DAMIR ĆAVAR, DUNJA BROZOVIĆ RONČEVIĆ
University of Zadar, Institute of Croatian Language and Linguistics

RIZNICA: THE CROATIAN LANGUAGE CORPUS

Introduction

The Croatian Language Corpus (CLC) is currently a spin-off product of sub-projects of the research program *Riznica* (Croatian Language Repository). Initially the Ministry of Science, Education and Sports (MZOŠ) of the Republic of Croatia granted the funds for the creation of the necessary infrastructure for the CLC. The goal of this particular grant was to organize the initial infrastructure for online language data and resources related to the Croatian standard language and dialects.

As an extension of some of the basic ideas from the initial project phase, in 2007 the research program *The Croatian Language Repository* (CLR) was granted by the MZOŠ (cf. Brozović Rončević and Ćavar, 2008). Being a research program (PI, Dunja Brozović Rončević) with numerous subsumed independent research projects that make use of the CLC, the corpus is mainly developed as a by-product of those research projects within the CLR, and in particular of the research activities conducted in the project *Semantic Nets and Computational Lexicology* (PI, Damir Ćavar). Related to the Semantic Nets research project, various technologies, automatic annotation tools and language processing components have been developed in cooperation with the Linguistics Department at the University of Zadar and funded by other independent research projects.

Previous work

Given a large number of corpus (development) projects in the world, mentioning all or just most of the dominant ones would exceed the scope of the current article. When focusing on Croatian corpus projects only, there are a couple of mentionable projects. Of particular interest, when comparing goals and methods with the CLC, are the "One million token corpus" and the Croatian National Corpus (see cf. Tadić, 2002).

The "One million token corpus" of Croatian literary language was compiled by Milan Moguš at the Institute of Linguistics at the University of Zagreb.[1] It covers texts from 1937 to 1978, and contains five sub-corpora with 200.000 tokens each that cover the genres: prose, poetry, drama, secondary school textbooks, and newspapers. Derived analysis results include and alphabetic dictionary and a frequency dictionary of types and lemmas, as well as concordance or keyword in context overviews based on tokens and lemmas.

The Croatian National Corpus (HNK) is a corpus of mostly contemporary Croatian text with approx. 101 mil. tokens. It is made available via additional software tools, see http://www.hnk.ffzg.hr/. As with the "One million token corpus", the HNK provides word-class information and lemmatized tokens for search and analysis. The HNK, on the other hand, contains mainly newspapers and journals, as well as online available texts from the late 20th century, thus linguistically it doesn't contain relevant texts that representative for changes and language development overviews.

The CLC provides on the one hand similar information as the two mentioned corpora, i.e. word-class tags and lemmas. On the other hand, the CLC aims at extending the annotation by other linguistic levels, e.g. including text sources that go beyond the sources found in the other corpora. Additionally, The CLC contains texts from the 19th century till now, as well as older texts that have been adapted to modern Croatian. Further, one of the technical goals of the CLC was to use exclusively standardized text and document encoding schemata (XML and TEI) and a uniformed common text-character encoding standard (e.g. UTF8). In addition

[1] See for more details the URL http://www.ffzg.hr/zzl/zzl-home.htm.

to the goal to maintain common format compatibilities, the linguistic annotation is designed to maximize interoperability and compatibility with other tools and corpora, and in particular with linguistic annotation standards. The morpho-syntactic features and annotation symbols are taken from the General Ontology for Linguistic Description (GOLD) (cf. Farrar and Langendoen, 2003), wherever possible. We have worked on mappings of the GOLD annotation schema to other annotation labels, e.g. MULTEXT(-EAST) (cf. Véronis and Khouri, 1995; Erjavec, 2010) in order to maximize the compatibility and interoperability with other corpora and corpus processing tools.

The Development of the CLC

The initial goals of the CLC were to serve the lexicographical projects at the IHJJ,[2] as well as other linguistic and language technology projects. For the purposes of the lexicographical project, it was desirable to include texts from all kinds of genres into the CLC, however, there was never an intention to balance it in any way. The notion of a balanced corpus seemed rather irrelevant for the respective research goals. Instead of one fixed balanced corpus, our goal was to create an annotated text corpus, as large as possible, that could be dynamically mapped on individual sub-corpora for specific research and development interests.

Currently the corpus contains more then 100 mil. tokens. The document collection that enters the CLC is extended on a daily basis. Depending on the complexity of the original text sources, the growth of the corpus could be approx. 1.5 million tokens a day. Some resources, in particular online available journals that are added to corpus, can be converted automatically to the required file format for manual checkup.

The current content of the literature sub-corpus of the CLC is shown in table 1:

	Fiction	**Specialized**
Proportion	28%	72%

Table 1: Content type proportions in the CLC

The number of books in the fiction portion of the corpus can be divided into sub-domains as shown in table 2:

Type		**Number**
Poetry		35
	Lyric	16
	Epic	19
Prose		184
	Novel	101
	Short Stories	83
Drama		13
Total		**232**

Table 2: Number of books in the fiction portion of the CLC

The number of books in the specialized text proportion of the literature sub-section of the CLC is shown in table 3:

Type	**Number**

[2] Some of the projects are the development of a Standard Croatian One-Volume Dictionary, a Croatian Verb Valency Dictionary, a Croatian Terminology Dictionary and Online repository (STRUNA), a Dictionary of Croatian place names, a Dictionary of Anthroponyms, and various bilingual dictionaries.

Non-fiction Prose	60
Studies, Polemics	36
Essays, Memoirs, Letters	24
Scientific Literature and Popular Science	35
Textbooks	27
Manuals, Instructions	8
Total	95

Table 3: Number of books in the specialized text portion of the CLC

From scientific literature and popular science so far we included textbooks from the domains of natural and technical sciences 63%, which are textbooks from medicine, biology, geography, informatics, chemistry and physics. Many of these texts are relevant for the STRUNA project, which is developing terminological dictionaries and online interfaces for them. From social sciences and humanities 37%. The majority of manuals thus far cover the law domain.

Besides this literature component, the CLC also contains a large number of articles from various newspapers and journals, among others Glas Koncila, Vjesnik, and Zarez. The amount of tokens in the CLC from these resources represents the largest proportion.

The CLC is a text corpus, annotated in XML (cf. Bray et al., 2008). It is using the Text Encoding Initiative (TEI) P5 standard (cf. Sperberg-McQueen 1994) for XML, as described in cf. Vanhoutte (2004) or Burnard and Bauman (2007).

The CLC is created using various text sources, among others:

- Online newspapers, books, articles
- Printed and published books and other printed hard copies
- Digital files of printed books made available by publishers[3]
- Transcriptions of collected data and recordings

Various resources with online available documents, newspapers, article and books are integrated in the corpus using digital archives provided to us by the publishers, or directly via mirroring of the web resources. These documents are mainly based on HTML and other more or less systematically structured file formats, that can be converted into the digital target format via manual conversion methods or automatic tools.

As far as books are concerned, some of them are made available to us in digital format. However, many of the books in the CLC, in particular the older ones, are created from hard-copy originals. Those books are scanned to TIFF images, and the resulting images are converted to text via OCR software.

Since some OCR tools identify font formats and other textual properties, an intermediate format that is also used in the corpus creation process is the Rich Text Format (RTF), which in this case does not contain annotations or meta-information.

The different digital file sources for the CLC come in various formats, e.g. InDesign, Pagemaker, Quark, Word, and PDF. These digital formats are either converted to Unicode text or directly to the basic TEI P5 XML format. For all standard document types that can be opened with OpenOffice, we use the TEI P5 XML export plugin (see XYZ) to export the files directly to the target XML format, minimizing additional manual annotation effort. For many other formats, the export had to be to an initial UTF8 encoded raw text format, which is manually annotated using TEI P5 XML standards and tags. In the worst case, with image

[3] The Institute of Croatian Language and Linguistics has contracts with some of the prominent Croatian publishers, among others Školska knjiga, and the Croatian Academy of Arts and Sciences. Those publishers provide the IHJJ with digital copies of their books for the purposes of the CLC. At the same time, the IHJJ is a publisher as well, and includes its own books in the corpus.

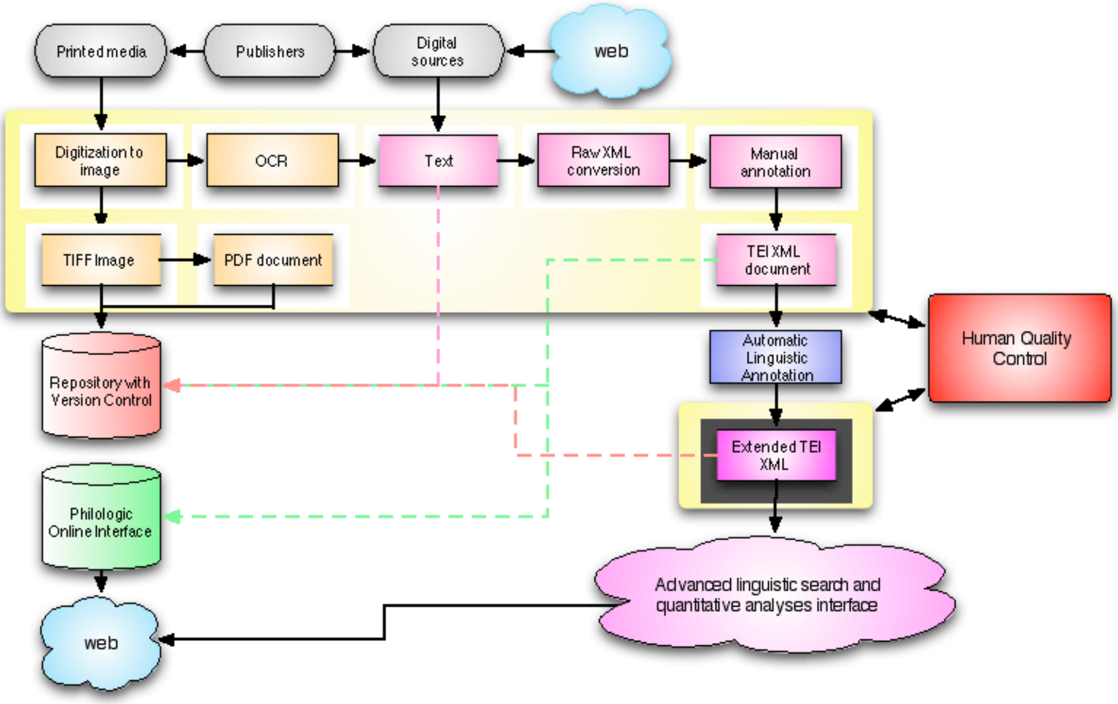
based file formats, e.g. image-based PDF files, the file content has to be converted to raw text via Optical Character Recognition (OCR) tools.

In an initial text-to-XML conversion steps the meta-information is added, using the TEI (Text Encoding Initative) P5 XML standard. The text chapters, sections and paragraphs are annotated automatically or manually, depending on the underlying conversion pipeline. Subsequently, formats of fonts, markup of section titles, page-breaks and page-numbers, footnotes and other document properties are performed.

All resulting XML files, annotations, and markups are manually checked, corrected, and all changes and corrections are kept track of. The resulting XML files are stored together with the underlying digital format and intermediate conversion formats in a file and document repository. The file management allows for the annotation of files being ready for inclusion in the final corpus, and indexing for the interactive web interface.

We make use of various tools for corpus analysis and processing. As the web-based front-end for the CLC we decided to use Philologic with various manipulations and adaptations. PhiloLogic™ was developed at the University of Chicago. It “is the primary full-text search, retrieval and analysis tool developed by the ARTFL Project and the Digital Library Development Center (DLDC)”[4]. In its version from 2006 it was able to generate an index and an interface for a TEI XML based corpus. Currently it is being extended and further developed as an open source project, and is available in its version 3.2 at the URL http://sites.google.com/site/philologic3/. The installation requirements for Philologic were minimal, including a running Apache web server, and basic interpreters like Perl and compilers like GCC. Besides the localization to Croatian of the frontend web interface, we have added for our purposes necessary additional functionalities to the 3.0 version of Philologic in our online interface.

The following diagram summarizes the corpus creation process as it is defined for the CLC:



Currently the search functionality includes token search with limited regular expression capabilities and a resulting visualization of text passages. It also allows for a

[4] See the URL http://philologic.uchicago.edu/.

keyword-in-context view on the hits, as well as left and right contexts with statistics and variable window sizes, and various further sub-specifications and limitations of the query.

One of the extensions includes the virtualization of the corpus-base and the resulting index, and front-end handling of multiple corpora on the same installation. That is, in out version of Philologic it is possible to generate indexes and a web-frontend for an individually specified corpus. A sample set of four individually specified corpora can be accessed at the CLC URL http://riznica.ihjj.hr/.

In its current version, the corpus annotation is aimed at including not just lemmatization and part-of-speech annotation, but also multi-level linguistic annotation that includes phonemic transcriptions, syllable structures, morpheme segmentations of tokens and specific morpheme annotations, as well as syntactic phrasing and parse tree markup. While most of these annotations can easily be integrated in one XML TEI file in a classical way, for the syntactic annotation a stand-off markup (cf. Bański and Przepiórkowski, 2009) of phrase structures and trees is necessary for the final version, although currently various problems exist (cf. Bański, 2010), in particular since to our knowledge indexing and retrieval across links to target URIs is not supported. Thus, for indexing, we need to have inline markup within one XML file. This limitation might be irrelevant with future indexing engines.

The stand-off annotation makes particular sense, if different theoretical frameworks for syntactic tree annotations are taken into account, as is the case for the CLC. A concrete tree annotation, however, does not yet exist for the CLC, although the respective framework and approach is defined, annotation formats have been defined, and basic parsers and shallow grammars have been developed.

Automatic Linguistic Annotation

Different research foci require different annotation levels. A particular goal for the CLC was to provide a multi-level annotation, which includes phonological, morphological, lexical, and syntactic information. A fundamental annotation of the CLC is supposed to include at least:

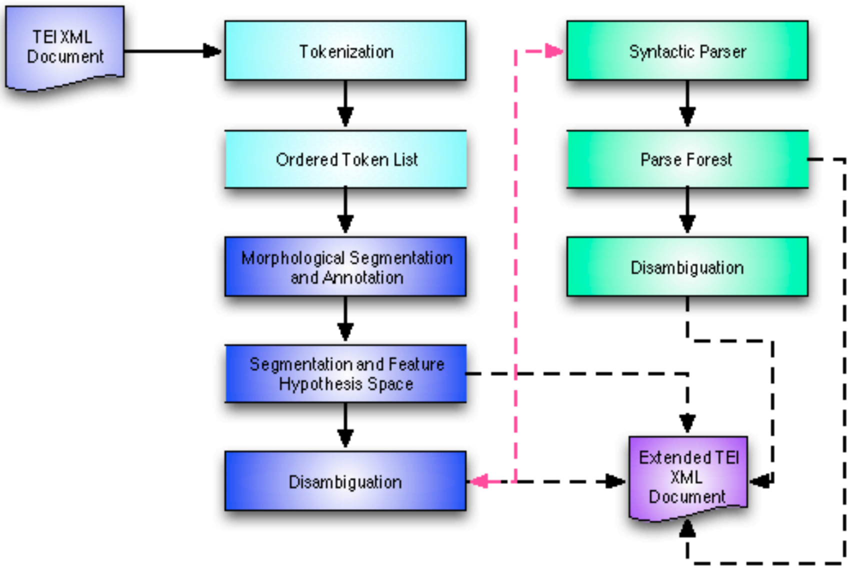
- Lemmatization
- Part of speech information
- Named entity information
- Shallow syntactic structures and phrases

In addition to these common corpus annotation levels, the extended annotation is supposed to include phonological and morphological levels as well, including at least the following:

- Phonemic transcription
- Syllable structure
- Morphological segments or morphemes and their features

In general, qualitative and quantitative data of morphotactics, phonotactics, and related distributional linguistic information from corpora is sparse for any language, and for Croatian in particular. For various reasons, such distributional models are highly desirable for the CLC related research projects. Thus, the activity focuses on the corpus annotation along these levels.

A manual annotation on various linguistic levels would be to costly and most likely very error prone. An automatic annotation might be error prone as well, however, more likely the errors would be systematic, and could be eliminated with additional rules, constraints and data in the grammars or language models of the particular NLP tools used in the markup process. Thus, one of the basic goals and concepts in the CLC development process was to also develop and use language-processing components to automatically annotate the corpus, and avoid direct manual annotation of the corpus at the linguistic level completely. The following graphics shows the concept of the linguistic processing pipeline in the CLC annotation process:


TEI XML Document
Tokenization
Ordered Token List
Morphological Segmentation and Annotation
Segmentation and Feature Hypothesis Space
Disambiguation
Syntactic Parser
Parse Forest
Disambiguation
Extended TEI XML Document

Numerous language and text processing tools have been made available for the specific XML annotation used in the CLC.

We have developed an extended tokenizer for Croatian text, using a finite state transducer approach. The tokenizer annotates not only tokens and orthographic symbols, but also basic tokens types and named entities, as well as sentence boundaries. The tokenizer consumes raw text and its output is wrapped token annotations and sentence markup using TEI P5 XML. Clause type markup is possible to some limited extend, as shown in the following example.

```
<s>
        <cl type="matrix">
                <w>Petar</w>
                <w>je</w>
                <w>vozio</w>
                <w>mračnom</w>
                <w>ulicom</w>
                <pc>,</pc>
                <cl type="subordinate">
                        <w>ali</w>
                        <w>nije</w>
                        <w>bio</w>
                        <w>upalio</w>
                        <w>svjetla</w>
                        <pc>.</pc>
                </cl>
        </cl>
</s>
```

A morphological segmentation application for the annotation at the sub-word level has been developed. It is based on a morpheme dictionary that was generated from data bases developed for lexicographic purposes at the IHJJ, and a set of morphological rules that are compiled into a binary finite state transducer (see Ćavar et al. 2009) for morphological segmentation of complex words and feature annotation of single morphemes. The same transducer is also used for rule-based lemmatization and morpho-syntactic feature annotation (without disambiguation). Example input and output structures can be seen in the following sample:

Input: *pročitamo*
Output:

```
<w type="verb" lemma="pročitamo">
    <w type="verb">
        <m type="prefix" baseForm="pro">pro</m>
        <m type="root">čita</m>
    </w>
    <m type="suffix">mo</m>
</w>
```

An automatic transcription tools for the conversion of Croatian text to the International Phonetic Alphabet (IPA) is made available online (see Ćavar and Ćavar, 2010). An extension of this component is used to wrap the corpus text in IPA transcription. A syllabification algorithm is being tested for Croatian as well, that consumes the phonemic transcription and morphological segmentation of each word.

While for the XML coding a more elegant technological solution would be to link specific tokens via URIs to external dictionary entries where phonemic transcriptions and other information is stored, as for example syllable or morpho-syntactic feature structure, due to limitations of current indexing engines, we have chosen the inline approach to annotate at least phonemic transcriptions and syllable structures, as shown in the following sample:[5]

```
<seg type="w">
  <w type="verb" lemma="pozvati">
    <w type="verb">
      <m type="prefix" baseForm="po">po</m>
      <m type="root">zove</m>
    </w>
    <m type="suffix">mo</m>
  </w>
  <fs type="phonology">
    <f name="pron">pɔzɔvɛmɔ</f>
    <f name="syll">po-zo-ve-mo</f>
  </fs>
</seg>
```

The above mentioned tools have been developed as standalone tools, and partially they are being integrated in the SNLK[6] (cf. Ćavar et al., 2011), together also with a shallow parsing component for elementary phrasing, that makes simple noun phrase annotation possible, but also more complex phrase structure markup. On the one hand, we make use of the PARC XLE parser (cf. Kaplan and Maxwell, 1993; Crouch et al., 2008) using Lexical Functional Grammar (LFG) (cf. Bresnan, 2001), aiming at the development of LFG grammar segments for standard Croatian, and eventually a complete grammar for it. On the other hand, currently we also implemented our own algorithms that are based on variants of agenda based chart parsing (cf. Earley, 1970) for context free grammars (CFGs), using feature augmentation extensions on symbols and unification over CFG rules (cf. Shieber, 1986).

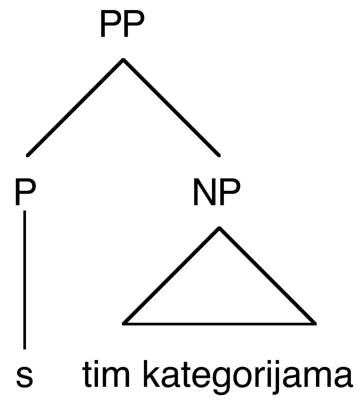
The primary goal of the syntactic annotation in the CLC is to be able to encode phrase structure trees, possibly with functional annotations (e.g. in the sense of LFG), for the purpose of visualization and computational processing (e.g. extraction of probabilistic grammars and other types of language models). The TEI P5 specification provides various possibilities to encode tree structures. The `tree` and the `eTree` tags, as described in chapter 19 of the TEI

[5] Thanks to Piotr Bański (p.c.) for annotation suggestions wrt. inline solutions for our indexing problem.

[6] See www.snltk.org for more details.

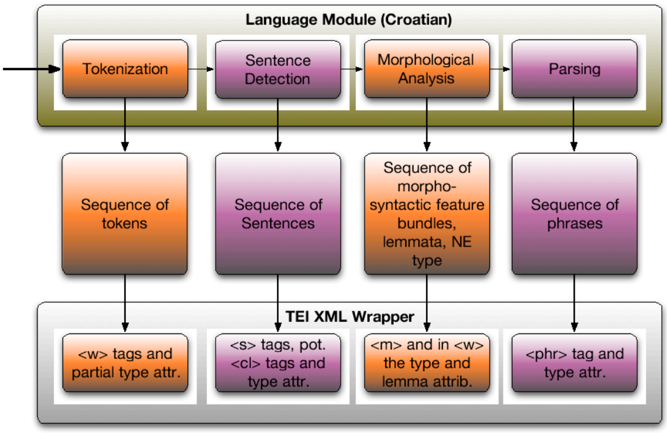
P5 guidelines, are ideal for a graph oriented ancoding and subsequent visualization of the corresponding phrase structure trees. Here is an example of such a tree encoding for visualization purposes:

```
<eTree n="2451">
  <label>PP</label>
  <eTree>
    <label>P</label>
    <eLeaf>
      <label>s</label>
    </eLeaf>
  </eTree>
  <triangle>
    <label>NP</label>
    <eLeaf>
      <label>tim kategorijama</label>
    </eLeaf>
  </triangle>
</eTree>
```



Alternatively, the tree encoding can be done using a strategy as suggested for the National Corpus of Polish (cf. Przepiórkowski, 2009). Since our goal is limited to the task of wrapping a parser output in the appropriate XML markup, allowing for visualization and search over tree structures primarily, we initially restrict ourselves to the graph oriented eTree encoding model of the TEI P5 standard, but remain open towards other annotation formats that allow the mapping of our parser output on TEI P5 compatible XML code.

The following graphics shows the different NLP components and the corresponding XML wrapping for the CLC:


Language Module (Croatian)
Tokenization
Sentence Detection
Morphological Analysis
Parsing
Sequence of tokens
Sequence of Sentences
Sequence of morpho-syntactic feature bundles, lemmata, NE type
Sequence of phrases
TEI XML Wrapper
<w> tags and partial type attr.
<s> tags, pot. <cl> tags and type attr.
<m> and in <w> the type and lemma attrib.
<phr> tag and type attr.

In general, the approach we have chosen in the CLC linguistic annotation is to generate the complete markup via NLP annotation tools, rather than human or expert annotation. Thus, corrections and changes of annotation mistakes have to be corrected in the corresponding NLP tools, and not manually in the corpus. This approach has the positive side-effect of improving a mainly rule-based NLP tool-chain, and generating a qualitatively annotated corpus for the improvement of quantitative NLP models.

Conclusion

The CLC is a corpus is not a final result or product, it is rather a project, being developed and extended continuously. Not only does the amount of text in it grow permanently, also the annotation depth and quality is extended over time, based on the needs and goals of related research projects.

Nevertheless, the CLC is made available online in all its phases and stages of size and annotation for the purposes of a broader linguistic community. We are continuously working on ways and new technologies to improve the open access to the corpus via online interfaces and other technologies.

Compared to other corpus projects, a specific challenge of the CLC is to include texts from a period covering at least 150 years, and thus involving a lot of manual digitization effort. The same standards, tools and methodology that were developed for the CLC are also implemented and used for the development of language corpora of earlier phases of the Croatian language, which are being currently prepared at the IHJJ within the Riznica research program. These corpora will be made available on-line in the near future.

Additionally, the CLC is a corpus that aims at a qualitative multi-level linguistic annotation that allows for analyses of linguistic properties, correlations and changes across linguistic domains, i.e. phonology, morphology and syntax. The necessary annotation technologies and language processing tools had to be created and adapted for this goal, and tools for the analysis and use of the encoded information had to be adapted, and even drafted and developed from scratch.

The CLC is not a corpus that is designed as a standalone general corpus for all kinds of potential purposes that are unknown to its creators. It is rather the result of concrete research goals, embedded in concrete ongoing research projects. Its properties, annotation formats, and encoding technologies are not oriented towards general compatibility or even international standards, whether those exist, or not. They are however kept in a format that allows for conversion and adaptation towards other common formats and linguistic annotation standards, being closely linked for example to the GOLD linguistic annotation approach, or using the TEI P5 compatible encoding.

References

Bański P. 2010: Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless. In Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, vol. 5.

Bański P., Przepiórkowski A. 2009: Stand-off TEI annotation: the case of the National Corpus of Polish. In Ide N., Meyers A. (eds.) Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009, Singapore, pp. 64-67.

Bray T., Paoli J., Maler E., Yergeau F., Sperberg-McQueen C.M. 2008: Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation, URL http://www.w3.org/TR/2008/REC-xml-20081126/

Bresnen J. 2001: Lexical-Functional Syntax. Oxford: Blackwell Publishers.

Brozović Rončević D., Ćavar D. 2008: Hrvatska jezična riznica kao podloga jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika. In: Vidjeti Ohrid, Zbornik radova XIV. međunarodnog slavističkog kongresa u Ohridu, Hrvatsko filološko društvo - Hrvatska sveučilišna naklada, Zagreb, pp. 173-186.

Burnard L., Bauman S. 2007: TEI P5: Guidelines for Electronic Text Encoding and Interchange (webpage). URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Ćavar D. 2008: The Croatian Language Repository: Quantitative and Qualitative Resources for Linguistic Research and Language Technologies. Paper presented at Eastern Michigan University, Institute for Language Information and Technology (ILIT), 31st of January, 2008.

Ćavar D., Gulan T., Kero, D., Pehar F., Valerjev P. 2011: The Scheme Natural Language Toolkit (SNLTK): NLP libraries for R6RS and Racket. In Proceedings of the 4th European Lisp Symposium, Hamburg University of Technology, 31st of March, 2011.

Ćavar D., Ćavar M. 2010: Phonemic IPA transcription and syllabification for Croatian. Web page: http://ling.unizd.hr/~schemers/phonemic/.

Ćavar D., Jazbec I.-P., Stojanov T. 2009: CroMo - Morphological Analysis for Standard Croatian and its Synchronic and Diachronic Dialects and Variants. In: J. Piskorski, B.W. Watson, A.

Yli-Jyrä (eds.) Finite-State Methods and Natural Language Processing. Frontiers in Artificial Intelligence and Applications 19, IOS Press, pp. 183-190.

Crouch D., Dalrymple M., Kaplan R., King T., Maxwell J., Newman P. 2008: XLE Documentation. Palo Alto Research Center (PARC), http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Earley J. 1970: An efficient context-free parsing algorithm, Communications of the Association for Computing Machinery 13 (2), pp. 94-102.

Erjavec T. 2010: MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proc. of the LREC 2010, Malta, 19-21 May, 2010.

Farrar S., Langendoen D.T. 2003: A linguistic ontology for the Semantic Web. GLOT International. 7 (3), pp. 97-100.

Kaplan R.M., Maxwell J.T. 1993: LFG grammar writer's workbench. Technical report, Xerox PARC.

Przepiórkowski A. 2009: TEI P5 as an XML Standard for Treebank Encoding. In: Passarotti M., Przepiórkowski A., Raynaud S., Van Eynde F. (eds.) Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8), pp. 149-160.

Shieber S. 1986: An Introduction to Unification-Based Approaches to Grammar. Stanford, CSLI.

Sperberg-McQueen C.M. 1994: The Text-Encoding Initiative, in: Current Issues in Computational Linguistics: Essays in Honour of Don Walker, (eds. A. Zampolli, N. Calzolari, M. Palmer), Giardini Editori e Stampatori and Kluwer Academic Publishers, Pisa and Dordrecht, pp. 409-427.

Tadić M. 2002: Building the Croatian National Corpus, In Proceedings of the LREC 2002 in Las Palmas, Vol. II, pp. 441-446.

Vanhoutte, E. 2004: An Introduction to the TEI and the TEI Consortium. Literary and Linguistic Computing 19, No. 1, pp. 9-16.

Véronis J., Khouri L. 1995: Etiquetage grammatical multilingue: modèle. Centre National de la Recherche Scientifique, URL: http://aune.lpl.univ-aix.fr/projects/multext/LEX/LEX2.html.