Maria Irena Szawerna

Lab 1: Chapter 1 report

For this lab I decided to compare Polish (my native language) and English. I downloaded the required Parallel UD treebanks and attempted to explore them using the gf-ud tool. It proved to be impossible to download and install it on my computer, and the eduserv installation kept throwing me errors even when the treebanks were downloaded directly onto it. Other students who I have asked did not run into those issues on the server, but nobody succeeded with a local installation. Instead I decided to write my own Python script to get the necessary information using the CoNLL-U Parser module. I have included the files for it in my course repository: the readme/setup instructions, the first-time setup script (which installs the necessary module/library), and the python script itself.

I proceeded to get the UPOS, XPOS, and DEPREL tag counts for each of the treebanks. Not all of these had enough variants to retrieve top 20 (there were fewer than 20 tags altogether). I retrieved the following counts:

- XPOS for the en_pud-ud-test.conllu file: 1. NN: 3000, 2. IN: 2716, 3. DT: 2121, 4. NNP: 1612, 5. JJ: 1435, 6. NNS: 1103, 7. ,: 1002, 8. .: 1000, 9. VBD: 875, 10. RB: 773, 11. VBN: 591, 12. CC: 576, 13. VB: 507, 14. PRP: 490, 15. CD: 460, 16. VBZ: 439, 17. VBG: 332, 18. TO: 267, 19. VBP: 259, 20. PRP$: 255
- UPOS for the en_pud-ud-test.conllu file: 1. NOUN: 4036, 2. ADP: 2493, 3. PUNCT: 2451, 4. VERB: 2156, 5. DET: 2086, 6. PROPN: 1741, 7. ADJ: 1530, 8. PRON: 1021, 9. AUX: 1014, 10. ADV: 849, 11. CCONJ: 576, 12. NUM: 455, 13. PART: 426, 14. SCONJ: 290, 15. SYM: 42, 16. X: 16, 17. INTJ: 1
- DEPREL for the en_pud-ud-test.conllu file: 1. case: 2499, 2. punct: 2451, 3. det: 2047, 4. nsubj: 1393, 5. amod: 1336, 6. obl: 1237, 7. nmod: 1076, 8. root: 1000, 9. obj: 876, 10. advmod: 852, 11. compound: 810, 12. conj: 634, 13. cc: 574, 14. mark: 555, 15. aux: 410, 16. nmod:poss: 365, 17. cop: 316, 18. advcl: 293, 19. aux:pass: 274, 20. xcomp: 271
- XPOS for the pl_pud-ud-test.conllu file: 1. interp: 2678, 2. part: 665, 3. prep:loc:nwok: 622, 4. conj: 575, 5. subst:sg:gen:f: 507, 6. subst:sg:gen:m3: 421, 7. fin:sg:ter:imperf: 391, 8. subst:sg:nom:f: 369, 9. subst:sg:nom:m1: 339, 10. prep:gen: 337, 11. comp: 335, 12. subst:sg:nom:m3: 334, 13. adv:pos: 297, 14. subst:sg:loc:m3: 286, 15. subst:sg:acc:f: 283, 16. prep:loc: 237, 17. prep:acc:

230, 18. prep:gen:nwok: 223, 19. subst:sg:acc:m3: 220, 20. subst:sg:gen:n:ncol: 214

- UPOS for the pl_pud-ud-test.conllu file: 1. NOUN: 4504, 2. PUNCT: 2658, 3. ADJ: 2358, 4. ADP: 2050, 5. VERB: 1633, 6. PROPN: 1326, 7. PRON: 639, 8. CCONJ: 575, 9. ADV: 535, 10. DET: 499, 11. AUX: 455, 12. PART: 379, 13. SCONJ: 335, 14. X: 230, 15. NUM: 193, 16. _: 49, 17. SYM: 20
- DEPREL: for the pl_pud-ud-test.conllu file: 1. punct: 2658, 2. case: 1993, 3. amod: 1431, 4. nsubj: 1076, 5. root: 1000, 6. obl: 975, 7. nmod: 942, 8. obj: 818, 9. conj: 714, 10. nmod:arg: 703, 11. cc: 558, 12. advmod: 509, 13. obl:arg: 429, 14. mark: 341, 15. flat: 315, 16. iobj: 290, 17. amod:flat: 273, 18. expl:pv: 271, 19. acl: 248, 20. advmod:emph: 225

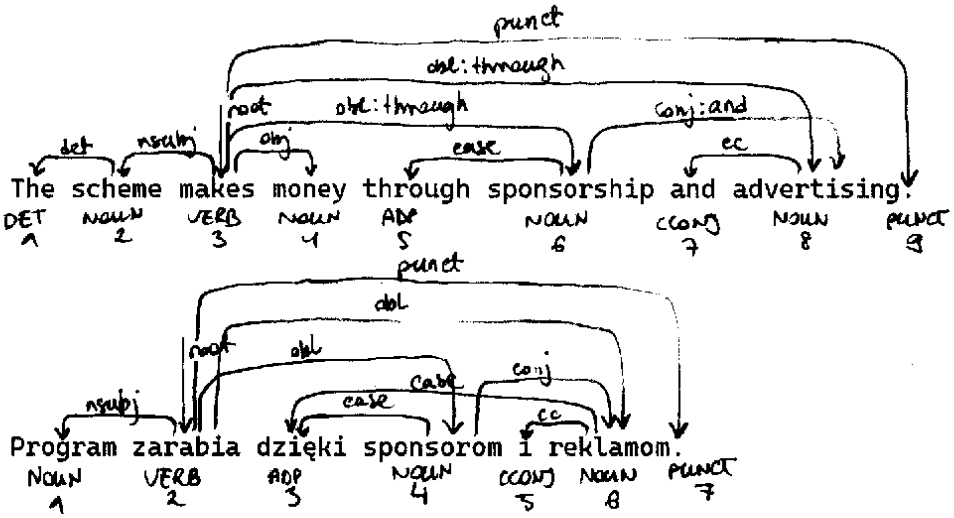
It is easy to notice that for the XPOS (language-specific) tags, different tags are used for English and Polish. The English ones seem to follow the Penn Treebank convention, while the Polish ones reflect the much broader spectrum of word forms that this language possesses. Thus, instead of just having categories for nouns (NN), proper nouns (NNP), plural nouns (NNS) etc., the Polish one has to include information such as noun, singular, genitive case, feminine (subst:sg:gen:f) or noun, singular, nominative case, masculine animate (subst:sg:nom:m1). Thus, the two are not really comparable 1:1 in terms of the distribution of the parts of speech.

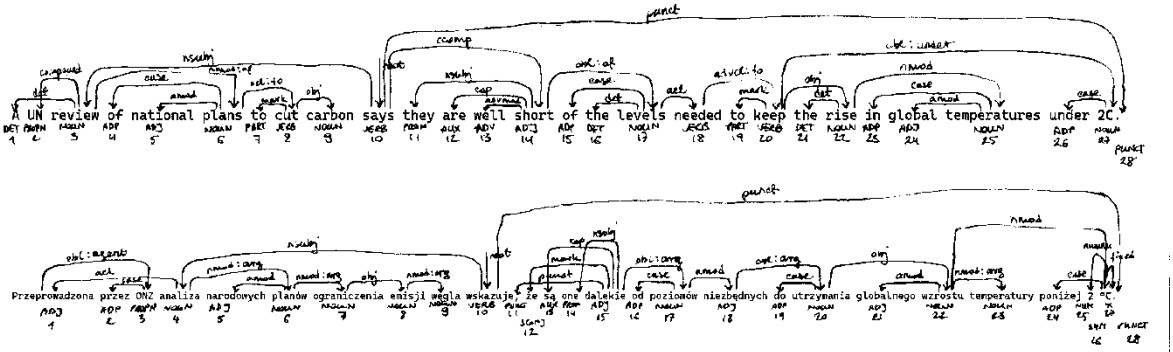
The UPOS (universal) tags give a much better comparison. Nouns are the most common part of speech in both languages; however, Polish seems to have more of then than English in this treebank. It also uses more punctuation, but fewer verbs. The lack of verbs in the Polish version could be attributed to the fact that the language does not make much use of auxiliary verbs; that, however, is a separate UPOS category (where, indeed, English has a higher count). The difference thus may stem from the fact that Polish prefers to use verb-derived nouns in place of some verbs. There is also a massive disparity in terms of determiners, as Polish is a language that does not have articles. This category also included only 17 tags.

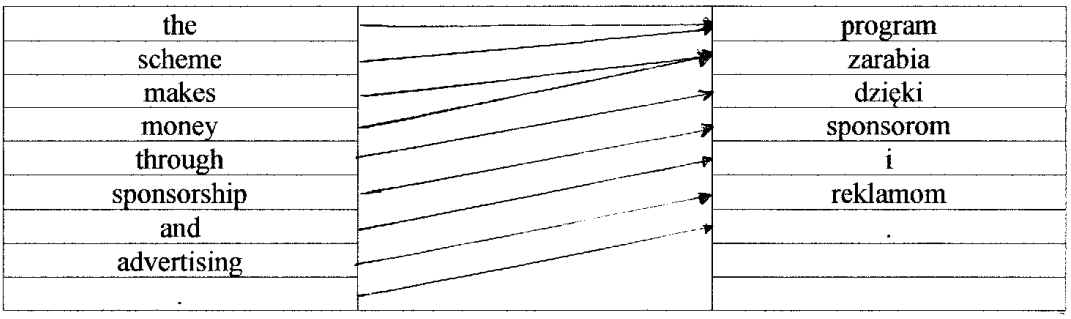
To some extent, the DEPREL tags can also be compared. The case and punct tags dominate both treebanks, but with a slightly different proportion. Naturally, both have the same number of root tags. Tags like amod, nsubj, nmod, obj, avmod are also high up on both lists. The differences in terms of the other tags also reveal a bit about the grammar of the two languages: for example, the 15th most popular DEPREL tag in English was aux, which does not appear in the top 20 of Polish, due to the reasons mentioned above; same holds true for the 3rd most popular tag in English, det.

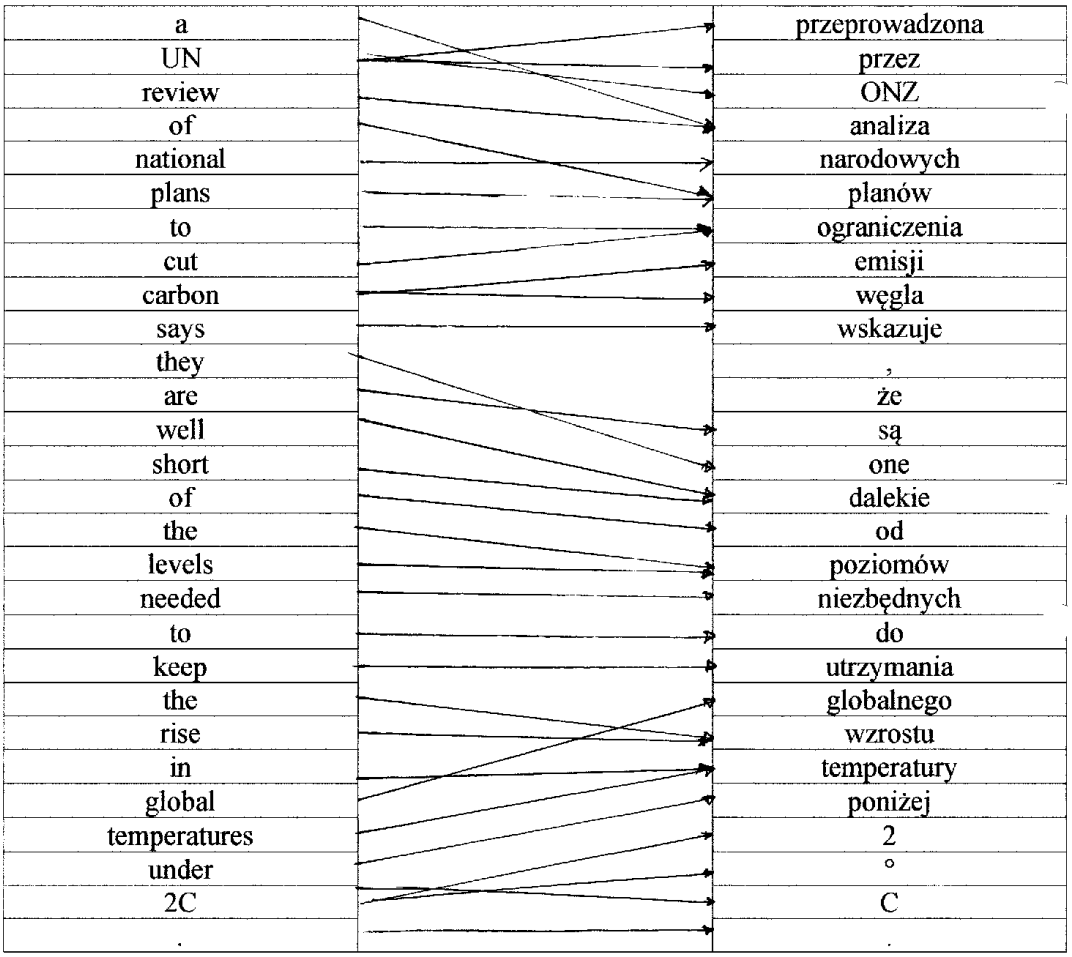
For the last two parts of the assignment, I selected the following sentences:

- English short: `The scheme makes money through sponsorship and advertising.`
- Polish short: `Program zarabia dzięki sponsorom i reklamom.` (sentence 39)
- English long: `A UN review of national plans to cut carbon says they are well short of the levels needed to keep the rise in global temperatures under 2C.`
- Polish long: `Przeprowadzona przez ONZ analiza narodowych planów ograniczenia emisji węgla wskazuje, że są one dalekie od poziomów niezbędnych do utrzymania globalnego wzrostu temperatury poniżej 2 °C.` (sentence 48)

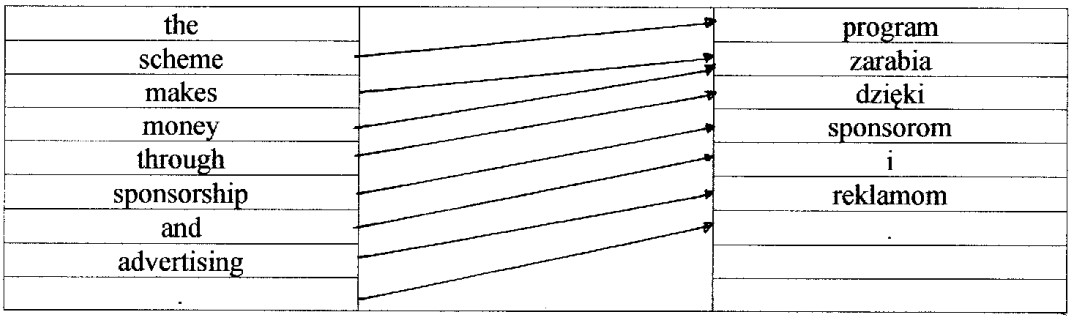


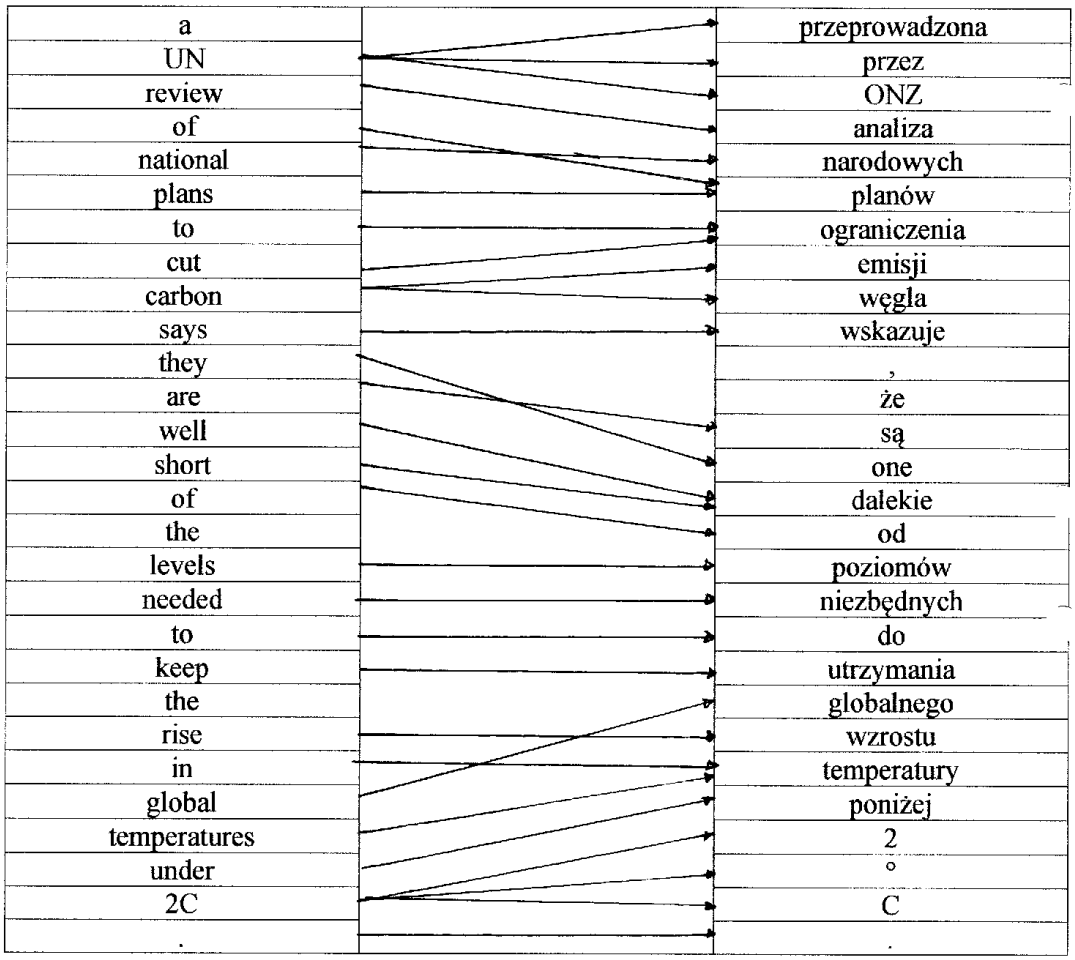



the
scheme
makes
money
through
sponsorship
and
advertising
.
program
zarabia
dzięki
sponsorom
i
reklamom
.


a
UN
review
of
national
plans
to
cut
carbon
says
they
are
well
short
of
the
levels
needed
to
keep
the
rise
in
global
temperatures
under
2C
.
przeprowadzona
przez
ONZ
analiza
narodowych
planów
ograniczenia
emisji
węgla
wskazuje
,
że
są
one
dalekie
od
poziomów
niezbędnych
do
utrzymania
globalnego
wzrostu
temperatury
poniżej
2
°
C
.

Alignment 1


the
scheme
makes
money
through
sponsorship
and
advertising
.
program
zarabia
dzięki
sponsorom
i
reklamom
.


a
UN
review
of
national
plans
to
cut
carbon
says
they
are
well
short
of
the
levels
needed
to
keep
the
rise
in
global
temperatures
under
2C
.
przeprowadzona
przez
ONZ
analiza
narodowych
planów
ograniczenia
emisji
węgla
wskazuje
,
że
są
one
dalekie
od
poziomów
niezbędnych
do
utrzymania
globalnego
wzrostu
temperatury
poniżej
2
°
C
.

Alignment2

The trees and alignment charts were done partly on computer, partly by hand mostly due to the fact that I needed to make one of the sentences very small and that it is easier to read the words in the table when presented in this way. I hope it is still okay, and all the connections were drawn by hand on printed copies and subsequently scanned. Initially (Alignment 1), I connected the articles and other elements that do not have a direct correspondence or are not featured in the other language to whatever was the closest in the meaning in the other language (e.g. connecting articles – which do not exist in Polish – to the nouns). Having consulted the teaching assistant, and having learned that both this and not connecting them to anything is valid, I opted for the latter option (Alignment 2), as it does a better job of showing what elements are missing from or are obligatory in which language.

From the alignment charts we can draw a couple of conclusions about syntactic differences between Polish and English. First of all, determiners are not present in Polish. Word order in Polish in general is similar to that of English, but in some cases, it can be much more free (as with the global temperatures and "they are"). Polish does not need as many ADP as English as the relations between the parts of the sentence are encoded in case endings (thus, many prepositions are connected to the nouns with appropriate case endings). Finally, Polish verbs can contain more information than their English counterparts ("make money" = "zarabiać"). Sometimes more than one word is needed to represent a concept from English, or, more accurately, a certain part cannot be dropped ("carbon (**emissions**)" = "**emisji** węgla", where in English "emissions" can be omitted, or "says (**that**) they are" = "wskazuje**, że** są one" – here "that" can be omitted). The noun and adjective order seem to usually follow the same pattern (adjective first), although there is a rule to it that is not shown in these examples, where using the adjective first indicates simply the quality of the item, and using the adjective last denotes some fixed classification (e.g. "brunatny niedźwiedź" means a brown bear and "niedźwiedź brunatny" means a bear of the species brown bear; this difference is similar to the English "black bird" vs. "blackbird", and the adjective-last forms are rather lexicalized).