Agata Dominowska, 014137590
Introduction to methods in digital humanities, autumn 2016
Final project

1. Dataset

The data is connected with my Master's research, where I analyse Finnish address forms used by foreigners who know the language on an intermediate level. The research data consists of recordings from the spoken part of the final test developed in Testipiste (http://www.testipiste.eu/FI/testipisteesta), a ESF-funded co-operation project on language testing, which took place at a few adult education institutions in 2011-2013. There are 4 different sets of tests.

The part of the recordings I am interested in is a task where the students react to a pre-recorded instruction. The instruction includes a written presentation of the situation on a test paper (e.g. *You are eating lunch in a restaurant. Something happens. Tell the waiter about it. Ask for help*., accompanied by a picture) and a spoken presentation on the tape (the same text is read and there is a sound suggesting what is happening: a sound of broken glass/plate). The person undertaking the test has 20 seconds to reply before moving on to the next situation. The situations that I chose for the thesis dataset are those where the hypothetical addressee is "unknown" to the speaker, which implies a degree of distance. The hypothesis is that foreigners speaking Finnish at an intermediate level are probably using more formal forms and more overt politeness in comparison with Finnish speakers (whom I am going to record later to collect the second dataset; I will also comment on my results using earlier research data based mostly on interviews/questionnaires, but this is not part of this project).

The dataset I collected for this final project is smaller than the thesis dataset. It is based on 57 recordings which I transcribed. The recordings are not available without consent from Testipiste, but the transcripts are available as attachment to the project. The principles of transcription are explained in more detail in part 2.

My goal for this project is to find and analyse the forms that indicate the speaker's choice of address form. The possible forms taken into account are:

- **second singular *you*** (a verb – indicative or imperative - or a personal pronoun), indicating the preference for less formal and more standard form, *sinuttelu* in Finnish;
- **second plural *you*** (also a verb in indicative or imperative, or a personal pronoun), indicating that the speaker uses forms considered more formal, distant, and nowadays used when addressing much older people or customers, but also possibly indicating respect – *teitittely* in Finnish;
- **name of occupation/position** of the addressee, such as *teacher, waiter*.

I also chose to search for overt politeness forms, such as *anteeksi* or *kiitos*, because earlier research indicates that Finns might be using less of them than foreigners and I want to include an analysis of that in the thesis.

There are recordings where none of these are used, but the speaker emphasizes their own perspective and/or avoids using an overt *you* form, only using 1st person singular verbs or pronouns. There are also recordings where no forms are used, but contain only a couple of words. Recordings where nothing is being said were discarded.

2. Data processing

Recordings: preprocessing and transcription

All recordings were first preprocessed manually using Audacity (http://www.audacityteam.org/). Each recording is long (20-25 minutes) and contains the entire speaking test, while I was interested only in some parts of one task, each lasting around 45 seconds (task description and 20s reply), so I cut out the ones that interested me.

Then I transcribed 3 batches of recordings. All transcripts are one-line .txt files. The typical content of a file looks like this:

```
voitko auttaa minua tarjoilija tarvitsen apua lautanen rikki
```

The transcription is simplified: it does not take into consideration any prosodic features, duration of breaks in speech, indication of hesitation, stuttering, or any other of the commonly used features (see e.g. http://www.helsinki.fi/hum/skl/ca/merkit.pdf). I also did not account for small pronounciation mistakes that do not affect the meaning (e.g. in the example above, that would mean pronouncing *rikki* as /ri.ki/), however, I wrote down obvious mistakes (e.g. *voitko auttaa *minulle*) and especially the forms where the speaker seems to confuse the pronouns or the situation is entirely unclear (like *tarjoilija voiko auttaa*, where it is very probable that the speaker meant *voitko* or, with less probability, *voinko,* but they clearly say *voiko*). All words are transcribed in small letters, and no punctuation marks are used with 3 exceptions: (1) the exclamation mark, which is used to indicate imperative mood, (2) *x* added at the end of the words ending with *t* and not being the 2nd person indicative verb, in order for the regular expression to find only the accurate forms, and (3) capital letters for the name (e.g. *hei olen A*). This clean form of data makes it more easily searchable, even though it would be possible to construct regular expressions in such a way that they return all forms of a word also if there would be transcription signs in between.

All recording files and all corresponding transcript files are named according to the following scheme: SCHOOL_COURSE_DATE_PACKAGE_TASK_PARTICIPANTid_PARTICIPANTgender.txt.

Python script

Using Python 3.5.2 (and the IDLE environment), I wrote a script that does the following things:

1. gets all text files from the directory where they are stored,

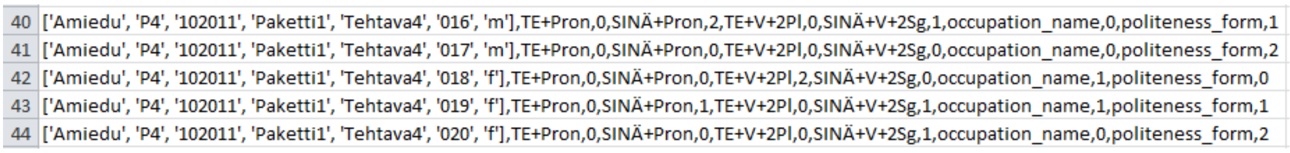
2. creates a list of all file names containing a full path and a second one with file names stripped of the path and of the extension, then refers them to each other (with Python's dictionary structure),
3. opens, reads and closes each file, one file at a time,
4. splits the file name at underscore,
5. parses through the content of each file using regular expressions and (mostly as a testing method) prints the exact forms, the corresponding file name split into elements, and the number of forms found,
6. appends the results to a list of results, where each line consists of the file name elements, categories of results and their corresponding frequencies of occurrence in the file (also when there are none):
   ```
   ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '021', 'm'],TE+Pron,0,SINÄ+Pron,0,TE+V+2Pl,0,SINÄ+V+2Sg,1,occupation_name,0,politeness_form,2
   ```
7. writes the results to a .csv file, with each element separated by a comma,
8. exports the content of all the files to one big .txt file (for later processing as a corpus).

A more detailed explanation of the script commands is provided in the script as comments.

Excel

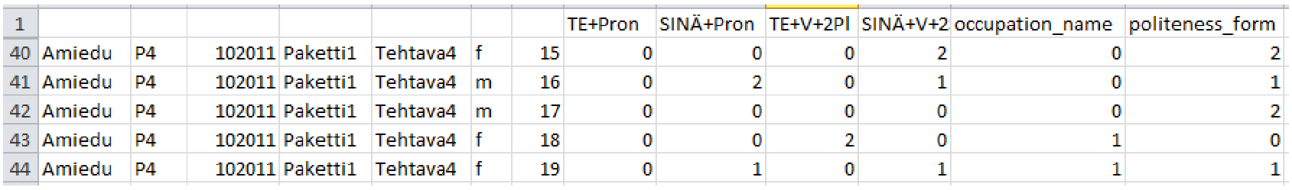
The data in the comma-separated .csv file was then reorganized. With the data-to-columns functionality, I separated the data into cells, then removed the commas and square brackets and moved the categories to the top row as headers, and saved the processed file in the .xlsx format.

The initial file looked like this:



```
40 ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '016', 'm'],TE+Pron,0,SINÄ+Pron,2,TE+V+2Pl,0,SINÄ+V+2Sg,1,occupation_name,0,politeness_form,1
41 ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '017', 'm'],TE+Pron,0,SINÄ+Pron,0,TE+V+2Pl,0,SINÄ+V+2Sg,0,occupation_name,0,politeness_form,2
42 ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '018', 'f'],TE+Pron,0,SINÄ+Pron,0,TE+V+2Pl,2,SINÄ+V+2Sg,0,occupation_name,1,politeness_form,0
43 ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '019', 'f'],TE+Pron,0,SINÄ+Pron,1,TE+V+2Pl,0,SINÄ+V+2Sg,1,occupation_name,1,politeness_form,1
44 ['Amiedu', 'P4', '102011', 'Paketti1', 'Tehtava4', '020', 'f'],TE+Pron,0,SINÄ+Pron,0,TE+V+2Pl,0,SINÄ+V+2Sg,1,occupation_name,0,politeness_form,2
```

And the processed file like this:



1								TE+Pron	SINÄ+Pron	TE+V+2Pl	SINÄ+V+2	occupation_name	politeness_form
40	Amiedu	P4	102011	Paketti1	Tehtava4	f	15	0	0	0	2	0	2
41	Amiedu	P4	102011	Paketti1	Tehtava4	m	16	0	2	0	1	0	1
42	Amiedu	P4	102011	Paketti1	Tehtava4	m	17	0	0	0	0	0	2
43	Amiedu	P4	102011	Paketti1	Tehtava4	f	18	0	0	2	0	1	0
44	Amiedu	P4	102011	Paketti1	Tehtava4	f	19	0	1	0	1	1	1

After that, I used Excel's sum function to add frequencies for plural *you* forms (TE+Pron, TE+V+2Pl) and singular *you* forms (SINÄ+Pron, SINÄ+V+2Sg), and to add frequencies for female/male speakers. I also used Excel to create charts.

Visualization is not the main goal of this project and there is lot to be said against using visual tools without a clear purpose, however, I also included word clouds from Voyant Tools (https://voyant-tools.org/). I removed stopwords such as *ja, on* etc. using search and replace in MS Word. I wanted to show that even a quick look at basic frequencies and frequency-based visualization can reveal the topics dealt with in the data.

3. Results and analysis

The results are only preliminary, as I did not have time to preprocess and transcribe all recordings (over 300). My data for this project includes 57 transcripts, which means that the analysis had to be carried out on a smaller scale than I had initially planned.

There are 3 different tasks included in this analysis. The first is about calling a language school to inquire about the language course (*You have received an invitation to a language course. Call the school*. *Say who is calling, ask at least two questions.*) , the second one involves asking a waiter for help (*You are eating lunch in a restaurant. Something happens. Tell the waiter about it. Ask for help*), and the third one involves calling about an ad concerning work (*There is a job ad in the newspaper. You are calling the number provided. Say who you are. Ask at least two questions*).

The most apparent result is that the second task ("ask the waiter for help") involves by far the most occurrences of all the forms. It may be seen on the first chart of overall frequencies (output_per_file_final.xlsx, "charts"-tab), which shows how the frequencies are distributed across the files and includes participant ID for the three tasks. The second task includes 28 out of 40 forms involving *you* as address forms (70%) and 28/32 politeness forms (87.5%). Also, all 5 address forms including occupation are located in files pertaining to task 2 (there is also one file in task 1 where the participant uses a first name given in the response-eliciting recording in order to address the school secretary, but I did not account for that).

When it comes to the distribution of singular vs plural *you* forms, it is interesting to see that singular *you* is overwhelmingly more popular (20/28 forms) and that the plural *you* is used only by female speakers (chart 2 and 3). Singular *you* forms are more popular than plural *you* forms overall.

Male speakers use only plural *you* forms in task 1, involving a call to the language school, but only singular *you* in task 2. The choice in task 1 might be influenced by the fact that they are addressing the institution rather than the fictional person-addressee, as in *what courses do you offer*, where *you* means the school. Female speakers seem to prefer singular *you* overall (it is the only form used in task 1 and task 3, and the more popular one in task 2). Moreover, only female speakers use the occupation name as address form (*tarjoilija*). It could be tentatively said that in the dataset female speakers use overt address forms more often than males.

Lastly, as an exercise in visualization and a bit of topic modelling, I wanted to see if it would be possible to guess the topic from the word frequencies and wordclouds based on them; I think it is very much so. Even the first 12 most frequent words in task 1 clearly reveal a topic connected to a course, where a student wants to register, and so does the wordcloud based on the 35 most frequent terms (https://voyant-tools.org/?corpus=6e0113900fe4d2e913fd125d73bd1fa1&visible=45&view=Cirrus):

Task 1		Task 2		Task 3	
minä	25	anteeksi	22	minä	8
kurssi	18	rikki	13	olen	4
olen	17	lasi	10	haluaisin	3
alkaa	15	apua	9	se	3
milloin	15	auttaa	8	tässä	3
haluan	14	minä	7	ei	2
se	11	mä	6	kysyä	2
minulla	9	tämä	6	lehdestä	2
mä	6	voisitko	6	milloin	2
terve	6	minua	5	minun	2
kielikurssille	5	minulla	5	tänään	2
missä	5	tarjoilija	5	töitä	2
tulla	5	vähän	5		
		voi	5		

For task 2 (https://voyant-tools.org/?visible=35&view=Cirrus&corpus=c38d569ce4c7245f94018e19daea37b5), the subject is probably less clear, as there is no mention of restaurant, and the waiter is mentioned only 5 times. It is, however, clear that the situation concerns a broken glass and the need for help.

The topic of the third task is not as clear due to the fact that the corpus is very small, but the most frequent words and the wordcloud could suggest that it involves asking for a meeting and concerns some job (https://voyant-tools.org/?visible=25&view=Cirrus&corpus=56c64a853503de084ffcd68f60172784).

This sort of modelling would be helpful in a larger dataset where only the transcripts would be available.

4. Further research

The time I spent on this project was mostly involved in preprocessing the data, improving my knowledge about Python and finding solutions to the problems I encountered along the way. The obvious first objective for further research is to transcribe the rest of the recordings and follow the same steps in the analysis. Also, I will rather compare relative frequencies than absolute ones in the case of female vs male speakers, as usually there are more females than males in a group.

As the analysis of this dataset is an important part of my Master's thesis, I want to undertake further steps to make the code more efficient (for example, now the script opens all transcript files twice, which is not ideal in any case). Other ideas that I have for the data are as follows:

- I want to look at the use of first person singular pronouns and verbs, because it seems to me that this is the most common usage in the recordings where none of the other strategies were used.
- It might be useful to include imperative as a separate category, as it is a form of direct address.
- Look at the possible first-name address, if it happens to occur further in the data (once until now).
- Possibly look for the files where both singular and plural *you* forms are used, which could indicate that the speaker's choice is not entirely conscious, and the hesitation could be attributed to the lack of skills.

I also want to apply some statistics, especially when I have the data for the Finnish recordings: I plan to use chi-square or Fischer's test to see if the differences between Finnish-non-native and Finnish-native datasets are significant.