Natural Language Processing of Kyrgyz Morphotactics based on Turkic Lexicon Apertium Toolkit

Ph.D., Dr. Aida Kasieva
Kyrgyz-Turkish Manas University
Seminar: 18.05.2021
Outline:
What is Morphotactics?
Features of Morphotactics
Language Variations
Parts of Speech Tagging (PoST)
Kyrgyz Corpus
Turkic Lexicon Apertium Toolkit
Examples of tagging
What is Morphotactics?
Morphotactics is a branch of morphemics that studies the patterns of morpheme combination in a word. Morphotactics determines how freely morphemes can be combined, what kinds of restrictions on morpheme combinability exist. Morphotactics also studies such a problem as the use of interfixes in the combination of morphemes. Morphotactics represent the ordering restrictions in place on the ordering of morphemes. Etymologically, it can be translated as "the set of rules that define how morphemes can touch each other".
In order to study the valency of morphemes, it is necessary to identify combinable properties that depend on the language system. It is worth distinguishing between phenomena that are generally unacceptable in the language system, and phenomena that are possible, but are not represented lexically in a given language. Morphotactics is one of the parts of traditional morphology. The specifics of morphotactics in the historically established system of the modern science of language, characterized by a departure from the morphological tradition and divergence of four related scientific disciplines descending from traditional morphology: morphemics, morphonology, word formation and morphology itself.
As Kemal Oflazer and Murat Saraclar (2018) note, Morphotactics is closely connected to Morphology and they function together in the language. Most of the agglutinative languages have a particular morphological tactics to produce new derivations and words.
Variations

In the process of working on the Morphotactics of the Kyrgyz Language and its processing/computization, my paper found various differences in Morphology between KG-EN languages, especially in word formations. There are distinct variations in Morphotactics in KL, comparing to other Agglutinative languages.
What is Parts of Speech Tagging?

It is generally called POS tagging. In simple words, we can say that POS tagging is a task of labeling each word in a sentence with its appropriate part of speech. We already know that parts of speech include nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories. Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, by a set of descriptive tags. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic.
Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken. This is not rare—in natural languages, a large percentage of word-forms are ambiguous. For example, even "dogs", which is usually thought of as just a plural noun, can also be a verb:

“The sailor dogs the hatch.”
“The sailor fastens the hatch.”

“Мен диванда жатам.”
«Мен үйге бара жатам.»
After long discussions it was decided to exploit standard toolkits of the Turkic Lexicon Apertium (toolkit used for Turkish language), an open-source machine translation platform. For annotating, each word is labeled and analyzed along with the process of tagging in closer observation. We attempt to make detailed analysis of Kyrgyz sentences extracted from the Kyrgyz corpus. We believe that this work will give impetus for further development and enrichment of the Kyrgyz corpus and attract students and linguists to get involved in this interesting process.
POST in Kyrgyz

Nouns
18 ^Ыслам/*Ыслам$ - <np> <ant> <m> <nom>
18 ^Шералини/*Шерали$ - <np> <ant> <m> <acc>
18 ^шоокум/*шоокум$ - <n> <sg> <nom>
18 ^Төштүктүн/*Төштүк$ - <np> <ant> <m> <gen>
18 ^оокумда/*оокум$ - <n> <abst> <nn> <sp> <loc>
18 ^өкүрүк/*өкүрүк$ - <n> <sg> <nom>
POST in Kyrgyz

Verbs

18 ^ийкегиледи/*ийкегиле$ - <v> <tv> <p3> <ifi>
17 ^куймөнүп/*куймөн$ - <v> <iv> <ref> <gna_perf>
16 ^чубалган/*чубал$ - <v> <iv> <ref> <gpr_past>
16 ^туталана/*туталан$ - <v> <iv> <ref> <gna_impf>
16 ^келгемин/*кел$ - <v> <iv> <p1> <ifi>
15 ^жетпеди/*жет$ - <v> <tv> <neg> <ifi> <p3>
POST in Kyrgyz

Adjectives
15 ^кызылала/*кызылала$ - <adj>
15 ^кардуу/*кардуу$ - <adj>
15 ^Карагаттай/*Карагат$ - <adj> <subst>
14 ^шамдагай/*шамдагай$ - <adj>
14 ^кичи/*кичи$ - <adj>
Example on how to search words in Kyrgyz Corpus:
http://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_2
After we open the word and in what context it is mostly used, we may start tagging
Tagging the word “эне”

Hereby, this word is used in many different contexts, meaning the word “mother” and its variations. “эне” is a noun, therefore, we tag it as follows: эне – <n> <aa> <sg> <nom>
**Conclusion**

As we see, languages have different Morphotactics, especially Kyrgyz language has various specific sides that are important to note and investigate. It is significant to research language on a good level in order to synthesize and process in computation. Thus, foreigners would be able to understand features of our native language and would be interested to learn it.
References:
retrieved from Research Methodology on Natural Language Processing; 2009