Improving Corpus Search via Parsing

Natalia Klyueva and Pavel Straňák
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
{kljueva,stranak}@ufal.mff.cuni.cz

Abstract

In this paper, we describe an addition to the corpus query system Kontext that enables to enhance the search using syntactic attributes in addition to the existing features, mainly lemmas and morphological categories. We present the enhancements of the corpus query system itself, the attributes we use to represent syntactic structures in data, and some examples of querying the syntactically annotated corpora, such as treebanks in various languages as well as an automatically parsed large corpus.

Keywords: corpus query, treebanks, parsed corpora, syntactic search

1. Introduction

Traditionally, corpus search engines like SketchEngine[1], IMS Corpus Workbench[2] and other search in a linear representation of a sentence as "attributed strings". They display results of the searches as concordances. In comparison to that treebank search engines like PML-TQ[3](Pajas et al., 2009) or INESS[4] or Tundra[5] allow sophisticated queries for the tree structures of treebanks and they typically display results as trees (one at a time). Their query languages are more expressive, but necessarily also more complicated. In this paper we present a compromise that we believe can be useful in many situations: adding limited capabilities for syntactic search to the corpus search engine Kontext[6] with its CQL language and adding display of syntactic trees of sentences to the concordance view of the search results in Kontext.

In the LINDAT/CLARIN centre[7] we use the Kontext search engine as our main tool to allow search in our corpora, with a long-term goal to provide search for all the corpora available in the LINDAT/CLARIN repository[8].

We also provide a specialised treebank search tool PML Tree Query (PML-TQ) and currently about 70 treebanks are available for search there. However PML-TQ as well as other similar tools do not display concordances, they cannot efficiently process parsed corpora orders of magnitude larger than manually created treebanks and so on. For these reasons we believe that adding syntactic search, even if limited, and display of syntactic trees to Kontext may provide useful complement to the PML-TQ functionality not only for large datasets, but also for treebanks like Prague Dependency Treebank (Bejček et al., 2013), HamleDT (Zeman et al., 2015) or Universal Dependencies (Nivre et al., 2015), for which it provides complementary interface and abilities.

[1] http://sketchengine.co.uk
[2] http://cwb.sourceforge.net
[3] http://hdl.handle.net/11858/00-097C-0000-0022-C7F6-3
[4] http://iness.uib.no
[5] http://weblicht.sfs.uni-tuebingen.de/Tundra/
[6] https://github.com/czcorpus/kontext
[7] http://lindat.cz
[8] https://lindat.mff.cuni.cz/repository/

2. Lindat corpora

In Lindat repository, there are 648 items of the type "corpus", but most of them does not contain files, just a metainformation. We feature 88 items that contain corpora, and our main aim is to load all of them to KonText with the exception of those that are already available in the Czech National Corpus via the same search engine. We plan to parse the corpora that do not have syntactic annotation in case we have the sufficient tools. We have a parser for Czech (will be described later), for English and for some other languages. As soon, as the parsers for different languages will be trained on the UD corpora[9], we will be able to parse texts in many more languages. So far we have parsed only one corpus of Czech, see Section 4.2.. Other syntactically annotated corpora from the repository have been annotated by the authors of the resources.

Currently, in KonText, there are 42 corpora annotated on the syntactic level, 13 of them belong to HamleDT Multilingual treebank and 21 – to the UD.

3. KonText UI and syntactic information

KonText (Machálek and Křen, 2013) started as a fork of the Bonito 2.68 python web interface[10] to the corpus management tool Manatee (Rychlý, 2000). It is developed by the Institute of the Czech National Corpus (http://ucnk.ff.cuni.cz/) and is widely used by linguists for querying monolingual, parallel and speech corpora mainly for Czech language or with regard to the Czech language in case of parallel data. We adopted this interface for querying corpora from the Lindat repository (http://lindat.mff.cuni.cz/services/kontext).

The concordance that matches the query is then displayed line by line, with KWIC (key word in kontext) colored in red. For registered users, there is a possibility to show attributes of either KWIC, or in the whole text, via "View

[9] See project page https://ufal.mff.cuni.cz/udpipe
[10] Also known as NoSke – https://nlp.fi.muni.cz/trac/noske

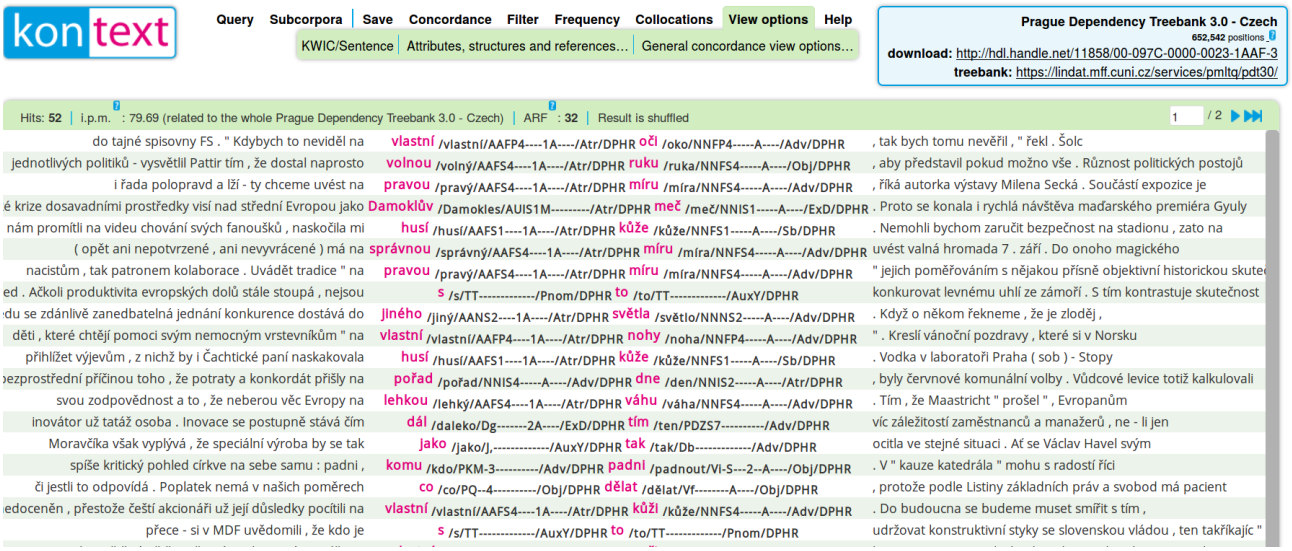

Figure 1: Concordance with displayed attributes for KWIC: word, lemma, tag, afun and functor.

options" button. In Figure 3., the concordance lines with lemma, tag and functor (deep syntactic relation) attributes are displayed.

3.1. Syntactic attributes

For syntactically annotated corpora from the Lindat repository, we introduced additional attributes of a word (node) reflecting syntactic information. This information comes from a dependency tree[11]. In addition to the original search options used in KonText – *form*, *lemma* and *tag* – we introduced the following attributes:

- attribute *deprel* – dependency relation (or *afun* - analytical function for corpora annotated in PDT style) which presents syntactic function of a word in a sentence;
- parent attributes. For each word, we added the four mentioned attributes for the parent of a node (*p_form*, *p_lemma*, *p_tag* and *p_deprel*);
- attribute *parent* specifying position of a parent with the respect to the node, e.g. [parent="-5"] means that the parent of a node stands 5 words to the left, whereas [parent="+3"] means the position three nodes to the right.[12]
- eparent attributes. Effective parent (eparent) is a special notation in the Prague Dependency theory when some non-dependency edges, such as coordinations or appositions, are skipped. For example, according to the PDT style, the conjunction is a parent of coordination members, but the 'true' parent is a node above the conjunction. For the Universal Dependency style, information related to eparent is not needed, because coordinating conjunctions are 'sisters' to the coordinated members and do not have to be skipped. The attributes that we consider in Prague-like annotation style are *eparent, ep_form, ep_lemma, ep_tag, ep_afun*.

3.2. Treex View

In addition to the default 'linear' representation of concordance, we adopted the functionality from PML-TQthat visualizes a tree and provides the information on the attributes of each node in the tree - **Treex View**[13] . Technically, a tree is generated from JSON file using Javascript library js-treex-view onsite.
An icon to display a tree is attached to each line of a concordance, the tree is visualized in PML-TQ format. When a node in a tree is being clicked, the attributes of this node are displayed, see the Figure 3.2.
So far, the functionality to view trees is enabled for some corpora, but we plan cover all syntactically annotated corpora from Lindat.

4. Examples of queries in syntactically annotated corpora

In this section, we will show the examples of querying for syntactic attributes in the three syntactically annotated corpora - Universal Dependencies, Web corpus of Czech and Prague Dependency Treebank. The queries are just illustrative and do not present any meaningful linguistic research.

4.1. Querying UD

Universal Dependencies (Nivre et al., 2015) is a project that provides the unified annotation for treebanks in 38 languages. The annotation scheme is based on Stanford dependencies (query attribute *deprel*), Google universal part-of-speech tags (attribute *pos*) and the Interset interlingua for

[11] We do not work with constituency trees, they should be processed in a different manner.

[12] It is necessary to escape the plus sign as it is evaluated as a regular expression. So the query for a parent that stands to the right will be "\+.*"

[13] https://github.com/ufal/js-treex-view/

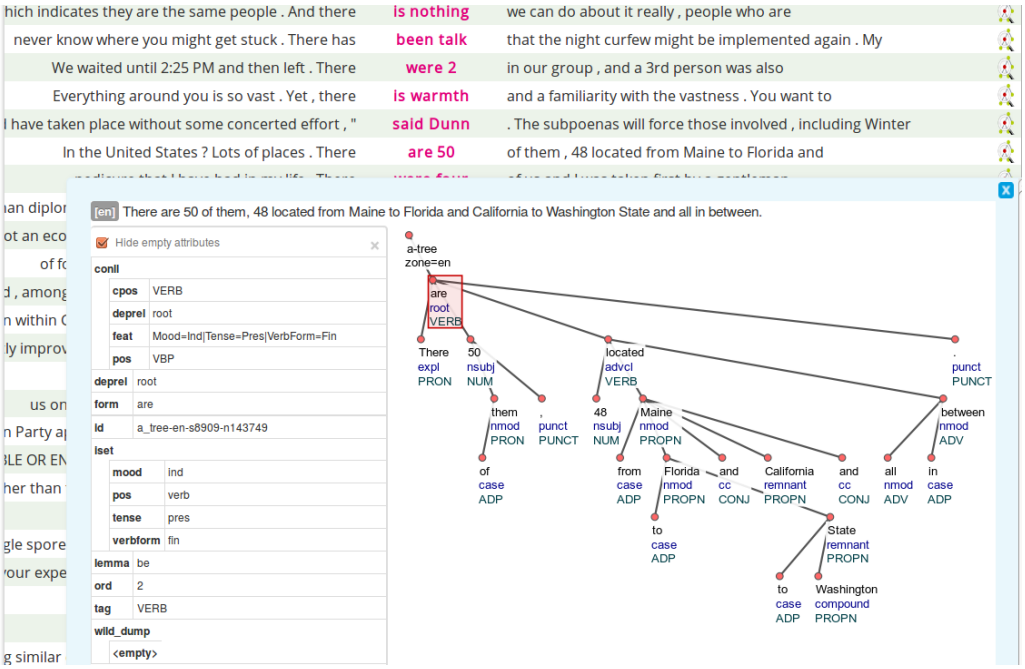

Figure 2: Concordance lines and tree visualization

morphosyntactic tagsets (attribute *ufeat*). In order to make some experiments in comparative linguistics, we compiled the joint treebank - 5,000 first sentences for each language from UD[14]. There is no sense in searching for some lexical issues, but the grammar attributes can be used to compare certain linguistic phenomena in several languages. The frequency distribution in the languages can be viewed with the function Frequency->Doc IDs (the user should be logged in to access this function), where Doc ID stand for a concrete language.

Following are some examples of queries over concrete UD treebanks.

Position of adjectives in the Romance languages. In the Romance languages, adjectives are generally placed after the noun they modify. Suppose, we want to know which adjectives precede the noun. Let us take as an example the French treebank. We put a query for tokens with part-of-speech (POS) value as **ADJ**, with a parent POS as **NOUN** that stands on the right from the adjective:

[pos="ADJ" & p_pos="NOUN" & parent="+.*"]

Then, we can make frequency analysis based on this concordance. The option 'Frequency->Node forms' (available for logged-in users) will show the list of adjectives that are used most frequently in preposition to the noun (*autres* – 'others', *premièr(e)* – 'first', *même* – 'same, similar', *grand(e)* – 'big' etc.).

The same query can be tested on other treebanks from the Latin group - Spanish or Italian - with similar results. Also, using the join treebank, we can see the frequency of the construction where adjective stands after the noun in various languages. The query for this construction will be slightly modified: we will search for the adjectives for which the parent noun stands on the left (order noun-adjective):

If we execute on this mixed treebank the slightly modified query as for French [pos="ADJ" & p_pos="NOUN" & parent="-.*"]], we will get many lines in various languages, which could be sorted according to the Doc ID (option Frequency->DocIDs). This will give the frequency distribution of postnominal adjective construction in several languages, see Figure 4.1. It is evident that Romance languages got the highest score, with Slavic and Germanic in the middle, and Finno-Ugric in the end.

Figure 3: Distribution of languages with adjective postposition.

[14]If there was more than one treebank for a language, we have chosen only one, like in case of Latin or Finnish

Word order Another query example concerns the word order. Let us examine the position of a subject and a direct object in relation to a predicate. The following query will find sentences with SVO order – where the predicate (*root*) expressed by a verb is placed on the right of a subject – *subj* and on the left of a direct object -*dobj*, provided that both subj and dobj are nouns. There is a number of words in between the arguments, that are not conjunction or punctuation:[15]

[deprel="nsubj" & pos="NOUN" & p_pos="VERB" & p_deprel="root" & parent="+.*"] [deprel!="conjlcc"]* [deprel="dobj" & pos="NOUN" & p_deprel="root"& p_pos="VERB"&parent="-.*"] within <s/>

By modifying the query above to SOV, VOS and other permutations, we calculated the number of sentences in languages matching the certain pattern and clustered them according to the language groups, see Table 1.

lang	SVO	SOV	VSO	VOS	OVS	OSV
hr	408	3	0	0	33	2
bg	296	1	2	4	22	1
cz	235	18	50	17	91	9
pl	683	19	28	25	185	30
sl	257	40	12	10	52	6
ch.sl	30	9	13	4	1	1
no	206	0	33	0	0	1
da	181	0	27	1	5	2
en	98	0	0	0	0	0
sw	319	0	61	0	8	5
du	155	1	32	1	125	0
de	157	49	66	0	7	3
got	23	70	11	3	4	1
fr	363	0	0	0	0	0
pt	249	0	0	8	3	0
es	333	0	0	4	3	0
it	285	0	0	5	2	0
ro	38	0	0	2	3	1
lat	30	0	21	19	14	29

Table 1: Surface word order in several language families in simple sentences (5,000 sentences from each treebank)

The table above should be taken with precaution, because some peculiarities in annotation schema could have led to false positives. However, the table proves the known fact that Slavic languages (especially Czech, Polish and Slovenian) have a relatively free word order in this specific case whereas Romance languages do not allow so many permutations of arguments.

On this example, we can see that introducing syntactic attributes enlarges our search options in comparison with just morphological search (lemmas and tags). On the other side, here we face the limitation of the CQL, as we can not specify the borders of the segment (clause) to search for a concrete word order pattern. Instead, we use some way around to specify there are no clause borders like punctuation in between the main constituents. Query in PML-TQ can deliver more accurate results as it supports the multi-layer search and allows us to select the clause in which we can search for a pattern.

4.2. Syntactically annotated Czech corpus CWC

One of the goals of Lindat-kontext project is to parse all the corpora from the Lindar repository that do not have syntactic annotation. Within the modular framework **treex**[16], we created the pipeline for parsing plain text corpora that includes pre- and post-processing, tagging with MorphoDiTa tagger(Straka and Straková, 2014) and parsing with MST parser[17]. Here, we describe a corpus that was processed in this ways – a large Web corpus of Czech – CWC (Spoustová and Spousta, 2011) with more than 627 million words. The attributes for the search are: form, lemma, tag and afun (analytical, or syntactic function, analogous to deprel) for a node, the respective attributes for a parent and an eparent are: ((e)p_form, (e)p_lemma, (e)p_tag, (e)p_afun). Next is the example of how we can search the corpus exploiting syntactic information. Nouns in subject position that are coordinated and that follow a verb might be found by the following query:

[tag="Vp.*"][p_afun="Coord" & ((tag="NNF.*" & afun="Sb")| tag="NNM.*" & afun="Sb")]

So far we have automatically parsed only one corpus, and we plan to parse all the corpora from Lindat that do not contain syntactic annotation.

4.3. Querying Prague Dependency Treebank in a linear manner

Prague Dependency Treebank (Bejček et al., 2013) was developed based on the Functional Grammar Description theory – FGD, it is annotated on several language layers, most important for us here are morphological, analytical (shallow syntactic) and tectogrammatical (deep syntactic) layers. We had to choose which information from PDT should be included in KonText search. First, PDT was searchable only for form, lemma, tag and afun (analytical function, e.g. Subject, Object, Predicate etc.) attributes. Then, we expanded the possibilities of the search so that we can look for the same attributes from morphological and syntactic layer as for the corpus CWC. The advantage of having PDT available via KonText, is that linguists familiar with CQL language can browse this corpus without knowledge of more complicated PML-TQ.

In addition to analytical layer, we also added some attributes from the tectogrammatical layer, but this can be disputable. First of all, on the tectogrammatical layer, the auxiliary nodes are collapsed, and some other nodes (like dropped personal pronouns) appear. This does not fit into KonText system because this query tool is more about 'surface' representation of a sentence. However, we added the following attributes to the node that can bring some additional value while querying corpora: t_lemma (lexical value of a word), functor (more semanticalized value of a syntactic relation - afun), grammatemes (semanticalized variants of morphological features), tfa – topic-focus articulation attributes (concern informational structure of a sentence),

[15] That is done to avoid greediness of the regular expression which can match words from other clauses.

[16] http://ufal.mff.cuni.cz/treex

[17] http://hdl.handle.net/11234/1-1480

sempos (semantic part of speech), and some attributes concerning coreference and discourse. Next, we will give some examples of the queries with tectogrammatical attributes.
Let us study a relationship between tfa values and the functors. The topic of a sentence – what is being talked about – can be searched via the query [tfa="t"] . After forming a concordance, we calculated the frequency distribution of functor values for the found topics. The most frequent was ACT (Actor), following were PAT (Patient) and RSTR (functor for free modification). As for the 'focus' words (the new information in the sentence), the frequency distribution of their functor attribute was a bit different. The most frequent was RSTR, then PAT and the third was PRED (Predicate), whereas in the topic position the PRED functor was not that frequent - only on the thirteenth position.
As for the attributes belonging to mostly extra-sentential level, like coreference or discourse, the possibilities of KonText are rather limited. It is impossible to reference between the nodes that stand far away from each other, sometimes not even in the neighbouring sentences. We added only several attributes that might be of some use, but generally it is better to use more appropriate corpus search engines like PML-TQ. So far, we can determine the type of discourse relation (query [discourse_type="reason"]) and make the frequency distribution of the most frequent lemmas (*být* – 'to be', *protože* – 'because', *že* – 'that' and functors (CAUS, PRED, CONJ) for this query.
Introduction of attributes from analytical and tectogrammatical layers into KonText will not substitute all the functions of PML-TQ, but will enlarge the possibilities of a linear search.

5. Conclusion and Future work

We have presented a small modification that allows KonText to query syntactically annotated corpora and display syntactic trees of the sentences in the results of queries. The expressive power of CQL query language is limited in comparison with full treebank search engines, but the query language is simpler and unlike most treebank search engines, Kontext can query efficiently even very large parsed corpora.
Because of its relative simplicity and also for the convenience of the concordance form of displayed results, we see Kontext also as a meaningful tool for searching treebanks, in addition to the more traditional treebank search tools.
Presently, we plan to parse all the corpora in the LINDAT repository that have no syntactic annotation (if we have a parser available for their language) and make them available for search in Kontext in this enhanced form.

6. Acknowledgment

This work has been supported by the LINDAT/CLARIN project No. LM2015071 of the MEYS CR.

7. Bibliography

Machálek, T. and Křen, M. (2013). Query interface for diverse corpus types. In *Natural Language Processing, Corpus Linguistics, E-learning*, page pp. 166–173. Lüdenscheid: RAM Verlag.

Pajas, P., Štěpánek, J., and Sedlák, M. (2009). *PML Tree Query*.

Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. Brno. Disertační práce.

Spoustová, J. and Spousta, M. (2011). *CWC2011*. URL http://hdl.handle.net/11858/00-097C-0000-0006-B847-6.

8. Language Resource References

Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., and Zikánová, Š. (2013). *Prague Dependency Treebank 3.0*. URL http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bosco, C., Bowman, S., Celano, G. G. A., Connor, M., de Marneffe, M.-C., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Erjavec, T., Farkas, R., Foster, J., Galbraith, D., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Gonzales, B., Guillaume, B., Hajič, J., Haug, D., Ion, R., Irimia, E., Johannsen, A., Kanayama, H., Kanerva, J., Krek, S., Laippala, V., Lenci, A., Ljubešić, N., Lynn, T., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., Mori, S., Nurmi, H., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Petrov, S., Pitulainen, J., Plank, B., Popel, M., Prokopidis, P., Pyysalo, S., Ramasamy, L., Rosa, R., Saleh, S., Schuster, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Simov, K., Smith, A., Štěpánek, J., Suhr, A., Szántó, Z., Tanaka, T., Tsarfaty, R., Uematsu, S., Uria, L., Varga, V., Vincze, V., Žabokrtský, Z., Zeman, D., and Zhu, H. (2015). *Universal Dependencies 1.2*. URL http://hdl.handle.net/11234/1-1548.

Straka, M. and Straková, J. (2014). *MorphoDiTa: Morphological Dictionary and Tagger*. URL http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0.

Zeman, D., Mareček, D., Mašek, J., Popel, M., Ramasamy, L., Rosa, R., Štěpánek, J., and Žabokrtský, Z. (2015). *HamleDT 3.0*. URL http://hdl.handle.net/11234/1-1508.