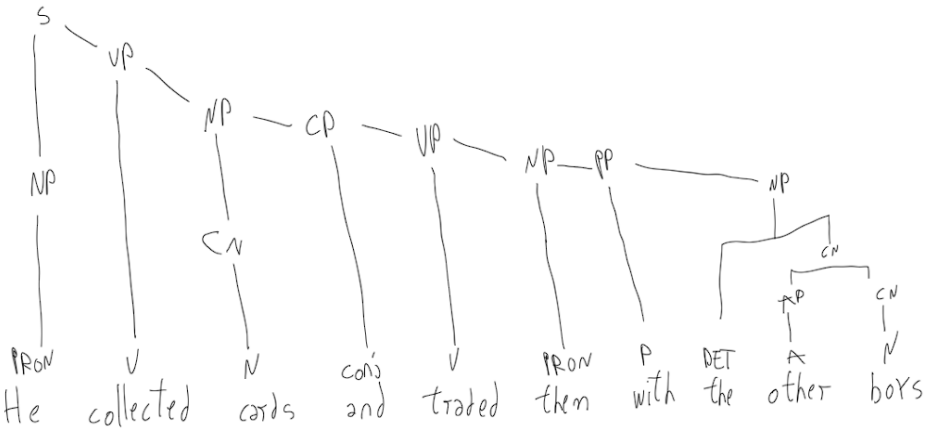
1. Phrase Structure Trees

English

Pre-annotated text: He:<PRON> collected:<VERB> cards:<NOUN> and:<CCONJ> traded:<VERB> them:<PRON> with:<ADP> the:<DET> other:<ADJ> boys:<NOUN>

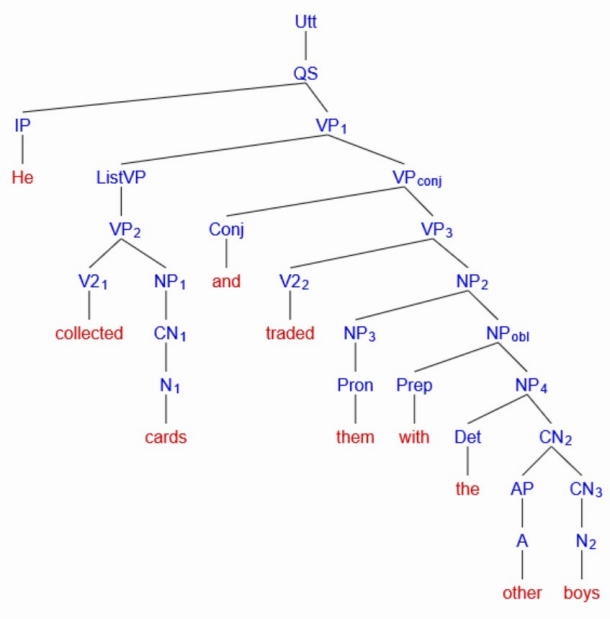
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Utt (QS (IP He) (VP (ListVP (VP (V2 collected) (NP (CN (N cards))))) (VP_conj (Conj and) (VP (V2 traded) (NP (NP (Pron them)) (NP_obl (Prep with) (NP (Det the) (CN (AP (A other)) (CN (N boys))))))))))

Graphical representation:



Comparison:

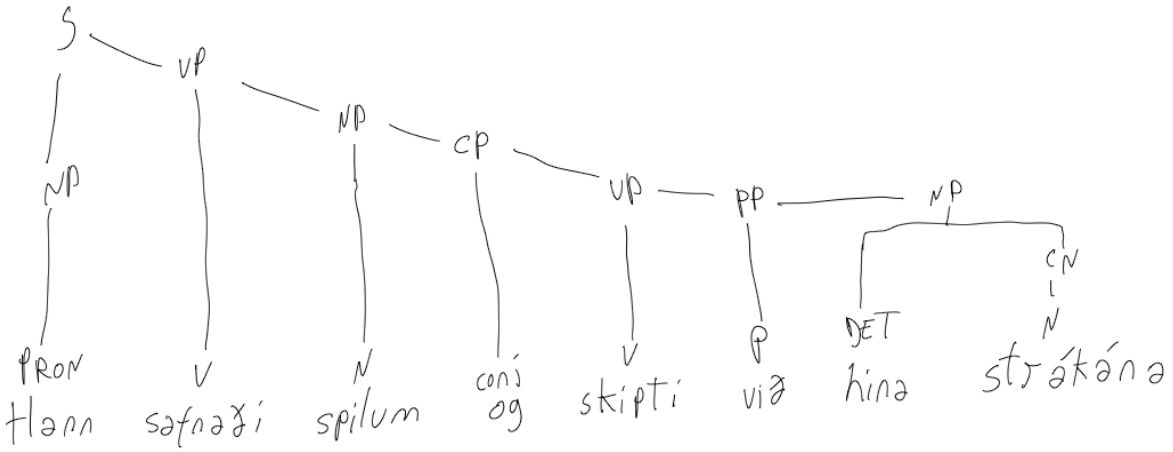
While the 2 phrase structure trees look mostly similar to each other, a few differences can be easily spotted. First of all, the main difference stems from GF's different terminology in labelling, such as the use of **QS (Quantifier Scope)** instead of S(Sentence), **IP (Inflectional Phrase)** (also worth noting is that it incorrectly labeled "he" as an IP, instead of a pronoun) instead of NP and **ListVP (List of Verb**

Phrases) instead of another VP. This is due to GF's specialized nature, as it conveys a higher degree of syntactic granularity in its analysis with its English dbnf file and its associated rules. Another difference that should be pointed out is a purely superficial one, regarding the use of a different term, the GF phrase structure tree uses the terms VP conj for a Conjunction Phrase (CP) and NP obl (**Oblique Noun Phrase**) for a prepositional phrase.

Icelandic

Pre-annotated text: hann:<PRON> safnaði:<VERB> spilum:<NOUN> og:<CCONJ> skipti:<VERB> við:<ADP> hina:<DET> strákana:<NOUN>

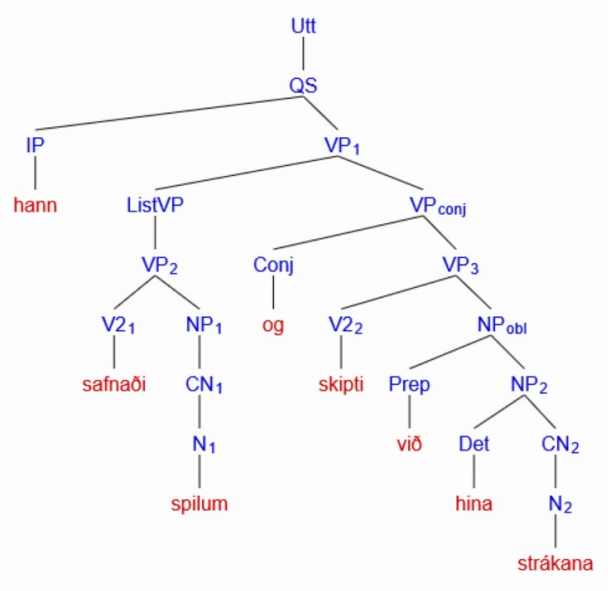
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Utt (QS (IP hann) (VP (ListVP (VP (V2 safnaði) (NP (CN (N spilum))))) (VP_conj (Conj og) (VP (V2 skipti) (NP_obl (Prep við) (NP (Det hina) (CN (N strákana)))))))))

Graphical representation:



Comparison:

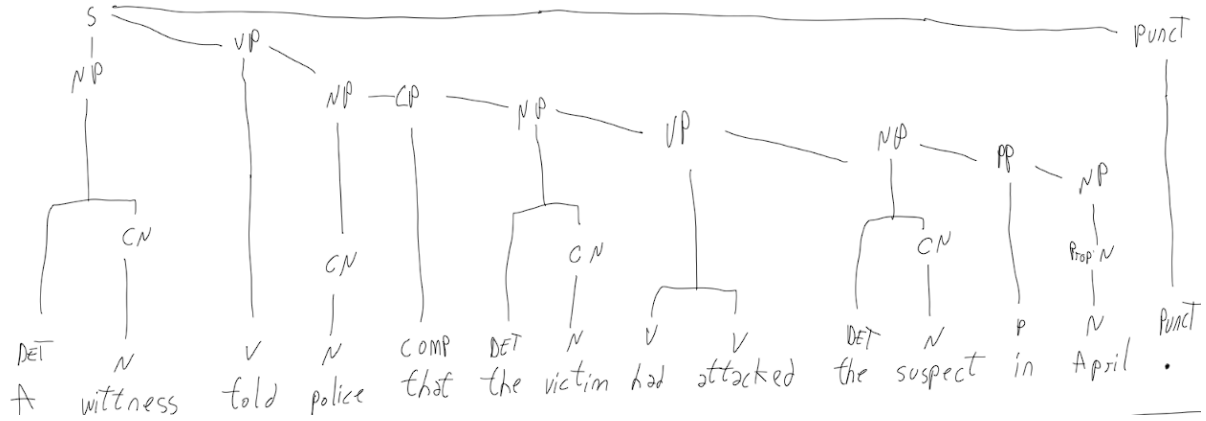
For this particular Icelandic sentence, it seems that GF managed to almost successfully create a phrase structure tree, which is very similar to my own. The differences consist of naming conventions, same as with the English sentence, which I have outlined above, but also repeating the same error, as classifying the pronoun “hann” as an IP directly, instead of a pronoun and then later as an NP.

English

Pre-annotated text:

A:<DET> witness:<NOUN> told:<VERB> police:<NOUN> that:<SCONJ> the:<DET> victim:<NOUN> had:<AUX> attacked:<VERB> the:<DET> suspect:<NOUN> in:<ADP> April:<PROPN> .:<PUNCT>

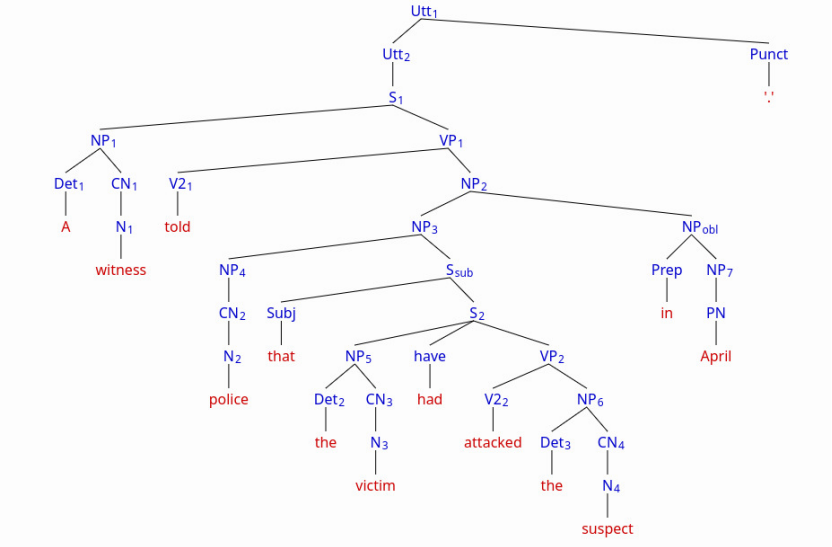
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Utt (Utt (S (NP (Det A) (CN (N witness))) (VP (V2 told) (NP (NP (NP (CN (N police))) (S_sub (Subj that) (S (NP (Det the) (CN (N victim))) (have had) (VP (V2 attacked) (NP (Det the) (CN (N suspect))))))) (NP_obl (Prep in) (NP (PN April))))))) (Punct '.'))

Graphical representation:



Comparison:

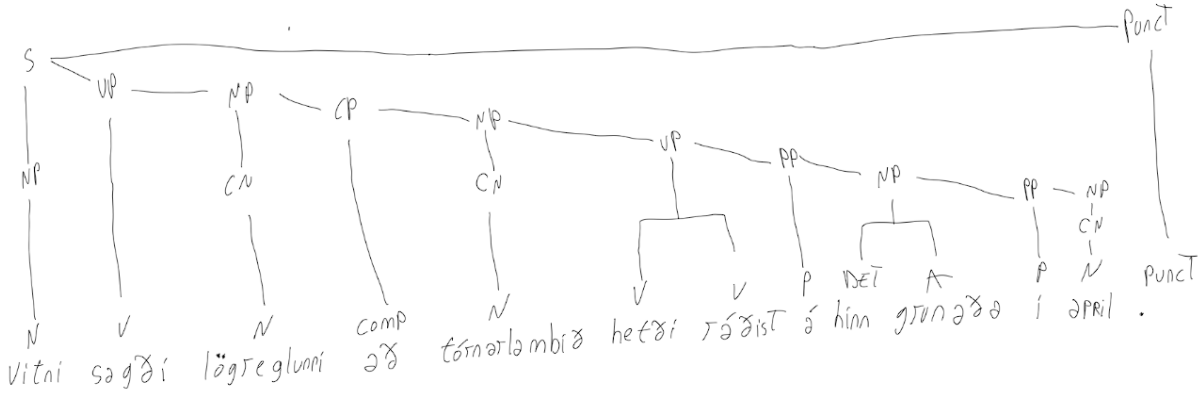
For this sentence, the English dbnf file does a great job at providing a detailed phrase structure tree, correctly identifying the subordinate phrase (and explicitly naming it). One significant difference from my attempt is that it did not correctly place the auxiliary verb “had” in direct relation to the verb “attacked”, classifying it incorrectly as “have” which is erroneous. Another difference is that it classified the prepositional phrase “in April” (named NP obl due to its internal naming conventions) as being separate from the nominal phrase “the suspect”. The rest of the differences being simply cosmetic (different naming conventions) or due to its higher granularity.

Icelandic

Pre-annotated text:

Vitni:<NOUN> sagði:<VERB> lögreglunni:<NOUN> að:<ADP> fórnarlambið:<NOUN> hefði:<AUX> ráðist:<VERB> á:<ADP> hinn:<DET> grunaða:<ADJ> í:<ADP> apríl:<NOUN> .:<PUNCT>

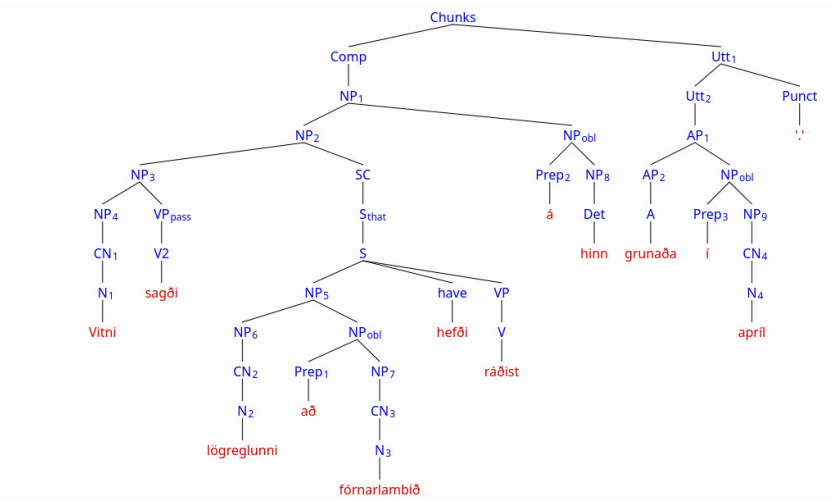
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Chunks (Comp (NP (NP (NP (NP (CN (N Vitni))) (VP_pass (V2 sagði))) (SC (S_that (S (NP (NP (CN (N lögreglunni))) (NP_obl (Prep að) (NP (CN (N fórnarlambið))))))) (have hefði) (VP (V ráðist)))))) (NP_obl (Prep á) (NP (Det hinn))))) (Utt (Utt (AP (AP (A grunaða)) (NP_obl (Prep í) (NP (CN (N apríl)))))) (Punct '.')))

Graphical representation:



Comparison:

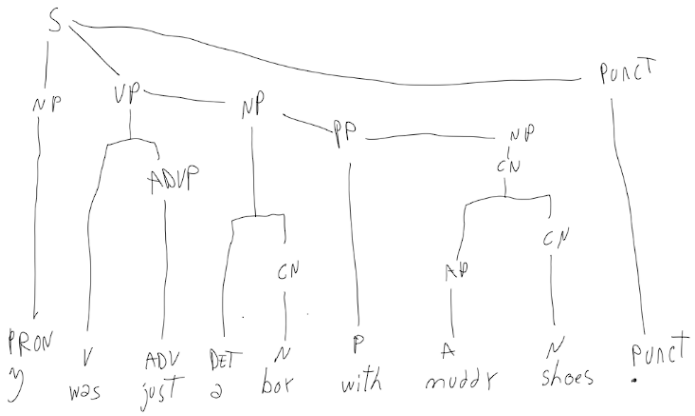
Regarding this sentence, GF's English dbnf's limits are particularly visible when it comes to creating phrase structure trees for Icelandic. It incorrectly classified "grunaða í apríl" as being the main phrase (signified by the Utt label). This is wrong as not only does "grunaða í apríl" mean something similar to "the suspect in April" but it is also an incomplete construction, as the full meaning is only conveyed by "á hinn grunaða í apríl", which is spread between the Complement Clause and the Main Clause. Besides this, I have to say that it also did not properly label the real Main Clause "Vitni sagði", instead placing in the Complement Clause, this way representing the entire phrase structure tree in an erroneous manner.

English

Pre-annotated text:

I:<PRON> was:<AUX> just:<ADV> a:<DET> boy:<NOUN> with:<ADP> muddy:<ADJ> shoes:<NOUN> .:<PUNCT>

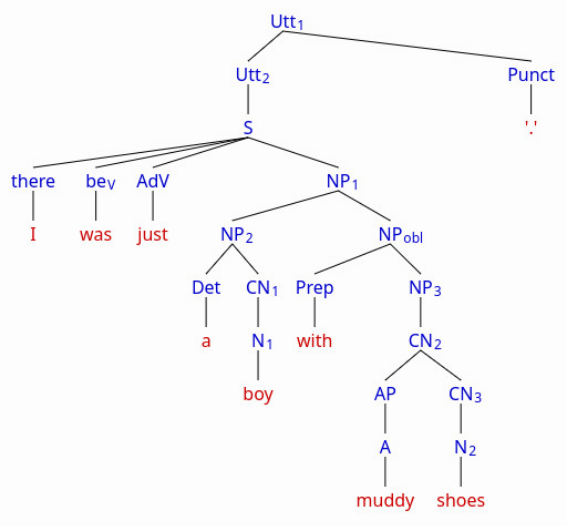
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Utt (Utt (S (there I) (be_V was) (AdV just) (NP (NP (Det a) (CN (N boy))) (NP_obl (Prep with) (NP (CN (AP (A muddy)) (CN (N shoes))))))))) (Punct '.'))

Graphical representation:



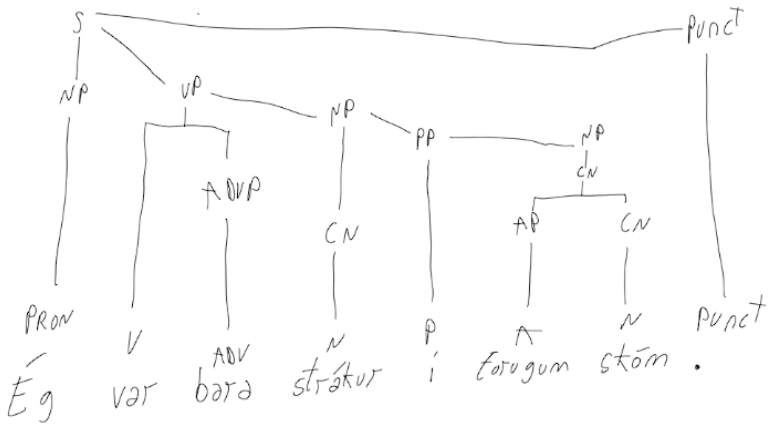
Comparison:

An seemingly simple phrase, which was parsed wrongly by the English dbnf rules. The problem lies with how the construction “I was just” was classified, as it is clearly an error due to the invalid “there” category for the “I” pronoun, which prevents it being classified correctly as part of a Noun Phrase. Not only that, but the Adverbial Phrase (of which “just” should be part of) is not properly categorized either, leading to not being able to classify the Verb Phrase pertaining to the verb “was” either.

Icelandic

Pre-annotated text: Ég:<PRON> var:<VERB> bara:<ADV> strákur:<NOUN> í:<ADP> forugum:<ADJ> skóm:<NOUN> .:<PUNCT>

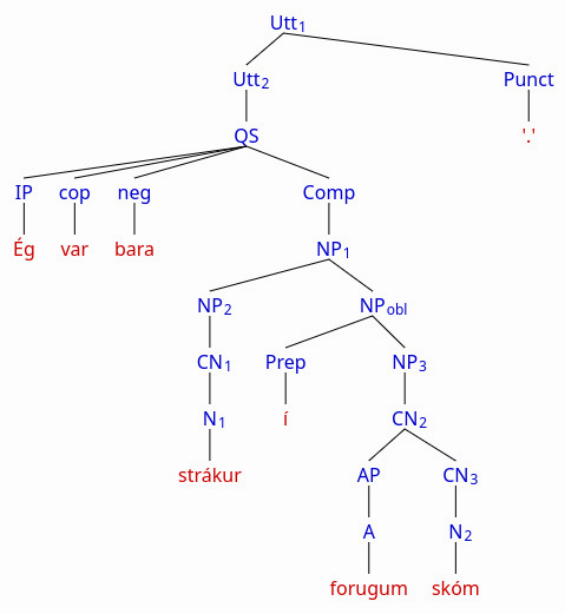
Manually created phrase structure tree:



Gf-ud automatically created tree:

(Utt (Utt (QS (IP Ég) (cop var) (neg bara) (Comp (NP (NP (CN (N strákur))) (NP_obl (Prep í) (NP (CN (AP (A forugum)) (CN (N skóm))))))))) (Punct '.'))

Graphical representation:



Comparison:

It seems that even with the Icelandic phrase, GF had a similar problem as with the English one, having catalogued the construction “Ég var bara” wrongly. “Ég” instead of being labeled as a pronoun, is directly attributed to an Inflectional Phrase, the verb “var” (past tense form of the “to be” að vera) is classified as a copula and the adverb “bara” is labeled as a negation. This mistake in classifying prevents the GF phrase tree from being an accurate representation, as such a major mislabeling leads to the rest of the phrase to be wrongfully classified as a Complement Phrase.

2. Testing the English Grammar on the Corpus

English Corpus:

UDScore {udScore = 0.5929797255884212, udMatching = 23, udTotalLength = 328, udSamesLength = 192, udPerfectMatch = 2}

It can be noticed that the overall score (udScore) has the value of 0.592, which would suggest that around 59.2% of the tokens were correctly parsed by using English.dbnf grammar. While parsing more than half the tokens properly, it is clear that there is still room for improvement. It can also be mentioned that there are 2 sentences which have been parsed completely identically by using the grammar.

Icelandic Corpus:

UDScore {udScore = 0.49281753946517887, udMatching = 24, udTotalLength = 302, udSamesLength = 151, udPerfectMatch = 1}

The overall score (udScore) has the value of 0.492, which would suggest that around 49.2% of the tokens were correctly parsed by using English.dbnf grammar. It is clear that slightly less than half of the tokens are parsed correctly, which is understandable, as there is a significant number of differences between the English and the Icelandic grammar.

It can also be mentioned that there is 1 sentence which has been parsed completely identically by using the grammar.

3. Adjusting the Grammar According to the Icelandic Grammar

The grammar was modified largely by changing its lexicons to better suit the Icelandic language, but also, through changes in its sentences, indirect questions, adverbials, coordination and noun phrases. The changes aren’t comprehensive, partially due to Icelandic’s similarities to English, but also due to my lack of in-depth knowledge of Icelandic grammar and special cases.

The results when using the new grammar are the following:

Icelandic

UDScore {udScore = 0.5264121107435736, udMatching = 24, udTotalLength = 302, udSamesLength = 157, udPerfectMatch = 1}

The overall score for Icelandic is now 0.526, which means that around 53% of the tokens are parsed correctly, a 3.4% increase of the previous grammar. While still having a noticeable number of errors, this grammar is a clear improvement compared to the English one regarding the parsing of Icelandic language.

English

UDScore {udScore = 0.5050941811811377, udMatching = 23, udTotalLength = 328, udSamesLength = 157, udPerfectMatch = 2}

Using the modified grammar on the English corpus once more shows that its performance has dropped by almost 10%, a significant amount, which showcases the significant linguistic distance between the 2 languages, despite being in the same greater (Germanic) branch.

Italian

UDScore {udScore = 0.22634484763190812, udMatching = 169, udTotalLength = 4790, udSamesLength = 876, udPerfectMatch = 3}

The overall score of Italian is 0.226, thus amounting for a correct parsing of just 22.6% of the

tokens, this is to be expected as Italian is part of a completely different branch of Indo-European languages.

Finnish

UDScore {udScore = 0.2930139988935566, udMatching = 213, udTotalLength = 3530, udSamesLength = 940, udPerfectMatch = 1}

The overall score of Finnish is 0.293, which makes sense, as despite not being an Indo-European language, Finnish has been greatly influenced by North Germanic languages throughout its existence.

Czech

UDScore {udScore = 0.37821145737603284, udMatching = 991, udTotalLength = 18488, udSamesLength = 6699, udPerfectMatch = 14}

The overall score of Czech is 0.378, thus 37.8% of the tokens being parsed correctly. This is a surprising result, which could be explained by the potential similarities between Czech and other Germanic languages, due to their geographic proximity but also due to their cultural (and linguistic) exchange.