Jacob Pawlak

LIN605 - Dr. Greg Stump

February 10th, 2017

Outline for Term Paper (Roughly Titled: A Computational Analysis on the Omission of Pronouns in Polish)

Knowing very little Polish before starting this project, I thought it would be hard to find a topic for a subsystem of its morphology. Since my first formal semantics course, I have been intrigued by the function and properties of pronouns, so I thought I would look at Polish pronouns. Consulting both my professor/advisor Dr. Jerzy Jaromczyk, a native Polish speaker, and my book *Polish: A Comprehensive Grammar*, I discovered that some personal pronouns get emitted from a sentence when used in subject position. After asking Dr. Jaromczyk about this, he told me that the omission is usually for style, and basically that it can be used or not used depending on the person who is speaking or writing. My project will be focused on examining the frequency that the pronouns are omitted in Polish poetry, music, and politics.

Before I outline the project, I should outline Polish. According to the Ethnologue website, there are approximately 41 million speakers of Polish, of which approximately 39 million live in Poland. Polish is classified as an Indo-European or Balto-Slavic language, which is mainly spoken in Poland, the Czech Republic, and Ukraine (when not spoken in small parts of America). Polish is typically a Subject-Verb-Object language, although Dr. Jaromczyk informed me that this is also mainly for style, and the ordering of words in a sentence can vary among text. Polish also has 7 case markings, compared to the 3 in English. The cases are as follows: nominative, accusative, genitive, dative, locative, instrumental, vocative. Since Polish has so many cases, it is

pretty easy to see that this additional information placed on words can allow deviation in sentence formation, which why (I believe) Dr. Jaromczyk has told me order is more for style.

With this project, I really want to focus on the computational aspect of this 'styling'. To start this task, I have selected a few dozen pieces of Polish literature, news articles, and music. With this large amount of data, I think I will have enough content to get a decent look at the Polish language and its use of pronouns, hopefully some omission. In terms of poetry, I have selected multiple authors, for diversity, from multiple time periods, for a better spread of any adaptations the language might have gone through. My first poet is Zygmunt Krasiński (born 1812 - died 1859), who wrote in the Polish Romantic Era. His work was mainly focused on emotional extremities and tragedy. Selections from Zygmunt include his most famous drama *Nie-boska Komedia*, *Agaj-Han*, and *Przedświt*. My second poet is Jan Kasprowicz (born 1860 - died 1926), a modernist writer, who unlike Zygmunt, had a poor upbringing that turned into a doctorate in literature. Jan wrote more on the philosophical plane, and a good amount of his works were original poetry. From Jan I have selected, *Poezje, Miłość,* and *Hymny*. These were for the most part written in a different century than Zygmunt's poetry, but I am not sure whether to expect any differences in the grammar.

For music I have selected a few popular albums: ŚWIATTŁO by Hades, KOMPONUJĄC SIEBIE by Sylwia Grzeszczak, OSTATNI KRZYK OSIEDLA by Paluch. I have not listened to these albums yet, but I am sure they are pretty good - I found them on the top 50 albums listened to in Poland from acharts.co. Hades is a Polish rapper who started his career in 2011. Needless to say his diction and word ordering should be vastly different than the two poets I have selected. I want to examine pop-culture music because it is far less formal than the modern political

literature. I have no way of knowing yet, but I suspect to find (especially in the rap songs) a more frequent omission of pronouns to fit syllable constraints in the songs. My next artist, Sylwia Grzeszczak, is a pianist whose career started in 2004. She seems to be kind of like the Taylor Swift of Poland, without all of the money. Besides being in a different genre than Hades, I think Sylwia's multiple awards will speak for herself. For an artist to be so popular and award winning, she must have good lyrics and good music. My hope is that the former will have a reduced amount of pronoun omission. Paluch is another Polish rapper, who interestingly enough has a 5 star rating on Discogs for his album Ostatni krzyk osiedla. With this album only being released in November of 2016, I am hoping to find some modern pronoun usage. I am not sure as of yet what will differentiate Paluch and Hades, but I can expect a difference in frequency from the rappers and the pianist.

Political literature should be really easy to come across, and I already have a few pieces selected. The first document is a report on a very recent (February 7th 2017) infiltration of several Polish banks. I have already read this report, translated of course, but its length and topic should make it a valuable article for this project. I also have some selections from the Polish Constitution, not that I expect to find many pronouns, but to use as more of a test bench for the program. Fro diversity, I have also selected the Gazeta Polska, a right-wing conservative magazine printed in Poland. I suspect that there will be some political and personal articles that should contain a healthy amount of pronouns. On the other end of the spectrum, is the Puls Biznesu, a business newspaper printed in Poland. I hope to find a higher level writing culture in this paper, and from this, hopefully a more structured style.

Now I would like to discuss the part of the project I am more familiar with - computational analysis with NLTK. For those unfamiliar, NLTK is the Natural Language Tool Kit library for the Python programming language. I have experience using this library, so I am not expecting too much trouble with its syntax, and having the program do what I want. Methods I have used before, and plan to use on this assignment are reading in text files to the program, parsing the text to look for tokens or indicators of tokens (possibly to be hand reviewed), and then making some statistical functions to give me frequency and placement of omission. Below I will describe some brief pseudo code for the project.

from nltk import * (and all external corpora I might need for Polish translation)

read in the text file to be parsed

set a variable for the number of omitted pronouns (looking at a pair of words that differ in some affix)

set a variable for the number of used pronouns (again looking at the same pair of words)

parse the text file, tokenize the words, tag for parts of speech, and group possible pairs

count the number of omitted pronouns

count the number of included pronouns

find the frequency of omission (omission / word_count)

find the frequency of inclusion (included / word_count)

generate a report based on the above to frequencies and store that in an external file

The report that will be generated from this program should answer my questions of frequency, and I think that with some extrapolation I can determine the style characteristics in

which both environments exist. I believe this project will be a fun way to explore computational morphology, and it will serve as a decent introduction into Polish grammar tactics.