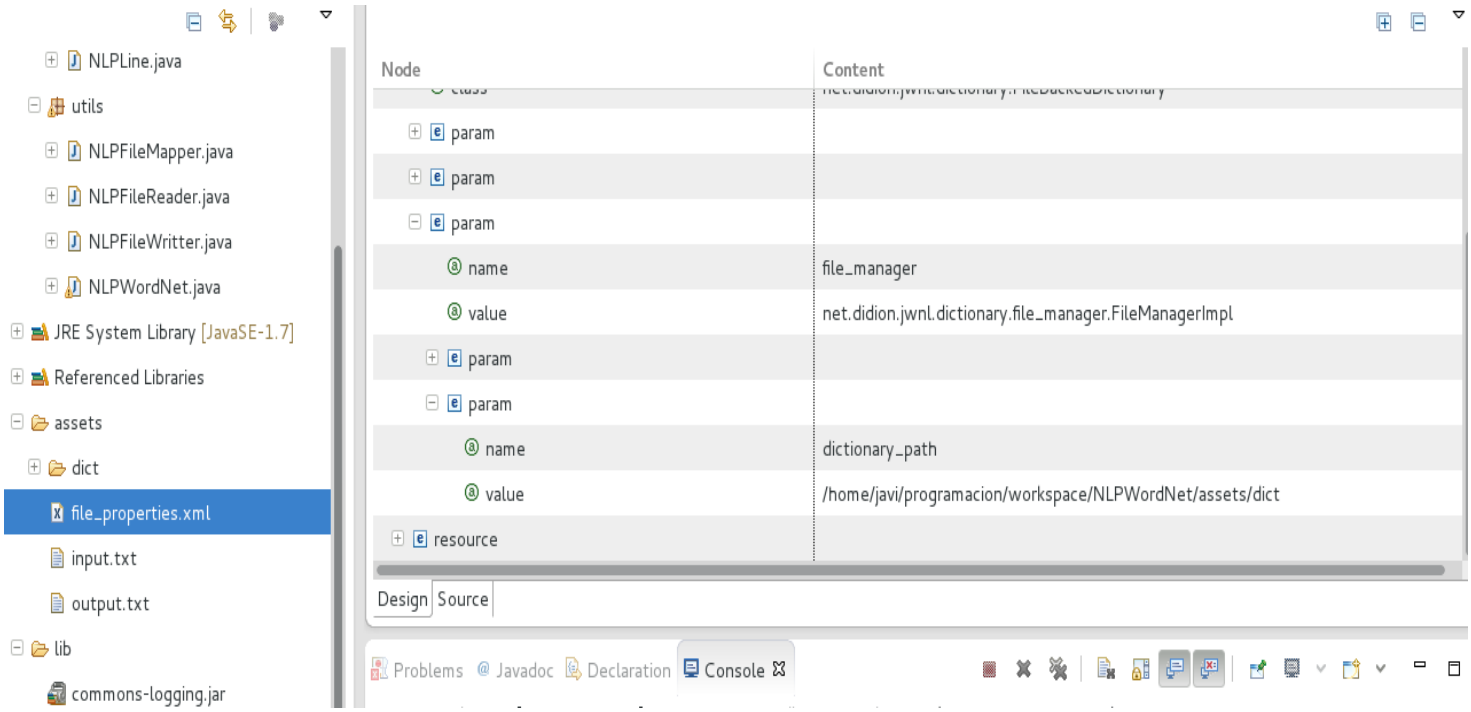
We can see its contents in the excel file attached.

	A	B	C	D	E	F
1	Token	Pitch	Intensity	Speaker	ialogueAc	POS
2	hi	133	60	bill	assert	NNS
3	i	135	44	bill	assert	PRP
4	am	124	65	bill	assert	VBP
5	bill	123	61	bill	assert	NP
6	pellowe	113	59	bill	assert	NP
7	nice	123	39	bill	assert	JJ
8	to	119	50	bill	assert	TO
9	meet	125	60	bill	assert	VB
10	you	100	56	bill	assert	PRP
11	nice	150	65	john	info-request	JJ
12	to	136	55	john	info-request	TO
13	meet	129	64	john	info-request	VB
14	you	133	61	john	info-request	PRP
15	too	133	61	john	info-request	RB
16	what	152	61	john	answer	NP
17	do	147	62	john	answer	VBP
18	you	161	66	john	answer	PRP
19	do	104	58	john	answer	VBP
20	i	153	65	bill	info-request	PRP

4. JAVA

We need to download the software Eclipse, JDK, libraries (there is a link in the references).

Configuration:

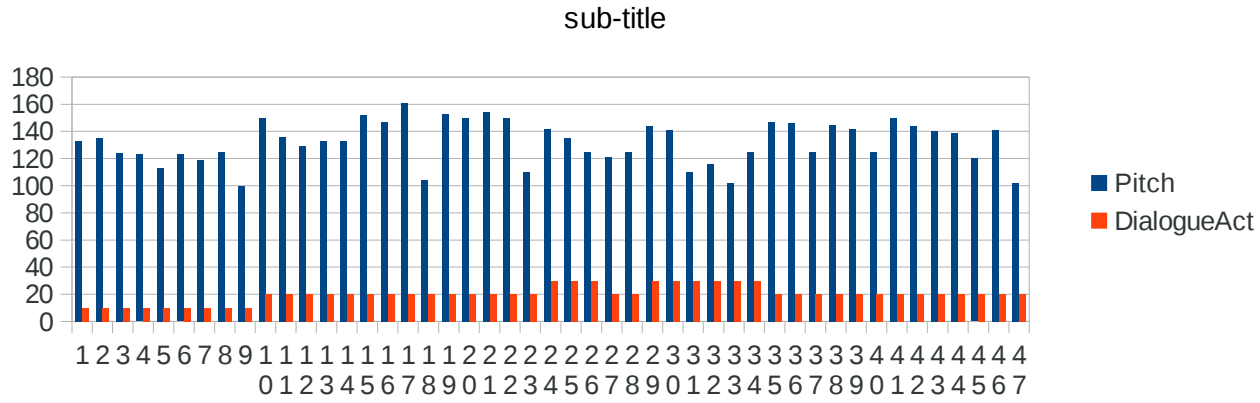


We need to edit the file_properities.xml and we need to modify the third parameter (value) in the dictionary path.
We can use a path absolute.

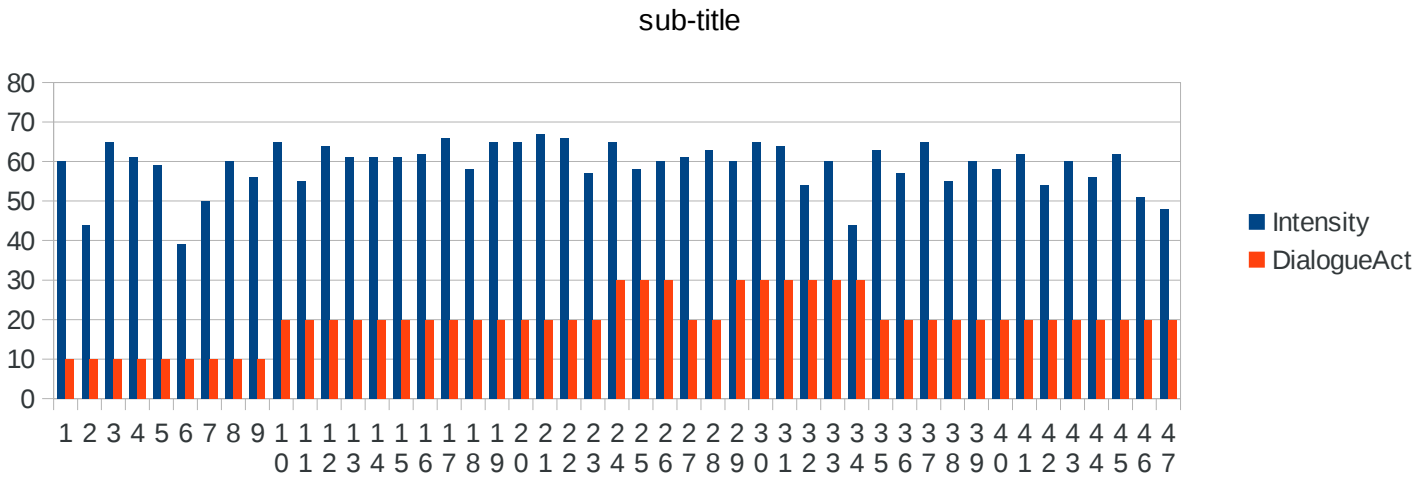
5. STATISTICS

This graph shows how the pitch changes regarding the different dialogueAct present in the conversation.

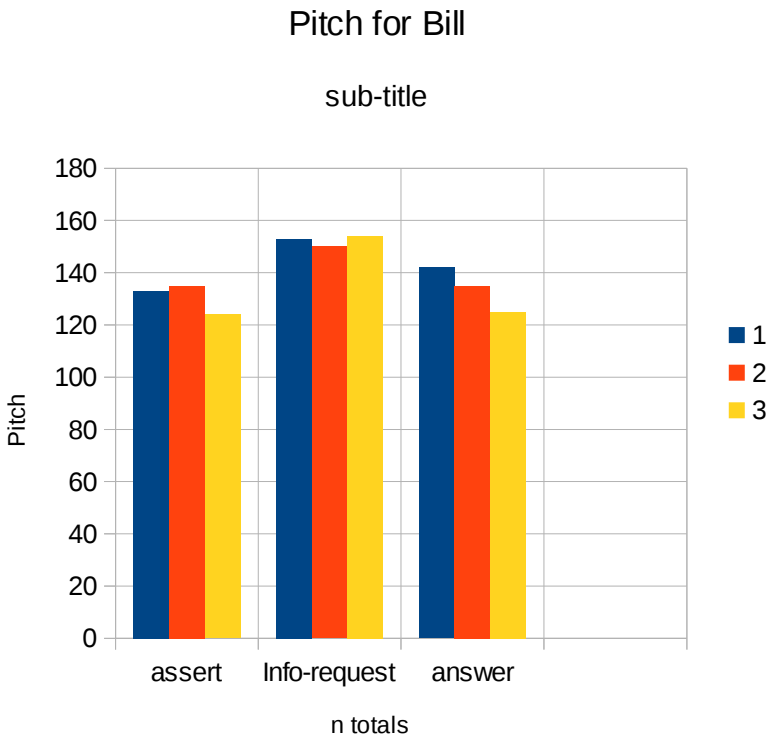
PITCH WITH DIALOGUEACT

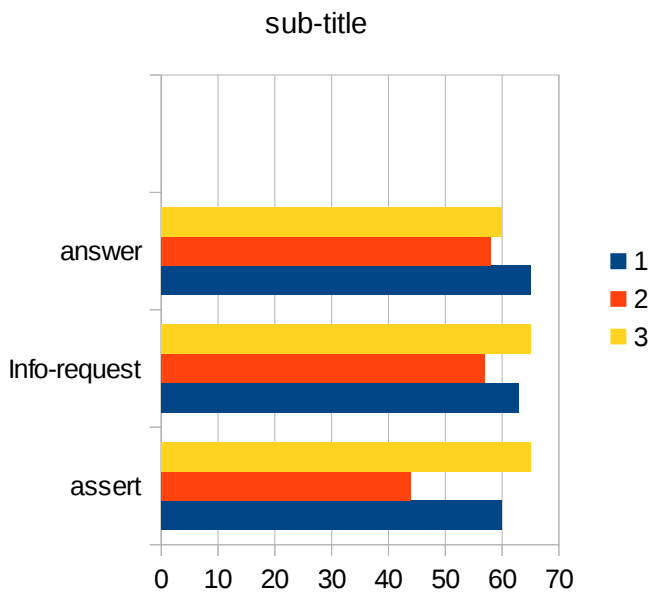


This is the common intensity ratio of the two speakers on the dialogueAct.

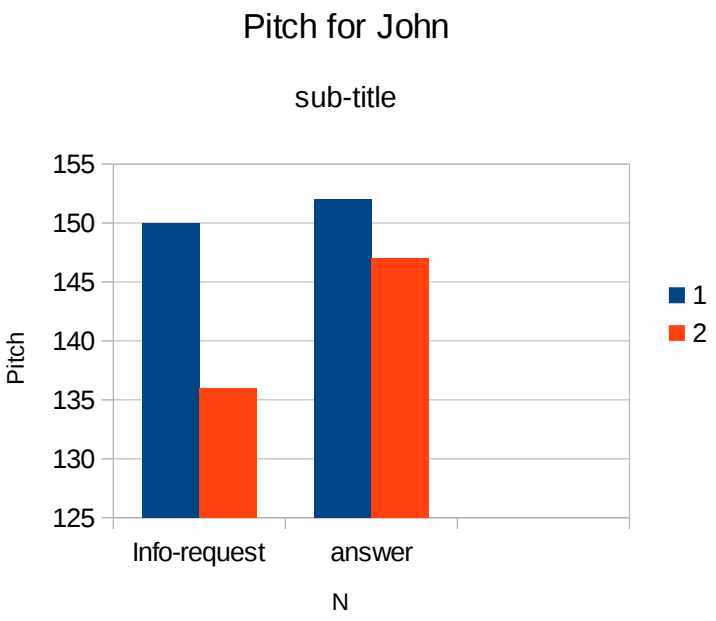


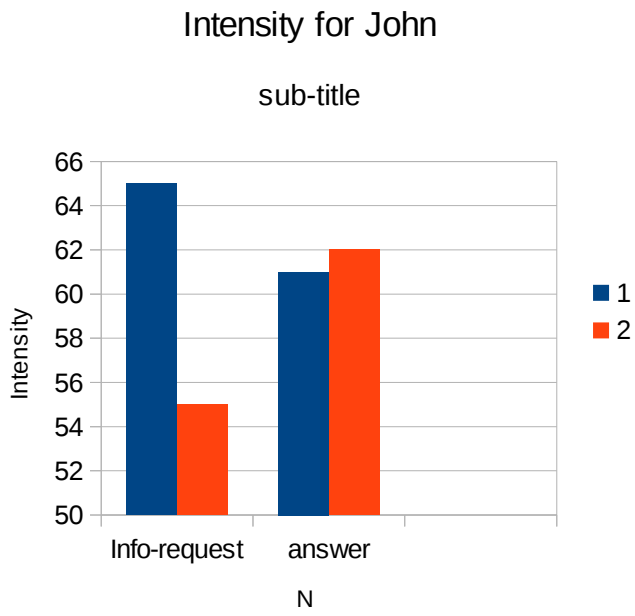
In this graph, we can see a comparison of the average pitch and average intensity per dialogueAct per speaker.





These are the STATS for John





We can see that the intensity changes Pitch John in answer field.

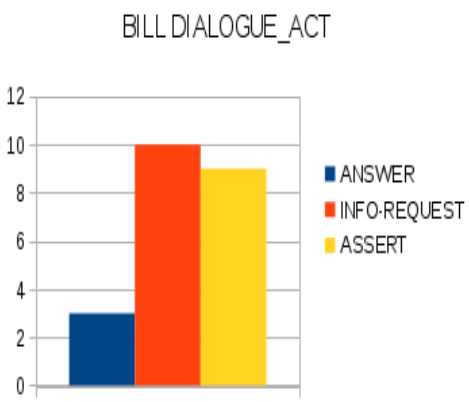
Here we can see the number of dialogueActs of each type uttered by each speaker.

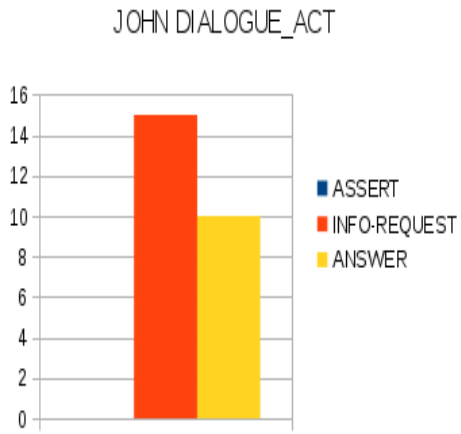
BILL

ANSWER	3
INFO-REQUEST	10
ASSERT	9

JOHN

ASSERT	0
INFO-REQUEST	15
ANSWER	10





6. REFERENCES

SPPAS
http://aune.lpl.univ-aix.fr/~bigi/sppas/download.php

AUDACITY
http://audacity.sourceforge.net/

PRAAT
http://www.fon.hum.uva.nl/praat/

POS TAGGING
Freeling
http://nlp.lsi.upc.edu/freeling/

WORDNET
Wordnet Download
http://wordnet.princeton.edu/wordnet/download/current-version/

WordNet with Java
http://shiffman.net/teaching/a2z/wordnet/

Foro
http://stackoverflow.com/questions/13881425/get-wordnets-domain-name-for-the-specified-word

Project that add domain at words with databases WordNet
http://wndomains.fbk.eu/

Library of WordNet para Java
http://sourceforge.net/projects/jwordnet/