Introduction to Methods in Digital Humanities, fall 2019
Dušica Božović

FINAL PROJECT: Computational analysis of the easy-to-read YLE news' language

1. Introduction

"With the term *easy-to-read* we refer to a simple form of language, used for communication with the three main user groups: people with intellectual or neurological disabilities, immigrants, and people with memory disorders. In Finland, *selkosuomi* (easy-to-read Finnish) and *lättläst* (easy-to-read Swedish) have been practiced since 1980's. Out of the 5.5 million people currently living in Finland, around 600 000 people are considered as potential users of easy-to-read." (Klaara-network)

The term "easy-to-read" in English is however a little bit misleading. Easy-to-read language has also its spoken form, so the Finnish term *selkokieli* ("clear language") is more precise. Even though I agree with this criticism, I will use the term easy-to read language in this report because it is the traditional way to call it and this is in the term that was in use at the Klaara 2019 Conference in Helsinki last year. In the end, I am primarily dealing with the written form of the language here.

Following the introduction of easy-to-read Finnish since the 1980s, the first PhD dissertation in Finland was published in 1998. All in all, there are 30-40 dissertations completed about it in Finland so far. Right now, there are 4 dissertations on this topic, and it seems that the interest in easy-to-read Finnish has grown in the recent years. (Vanhatalo) If we look into the previous research topics, it seems that the most popular kind of focus is on literature, mostly comparing "general" (*yleiskieli*) and easy-to-read version of a literary piece. However, I have not noticed a lot of computational methods using in this field.

Easy-to-read Finnish can be a translation from the general language or texts can be originally written in easy-to-read language. YLE news in easy-to-read Finnish is a form that is written based on a general language text. YLE news in easy-to-read Finnish have been made since 1992, and they were originally made for Finnish speakers born abroad, who had difficulties following regular program on the radio. Nowadays their target group is

immigrant population and that affects some language choices, such as greater use of loanwords. (Kielikello)

Finnish Center for Easy Language Selkokeskus has different services, one of which is granting the SELKO symbol (Finnish Centre for Easy Language 2018) on different texts in order to recognize that the text was done according to their standards. They have a set of recommendations for writing texts in easy-to-read Finnish. They also published the Selkomittari (Measure for easy-to-read language) in 2018 (Kehitysvammaliitto). Its purpose is to evaluate already made text and not to serve as guidelines, although it also reflects what is desirable to do when writing an easy-to-read text. Both writing and assessment of texts is normally done by humans even though the Measure uses quantitative methodology and could easily be digitalized and automated.

The first aim of this project is to assess how well are recommendations for writing easy-to-read Finnish implemented in YLE news in easy-to-read Finnish. In particular, I want to understand absolute measures, but also to compare some metrics for YLE news in easy-to-read Finnish and the news in a general language. The second aim is to identify how is the target group (i.e., immigrants) reflected in the news topics selection in easy-to-read Finnish. For example, I would like to understand the presence of some words, such as “immigrants”, or if there are particular political parties that are more included than others.

2. Data

First, I downloaded the corpus of YLE news in easy-to-read Finnish from Kielipankki, consisting of news from 2011 to 2018. The corpus is stored in .json format, which I was not familiar with, but I managed to manipulate them in bash. There were five files, so I had to concatenate them, ending with the file that looked like this


```
            "content": [
                {
                    "level": 1,
                    "text": "Tiistai 30.9.2014",
                    "type": "heading"
                },
                {
                    "text": "Talvivaaran metallikaivos halutaan pelastaa. Suomalaiset eivät\
 enää tuhlaa rahaa. Suuri Araljärvi on hävinnyt kokonaan. Liian harva käyttää heijastinta."\
,
                    "type": "text"
                },
```

By using egrep command, I extracted the lines which included “text”, but that still required some additional cleaning. I had to remove the lines such as “type” and “text”, then extra spaces, quotation marks at the beginning of each paragraph, and quotation marks and commas at their end. I also removed lines that were just about day of the week and the date. Finally, I also removed information about who read the news. My command pipeline looked like this in the end:

$ cat yle_selkokieli_text.txt | tr -s " " | sed '/^ \"type\"/d' | sed 's/^ \"text\": \"//' | sed 's/\",$//' | sed '/[1-8]$/d' | sed '/[1-8] $/d' | sed '/^(Lukija/d' > selkokieli.txt

An example of the end result was this:

```
Talvivaaran metallikaivos halutaan pelastaa. Suomalaiset eivät enää
Talvivaaran kaivos halutaan pelastaa
Kaivosfirma Talvivaara voidaan pelastaa. Selvitysmies Pekka Jaatine
Talvivaaran talous on huonossa kunnossa. Firma teki viime vuonna 70
Talvivaara tarvitsee lisää rahaa omistajilta. Valtio voi ehkä antaa
Suomalaiset eivät uskalla kuluttaa
Suomalaiset ovat entistä tarkempia rahojen käytössään. Syynä on hei
Suuri Araljärvi kuivui kokonaan
Yksi maailman suurimmista järvistä on kuivunut kokonaan. Satelliitt
Heijastin puuttuu liian monelta ihmiseltä
Heijastimia käytetään liian vähän. Liikenneturva kertoo, että vain
Armeija saa ehkä lisää rahaa. Pääministeri Stubb tapaa Euroopan joh
Armeija saa ehkä lisää rahaa
Eduskunnan mielestä Suomen armeija tarvitsee lisää rahaa. Eduskunta
Eduskuntapuolueet tekivät armeijan taloustilanteesta yhdessä raport
Suomen armeijan johtajat ovat sanoneet usein, että armeijalla ei ny
Armeijalle halutaan nyt antaa lisää rahaa, koska monet ajattelevat,
Pääministeri Stubb tapaa Euroopan johtajia
Suomen pääministeri Alexander Stubb on tavannut maanantaina Saksan
Pääministeri Stubb tapaa tällä viikolla monta eurooppalaista johtaj
Hongkongissa on isoja mielenosoituksia
Hongkongissa kymmenet tuhannet ihmiset osoittavat mieltä, koska he
```

At this point I have only corpus based on YLE news in easy-to-read Finnish. I did not manage to form a corpus for the general language because that corpus has a lot more complex organization. For example, there are multiple subfolders while end files have identical naming. Considering my time limits, I could not complete this part. However, if I get more time, I will gladly finish this part of the project, too.

Since the corpus of YLE news in general language is a lot bigger, even though both corpora cover the same years, my plan is to take the same amount of words as we find in easy-to-read news in order to balance them. More on this can be found in the section Discussion.

3. Analysis of the easy-to-read news according to the recommendations

Recommendations for easy-to-read language normally demand shorter words and sentences in order to make it easier to read and understand. This is why I calculated the following measures for easy-to-read language in YLE news:

- word length,
- words per sentence.

There are also recommendations about what morphological and syntactic forms should be avoided. To test how well this is achieved, I ran the text in Omorfi and got it morphologically tagged to get information about parts of speech and case, tense, person etc. My intention is to see how many "forbidden forms" there are in this corpus – meaning forms that are suggested to be avoided, such as complex verbal forms (e.g., pluperfect) or rare nominal forms (e.g., abessive). I will concentrate on nominal forms now, but my intention is to work on other forms in the future, too.

3.1 Word length

The corpus has 516,874 words and 4,540,925 characters, which means that the average length of the word is 8,8 characters. I have decided to include stop words, too, which are normally short, but the average word length is still quite high for my expectations.

My next step was producing a wordlist from the whole text. I also had to remove punctuation, so that it would not affect the word length. The code for that was:

```
cat selkokieli.txt | tr -s "\n\r\t " "\n" | tr -dc "A-ZÖÄÅa-zöää\n" > selkokieli_wordlist.txt
```

I had to stop here because of time limits, but my intention was to see how many long words there are in this wordlist. For example, words longer than 20 characters, as these long words should be avoided according to the recommendations. Right now, I have built the list of characters per word, but I will not go into further analysis.

3.2 Words per sentence

I am going to eliminate titles from this analysis, as well sentences ending in exclamation and question marks, since they all have specific syntax, and should be treated as separate cases for analysis. Therefore, I will extract all the sentences ending in full stop.

cat selkokieli.txt | egrep "\.$" > selkokieli_fullstops.txt

The average sentence length compared to the general language would show us if the recommendation for a shorter sentence is achieved in easy-to-read language.

3.4 Morphological features of easy-to-read language

Selkokeskus has a very detail list of what forms should be avoided in easy-to-read language. The corpus I got was not morphologically tagged, so I used Omorfi to get information about morphological features of the words in text. In this way I can compare the recommendations with their realization. Possible list of forbidden forms which are still in use can be used for further analysis and improvement of recommendations in the future. For example, we can get a list of forms that are supposed to be avoided and analyze why they were used.

This is the example of morphological analysis after using Omorfi:

```
# text = Talvivaaran metallikaivos halutaan pelastaa. Suomalaiset eivät enää tuhlaa rahaa. Suuri Aral
1       Talvivaaran     Talvivaara      PROPN   N       Case=Gen|Number=Sing    _       _       _
2       metallikaivos   metalli#kaivos  NOUN    N       Case=Nom|Number=Sing    _       _       _
3       halutaan        haluta  VERB    V       Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Pass     _
4       pelastaa        pelastaa        VERB    V       InfForm=1|Number=Sing|VerbForm=Inf|Voice=Act
5       .       .       PUNCT   Punct   _       _       _       _       Weight=0.0
6       Suomalaiset     Suomalainen     PROPN   N       Case=Nom|Number=Plur    _       _       _
7       eivät   ei      AUX     V       Number=Plur|Person=3|Polarity=Neg|VerbForm=Fin|Voice=Act
8       enää    enää    ADV     Adv     _       _       _       _       Weight=0.0
9       tuhlaa  tuhlata VERB    V       Mood=Imp|Number=Sing|Person=2|VerbForm=Fin|Voice=Act    _
10      rahaa   raha    NOUN    N       Case=Par|Number=Sing    _       _       _       Weight=0.0
11      .       .       PUNCT   Punct   _       _       _       _       Weight=0.0
12      Suuri   suuri   ADJ     A       Case=Nom|Degree=Pos|Number=Sing _       _       _       dETIT
13      Araljärvi       Araljärvi       PROPN   N       Case=Nom|Number=Sing    _       _       _
```

The command that was used is to get these results is:

omorfi-conllu.bash -X dusica/selkokieli.txt > dusica/selkokieli-ud.txt

3.5. Comparison of easy-to-read to general language forms

The numbers I get analyzing easy-to-read language tell us about some features of that language, but only if we compare them with the corresponding numbers for general language, we can get a better idea about how big intervention was made in easy-to-read language. For example, it can happen that some forms are equally rare in both varieties.

4. Lexical analysis

As I was preparing for this project, Korp has published corpus of YLE news in both easy-to-read language and general language for years 2011-2018. It was announced on 27 December 2019. (Uusia ainestoja (beta)). My plan to analyze lexical choices in easy-to-read Finnish and compare them to the general language are now a lot easier to achieve. The tool is very handy, but its problem that it takes 8 years of YLE news in general language as 8 separate corpora. The other problem is that general language corpus includes also news in easy-to-read language, so it is not recommended to use both corpora at the same time.

The other option for this part of analysis would be to use topic modelling in order to see what topics are selected for the immigrant population.

In any case, the main aim of this analysis would be to see both what choice of words and what choice of topics are used when targeting immigrant population.

5. Discussion

Traditional assessment of easy-to-read language is based on human estimation how well the recommendations for writing in easy-to-read are achieved in a particular text. The research project I outlined here would a try to use digital methods in assessment of easy-to-read language. The main idea would be to learn about features of this variety, but it can also be used to improve recommendations for producing it.

In order to learn about easy-to-read language I was planning to compare its features to the corresponding ones related to the general language. There are certain things to discuss about this comparison. First of all, the general language corpus is a lot bigger that easy-to-read language. In order to get comparable results, I have two options: to take the same amount of words from the general language as there is in easy-to-read corpus or to take the whole general language corpus and scale the result according to the difference in size. The second problem is that general language corpus includes easy-to-read new, so that deviates the results.

It was also stated that the corpora cover 8 years of news. There is space for thinking how this time span affects the results. It would be good to compare results between the years and see if there are some changes.

When analyzing the language of easy-to-read news, I did not include the name of the reader of those news (they also have audio form). Of course, the name of the reader does not belong to the language, but this information can be valuable in a possible study which would examine if there are some correlations between the reader (who is also the journalist writing the text) and linguistic and topical features of news.

So far, I have not analyzed the results of automatic morphological tagging, but my quick glance on that file showed me that tagging is somewhere wrong, which can affect the quality of results. For example, it marks “vuotta” as abessive and not as partitive, as it should.

This kind of research, regardless of some shortcomings, can significantly complement the traditional way of production and assessment of easy-to-read language. We can not only learn about its features, but also improve recommendations and assessment tools. Together with the human factor, it can bring more precise results and better quality of the text in easy-to read language.

6. References

Finnish Centre for Easy Language. (2018, September 25). Retrieved from https://selkokeskus.fi/in-english/the-finnish-centre-for-easy-to-read/

Klaara 2019. Conference on easy-to-read language research. Retrieved from https://www.helsinki.fi/en/conferences/klaara-2019-conference-on-easy-to-read-language-research

Klaara-network. Retrieved from https://blogs.helsinki.fi/klaara-network/in-english/

Leskelä, L. (2019). Selkokielen vaikeustasot [PowerPoint slides]. Retrieved from https://moodle.helsinki.fi/pluginfile.php/2794703/mod_resource/content/1/10.%20luennon%20diat.pdf

Maamies, S., Moilanen, R. (2010). Selkouutiset – uutisia selkeästi ja yksinkertaisesti. Retrieved from https://www.kielikello.fi/-/selkouutiset-uutisia-selkeasti-ja-yksinkertaisesti

Uusia ainestoja. (2019, December 27). Retrieved from https://www.kielipankki.fi/uutiset/uusia-aineistoja-beta-ylen-suomenkielinen-uutisarkisto-2012-2018-ja-ylen-suomenkielisen-uutisarkiston-selkouutiset/

Vanhatalo, U. (2019). Näkökulmia selkokielen tutkimukseen [PowerPoint slides]. Retrieved from https://moodle.helsinki.fi/pluginfile.php/2792364/mod_resource/content/1/9.%20luennon%20diat.pdf