Integrating TEITOK and KonText/PMLTQ at LINDAT

Maarten Janssen
Institute of Formal and Applied Linguistics
Charles University, Czech Republic
janssen@ufal.mff.cuni.cz

Abstract

In this paper we describe how the TEITOK corpus platform was integrated with the KonText and PML-TQ corpus platforms at LINDAT to provide document visualization for both existing and future resources at LINDAT. TEITOK is an online platform for searching, viewing, and editing corpora, where corpus files are stored as annotated TEI/XML files. The TEITOK integration also means LINDAT resources will become available in TEI/XML format, and searchable in CWB on top of existing tools at the institute. Although the integration described in this paper is specific for LINDAT, the method should be applicable to the integration of TEITOK or similar tools into an existing corpus architecture.

1 Introduction

LINDAT/CLARIAH-CZ is a Czech centre for data providing certified storage and natural language processing services. The LINDAT repository provides a direct implementation of the core objective of the CLARIN ERIC to advance research in humanities and social sciences by giving researchers unified single sign-on access to a platform which integrates language-based resources and advanced tools. But although the repository makes the raw data accessible, without an online interface for searching, that only makes the data usable for a limited group of researchers. That is why LINDAT aims to gradually make as many of the corpora in its repository as possible available via KonText (Josífko, 2014) and, in the case of treebanks, PML-TQ (Pajas et al., 2009), two corpus search interfaces.

However, many of the corpora in the LINDAT repository, as well as many (Czech) corpora that are not currently in the repository, are corpora with a solid footing in the digital humanities, such as historical corpora and learner corpora. And for such corpora, a good part of the users will be more interested in visualizing the individual documents in the corpus in a readable form, than they are in KWIC lists or search statistics. Since KonText and PML-TQ do not provide a graphical interface to view entire documents, but only a view on the direct context in terms of corpus tokens, the decision was made to integrate TEITOK (Janssen, 2016) as a way to make corpora available in a manner more fit to documents for the digital humanities. In TEITOK, corpus files are stored in the TEI/XML format, and adopting this well-established standard will further improve the interoperability of the LINDAT corpora. In this presentation, we will show how the integration between TEITOK an KonText, as well as PLM-TQ, was designed at LINDAT. The combined interface can be found online at: http://lindat.mff.cuni.cz/services/teitok/.

The two integrations take a different approach, and beyond explaining how the integration is done at LINDAT, we hope this paper illustrates how different corpus tools can be integrated within a single corpus architecture, and how TEITOK could be used by virtually any corpus tool for document visualization.

2 TEITOK, KonText and PML-TQ

This section will first give a short description of the three corpus tools involved in the integration, before turning to a description on how they were integrated. The description of the integration avoids too much technical detail, and attempts to use terms applicable to any corpus environment.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Maarten Janssen 2021. Integrating TEITOK and Kontext/PMLTQ at LINDAT. *Selected papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180: 180 104–110.

2.1 KonText

KonText is an advanced corpus query interface and corpus data integration platform built around the open source version of the corpus search engine Manatee (Rychlý, 2007), maintained by the Institute of the Czech National Corpus. It allows searching corpora in a (slightly modified version of) the Corpus Query Language (CQL), which was initially designed for the Corpus WorkBench (Evert and Hardie, 2011), and in addition several simplified search options are provided. It has the same overall design as most other online corpus interfaces: it has a landing page where you can select a corpus. And once a corpus has been selected, you can run search queries to obtain lists of results. The default visualization is as a KWIC list, while other views are available as well. And in the KWIC list, you can click on results to see the data about the document the result was found in, or the direct context for the result.

For a query result, you can then obtain various statistical analysis, including a list of collocates and various types of frequency lists. Queries can be stored and shared, and the search results can be downloaded in various file formats for off-line treatment. And there is support for both spoken corpora and parallel corpora in KonText.

2.2 PML-TQ

The PML-Tree Query (PML-TQ) is a powerful open-source search tool for all kinds of linguistically annotated treebanks. The tool works natively with treebanks encoded in the Prague Mark-up Language (PML) data format. The tool can be queried using a search API, and there is also an online search interface at LINDAT.

The online interface for PML-TQ has the same overall design as KonText: there is a landing page where you can choose a dataset, after which you can query that dataset using the PML-TQ Query Language. The result can be either a list of frequencies or a list of results, depending on the query. Result lists are not shown as KWIC lists like in KonText, but rather the first resulting sentence is shown, with the dependency tree for it below it, and a button to go to the next result. The dependency trees are drawn by a web service that produces SVG images for any sentence.

PML is a multi-layered format, which can contain multiple tree layers. For instance for the Prague Dependency Treebank (the latest version is PDT-C, (Hajič et al., 2020)) there are two different layers: the surface syntax layer (a-layer, containing dependency trees), and the deep syntax layer (t-layer). In the query, you always have to indicate the layer you want to search in, and the results show the trees from the relevant layer.

2.3 TEITOK

TEITOK is an online platform to view and edit corpus files stored in the TEI/XML format. The base TEI/XML files can be viewed in a range of different ways depending on the type of XML file, including facsimile-aligned text views, original or normalized text rendering, wave-form aligned views, etc. Dedicated visualization modules can be added when needed. For an explanation of the visualization features of TEITOK, see for instance Janssen (2016) or Janssen (2018).

Corpora in TEITOK are made searchable using the Corpus WorkBench (CWB). TEITOK uses a dedicated program to directly write CWB files from the TEI/XML files. During that CWB export, TEITOK also writes the filename of the XML file as an attribute, and keeps the byte-offset of the token in the original file in a separate file. With these byte-offsets, the CWB corpus becomes an index over the XML files, where the CWB results can be used to directly look up the corresponding XML fragments. The CWB index is regenerated frequently to make sure the byte-offsets reflect any possible changes in the XML files.

How a TEITOK corpus is exported to CWB is defined in the corpus settings - a central configuration file for the corpus in TEITOK that defines, apart from the CWB export, a myriad of other things like which token attributes should be editable, which items should appear in the menu bar, what the fixed fields in the teiHeader are, etc. The export settings are explicitly not kept in the teiHeader, since they belong to the corpus, and not to the XML files; and also because the same XML file is sometimes used in multiple corpus projects, each with its own settings. Since TEI files are edited in TEITOK, keeping

copies of the same file with a different header would lead to inconsistencies.

The TEITOK search results are rendered as an XML fragments: the CWB results are used to lookup the corresponding XML using the byte-offsets written during the export. Therefore, all the information in the XML file is present in the search result - including information that is not in the CWB corpus. This can include XML regions that were not exported to the indexed corpus (such as bold face or italics); elements that fall between tokens and as such cannot be exported to CWB (such as line breaks or notes); elements that fall below the level of the token and as such cannot be exported to CWB (such as morphemes); and elements that should not be exported as tokens since that would interrupt the token sequences if they were (such as deleted words). All such elements are often highly relevant to correctly interpret the context. Also, the XML fragment will represent contractions as contractions, and not split into grammatical tokens as they are in CWB. So a text containing *wanna* or the Spanish *del* (of+the), will render the original text faithfully, instead of having it changed into *want to* and *de el*. This is important for linguists who want to use corpora to find example to copy-paste into an exercise or article.

2.4 Combining TEITOK and KonText

As mentioned before, the decision was made at LINDAT to adopt TEITOK for document visualization. KonText and TEITOK are comparable frameworks in the sense that both are online platforms to search corpora using the Corpus Query Language (CQL). And intuitively, it would seem to be the ideal solution to fully integrate the two platforms in a single platform by either incorporating the relevant parts of TEITOK into KonText, or by rewriting TEITOK to use the search engine using in KonText, Manatee, as a backbone instead of CWB. But the problem with both these options is that they constitute a major overhaul of either TEITOK or KonText, since they share their query language, but not their general set-up or storage format. And neither of which is used exclusively for LINDAT, but used in a growing number of projects around the world. And a major rewrite would inevitably lead to compatibility problems at other projects using these tools. The only feasible option would be to drop KonText altogether in favour of TEITOK, but not only are most users of LINDAT more used to the KonText interface than they are to their TEITOK counterparts, but also there are various projects at LINDAT that rely on the KonText interface.

Therefore, a more modest integration was selected in which the two platforms are kept as independent interfaces, with links leading from one to the other. That set-up has the added advantage that it allows users that are more familiar with the CWB flavour of CQL than they are with the KonText implementation have the option to use that by doing their searches using TEITOK rather than KonText. Because although the base query language is the same, they start to diverge for more complex queries, such as queries involving global restrictions or multi-valued attributes, as well as small differences such as the names of structural attributes (`text_year` vs. `text.year`).

The way combined corpora are created is as follows: TEITOK is used to directly create a CWB corpus from TEI/XML files, using the standard TEITOK set-up. Once complete, the CWB corpus is exported to VRT (the one-word-per-line format used in both CWB and Manatee) using the CWB tools, after which it is loaded into Manatee. The registry file for both platforms is written by TEITOK using the aforementioned corpus settings. All this is done by a single script that can be run from within the TEITOK interface.

In order to link the two platforms, a small addition was made to both platforms: in KonText an option was added to allow the context of a token to be provided by an external REST service. And in TEITOK a module was added that can render the XML context of any corpus token as an HTML page. Combining these two modules makes it possible to click on a token in the KonText search result, and in the pop-up window that shows the context see the TEITOK rendering of the original XML with all its attributes. The fragment also comes with a link to the visualization of the full document in TEITOK. An example of a TEITOK context from a Czech text from Skript 2015[1] is given in Figure 1, where the word *cvičího* (practicing) is deleted in the original and hence not present in the CWB or Manatee corpus, as can be seen in the KWIC line. It is visible in the TEITOK context since it is in the original XML file.

[1] http://lindat.mff.cuni.cz/services/teitok/skript2015/

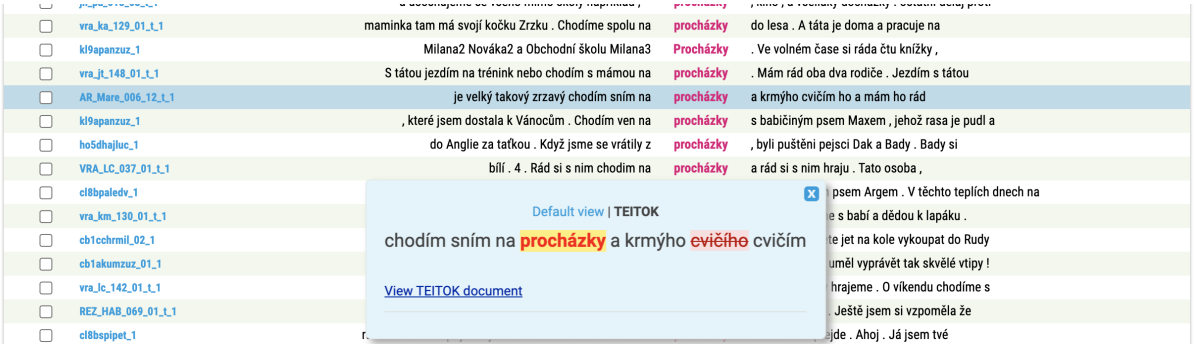

Figure 1: Example of a TEITOK context in KonText

Apart from the search integration, some additions were made or are being made to TEITOK to make it more compatible with an infrastructure like LINDAT. Two improvements that stand out in this respect are the following: (1) the inclusion of server-wide settings: in its original set-up, TEITOK corpora are completely independent, and each corpus has to define its own characteristics. But for an infrastructure with many corpora, it is necessary to be able to define a shared set of definitions and styles for all corpora. And (2) the option to generate static versions of TEITOK corpora: since TEITOK offers the option to edit corpora, TEITOK corpora are by default not fixed sets of data. But for a repository like LINDAT, fixed datasets are needed to be able to attribute them an object handle, and to allow users to quote reproducible data. To account for this, TEITOK now offers the option to create named corpus versions for inclusion in the repository.

2.5 Combining TEITOK and PML-TQ

In principle the same integration as was implemented from KonText could also be used for PML-TQ: have the web interface of PML-TQ display an externally retrieved context, which can then hence show the HTML output of TEITOK for the result sentence and a link to the textual context. The procedure for PML-TQ would have to be a bit more complex than that for KonText, since the PML-TQ corpus is not generated from the same TEI/XML files (and hence do not share their token IDs), but still possible since both the TEITOK and the PML-TQ corpus are generated from the same PML files. However, for the moment this has not been done, and we will see how people use the different tools whether this integration would be useful.

However, since the brunt of the work in PML-TQ is done by a web service, a deeper integration was made: in the TEITOK interface, you can type in a PML-TQ query. Upon submitting, that query is sent to the web service to retrieve a list of matching sentences (or a table of results). And rather than showing trees one by one, the system looks up the XML fragments for sentences in TEITOK that correspond to the first 100 results, and displays them as a KWIC list, with all the mouse-over token information TEITOK typically provides. From that KWIC list, it is possible to see the dependency tree generated by the PML-TQ API (the printserver), or to jump to the text display in TEITOK for that sentence (see Figure 2). For PDT corpora, there is furthermore a link that show the raw data from the PML files from which the corpus was generated, containing information from all the layers for the result sentence.

This approach of course needs a dedicated TEITOK module to deal with the search API of the target tool, in this case PML-TQ. But such a module is not very involved, and it makes for a full integration of the two tools, where the PML-TQ and CQL query languages work in very similar ways within TEITOK.

3 Adding resources

Having an integration between the three platforms is not sufficient for an infrastructure: it is also needed to get corpora into the hybrid system. When adding corpora to the integrated TEITOK/KonText infras-

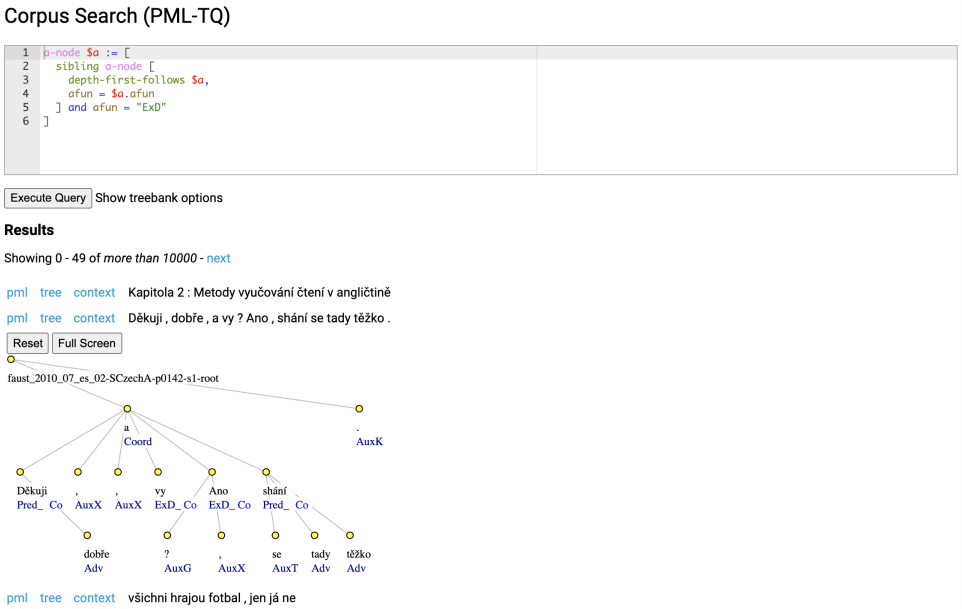

Figure 2: Example of a PML-TQ search in TEITOK

tructure, one has to distinguish between corpora that are already in KonText, existing corpora that are not yet in KonText, and newly planned corpora.

The corpora that are already in KonText can be converted automatically, although they have to be regenerated in order to create the byte-offset files used by TEITOK. The existing corpus is exported to VRT, from which it is converted to a collection of TEI/XML files, one for each document. Some corpus specific action is needed, for instance, it is necessary to indicate the correct XPath according to TEI for each of the metadata in the corpus in order to end up with correct TEI/XML documents; but once these settings are provided, the process is fully automatic. The result is not a full-fledged TEITOK corpus, since information that was not in the KonText corpus will not be available in the TEITOK corpus either. This typically includes typesetting, as well as word spacing and contraction decomposition. But it is a direct and automatic way to add corpora.

For the class of existing corpora that are not yet in KonText, TEI/XML files are generated from the raw data (in whichever format the corpus came in), keeping as much of the original information as possible. In order to facilitate this, we are working on a set of conversion tools from popular corpus formats including ELAN, FoLiA, PagesXML, etc. Even for some of the corpora that were already in KonText, the corpus is being recreated from source, since some of the information was lost in the conversion. An example of this is the Prague Dependency Treebank, where the original PML files contain more morphological information than the KonText corpus that was previously created from it. The richer TEI/XML structure of TEITOK makes it possible to incorporate all this information in the hybrid framework.

And finally, there are those corpora that are still under development or planned for the future. For such corpora, the various options provided by TEI/XML and TEITOK make it possible to encode richer information in the corpus than a traditional corpus based on a one-word-per-line architecture would allow. An example of this is the upcoming ParCzech corpus (Hladká et al., 2020) which was designed from the start as a spoken corpus in TEITOK, and it searchable as a KonText corpus using the hybrid TEITOK/KonText infrastructure. In some cases, existing corpora are being re-planned to make use of the new options. An example of that is CzeSL (Štindlová et al., 2012), a learner corpus where not all the information present in the source material was encoded, but which is now in part being redone to include the deletions, corrections, normalizations, etc. that were previously impossible to include.

TEITOK provides a wide range of visualization options. Therefore, for all corpora, whether they are

automatically converted from KonText, rebuilt from the source, or completely planned in TEITOK, it is worth while to see whether the options provided by TEITOK cannot make more use of the corpus data. For instance, the ACL RD-TEC corpus (QasemiZadeh and Schumann, 2016) is a corpus providing manually annotated terms. This corpus is much better served by adding an interface that presents the collection of terms from the corpus as a terminology.

3.1 Unresolved issues

The integration between KonText and TEITOK works seamlessly for basic corpus features. But on features that are added on to the base corpus architecture, different tools take a different approach. A case in point is the treatment of dependency trees: as the value for a head attribute, KonText expects a relative ordinal number: -2 to indicate that the head occurs 2 positions to the left. While TEITOK expects the ID of the head token. In this particular case, the decision was made to extend KonText to accept different types of head attributes, so that KonText can draw dependency trees from the data provided by TEITOK.

But there are other advanced features where we are still working on the best way to handle them, most crucially speech alignment and parallel corpora. These issues of course only arise because we want the same corpus to work in both systems. If TEITOK were only used for text visualization, these would not be an issue.

For speech corpora, the problem is similar to that for dependency trees: KonText expects the aligned segments in the corpus (typically the utterance) to have an attribute that names the sound-file containing the corresponding speech segment. While TEITOK uses entire sound files marked out higher up in the document (typically in the metadata), and the aligned segment is expected to have attributes marking the start and end time of the corresponding speech segment in that sound file.

For parallel corpora, KonText uses different corpora for different languages, with a table linking two corpora indicating which are the corresponding elements. While TEITOK uses several strategies for parallel corpora, typically with all languages in a single corpus, and using shared attributes to mark out corresponding elements.

For such advanced features, it is necessary to adapt at least one of the two integrated tools to accept the solution provided by the other, or alternatively to have the corpus contain both solutions. For both speech corpora and parallel corpora, we are still exploring all available options.

4 Conclusion

In this paper, we have shown how TEITOK was integrated in the existing workflow based on KonText at LINDAT. The additional options of the combined TEITOK/KonText workflow make it possible to provide a richer document view and context view that are more in line with the requirement of the digital humanities. The TEITOK platform has proven its appeal to the DH community by a growing number of projects of diverse nature, including historical corpora like PostScriptum (CLUL, 2014), spoken corpora like NURC (Oliviera Jr, 2016), and learner corpora like COPLE2 (Mendes et al., 2016). The TEI/XML based design of TEITOK allows those corpora to maintain their domain specific characteristics while at the same time adhering to a standard NLP pipeline.

It is our hope that the combined workflow will attract more corpora into the LINDAT infrastructure that would otherwise not have been made available as a CLARIN resource, with the preliminary indications looking promising in newly established collaborations towards this end. The combined TEITOK/KonText workflow provides a lot of potential for the future, and the fact that TEITOK is now maintained at LINDAT makes it much easier to expand the framework to account for newly arising demands, such as making the corpora accessible via the Federated Content Search (FCS), something that is relatively easy given that there are existing solutions for CWB corpora.

The conversion of the existing corpora in KonText to the TEITOK/KonText hybrid has not been completed yet, not because a straight-forward port would be difficult, but because in many cases it seems that by rebuilding the corpus and thinking of additional interface options, more value could be extracted from the source data. And the spoken corpora as well as the parallel corpora have not yet been ported since the discussion on how to best deal with those in the TEITOK/KonText hybrid has not been finalized, but

is expected to come to a conclusion soon.

It is still too early to say much about the user appreciation of the hybrid solution, but the internal reception for all the corpora added to the hybrid solution has been very positive.

The integration described of TEITOK with an established corpus workflow is not specific to KonText: the changes in KonText are minimal, and the XML context provided by TEITOK can be called from any corpus tool - the only requirement is to have shared identifiers between the different versions of the corpus. And this easy to establish integration could add three important things to any corpus environment: a rich XML view on the context from the environment, a link to a full document visualization, and the option to allow corpus editors to easily maintain and edit their own corpora.

References

CLUL. 2014. P.S. Post Scriptum. arquivo digital de escrita quotidiana em portugal e espanha na Época moderna.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Corpus Linguistics 2011*.

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Romportl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2020. Prague dependency treebank - consolidated 1.0 (PDT-c 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Barbora Hladká, Matyáš Kopp, and Pavel Straňák. 2020. ParCzech PS7 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Maarten Janssen. 2018. Adding words to manuscripts: From PagesXML to TEITOK. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, *Digital Libraries for Open Knowledge*, pages 152–157, Cham. Springer International Publishing.

Michal Josífko. 2014. KonText web demo. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Miguel Oliviera Jr. 2016. NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc). *CHIMERA: Romance Corpora and Linguistic Studies*, 3(2):149–174.

Petr Pajas, Jan Štěpánek, and Michal Sedlák. 2009. PML tree query. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pavel Rychlý. 2007. Manatee/Bonito - a modular corpus manager. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, pages 65–70.

Barbora Štindlová, Svatava Škodová, Jirka Hana, and Alexandr Rosen. 2012. CzeSL – an error tagged corpus of Czech as a second language. pages 21–32, 01.