https://dx.doi.org/10.17651/POLON.42.9

POLONICA XLII
PL ISSN 0137-9712

SERHII FOKIN
Taras Shevchenko National University of Kyiv
ORCID: 0000-0003-3920-1785

Analytical grammar forms extraction as a new challenge for corpora (Case of conditional mood in Polish and Ukrainian)*

1. Introduction

The conditional mood is an analytical grammatical category, which is why it cannot be morphologically tagged at the token level. Most modern corpora lack analytical categories annotation, although there are some outstanding exceptions, such as the Helsinki Annotated Russian Corpus HANCO (Hanco, 1999–2018). Such an annotation allows searching for conditional mood, analytical future tense, and analytical comparative degrees, accessible through a rich user interface. Nevertheless, in most cases, a user needs to deal with analytical phenomena by their own means; given that most analytical categories possess specific markers (typically, their auxiliary parts), they can serve as necessary (although not sufficient) conditions and a reasonable starting point for the search. In many Slavic languages, the particle *by* followed or preceded by a verb in the past form, for example, is quite a univocal marker for mining conditional mood forms, whose strengths and weaknesses are explored further in this paper.

With respect to the formal aspect of analytical forms, some researchers mention the technical difficulty of such selection. M. Alexandrov, X. Blanco, O. Mitrofanova, and M. Zakharov point out the issue of tagging broken-off analytical grammar forms (Alexandrov, 2007, p. 14). T. Jelínek, B. Stindlová, A. Rosen, and J. Hana report errors in tagging analytical forms, modal verbs, and copulative verb predicates (Jelínek et al., 2012, p. 133). A. Rosen, J. Hana, and B. Štindlová also observe that the tagging of compound forms

* The material reported in this paper was presented at the SlaviCorp panel during the 8th Grammar & Corpora Conference 2020 organised by the Institute of Polish Language, Polish Academy of Sciences, Krakow, Poland (25–27 November 2020).

usually leads to errors (Rosen et al., 2014, p. 12). In the UD morphological tagset, 14 possible moods, including the conditional, are provided with tags; although its analytical nature is not on the focus of attention (Mood tagset, 2022).

Finally, S. Fokin explores the possibility of using query languages for extracting analytical future tense forms in several Slavic languages corpora (Fokin, 2021).

First, it is essential to provide the researcher with the tools to select other analytical grammar forms. On the other hand, it is worth exploring some successful practices in this respect and trying to achieve a better outcome by modifying the sampling parameters as well as performing an accuracy measurement.

The modern, highly developed query languages allow for a much more precise search involving multiple parameters and conditions, which can help rule out unnecessary examples. Some query languages allow the user to specify the number of interrupting words, their grammar, and semantic features. Furthermore, with a query language, a search can be framed in a specific way, including several grammatical filters at a time, as is the case when extracting passive conditional mood forms or singular ones. In contrast, a user interface, however comfortable it may be, imposes pre-existing options and lacks such flexibility. For example, the use of semantic tagging in a query allows the user to search specifically for the conditional mood of only action verbs, or to limit the search to other semantic groups. Thus, it is evident that query languages provide users of corpora with a more powerful tool than a user interface can do.

Our intention is, hence, to prepare, perform, describe, and evaluate the query language potentiality for extracting the forms of conditional mood in Polish and Ukrainian by means of the Poliqarp Query Language (Przepiórkowski, 2011) and Corpus Query Language (*CQL*). Although these languages are different, the discrepancies basically concern the tags and attribute naming; therefore, it is appropriate to consider them concurrently. Since in this case, the corpus query is prepared and framed by a human user and based upon logical reasoning, the method is to be considered as a rule-based approach. Conversely, many modern-day corpora are automatically tagged by trained neural models. The rule-based approach is more likely to outperform the neural one due to a greater control over all the stages and parameters of sampling.

2. Conditional mood structural properties relevant for the Query

We should start exploring a number of debatable questions, including that of the conditional mood categorical semantic and structure, which is to be examined before formalizing it in a query language.

In most cases, Polish and Ukrainian grammarians describe the conditional mood as a structure including a past verbal form with the particle *by/би.* These structures may denote potential conditional mood (past + *bym*, *byś*, *by*, *byśmy*, *byście/би*) and unreal conditional mood (past + *byłbym*, *byłbyś*, *byłby*, *bylibyśmy*, *bylibyście*, *byliby/був би*). At the same time, given that these forms in Polish are often parts of more complex structures, especially conjunctions, some debatable questions arise regarding the classifications of these cases. As M. Zaleska states:

> In Polish one can delineate two series of conjunctions whose formal differentiating element is by, for instance *że* VS *żeby* "so that", "in order that"; *gdy* VS *gdyby* "if", "when if"; jak "how" VS jakby "as if". An identifying mark of these "particular uses of conditional" (as opposed to the normal use) is the agglutination of personal and number markers to the conjunctions. The main point of scholarly discussions is whether (an inflected construction such as żebym robil "so that I do" should be described as że + *bym* (i.e., as the conjunction agglutinated to a verbal morpheme of the conditional which carries modal values) or rather as żeby + *m* (i.e. as conjunction modelized by *by* and agglutinated to the verbal morpheme of the preterite) (1999, p. 140).

In other words, the fact that *bym*, *byś*, *by*, *byśmy*, *byście* appear united with the conjunction still does not deprive the particle of its functions, although from the formal standpoint, że + *bym*, *byś*, *by*, *byśmy*, *byście* qualify as conjunctions.

The component *by* is considered in some cases as a "floating inflection, particularly, as the morpheme bearing information on person and number can be attached to the verb itself […] or to some other word within the sentence […]. For that reason we always consider such a 'floating inflection' morpheme as a separate segment" (Przepiórkowski, 2003).

It turns out that, in this case, following the approach of Janusz S. Bień and Zygmunt Saloni, especially suitable for the formal POS-processing of tokens, the conditional verbal form, being totally coincident with the past participle, would perfectly qualify as a past tense form. Which is why the conditional mood remains out of the scope of morphological annotation. Many corpora provided with morphological annotation lack a syntactic one, which is why in most cases no tags are provided for this category. D. Zeman's proposal to annotate only the particle *by* (in the author's words, "the conditional mood (both present and past) is formed periphrastically using the active (*l*-) participle of the content verb and a special form of the auxiliary verb to be. The auxiliary form is annotated Mood=Cnd, the participle is not)" (Zeman, 1999, p. 169).

Although this proposal sounds innovative, to work properly, it should involve the disambiguation of the component *by* in a given syntactic context, either as per a rule-based approach, or by using neural models, since this particle can be associated with non-conditional semantics.

A deeper semantic view of the conditional mood forms can shed more light to this issue:

> A criterion to question the conditional nature of structures with żeby and similar, i.e. the purpose clauses, could be that of the absence of the opposition between real and unreal conditions in the purpose clauses […]. From the semantic point of view, in the purpose clauses, the opposition between real and unreal conditions is not possible, like in the example *Podkreślił ważniejsze kwestie, żeby się nad nimi zastanowiła* (Gramatyka współ., 1999, p. 186).

The question becomes even more debatable once we consider the case of *gdyby*, which introduces conditional clauses much more closely resembling conditional mood semantics than purpose ones. What is more, some scholars do recognize the purpose clauses predicates preceded by *że + by…* as conditional mood (Szober, 2022). M. Gaszyńska-Magiera also refers to works qualifying the compound forms of the particle *by* with conjunctions in the purpose clauses as a part of conditional mood (Gaszyńska-Magiera, 1998, p. 56).

Ultimately, both approaches (i.e., whether either qualifying purpose clause predicates as conditional mood forms or not) could be valid. However, the goal of the experiment being conducted here is to demonstrate the possibilities of a query language and performance of the query in terms of accuracy and precision for automating the extraction of conditional mood forms in Polish and Ukrainian (both merged and broken forms) and clarify the factors which influence the accuracy and precision metric. Since the case of purpose clauses remains a debatable question, we limit our queries of the pure conventional forms of the conditional mood; that is, we exclude the purpose clauses from our sampling, but include the conditional clauses introduced by *gdy* + particle conjunctions.

From the morphological point of view, the conditional mood in Polish and Ukrainian consists of a past verbal form and the particle *by/би*, the latter in most cases used separately from the verb. The particle *by/би* can either follow or precede the verb; both structural parts of the query can be separated by other tokens (here and after called "separators", see Schema 1). All their structural components must occur within the boundaries of a sentence:

verb in past + [SEPARATORS] + particle *by/би*

WITHIN A SENTENCE

OR

particle *by/би* + [SEPARATORS] + verb in past

WITHIN A SENTENCE

Schema 1. Conditional Mood Structure

As a grammatical structure, the conditional mood is not a mere set of forms: to be considered as such, it should be used in compliance with its functions. For example, in the sentence "Gdyby wrócili mieszkali by tutaj" (NKJP: Maria Nurowska, 1993) there are two past participles in the nearest context of the particle *by*, and only one of them (*mieszkali by*) is a part of the conditional mood. Such examples illustrate that a functional analysis performed by researchers is needed to rule out false positive results yielded by the query. To perform a more refined search of the conditional mood forms, it is crucial to consider the boundaries of the sentence where it appears.

For this reason, we are constrained to consider any complete or incomplete sentence in the corpus as a candidate for a true positive result, so long as it contains the aforementioned structure. Thus, the main formal features to be taken into account by the query are the particle *by* along with its paradigmatic variants (alternatively, some merged conjunctional forms such as *gdyby*) and its respective past participle form. They serve as the main hooks to fish the conditional mood forms out of a corpus not provided either with a specific syntactic nor a morphological annotation for this category.

3. Sampling and Methodology

Sampling through corpus queries based on the conditional mood structural features, data volume normalization and accuracy measurement constitute the core methodology followed in this research.

Let us now provide a step-by-step explanation of how a query is built in *Poliqarp* and *CQL* languages. Each pair of square brackets in *Poliqarp* (as well as in *CQL*) describes a token. A particular feature of a token is expressed by the query attribute, while the value is assigned to this attribute by means of the "=" sign. Thus, Query 1 would cover any sequence of a verbal past form followed by the particle *by*. Since the past tense is a morphological feature, it is denoted in the first token by using the attribute *tag* and its value *praet* (i.e., "past time"), whereas the second one refers literally to the lexeme *by*, which is assigned to the *base* attribute. Since a *tag* may include multiple values besides *praet*, it is essential to account for this possibility by the use of regular expression syntax: the sequence ".*" (dot and asterisk) in the queries stands for the sequence of any possible symbol.

Query 1

[*tag=".*praet.*"*][*base="by"*]

Query 2 covers the cases of possible inversion of the past participle and particle. The latter may follow or precede the past form. The vertical slash in the middle of Query 2 represents the disjunctive logical operator OR:

Query 2

[*base="by"*][*tag=".*praet.*"*]|[*tag=".*praet.*"*][*base="by"*]

Both parts of the conditional mood can be interrupted by other tokens, *e.g.*: "*Bardzo* ***by*** *nam się* ***przydał***" (NKJP: Rosław Figura), where *nam się* are the tokens that are not part of the conditional mood form. The query snippet (3) describes possible separators of the participle and the particles. A separator should not be a past form, particle *by* nor a punctuation mark (exclamation signs stands for negation in the query, logical "No"). Although this assumption might seem quite polemic, we reconsider it at the analysis stage of false negatives. The numbers in parentheses stand for the range of separators numbers (i.e., from 0 up to 7 tokens). This number may change depending upon the needs of the query; the choice of 7 tokens as a maximum is argued in Section 5.

Query 3

[*tag!=".*praet.*"&tag!=".*interp.*"&base!="by"*]{*0,7*}

We should also assume that all the structural parts of a conditional mood form should occur within the boundary of a sentence. This assumption is indicated with the operator "*within*" as follows: "*within s*". It is worth noting that the separators of the conditional

mood forms should not repeat the structural parts of the conditional mood itself, i.e., they should not be verb past forms nor particle *by*; we also assume that they are not likely to contain punctuation marks within the structure. Let us specify these conditions inside the brackets of the separator tokens through an exclamation mark in the expression "!=" ("not equals"):

Query 4

*[base="by"][tag=".*praet.*"]|[tag=".*praet.*"][tag!=".*praet.*"&tag!=".*interp.*"&base!="by"]{0,7}[base="by"] within s*

The respective Query 4 in the *CQP* language, "translated" into *CQL* and applicable to the GRAK corpus of the Ukrainian language, would yield 801,764 results, which would read as follows:

Query 5

*[word="би|б"][tag!=".*past.*"&tag!=".*punct.*"&word!="би|б"]{0,7}[tag=".*past.*"]|[tag=".*past.*"][tag!=".*past.*"&tag!=".*punct.*"&word!="би|б"]{0,7}[word="би|б"]*

3.1. Merged Particle Forms

The particle *by/*би in both languages can be merged with some tokens (verbs or conjunctions). It can be observed that the corpus developers have accounted for this feature; in some cases (*e.g.*, żeby, *якби*), the particles are processed as separate tokens, although some tokens (*e.g.*, *gdyby*, *щоб*) are not separated from the stems. The latter being treated in the corpus as a whole token (see Fig. 1), we need to extend the query with the alternatives of the particle *by* such as *gdyby*. The extended query would appear as follows:

Query 6

*[base="by|**gdybym**|**gdybyś**|**gdyby**|**gdybyśmy**|**gdybyście**|**gdyby**| |jeśliby|jeślibym|jeślibyś|jeślibyście|jeślibyśmy"][tag!=".*praet.*"&tag!=".*interp.*"&base!="by|gdyby|jeśliby|jeślibym|jeślibyś|jeślibyście|jeślibyśmy"]{0,7}[tag=".*praet.*"]|[tag=".*praet.*"][tag!=".*praet.*"&tag!=".*interp.*"&base!="by"]{0,7}[base="by"] within s*

Fig. 1. The particle *by* merged to other wordforms in the National Corpus of Polish (NKJP)

The respective query for Ukrainian (see Query 7) in GRAK-12 yields 1,323,674 results:

Query 7

[*word="би|б"*][*tag!=".*past.*"&word!="би|б"*]{*0,7*}[*tag=".*past.*"*]|[*tag=".*past.*"*] [*tag!=".*past.*"&word!="би|б "*]{*0,7*}[*word="би|б"*] *within <s/>*

Query 7a

[*word="би|б"*][***tag!=".*punct.*"&***tag!=".*past.*"&word!="би|б"*]{*0,7*}[*tag=".*past.*"*]|[*tag=".*past.*"*][***tag!=".*punct.*"&***tag!=".*past.*"&word!="би|б "*]{*0,7*}[*word="би|б"*] *within <s/>*

In the case of including the purpose clauses' conjunction, the query would read as follows:

Query 8

[*word="би|б|щоб|абы"*][*tag!=".*past.*"&word!="би|б|щоб|абы"*]{*0,7*}[*tag=".*past.*"*]|[*tag=".*past.*"*][*tag!=".*past.*"&word!="би|б|щоб|абы"*]{*0,7*}[*word="би|б"*] *within <s/>*

Given that parts of queries contained within quotation marks in the *Poliqarp* language and *CQL* are processed as regular expressions, Query 6 can be substantially reduced by using regular expressions for covering similar forms simultaneously with one query. Since the difference is in the endings of these forms while the beginning of conjunctions is identical, the usage of wildcard characters seems reasonable (see the dots and asterisks in the Query 8).

Query 9

[*base="****by|gdyby.*|jeśliby.****"*][*tag!=".****praet.****"&tag!=".*interp.*"&base!="****by|gdyby.*|jeśliby.****"*] {*0,7*}[*tag=".****praet.****"*]|[*tag=".****praet.****"*][*tag!=".****praet.****"&tag!=".*interp.*"&base!=" ****by|gdyby.****|* ***jeśliby.****"*]{*0,7*}[*base="****by|gdyby.*|jeśliby.****"*] *within s*

Including the purpose clauses' conjunction, the respective query for Polish would be as follows:

Query 10

[*base="****by|gdyby.*|jeśliby.*|żeby.****"*][*tag!=".****praet.****"&tag!=".*interp.*"&base!="****by|gdyby.*|jeśliby.*|żeby.****"*]{*0,7*}[*tag=".****praet.****"*]|[*tag=".****praet.****"*][*tag!=".****praet.****"&tag!=".*interp.* "&base!=" ****by|gdyby.*|jeśliby.****"*]{*0,7*}[*base=" ****by|gdyby.*|jeśliby.****"*] *within s*

The parts selected in bold represent specific grammatical category parameters, whilst the rest of the query functions merely as a wrapper for them to indicate the possible inversion of components, and to signal a distance between the two components in case the conditional mood is a broken-off form. Hence, the model of Query 10 can be extrapolated to other analytical categories, *e.g.*, passive voice. This can be achieved by indicating the infinitive of the verb *to be*: *być|zostać*, instead of the value of the attribute base,

selected in bold, *by|gdyby.*|jeśliby.** and the past passive participles *.* ppas.** instead of the past verbal forms *.* praet.**:

Query 11

*[base="być|**zostać**"][tag!="**.*ppas.***"&tag!=".*interp.*"&base!="być|**zostać**"]{0,7}[tag="**.*ppas.***"]|[tag="**.* ppas.***"][tag!="**.* ppas.***"&tag!=".*interp.*"&base!="być|**zostać**"]{0,7}[base="być|**zostać**"] within s*

3.2. Custom Query Tuning

Within the same query language, some corpora may possess their own specificity in terms of usage. This specific usage is comparable to a dialect, in the sense that some rules of language usage in their context are slightly modified. For example, in the KORBA corpus (Korpus barokowy, 2013–2018), Query 10 would undergo several changes to be successfully performed since the exclamation mark, which stands for logical negation in the queries, precedes the attribute name instead of following it:

Query 12

*[base="**by|gdyby.*|jeśliby.***"][!tag="**.*praet.***"&!tag=".*interp.*"&!base="**by|gdyby.*|jeśliby.***"]{0,7}[tag="**.*praet.***"]|[tag="**.*praet.***"][!tag="**.*praet.***"&!tag=".*interp.*"&!base="**by|gdyby.*|jeśliby.***"]{0,7}[base="**by|gdyby.*|jeśliby.***"] within <s/>*

3.3. Defining Sub-corpus Constraints

The accuracy measurement imposes some objective criteria, which must be maintained for the purpose of achieving reliable conclusions. For objectivity's sake, we need to scale and normalize the data, so it may be both manually reviewable and statistically persuasive. Query 7 and Query 7a yield millions of examples out of the GRAK corpus, which is beyond human capability to process. The queries in the National Corpus of Polish have a constraint of retrieving exactly 1,000 results, therefore, the GRAK corpus results will be used as the main statistical input.

The Sketch Engine interface, implemented for this corpus, allows for creating sub-corpora of smaller volumes. This feature comes in handy to delimit a corpus of interest so that it may contain statistically representative data. In our case, constraining ourselves to the order of 1,000 results is manageable; it is short enough to be manually processed whilst still being quite representative from an illustrative point of view. To accomplish this, we start limiting its sections until we reach the optimal approximation to the order of around 1,000 results. In the GRAK-12 corpus, by filtering only academic texts from the year 2004, we finally retrieve from 718 up to 1,172 results depending on the separators number, which the aspired sample volume.

3.4. Optimal Separators Number

Before establishing the optimal number of separators empirically, let us explore the pertinent theoretical background. Our operative memory limits allow for keeping around 7±2 items (items can correspond to words or other sort of units), as the notable American psychologist George A. Miller established (1956, p. 94). As applied here, these observations mean that if there is a gap of more than seven items between both structural parts of a grammatical form, a reader or listener might forget the first part before reaching the second one, which should not occur in effective communication. Accordingly, in our queries we have limited the number of separators of both parts of the analytical category in question to seven tokens. Raising the number of separators might significantly increase the frequency of false positives, which is why the indicated number seems just a sort of Occam's razor to segregate potentially useful and wrong results. Therefore, we aim to put this hypothesis to the test, argue for it, and possibly decline it.

The preliminary raw results seem quite contradictory, since Query 7 keeps yielding more and more results as the number of separators rises, as shown in Table 1, whereas Query 7a does not:

Distance between verbal past form and particle	Number of examples extracted
0	718
0 – 1	769
0 – 2	808
0 – 3	972
0 – 4	983
0 – 5	991
0 – 6	994
0 – 7	982
0 – 8	1000
0 – 9	1026
0 – 10	1048
0 – 11	1062
0 – 12	1075
0 – 13	1092
0 – 14	1101
0 – 15	1110
0 – 16	1121
0 – 17	1133
0 – 18	1136

Distance between verbal past form and particle	Number of examples extracted
0 – 19	1140
0 – 20	1146
0 – 21	1149
0 – 22	1154
0 – 23	1160
0 – 24	1163
0 – 25	1163
0 – 26	1165
0 – 27	1168
0 – 28	1171
0 – 29	1172
0 – 30	1172

Table 1. Correlation between the number of separating words in the query and number of results as per Query 7

Distance between verbal past form and particle	Number of examples extracted
0	718
0 – 1	769
0 – 2	808
0 – 3	822
0 – 4	837
0 – 5	844
0 – 6	845
0 – 7	848
0 – 8	848
0 – 9	851
0 – 10	852
0 – 11	852
0 – 12	852

Table 2. Correlation between the number of separating words in the query and number of results as per Query 7a

As seen in Tables 1 and 2, including punctuation marks as possible separators yields 137 more results. Our fundamental assumption here is that a single grammatical structure

like conditional mood is a single sentence constituent, and it is not likely to be interrupted by punctuation marks. Hence, we focus on processing the results of Query 7, i.e., excluding the punctuation marks from the query with up to seven separators (848 examples totally yielded). If this assumption is wrong, the number of false negatives (i.e., useful examples missed by the query) will be significant.

3.5. Exploring Cases with over Seven Separators

Among the three examples of broken-off conditional mood forms with a distance higher than seven, only false positives were found. In two out of three examples, the verbal predicates were used in different clauses placed side by side, united through a copulative conjunction and not separated by punctuation marks, in accordance with the punctuation rules commonly accepted in Ukrainian for such cases. The particle *by* referred to previous conditional mood form; thus, there was a boundary between two conditional forms:

> *Це уможливило б поступове перетворення української мови на мову загальнотериторіального поширення і зробило б її таким чином загальнодержавною* (GRAK, 2017–2022: Орест Ткаченко, 2004, Українська мова і мовне життя світу).

> *А потім з часом, шляхом дедалі більшої українізації російської мови східноукраїнських міст і деякого зближення з цією мовою української мови в Західній Україні, виник би синтез цих двох українських мов і на цій основі утворилася б нова українська мова* (GRAK, 2017–2022: Орест Ткаченко, 2004, Українська мова і мовне життя світу).

In the third example, the false positive result owed to a wrongly parsed sentence: two sentences were subsequently merged by mistake, the particle and the verb were originally in different ones and did not constitute a single conditional mood structure:

> *Очевидно, якби загроза витіснити англійську мову з боку французької тривала б і надалі, ставлення до французьких елементів в англійській мові було б зовсім іншим, набагато стриманішим, а можливо, і зовсім протилежним, різко негативним, таким, яке б зовсім виключало прийняття до мови елементів У такому разі їх би, очевидно, з мови викидали і замінювали словами германського походження, тим більше, що англійська мова мала б цілком у своєму розпорядженні цю можливість, – для цього можна було б використати як багатющі запаси давньоанглійської мови, так і моделі споріднених германських мов, на підставі яких, шляхом їхньої англізації, можна було б створити, замінивши ними французькі запозичення, безліч суто германських слів* (GRAK, 2017–2022: Орест Ткаченко, 2004, Українська мова і мовне життя світу).

Another factor inducing a false positive can be observed in Polish: the particle *by* may be homonymic to the conjunction of purpose *by*, so that the query will also yield this result:

> *Trzej mężczyźni, którzy nim* ***podróżowali, zażądali, by oddał*** *im telefon komórkowy* (NKJP: MAW).

Apart from being an erroneous example, a closing comma after a relative sentence is missing in this fragment.

As stated earlier, qualifying this sort of structures with purpose conjunctions as conditional mood is debatable.

3.6. Accuracy Measurement

Traditionally, four parameters are measured to evaluate model performance:

1) true positives (TP, truly detected samples);
2) true negatives (TN, correctly ruled out samples);
3) false negatives (FN, erroneously left out samples); and
4) false positives (FP, wrong samples considered as true by mistake).

Accuracy is defined as the overall ratio of correctly detected samples, i.e.:

$$\text{ACCURACY} = TP + TN/(TP+FP+TN+FN)$$

In the research performed here, it seems implausible (or at least extremely tedious) to calculate accuracy, since true negatives are difficult to count; it implies manual analysis throughout each clause of the sub-corpus for possibly missing cases of the conditional mood.

On the other hand, with Sketch Engine queries, the number of sentences can be calculated. Although this number does not exactly cover the number of clauses, it may correlate with them quite closely. The latter could be calculated in case the corpus was fitted with syntactic dependency tags; therefore, a much more finely tuned calculation of the true negative samples will be available in the future. The sub-corpus chosen contained 39,150 sentences.

The selection of all true values constituted a methodological issue as well, since no etalon set of sentences with the conditional mood could have been used for comparison. To bridge this gap, we fulfilled a query of all the particles *б*/*би*; this yielded 1,081 cases across the entire sub-corpus. Then, we manually mined out of the concordance generated all the sentences containing examples of the conditional mood. We finally obtained 852 such examples qualified here as all the true examples. Since Query 7a yielded 848 results, the true negatives could be roughly evaluated by the subtraction of the number of positives from the overall number of sentences in the sub-corpus, i.e., 38,302 examples.

At the same time, it was relatively easy to calculate the FPs, and FNs in the explored sub-corpus, which are represented in Table 3:

Score	Value
Results yielded by query 7	848
True positives	845
False positives	3
False negatives	7
True negatives	38,302

Table 3. Accuracy scores of the query for extracting conditional mood forms from the GRAK 9 corpus

The figures indicated allow us to calculate the following features:

ALL FOUND SAMPLES = TRUE POSITIVES + FALSE POSITIVES = 845 + 3 = 848

This number, in fact, corresponds to the size of concordance yielded by the query:

ALL CORRECT SAMPLES = TRUE POSITIVES + FALSE NEGATIVES = 845 + 7 = 852

Once obtained these data, we can proceed to calculate the precision and recall:

Precision = TP / (TP+FP) = 845 / (845+ 3) = 99.6%

Recall = TP / (TP + FN) = 845 / (845+ 7) = 99.1%

Accuracy is calculated in the following way:

Accuracy = 845 + 38,298 / (845 + 38,298 + 7 + 3) = 99.97%

Unfortunately, such high accuracy still neither confirms nor negates the model's validity, since the occurrence of sentences with conditional mood turns out to be lower than 1%. It can be compared to a straightforward model for diagnosing diabetes: even if it simply negates the diagnosis in all cases without accounting for any particular patient parameters, it will be nearly 90% accurate, since the occurrence of diabetes is around 10% worldwide. An awareness of this polemic leads us to focus mainly on *precision* and *recall* for the purpose of evaluating model performance.

4. Results and Discussion

4.1. Interpretation of the Score

A precision rate higher than recall means that the selection contains little "rubbish" (false positives). At the same time, there must be some correct samples missing (false negatives). Further fine-tuning of the query should depend on a general goal: for exhaustive sample extraction, 100% recall is needed. This is theoretically achievable by reducing the filters (constraints) concerning the separators' characteristics. One may query whether such high recall is really needed. Our answer would tend to be negative. Dealing with a large volume of data, we often need to quickly evaluate the general tendencies in relative terms, for example, such as to outline the contrastive feature. In this case, precision of 99.6% and recall of 99.9% seem persuasive.

A precision rate above 90% seems acceptable for many models. For example, the models for corpora tagging elaborated by the Institute of Formal and Applied Linguistics, Charles University and Faculty of Mathematics and Physics (Czech Republic) are precise within a range of 85% (or even lower for some language levels in low-resources languages), and 98% for grammatical features (Institute of Formal and Applied Linguistics, Charles University 2022). On the other hand, since we have applied a straightforward rule-based approach, it should come as no surprise that it outperforms the neural networks'

scores. Thus, we should not stop elucidating the reasons for the errors and looking for fixes. For example, R. McDonald and J. Nivre observe that the accuracy of syntactic dependencies parsing starts dropping drastically with structures of at least ten words in length: "Having analyzed the FP and FN, we can conclude that errors do happen in case of large distances, as J. McDonald points out, the parsing model achieved accuracy (84%) starts dropping from 10 tokens distance in dependencies" (McDonald, et al. 2007). Thus, by increasing the distance, we are more likely to overfit our model, which would include more false positives, as we have confirmed with the present survey.

4.2. Reasons for False Positives and False Negatives

In some examples, an unresolved POS-ambiguity can lead to a mistaken (false positive) result: in the following example, the POS-ambiguity of the noun in genitive case виправ (виправа "exercise") has been tagged in the corpus as the verb випрати ("to wash") in the past participle form:

Були висунуті антистаршинські гасла, опозиція не сприйняла і спроб Виговського налагодити військовий союз із Кримським ханством, що, як і за часів Хмельницького, стало б на перешкоді морських виправ запорожців (GRAK, 2017–2022: Юрій Мицик, 2004).

Here, another example illustrates how a distance filter of six tokens leads to a false positive result, since the particle and the verb belong to different conditional mood forms:

У цьому випадку неросійські автономні утворення, що входили б до складу Співдружності, могли б або залишитися під спільним дахом з етнічною Росією, або встановлювати щодо неї вищий ступінь незалежності, але це не перешкоджало б нормальним економічним **і** *культурним зв'язкам* **і** *не потребувало таких кровопролить, які відбуваються на Кавказі і які ще можливі в інших місцях пострадянського простору* (GRAK, 2017–2022: Орест Ткаченко, 2004).

Let us cursorily notice that unless we had excluded punctuational tokens from possible separators (i.e., applying Query 7 instead of Query 7a), several cases of the letter *б* used as numeration were selected by the query and would have caused additional false negative results:

Поточна процентна ставка дорівнює 6% у США і 4% у Японії. а) Якби фірма США обміняла долари на йєни сьогодні та інвестувала єни в Японії на один рік, скільки доларів їй потрібно сьогодні? ***б*** *Якби фірма| США* ***вдалася*** *до ф'ючерсного контракту* (GRAK, 2017–2022: Глен Габбард, 2004).

Since parentheses are punctuational marks, similar cases of numeration are filtered out by Query 7a.

Some of the reasons indicated can lead to false negatives as well. One more cause of the false negatives' occurrence is the possibility of punctuational separators within the conditional mood. Two of the false negative results were caused by two missed examples of a conditional mood separated by a comma, e.g.:

Політика ця призвела до такого становища, яке Пантелеймон Куліш у«Листах з хутора» окреслив роздумом-спожалінням:«Нехай би через науку, через освіту простого нашого люду не меншало (GRAK, 2017–2022: Іван Ільєнко, 2004).

Another FN was due to a misspelling of the verb as a consequence of a scanning mistake and, consequently, incorrect tagging:

Наприклад, замість обчислення поточної вартості платежів, які буде отримано внаслідок придбання облігації Державної скарбниці зі строком погашення 30 років, обчислюють процентну ставку, за якої заплачені за облігацію гроші можна бу-ло б інвестувати на 30 років і одержати таку саму поточну вартість (GRAK, 2017–2022: Глен Габбард, 2004).

Назвіть кілька потенційних інформаційних проблем на цих ринках, наявність яких ви-правдовувала б втручання кредитора останньої інстанції (GRAK, 2017–2022: Глен Габбард, 2004).

Така критика поставила ФРС перед дилемою: якщо б вона відкрито роз-ширила допустимі межі коливання грошових агрегатів для стабілізації процентних ставок, їй довелось би визнати, що вона до певної межі використовувала у ролі поточного орієнтиру федеральну резервну ставку, що вона до цього обіцяла не робити (GRAK, 2017–2022: Глен Габбард, 2004).

With respect to the overall false negatives score, we can summarize that those seven missed results, when compared to potentially 137 false positives have proved to be the correct choice. In other words, Query 7a has caused only seven useful results to be missed, whereas Query 7a would have increased the number of wrong results by 137 items, 135 of which would be mistaken, without solving the problem of misspelled scanned texts. Furthermore, this fact also confirms the ground hypothesis: the conditional mood is not likely to be separated by more than 7±2 tokens.

4.3. Syntactic Annotation Perspective

The accuracy achieved of 99.97% based on the rule-based approach being subject to the test can be extrapolated onto the creation of syntactic annotation for other analytical grammar forms. The parent-child relation between the past participle and the particle *by* within the same predicate group could be established as per the following rules:

1) the distance between both elements does not exceed 10 tokens;
2) the separating tokens are not cases of *by* or past verb forms.

As a promising domain for research, some false-positive results appear on the boundary of two conditional mood forms, where the past form belongs to the previous analytical form and the particle to the next one or vice versa, include copulative conjunctions, punctuation marks such as commas, and possibly other markers which help rule out many FP. Separating tokens containing copulative conjunction as an excluding condition is still to be subject to experiment.

As per D. Zeman's idea, the attribution of the respective categorical tag to the particle *by* (or, in the author's terms, to a form of the auxiliary 'to be') would force one to tag the participle as the past tense. Although it is logical from the structural standpoint (as the conditional mood does not seem a clear candidate for a morphological category), in this case, where several past participles are present in the context, an issue may arise. Unless *by* and the past participle are linked through a sort of child – parent syntactic dependency (in the Universal Dependencies tagset, there is an aux:cnd annotation for

this specific case (Conditional Marker Auxiliaries), in some infrequent cases this might lead to the corpus users' confusion, implying the need to perform their own analysis of the example yielded. For example, *by* may be used in purpose clauses. And, in many cases, this lack of such annotation would technically prevent the search engine from highlighting both structural parts in the concordance generated for the users' convenience. In the absence of such a syntactic dependency annotation, a query language can help solving the above-mentioned issues.

5. Conclusions

Although most modern textual corpora are not provided with particular tags for analytical categories, specific corpus query languages allow performing queries to mine examples of analytical categories, as shown in the case of the conditional mood, with a relatively high precision and recall of over 99%. A precision rate reaching almost 100% may mean that the query does not yield unnecessary erroneous results, while a recall slightly lower than the precision indicator means that some useful examples are not covered by the query. Nevertheless, these missed examples are mostly due to misspelled and, subsequently, mis-tagged tokens; in some rare occasions, the conditional mood form is interrupted by a punctuation mark. On the other hand, removing the punctuation mark filter from the query increases by around fifty-fold the number of erroneous results, i.e., false positives. The main reason for false positives was the selection of the boundary of the previous and the next analytical forms, mostly in queries allowing for a significant distance between the components.

To achieve the desired results, a query of an analytical grammar form, such as conditional mood in Polish and Ukrainian, should account for the factors that may influence the accuracy, precision and recall, as well as possible variance of the structural parts of the analytical category in question (the past participle form along with the particle *by* as its immediate or rather distant "satellite"), their sequence and possible inversions, as well as the number and nature of separators. The experiment confirms that most examples of broken-off analytical forms do not exceed Miller's number (7±2 items). The forms of the conditional mood can be separated by any word or even a punctuation mark. It seems promising to hold a similar experiment for mining other analytical categories, such as the passive voice, adjective and adverb comparative degrees, and future tense forms. Another interesting perspective which may arise from the observation provided is to argue for the competitivity among rule-based and neural models in terms of the metric scores yielded, although some early observations as per current surveys suggest that the precision metrics of the currently used trained models vary from 85% up to 98%.

References

Alexandrov, M., Blanco, X., Mitrofanova O.M., & Zakharov, V. (2007). Nooj Applications for Document Clustering and Corpus Linguistics. In X. Blanco, & M.Silberztein (Eds.), *Proceedings of the 2007 International NooJ Conference* (pp. 6–19). Newcastle: Cambridge Scholars Publishing. https://www.cambridgescholars.com/download/sample/60082

Conditional Marker Auxiliaries, https://universaldependencies.org/pl/dep/aux-cnd.html (accessed: 28.10.2022).

Conditional Mood Tagset, https://universaldependencies.org/u/feat/Mood.html (accessed: 28.10.2022).

Corpus Query Language (n.d.). *Sketch Engine.* https://www.sketchengine.eu/documentation/corpus-querying/ (accessed: 28.10.2022).

Fokin, S.B. (2020). Estructura de consultas para la selección automática de formas gramaticales analíticas del tiempo futuro en lenguas eslavas. *Mundo Eslavo*, *19*, 25–38.

Gaszyńska-Magiera, M. (1998). Tryb przypuszczający w nauczaniu języka polskiego jako obcego. *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców*, *10*, 51–60.

GRAK, General Regionally Annotated Corpus of Ukrainian. (2017–2022). Генеральний Регіонально Анотований Корпус Української Мови, http://www.parasolcorpus.org/bonito/run.cgi/first_form (accessed: 28.10.2022).

Grzegorczykowa, R., Laskowski, R., & Wróbel, H. (1999). *Gramatyka współczesnego języka polskiego* (t. 1). Warszawa: PWN.

HANCO. Helsinki Annotated Russian Corpus (1999–2018). ХАНКО – Хельсинкский аннотированный корпус русского языка. http://h248.it.helsinki.fi/hanco/ (accessed 2.02.2022).

Institute of Formal and Applied Linguistics Charles University, Czech Republic Faculty of Mathematics and Physics. (2022). *UDPipe 1 Model*s. https://ufal.mff.cuni.cz/udpipe/1/models (accessed 2.02.2022).

Jelínek, T., Stindlová, B., Rosen, A., & Hana, J. (2012). Combining manual and automatic annotation of a Learner Corpus. In P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), *Text, Speech and Dialogue – Proceedings of the 15th International Conference*. *TSD 2012* (pp. 127–134). Brno: Springer Verlag.

Korpus barokowy (2013–2018). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.). https://korba.edu.pl/query_corpus/ (accessed: 2.02.2022).

Haitao, L., Chunshan, X., Junying, L. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.

McDonald, J. (2007). Characterizing the Errors of Data-Driven Dependency Parsing Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 122–131). Prague: Association for Computational Linguistics.

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

NKJP. Narodowy Korpus Języka Polskiego (2008–2010). *Poliqarp search engine for NKJP data*.

Polish Newscrawl (Leipzig Corpora Collection), http://cql.corpora.uni-leipzig.de/bonito/run.cgi/first?corpname=pol_newscrawl_2011 (accessed: 2.02.2022).

Przepiórkowski, A., & Woliński, M. (2003). *A flexemic tagset for Polish. ACL anthology*. https://aclanthology.org/W03-2905.pdf (accessed: 28.10.2022).

Przepiórkowski, A., & Wil, J. (2011). *Poliqarp Query Language.* http://nkjp.pl/poliqarp/help/ense3.html#x4-50003 (accessed: 2.02.2022).

Rosen, A., Hana, J., Štindlová, B. et al. (2014). Evaluating and automating the annotation of a learner corpus. *Lang Resources & Evaluation*, *48*, 65–92.

Szober, S. (2022). *Nauka o języku. Dla klasy trzeciej gimnazialnej*. Warszawa: Wydawnictwo M. Arcta w Warszawie.

Zaleska, M. (1999). The Irrealis in the Polish Language: A question of verbal moods, conjunctions or the modal particle *by*? In L. Mereu (Ed.), *Boundaries of Morphology and Syntax* (pp. 137–156). Roma: John Benjamins Publishing Company.

Zeman, D. (2016). Universal Annotation of Slavic Verb Forms. *The Prague Bulletin of Mathematical Linguistics*, *105*, 143–193.

SUMMARY

Keywords: conditional mood, Slavic languages, analytical grammar category, corpus query language

A particular challenge for modern textual corpora is the tagging of analytical grammar categories. The components of these categories may be separated in certain contexts by other words or may even be inverted. A particular interest regarding the selection of analytical grammatical forms is centred around the conditional mood in some Slavic languages, as expressed by means of two words: a past verb form and the particle *by*/*б*/*би*/*бы*, which is why in most modern corpora, this category lacks a specific tag for these compound forms. The case of Polish is particularly complicated because the particle *by* may either be merged with the participle or used separately; furthermore, its separated form may contain a personal verb ending. Specific queries subject to experiment on Polish and Ukrainian corpora allow selecting the analytical forms in question.

STRESZCZENIE

Wydobywanie form gramatycznych analitycznych jako nowe wyzwania korpusowe (na materiale form trybu przypuszczającego w języku polskim i ukraińskim)

Słowa kluczowe: tryb przypuszczający, języki słowiańskie, kategoria gramatyczna analityczna, język zapytań korpusowych

Szczególnym wyzwaniem dla współczesnych korpusów tekstowych jest tagowanie kategorii gramatycznych analitycznych. Składniki tych kategorii mogą być w pewnych kontekstach oddzielone innymi słowami lub nawet odwrócone. Szczególne zainteresowanie wyborem form gramatycznych analitycznych budzi tryb przypuszczający niektórych języków słowiańskich, wyrażany za pomocą dwóch słów: formy czasownika przeszłego i cząstki *by*/*б*/*би*/*бы*, dlatego w większości współczesnych korpusów kategoria ta nie ma specyficznego tagu dla tych form złożonych. Przypadek języka polskiego jest wysoce skomplikowany, ponieważ cząstka *-by* może być albo połączona z imiesłowem, albo użyta oddzielnie, ponadto jej oddzielna forma może zawierać końcówkę osobową czasownika. Specyficzne zapytania korpusowe, testowane na korpusach polskim i ukraińskim, pozwalają na wyselekcjonowanie omawianych form analitycznych.