ACT - Computer Processing of Written Cultural Heritage Sources

Kiril Ribarov, Jiří Bubník, Jiří Čelák, Vojtěch Janota,

Alexandr Kára, Václav Novák, Tomáš Vondra

Charles University - Institute of Formal and Applied Linguistics, Czech Republic

ribarov@ufal.mff.cuni.cz

INFORUM 2004: 10th Conference on Professional Information Resources

Prague, May 25-27, 2004

Abstract

The aim of this work is to introduce the integration aspects of the Annotated Corpora of Text (ACT) package: software tools for lexical and corpus processing of written cultural sources.

ACT is suitable for manipulation and capturing of rich language variability on word and sentential level. It is not the word-form, but its understandings that become central processing units, which can be assigned morphology distinctions, headwords (including recensional), translation equivalents, complexes (multi-word units), and correlations to other sources. The whole annotation process is automated and private sorting orders and morphology tags structures can be defined. ACT incorporates modules for: complex searches on one or more sources; creation of various ready-to-use documents (in various output formats) as index verborum, retrograde index, index of concordances, frequency lists and others, from one or more sources; on-line web-based queries for text and image access; incorporation of lexical card-files into a corpus.

Introduction

Processing of old documents in the complexity of the whole societal and cultural environment, therefore processing of written cultural heritage sources, is a non-easy task. There is current experience only on separate processing tasks, some of which are cataloguing, digitization, text processing and historical studies. First attempts to incorporate computers for these (sub-) tasks is characteristic for the past decade; as it is usually the case, many problems have been solved but also many new ones, unknown before, arose. Let us take these new problems as challenges, which highlight new possibilities. These challenges have qualitatively a completely different nature than the tasks we had to solve in the past years since we aim towards a complex environment with self-integrating nature. We believe that the character of self-integration is inherent to a well-designed and fully computerized problem network. We aim towards a framework for the above stated subtasks, which will result in an open and functional cultural heritage network.

To fulfill such aim, one needs to concentrate on the subtasks and present them within a framework of an open character. In this paper, relevant to the subtask of textual processing of Old-Church Slavonic (OCS) documents, we will present a newly released tool language

independent[1] software tool for lexical and corpus processing of written cultural sources: the Annotation Corpora of Text[2] (ACT) package.

ACT places the written cultural sources in an electronic contextual (e-context) field with two major connecting elements:

a) source image along with language based contextual structure of the word mass present in the sources;

b) connections (inner and outer links) among various types of written cultural sources within a wider cultural environment.

Such framework incorporates technologies and tools necessary for large-scale activities aimed towards multi-aspectual presentation of written cultural heritage in a highly distributed manner.

The reasons for implementation of certain solutions are motivated by OCS language characteristics. These characteristics (taken as representatives for processing of also other language sources) from the point of view of their computer processing are summarized, e.g. in (G. Camuglia, M. Camuglia, K. Ribarov 2003).

In the sequel, we will try to present elements of ACT, which underline the integratory capabilities of this tool.

Overview of ACT

ACT is a complex tool designed for processing of large text corpuses along with linguistic annotation. The annotation level is mainly lexicographic and morphological including framework-free basic syntactic and (lexical) semantic information.

It comprises of the following modules: ACT Server, ACT Client, ACT Distiller, ACT Client Light and ACT Web. The system allows users to work on different remote workstations while sharing all created data. History-based modules record changes (on-line) made by multiple users working simultaneously. ACT is a network application, but with a possibility to work also off-line, which is ensured by the ACT Client Light module.

The following is a list of ACT main features (more details on specific functions of ACT can be found in (K. Ribarov et al. 2004)[3]):

- rendering of oforms of a manuscript with necessary freedom to include more variants and more rforms per variant,
- user-editable regular expressions for rendering rules for automation of the rendering process,
- context display and basic editing within a context,
- assignment of more than one identification headword,
- recension headwords and work with multiple recensions,
- multiple morphology tags with configurable morphology assistant,
- flexible-context translation,

[1] Within the current software version the language independence is restricted to linearizable, left to right languages.

[2] ACT is accessible at http://prometheus.ms.mff.cuni.cz/act. ACT has been developed as a student project (at the Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic) lead by Kiril Ribarov. The programmer team consisted of the authors of this work and D. Linh, who helped mainly with the documentation collation and transformation of the morphological tags into their positional form.

[3] Another work related to ACT is (K. Ribarov 2004).

- history-based semiautomatic procedures[4] for rendering, headword assignment, morphology tagging, translation,
- user rights and access control,
- searching using equivalences,
- non-standard fonts support,
- client (ACT Client) - server (ACT Server) architecture with GZIPped connection,
- platform independence,
- multilingual configurable GUI with wide variety of settings,
- export and import to/from XML with native XML format,
- import/export from/to RTF/plain text; export also to HTML,
- complex query assistant,
- creation of ready to use documents as index verborum, retrograde indexes, concordances, list of headwords, list of forms,
- creation of basic statistic outputs, including bi-gram statistics,
- incorporation of corpus and card-file techniques,
- web access.

ACT has a tendency to unite existent technologies of contemporary language processing and characteristics of processing of old sources. Where possible, ACT incorporates components, which although newly programmed are similar to those used in current NLP, an example of which could be the customized positional morphological tag system extensively used in processing of Czech.

Besides many similarities there are as well many differences, which need to be respected and which require new solutions. The processing of old sources has a unique richness of varieties of problems it comes along with. Those are of technical nature (as problems of formatting, coding, mark-up, digitization) and subtask-inherent nature (as, e.g., processing of language varieties on all levels of the language, or processing of citation and commentary networks).

On some characteristic features of ACT

In order to make various interpretations of the manuscripts' texts possible, we distinguish a surface presentation of a word form and its understanding. The surface presentation is taken to be a sequence of characters called original form (oform). Each oform may have various understandings/variants, each of which may consist of one or more rendered forms (rform). Schematically this is presented in Figure 1.

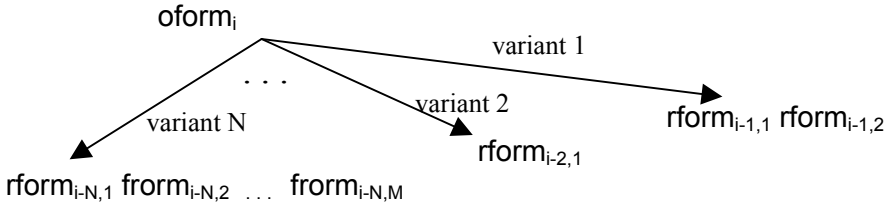

Figure 1

Each rform is a central processing unit, which can be annotated in ACT Client. Annotation is the process of assigning values to a set of types of attributes, as:

- context (collocation),
- headword(*) with recension specific headword; references to headwords,

[4] Such mechanisms also preserve the homogeneity of the annotation.

- morphology(*),
- various types of complexes,
- translation equivalents(*),
- correlation mark to other sources,
- unique document ID, i.e. location for the purpose of a unique identification of the oform and rform.

Apart the fact that each oform can have various rforms, the (*) sign determines a possibility for multiple input (variant recording) of the corresponding types of annotation. Therefore, each rform may have any number of morphological tags (mt) and any number of identification headwords (ihw), as presented in Figure 2.

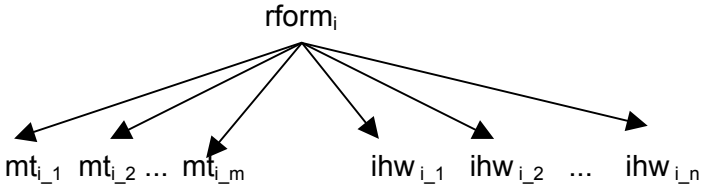

Figure 2

The positions of the morphological tags can be set by the user, including user specific setting of their meanings.

Each manuscript can be a part of a specific recension. Each recension has its own list of headwords which are linked to a unique identificational headword.

The processes of rendering, morphological annotation, headwords assignment, and translations assignment are history based - while working in ACT the user has a complete control over the previous work (including the work of her/his colleagues in the network). In many cases the annotation/assignment process is as simple as selecting a member of a list.

The concept of a complex, any type of multi-rform unit, deserves a special attention. A complex is a relation of specific type (the type is determined by the user) between any number of rforms, which do not necessarily need to be adjacent. A complex can be (the user has a freedom of defining own meaning of complexes), e.g.: analytical form, prepositional phrase, clause, sentence, discourse mark, textological mark. As it will be shown later in the text, the complexes can be also used for XML structure representation.

Complexes are also used for assignment of translation equivalents: translation pairs are established over two complexes of specific type, one in the original document, while the other in the target document. This allows recording of one-to-one, one-to-many, many-to-one and many-to-many relations.

ACT uses catalogues for organization of documents. The catalogue records basic information about the document (as name, identification, date and place of creation, recension, language), technical characteristics as font specifics, location of the text file and location of the scanned pages of the document. Processing of the document can therefore be accompanied by views on images of the original document.

ACT Web

The document material presented in a form of scanned collections of pictures, pages of rewritten texts, and annotated rforms can be accessed via the ACT Web module, http://prometheus.ms.mff.cuni.cz/act/www. Any document processed in ACT Client can be automatically accessed via ACT Web.

With its 700,000 word forms[5], most of which lemmatized with assigned POS, available also in a form of a text and some of them as graphical images, the ACT Web collection is a unique one and the biggest of its kind accessible in electronic form via Internet.

The ACT Web module allows a user to:

- select a manuscript or a subset of manuscripts,
- perform a search on a part of a word-form, morphology tag, headword,
- display results with concordances and frequences,
- display manuscript text and picture if available.

Usage of complexes for non linguistic annotation

A base XML tree, a small example of which is presented in Figure 3 can be input into ACT in various ways, depending on the assigned meaning of the XML tags.

<A> *w1*
<B> *w2 w3*
<C> *w4*
</C>
w5
</B>
</A>

Figure 3

The main idea to follow is the possibility to interpret each XML tag as a complex in ACT. Therefore, the sequence of five words *w1 w2 w3 w4 w5* (from Figure 3) can be understood as in Table 1 (assuming that the complex type is implicitly taken to be the XML tag name).

Surface word: oform	Rform	List of complexes the rform belongs to
w1	*w1*	A
w2	*w2*	A, B
w3	*w3*	A, B
w4	*w4*	A, B, C
w5	*w5*	A, B

Table 1

For the purpose of word form identification, ACT has four levels of mark-up: foglio, position on foglio (as r, v, or a, b, c, d), line number (or paragraph) and position on line (or paragraph). This permits that certain XML tags determining document's structure do not necessarily need to be imported as complexes of certain type but can be imported as page identification or paragraph identification thus allowing the user to focus on more specific tags.

Query assistant (a part of ACT Client) allows a user to search any subset of rforms and/or oforms and/or complex types, to create concordances from various sources e.g. restricted only to certain complex type, to create frequency list and to create many other output forms definable by the user. Thus, if marked up in the XML file, one may search only in the heads of the articles, or only in the bibliographical notes, compare lists of citations from various documents, etc.

Assuming that formally oform can also be an "empty" string, the import presented above can also be as in Table 2 (one can easily image also other possibilities).

[5] In terms of distinct word-forms 163,607 were recorded, with 15,941 distinct lemmas.

Surface word: oform	Rform	List of complexes the rform belongs to
w1	*w1*	A
w2	*w2*	A, B
w3	*w3*	A, B
EMPTY	*w4*	A, B, C
w5	*w5*	A, B

Table 2

According to Figure 1, each variant of an oform can have any number of rforms, therefore the "empty" oform can have more than a single word associated to it.

The advantage of such or similar import strategies would be to separate e.g. editorial remarks inserted into the text, from the text itself. In this way the text would not be polluted by citations, bibliographical or other types of notes; the text to read would be *w1 w2 w3 w5*.

For cross-reference (citation type) mark-up, ACT offers the correlation to other sources attribute, by which any sequence of words can be marked as belonging also to other sources.

ACT distiller

Another module with strong integration character is the ACT Distiller, a module for incorporation of card-files into a corpus. To our best knowledge, ACT Distiller is the first module of its kind.

By a card-file, a lexicographic card-file is understood, e.g. card-file with some subset of the following information:

- lemma (headword),
- additional lemma (serves for more specific definition of the lemma, usually in multi-word components),
- word-form (obligatory),
- morphological identification of the word-form,
- word-form ID, location in the manuscript (obligatory)
- correlation of the word form to other sources,
- context of the word form (obligatory),
- translation of the word form, including the context of the translated part.

ACT Distiller permits the user to:

- view scanned card-file cards
- rewrite the obligatory parts of the cards.

We understand a card-file catalogue as a static view on the original text data. Card-files present the old "technology" of processing of texts, suitable for manual search. Besides their advantages, card-files have serious drawbacks as practical impossibility of verification of their contents and of their completeness. If those texts are part of an annotated corpus that overcomes the mentioned drawbacks, card-files can be easily and automatically derived.

Current card-files catalogues gather immense information, which we believe should be used and incorporated in modern systems. Simple manual rewriting of the contents of millions of card-files creates practically unreachable amount of work. ACT Distiller gives the opportunity to do a first-phase partial rewriting (rewriting only of obligatory parts) according to the scheme as in Figure 3.

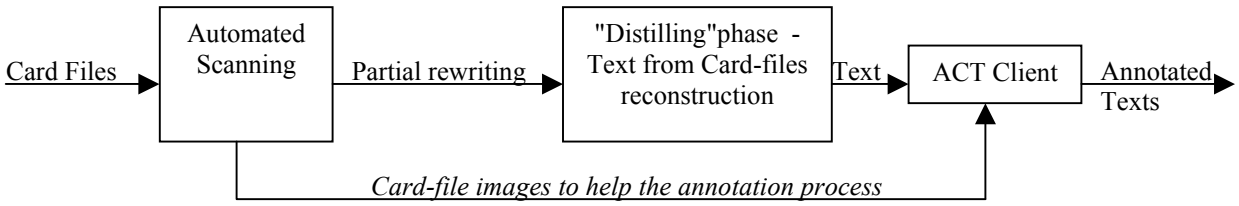

Figure 3

To ease manual check-up, ACT Distiller incorporates a semi-automatic context binding tool and a comparative tool that visualizes possible overlaps, mistakes, and differences while text from card-file reconstruction. Once the location is inserted, the context, if inserted earlier, is displayed and is not repetitively inserted.

The obligatory word-form location is inserted manually. Serious studies should be made for application of OCR system for automatic location identification, since any second of saved time per card processing results in saving of months of work for, e.g., 10-million card-file catalogue.

Concluding note

As designed we believe that ACT contributes towards contextual and intelligent heritage Information Technology framework. ACT is currently used for processing of mediaeval Slavonic manuscripts.

Acknowledgement

This work was supported by the Center of Excellence, Center for Computational Linguistics, project number LN00A063 of the Czech Ministry of Education.

References

G. Camuglia, M. Camuglia, K. Ribarov (2003). "Computer Processing of a Clopen Language: Old Church Slavonic", In Linguistica Computazionale, Volume XVI-XVII, Special Issue, Editors: A. Zampolli, N. Calzolari, L. Cignoni. Instituti Editoriali e Poligrafici Internazionali, Pisa-Roma.

K. Ribarov (2004). “Towards Intelligent Written Cultural Heritage Processing - Lexical processing. To be published in: Proceedings of LREC 2004, Portugal.

K. Ribarov *et al.* (2004). "We present the ACT Tool", To appear in: Scripta & e-Scripta, Volume 2, Institute of Literature, Bulgarian Academy of Sciences, Sofia.