Maria Irena Szawerna

**Lab 1: Chapter 4 report**

1. Phrase structure trees

2. Test the grammar

For testing the grammar quantitatively, I first tries to use non-POS tagged sentences, which turned out to be a disaster, with hardly anything properly marked and attached, as the grammar does not include a lexicon that would account for most of the vocabulary in the sentences. Instead, I used the files that had the POS tags (which I luckily annotated in the text files with the translations, before transforming them to conllu files in the previous week’s assignment). All of the gfud elements were done on eduserv, as I am unable to get a working version of the tool locally; I also had to use the solution that I previously shared in the program Discord so as not to get errors caused by encoding.

I first used gfud to create a conllu file containing the parses based on English.dbnf. This meant running the following command: *cat comp-syntax-corpus-english.txt | gfud dbnf English.dbnf Utt > my\_english.conllu* and then evaluating the created file against my hand-annotated conllu file from last week: *gfud eval macro LAS english.conllu my\_english.conllu.* I obtained the following results: ***UDScore {udScore = 0.6590285318546188, udMatching = 20, udTotalLength = 289, udSamesLength = 184, udPerfectMatch = 1}***, which seems to be quite good, granted that the same grammar only got a udScore of ~0.5 when tested in the lecture. Naturally, the grammar must be lacking some of the structures that appear in the corpus (which contains some really complex sentences), and it is also possible that my annotation is faulty at times (meaning that the grammar analyzed the sentence better than I did, but the results do not match up and my annotation is treated as the golden standard regardless of its quality).

I then continued to do the same for Polish: the problem here though was that gfud refused to read my conllu file with any sorts of comments in it, stating *gfud: ERROR: ﻿# sent\_id = 1 incomplete UDWord*. I solved that by creating an additional file, polish-nocomm.conllu, where I removed all the comments (so the original sentences and sentence numbers, as well as my comments pertaining to the difficulties I encountered when annotating). Thus, the commands I ran looked like this: first *cat comp-syntax-corpus-polish.txt | gfud dbnf English.dbnf Utt > my\_polish.conllu*, and then *gfud eval macro LAS polish-nocomm.conllu my\_polish.conllu*. I obtained the following evaluation: ***UDScore {udScore = 0.4356202620908503, udMatching = 17, udTotalLength = 186, udSamesLength = 78, udPerfectMatch = 1}***. This is a significantly worse score than the one obtained for English. While the same may hold true as for the English evaluation (that my mistakes in the annotation may influence it), the magnitude of the difference indicates that the grammar rules that work for English do not work all that well for Polish. The reason behind it may be that Polish has a much more flexible word order; while sentences usually follow the SVO order, inversion can be used for emphasis and subject can be dropped. Adjectives can also modify nouns both by preceding and following them, and auxiliaries can follow verbs instead of preceding them as well; there are more differences, as per my submission for Chapter 2, but these I remember to occur in the corpora.

3. Modify the grammar, test on UD treebanks