Towards Intelligent Written Cultural Heritage Processing - Lexical processing

Kiril Ribarov

Research fellow
Center for Computational Linguistics
Faculty of Mathematics and Physics, Charles University
Malostranske nam. 25, Prague 1
ribarov@ufal.mff.cuni.cz

Abstract

Through ACT (Annotated Corpora of Text) software package for lexical and corpus processing of European written cultural sources (currently used for processing of mediaeval Slavonic manuscripts) this work presents another step forward towards a contextual and intelligent heritage Information Technology framework. ACT is suitable for capturing characteristics of old written sources including 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, head-words (including recensional), translation equivalents, multi-word units, and correlation to other sources. The whole annotation process is automated, and individual 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, web text and image access, incorporation of lexical card-files into a corpus, and text-from-card-files reconstruction.

1. Introduction

Intelligent heritage IT framework 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.

Applied on mediæval Slavonic written cultural heritage in accordance with the above stated intelligent heritage framework, this work is aimed as an outline of:

(1) the main functions of Annotation Corpora of Text [1] (ACT), a language independent[2] software tool for lexical and corpus processing of written cultural sources;

(2) the language specifics implemented in ACT;

(3) the first release of lemmatized and POS-annotated Old-Church Slavonic (OCS) language resource (LR)[3].

This work is another step, hopefully forward, in series of continuous efforts in computerized language processing of Old-Church Slavonic (OCS) manuscripts, the most recent papers of which are (G. Camuglia, M. Camuglia, K. Ribarov 2003), (M. Camuglia, K. Ribarov 2003), and (K. Ribarov, M. Camuglia 2003).

2. On language specifics

Apart from contemporary languages the old sources are characterized with problems relevant, among others, to the development of the language (synchronic, diachronic and diatopic characteristics), low presence of language spelling norms, and influences from frequently used translations from other languages. Thus, the language problems to resolve exhibit particularities which make the usage of current lexicographic stations or corpus managers impossible. The most important of the distinctions (particularities) are:

- scriptum continuum,
- variants at various levels of the language,
- abbreviations,
- damaged and unknown parts,
- correlation to other sources,
- multi-lemmatization (due to existence of various recension centers and high level of variability, and/or due to lack of material, usually, lemmatization means assignment of more than a single lemma),
- existence of translation equivalents important for, e.g. contents reconstruction and variability resolutions.

Along with the OCS resources the ACT system is presented as a framework capable of manipulation and capturing of the high-level language variability on word and/or sentential level.

2.1. Some examples

A simple example[4] on surface variability due to scriptum continuum would be

и‖єгожє‖видиши‖плода‖сє‖сътвори‖въ‖мнѣ
(*and the fruit you see created in me*),

[1] ACT is accessible via http://ckl.ms.mff.cuni.cz/~ribarov (further ACT web page). 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: Jiri Bubnik, Jiri Celak, Vojtech Janota, Alexandr Kara, Vaclav Novak; the web interface was developed by Tomas Vondra.

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

[3] For web access to the OCS material visit http://ckl.ms.mff.cuni.cz/~ribarov

[4] The example is taken from the *Povest o Varlaam i Joasaf*, an unpublished manuscript stored at the Rila Monastery (Bulgaria) under the signature 3/14.

where the string сътвори||въ||мнѣ could also be divided as сътворивъ||мнѣ (where сътворивъ is the past participle - active mood of *create*), so that both are grammatically correct, but the correct reading can be found only in a wider context. Such wider context is not always available.

Abbreviations of various types, damaged or unknown parts are very frequent and as such they introduce higher level of variability in interpretation and understanding. In order to process them, they need to be rendered, e.g.: (с҃нъ → с[ы]нъ *son*), (гл҃ите → гл[агол]ите *say*), (г҃ь → г[оспод]ь *God*), (цⷭ҇рь → ц[ѣ]с[а]рь *King*), (реⷱ҇ → ре[ч]е *say*), (придох → придох[ъ] *come*).

Although for processing of the contemporary languages it is taken as granted that the main unit to process is either a word-form or a sentence (e.g. for parsing) such a priori certainty is not possible for, e.g. OCS: scriptum continuum eliminates punctuation signs[5] and surface sentence is impossible to capture; some uncertainties in word-form boundaries were stated above.

We suppose that other old language documents, as well as the OCS ones exhibit not only orthographic variability, but also morphological or syntactic one. We stress the need to design systems capable of recording variabilities on various levels - due to the closeness of the corpuses of dead languages any disambiguation process lacks the support of a wider language context or living language evidence in order to approve disambiguation choices.

3. ACT solutions

It this part, only the most characteristic solutions will be pointed out. Those are in close relation to variability resolutions. We will present that the main processing unit is not the surface word-form, but its understandings; we will also present that the main "syntactic processing unit" is not a sentence but a set of any type of multi-word units which aim towards an understanding of a sentence.

3.1. Set of rendered word-forms

In order to resolve the word-level variability, a word-form is understood as a pair (original form, set of rendered forms). The sting of characters identified as a part of an image or as a part of a text (e.g. scriptum continumm) delimited by the user or word-segmentation algorithm, represents the original-form (e.g. сътворивъ). The understanding, or the set of possible understandings of the original form is a set of rendered forms (variant 1: сътвори въ, variant 2: сътворивъ). A single original form may have various rendered forms in two levels:

- horizontal: the original form is identified as series of neighboring rendered forms (as in variant 1, two rendered word-forms exist: сътвори въ)
- vertical: the original form exhibits variants of the rendered forms, which are listed as alternatives such that each of them can become a part of a(n) (alternative) context.

A rendered form (further word-form, word) becomes a main processing unit, which is further:

- assigned a morphology distinction (or a set of morphology distinctions in case of an unresolved variant)
- assigned a head-word (disambiguated lemma accompanied by basic dictionary information and/or inter head-word's links) or a set of possible head-words in case of a variant; a head-word is further placed within a specific recension and linked within a network of equivalent recension head-words,
- assigned a translation equivalent (or a set of possible equivalents), if any,
- correlated to other sources, if any,
- assigned a complex (or a set of complexes[6], *see later*).

Within user-friendly environment, assignment of morphology, of head-words and of translations links is automated in order to speed up the manual parts of annotation and lexical work as much as possible. The process of rendering, that is assignment of rendered form to an original form, is also automated through creation of ordered lists of re-writable rules based on regular expressions.

3.2. Complexes

Any kind of multi-word unit is called a complex. The term complex is used because of the freedom to assign any kind of liberally distant link between any two (or among a set of) words. ACT supports user definable complexes, therefore complexes of various types. Each rendered form can become a member of a complex.

The possibility to determine various complex types allows the user to study the texts on various levels, and to resolve phrasal, idiomatic, and/or sentential variabilities. Starting from the simple ones, one may define complexes of, e.g. the following types:

- analytic verb form,
- reflexive particle,
- noun phrase,
- prepositional phrase,
- a whole sentence, if identifiable,
- discourse relation,
- idiom,
- citation,
- date, etc.

This possibility permits to treat the text as string of words with various stand-off structures above it, not restricted to spelling or other norms.

3.3. Complexes for translations and processing of other languages

The set of documents processed in ACT are organized in catalogues, a folder of documents with given language specifics. Various instances of a catalogue can be created, each of them, if needed, with different language specifics as character set coding, sorting order, and morphological tag structure.

Assuming that manuscripts were frequently rewritten in the past or translated from other languages (OCS are often translations from Ancient Greek or Latin) marking translation equivalents is needed for correct understanding of the, e.g. damaged part of the original document.

[5] Punctuation marks are more frequent in newer documents and may characterize tendencies of creation of, originally missing, spelling norms.

[6] Any type of multi-word unit.

ACT allows establishment of translation links between documents of two different catalogues. These links are established between complexes, assuming that:

- a complex of `translation` type is defined,
- each word-form is a complex,
- for many-to-many translation relation the corresponding group of word-forms are marked as complexes of the required `translation` type.

During translation equivalents' assignment, ACT builds a translation memory, which is further used for automatic suggestion of translation pairs.

3.4. The DTD

During the last two years, significant developments of the original STINO, now ACT system were made in the stream of the already performed or announced changes, as in (Ribarov 2002). The whole original system has been reprogrammed and new data formats have been introduced[7]. Besides others, newly, XML format has been introduced[8] with the below-presented DTD. This DTD is included at this point also as an implicit and more specific description of the ACT annotation span.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT bindkeyword (keyword)>
<!ELEMENT complex (#PCDATA)>
<!ATTLIST complex
        complex_group_refid IDREF #REQUIRED
        position CDATA #REQUIRED
>
<!ELEMENT complex_group (#PCDATA)>
<!ATTLIST complex_group
        complex_type_refid CDATA #REQUIRED
        refid IDREF #REQUIRED
        note CDATA #IMPLIED
>
<!ELEMENT complex_groups (complex_group+)>
<!ELEMENT   document   (pages,   originalform+,
complex_groups)>
<!ATTLIST document
        created CDATA #IMPLIED
        notes CDATA #IMPLIED
        place CDATA #IMPLIED
        scanedmanuscriptdir CDATA #IMPLIED
        documentAbbreviation CDATA #REQUIRED
        date CDATA #REQUIRED
        idsorting CDATA #REQUIRED
        idredaction CDATA #REQUIRED
        dateofcreationupper CDATA #REQUIRED
        idtranslation CDATA #IMPLIED
        manuscriptfont CDATA #IMPLIED
        dateofcreationlower CDATA #REQUIRED
        name CDATA #REQUIRED
        exportType CDATA #REQUIRED
        typization CDATA #IMPLIED
>
<!ELEMENT keyword (#PCDATA)>
<!ATTLIST keyword
        partOfSpeech CDATA #IMPLIED
        id_ident IDREF #IMPLIED
        lemma CDATA #IMPLIED
        paradigm CDATA #IMPLIED
        homonym CDATA #IMPLIED
        refid IDREF #IMPLIED
        idredaction CDATA #REQUIRED
>
<!ELEMENT morphology (text)>
<!ATTLIST morphology
        keyword_refid IDREF #IMPLIED
>
<!ELEMENT originalform (text, renderedform)>
<!ATTLIST originalform
        form_image_url CDATA #IMPLIED
        row IDREF #IMPLIED
        positioninrow IDREF #IMPLIED
        page IDREF #IMPLIED
        external_id CDATA #IMPLIED
>
<!ELEMENT page (#PCDATA)>
<!ATTLIST page
        user_page_part CDATA #IMPLIED
        page IDREF #IMPLIED
        page_image CDATA #IMPLIED
        user_page IDREF #IMPLIED
>
<!ELEMENT pages (page+)>
<!ELEMENT   renderedform   (text,   morphology?,
complex?, bindkeyword?)>
<!ATTLIST renderedform
        variantnumber CDATA #IMPLIED
        colocationright CDATA #IMPLIED
        otherSource CDATA #IMPLIED
        colocationleft CDATA #IMPLIED
        renderedForm CDATA #REQUIRED
>
<!ELEMENT text (#PCDATA)>
```

3.5. On inputs and outputs

ACT inputs can read RTF, TXT, and XML file formats. The RTF and TXT format may include characters with special meaning (mark-up characters). Any type of user defined search becomes an output written as a file or displayed on the screen. Output file formats are: HTML, RTF, TXT, XML.

The user defined searches can search for any kind of information subset relevant to a word-form (wildcard characters for any attribute values can be used), as e.g.:

- word-forms that initiate, include or end on some character,
- word-forms with some morphological features
- all word-forms of a lemma (head-word),
- word-forms of a given complex type,
- word-forms in which vicinity another word-form occurs,
- word-forms with specific translation, etc.

Any type of searches can be performed on one or more than one document, within a single catalogue. Any type of searches (including complete lists of all word-forms) can be, according to user selection, presented in a form of:

- a list
- index veborum
- retrograde index
- concordance index

[7] All of these changes are in compliance with the basic framework principles published in my earlier works.

[8] For technical specification, system design or other questions visit the ACT web pages.

- frequency list.

Any of the outputs can be sorted according to various sorting criteria. The output are also basic statistic-oriented outputs, as frequencies and bi-gram lists.

The searches are implemented via a query assistant, which is adaptable and can be defined by user needs.

4. ACT Web

The document material presented in a form of scanned collections of pictures, pages of rewritten texts, and annotated corpus can be accessed via the ACT-Web module, accessible at the address as stated in the introduction of this paper.

With its 700,000 word forms[9], most of which lemmatized with assigned POS, available also in a form of a text and some of them scanned, 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, head-word,
- display results with concordances,
- display manuscript text and picture if available.

5. ACT for Card-Files

In accordance with (Ribarov 2002) and (Ribarov, Camuglia 2003) ACT module, called Distiller, is, up to my knowledge, the first module for incorporation of card-files into a corpus.

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

- lemma (head-word),
- 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.

Rewriting the obligatory parts of the card-files follows the following steps:

1 The word-form location is inserted manually (as a part of further considerations a design of OCR system for automatic location identification is planned; for notes on card-file structure see (Ribarov, Camuglia 2003)).

2 Relative to the inserted notation closer and wider contexts are displayed:

   i. if the word-form to be inserted is already in the context the user is only expected to verify the information,

   ii. if the word-form is missing, the word-form is added together with the parts of the missing context.

The other card-file information is filled in as a part of an annotation process within the ACT main module; in this case the word-form to process (lemmatize, tag) is accompanied by the card-file image.

To ease manual check-up, ACT-Distiller incorporates a context binding tool and a comparative tool that visualizes possible overlaps, mistakes, and differences.

6. Conclusion

Let us, therefore, conclude that: ACT integrates tools necessary for state-of-the-art linguistic processing and presentation of written cultural heritage sources, demonstrated on mediaeval Slavonic written cultural heritage sources. It contributes towards a creation of adequate and innovative intelligent heritage Information Technology framework for addressing digital presentation of written cultural sources. In general, the ACT framework does not neglect the possibilities for link establishment to other (e.g. European) written cultural sources. Along with the presented OCS LR, ACT fills in the currently existing gap in the European e-space where mediaeval Slavonic cultural heritage is presented in scattered and non-unified manner.

7. Acknowledgements

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

8. References

K. Ribarov (2002). "Old Sources and Modern Procedures". In: Proceedings of LREC 2002, Spain.

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.

M. Camuglia, K. Ribarov (2003). "Old-Church Slavonic in Codes", In: Computational Approaches to the study of Early and Modern Slavic Languages and Texts - Proceeedings of the "Electronic Description and Edition of Slavic Sources", Pomorie, Bulgaria. Sofia.

K. Ribarov, M. Camuglia (2003). "Incorporation of Old Church Slavonic Card Files into a Corpus", In: Scripta & e-Scripta, Volume 1, Institute of Literature, Bulgarian Academy of Sciences, Sofia.

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