Analytical grammatical forms extraction as a new challenge (case of conditional mood in Polish and Ukrainian)

A recent boom and development of grammatically annotated corpora has opened a wide range of possibilities, and, at the same time, it has made them more difficult to deal with. Since the words in corpora are marked up with a rich set of tags indicating the morphological, syntactic and even semantic features, to access them a researcher might need to be instructed in using a specific interface or, which is often more complicated, to build up a query in a specific query language, such as *CQL*, which can include information about the lexeme, its grammar and other parameters, as well as a sequences of all this data. Since grammar level is highly formalized, such an approach seems to allow providing exhaustive data about grammatical characteristics of a given word. Nonetheless, some categories are expressed by analytical forms, i.e., are composed of two or more physical words, such as the Present Perfect tense in English. The components of these categories do not necessarily follow each other, i.e., in certain context might be separated by other words or even might be inverted. A particular interest with regard to both annotating and selecting analytical grammar forms may be generated by the conditional mood in some Slavic languages (Belorussian, Check, Polish, Russian, Slovak and Ukrainian) as expressed by means of two words: a past verb form and the particle “б/би/бы/by”, which is why in most modern corpora this category lacks a specific tag. The case of Polish is particularly puzzling because the particle “by” may be either merged with the infinitive or used separately, and, on top of this, it contains a personal verb ending.

Our goal is hence to show the method of building up queries for extracting analytical grammar forms and to illustrate their usage by counting the possible distance between its components. This distance presumably would rarely overpass 7 words, assuming that a speaker should keep in memory both words, and the operative memory limits revolve around 7±2 units, as famous George A. Miller found out (Miller 1956, p. 94), though the empirical data shows tha in conditional mood the distance between the particle and the verb form can sum up to 14 words, according to the data of *NKJP* and *GRAK* ; the selection and revision in a grammatically annotated corpus can be performed with a *CQL* query which takes into account the base form of the particle, the verb form (either participle or infinitive), while in the intermediate words the punctuation marks, another participles and particles, as well as infinitives should be excluded.

Besides the conditional mood, a new tough challenge arises due to word-oriented tagging, which is predominant in modern corpora. Not only does this posit the problem of other analytical grammar forms, such as Future Tense, comparative degrees of adjectives and adverbs, but also that of a bunch of analytical lexical-grammar structures, where one of the words can be used as an auxiliary, preceding or following the notional component, used both separately or adjacently to it.

George A. Miller. (1956). The Magical Number Seven, Plus or Minus Two. *The Psychological Review*, vol. 63, 81-97.

NKJP, Narodowy Korpus Języka Polskiego http://nkjp.pl/poliqarp/nkjp300/query/

GRAK (ГРАК), Генеральний Регіонально Анотований Корпус Української Мови [General Regionally Annotated Corpus of Ukrainian] http://www.parasolcorpus.org/bonito/run.cgi/first_form