R. Meyer

THE REGENSBURG DIACHRONIC CORPUS OF RUSSIAN

1. Introduction

While corpora are only *one* source of data for synchronic linguistic research, they are – in one form or the other – the *only* source for diachronic research. Recently, several research projects have started working on *electronic* historical Russian corpora[1], and impressive first results have been achieved. The goals of the Regensburg Corpus overlap with those of the above projects, but are also to a certain extent distinct: We aim at a *diachronic* (rather than historical) corpus, which is going to be used specifically for *linguistic* research. The present paper is organized as follows: In section 2, we give an overview of the tasks following from the idea of diachronic linguistic corpus, and the solutions taken in the best currently available electronic Slavonic diachronic corpus, the diachronic part of the Czech National Corpus. The current state of the Regensburg corpus is presented in section 3. Section 4 provides a summary and outlook.

2. Linguistic and Technical Tasks for a Diachronic Corpus

The digitization of old documents may serve a number of goals, an important one being the preservation of valuable sources as part of a language's cultural heritage. This *document preservation approach* naturally has to strive for maximal precision, in order to virtually recreate a historical monument, with all the information attached to it. Work in this domain is usually directed towards an open community of users who access the digital source under diverse research interests.

[1] Cf. the talks: *Baranov V.A.* Information Retrieval System «Manuscript» – a Specialized Tool for Complex Researches of Slavonic Hand-Written Heritage; and *Azarova I.V., Alexeeva E.L., Zakharova L.A.* Tagging of Text Fragments in the Agiographic Texts Corpus SKAT.

An encoding standard guarantees the reusability and reliable interchange of documents. The most prominent project under this umbrella is the text encoding initiative, which follows the rationale to «support the encoding of all kinds of features of all kinds of texts studied by researchers» (TEI-P4[1]). Linguists working on the historical development of a language have some more specific and additional needs. They require general *textological* features, *structural* features of the text, as well as annotation at the *token level* and *across* several *tokens*. The necessary textologial features are readily covered by the TEI recommendations, which provide XML elements and attributes for the author, scribe, state of the document, the text history, and the various aspects of coding resonsibility. Equally rich is the TEI selection for text-structural markup, as e.g. front and back matter, pagination, sectioning, headings, graphics, marginalia, glosses and notes, added material and gaps. When it comes to the token level, however, the TEI recommendations leave a lot to be added. For research into orthography, and for transparency, it is desirable to provide a close rendition of the original orthography. Secondly, we need a normalized form of each token, representing how the original graphic form is being «understood» and fed into further linguistic analysis. The relation of the graphic unit in the original document and the understood tokens in the corpus may be manifold: There may be various ways of splitting up a graphical string into tokens, and each part may be understood in various ways. To take a typical example of the latter, the token оубилъ in (1) may enter the «understood text» as оуби or as оубил, depending on whether the later added gloss is taken into account or not, which can lead to very different linguistic analyses.

(1) ... от своих снѣвъ е с мѣчи оубилъ (Flavius)

[1] **http://www.tei-c.org/P4X/index.html**

Thirdly, especially for lexicological research, we want to access all forms of a lexeme together. The actual «dictionary form» assigned to the lemma is relatively unimportant; in fact, there is little need for a text-specific lemma apart from the hyperlemma which bundles all forms of the lexeme, if queries can be properly restricted to texts and times. Fourthly, part-of-speech (POS) and morphosyntactic information is needed for each token. While most modern corpora (e.g. the BNC, the ČNK, the German DWDS Corpus) provide only *one* analysis, the IPI PAN corpus of Polish[1] rightly acknowledges ambiguous morphosyntactic tags. In the case of a diachronic corpus, this point, although generally overlooked, is important: developments in historical syntax often proceed across «bridges», ambiguous forms which are later reanalyzed. E.g., the case of *desjat' krestov* in (2) (cited after Billings & Maling 1995[2]) is not at all obvious – we might be facing a personal passive with a nominative (as in modern standard Russian) or a early *no-*/*to*-impersonal with an accusative (as in modern standard Ukrainian). A decision between these two annotations is not conditioned by the data, but it influences analyses of the development of the *no-/to*-construction in East Slavonic. In the ideal case, *both* a query of nominative *and* one of accusative would output the example, together with an indication of its ambiguity.

(2) ... *da k vam že poslano desjat' krestov* (Avvakum)

Fifthly, the technology used for a diachronic corpus should be flexible enough to include semantic and syntactic annotation. The XML format (together with a schema or DTD) ensures consistency and exchangeability; but since XML disallows partially overlapping elements, it is difficult to tag discontinuous stretches of tokens as a unit. To give an example, it is desirable to tag reflexive verbs as one item, even if сѧ occurs separately (as is often the case in Old Russian).

[1] *Przepiórkowski, A.*: Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version. IPI PAN, Warszawa, 2004.

[2] **ftp://ftp.pitt.edu/dept/slavic/jsl/notobibl.ps.zip**

There are various ways to get around this problem, the conceptually clearest one being a stand-off annotation, in which a pointer to the respective annotation unit is set for every token belonging to it. The ACT system[1], which is used for the Regensburg corpus, allows for a stand-off annotation of arbitrary selections of tokens («complexes») assigned to freely chosen types. The diachronic part of the Czech National Corpus[2] follows a somewhat radical, but extremely fruitful approach. Only for the very early period are full texts included; after 1500, the corpus aims at a balanced collection of text samples. Following a specific Czech tradition, only unusual original orthography is kept in the annotation, and all tokens are orthographically standardized. The structural annotation of the texts is restricted, only headings, notes, citations, crossed-out material, glosses and marginalia, and unreadable parts are marked up. All texts are searchable via the same interface as the synchronic Czech National Corpus[3] – a user-friendly solution, which guarantees a high standard of query options: the powerful query language (essentially the same as in the Stuttgart CWB[4]) allows for regular expressions over all levels of annotation; lines of development can easily be retrieved via frequency distributions over single texts, or over time-sliced subcorpora. In the Regensburg corpus, we would like to follow the sampling approach, at least for the Middle Russian period, because of resources, but also in the light of modern corpus linguistic methodology. As mentioned

[1] *Ribarov K. et al.* ACT – Computer Processing of Written Cultural Heritage Sources // Proceedings of INFORUM 2004 conference, Prague. 2004. URL: **http://www.inforum.cz/inforum2004/pdf/Ribarov_Kiril.pdf**

[2] **http://ucnk.ff.cuni.cz/diakorp.html**; *Kučera K*. The Czech National Corpus. Principles, Design, and Results. Literary and Linguistic Computing 17/2, 2004. P. 245–257

[3] *Rychlý P*.: Manatee, Bonito and Word Sketches for Czech // Proceedings of the 2nd International Conference on Corpus Linguistics. St.-Petersburg State University Press, St.-Petersburg, 2004. P. 124–132.

[4] **http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/**

above, our primary goal is not the digital conservation of monuments, but diachronic linguistic research. However, following Russian traditions, we want to be able to restore as much as possible of the original orthography. Furthermore, given the above arguments, we need some more flexibility in the annotation: multiple normalizations and lemmata, and ambiguous morphosyntactic tags.

On the technical part, a user-friendly diachronic corpus would require at least the features of a good synchronic one (see e.g. the ČNK): a concordancer for querying tokens, lemmata and morphosyntactic annotation with subcorpora management, sorting, and a statistical module for frequency counts and retrieval of collocations. Additionally, retrieval of complex annotations (across tokens and within a token), and precise source information for each hit, is called for. The retrieved text should be rendered in a form graphically close to the original. Moreover, especially old Russian texts require a comparative view on Greek sources, for which aligned stretches of text (as in a parallel corpus) must be available. For a truly diachronic corpus, the most important point seems to be the retrieval of frequency distributions for arbitrary queries across texts and time slices, in order to trace historical developments.

3. The Current State of the Regensburg Diachronic Corpus

The Regensburg Diachronic Corpus of Russian is an ongoing research project; at the current stage, it contains several long prose works of the 14th and 15th century (Flavius' *Iudejskaja vojna*, the *Šestodnev*, *Nestor's chronicle* in the *Ipat'evskaja* and *Lavrent'evskaja letopis'*), and some shorter texts originating from the 12th century (*slova* and *pritči* Kirilla Turovskogo, a January *minea* in prep.). Part of the *Šestodnev* has already been annotated with (hyper)lemmata, a procedure which is going to be subsequently extended to all texts of the corpus. The other achievements so far concern the technical setup: After experiments with a native XML database, we decided to switch to the ACT system. ACT consists of a relational database to store the

documents, a convenient Java client (and server) for annotating the documents[1], and a PHP client to query the database directly over the internet. The database and web interface was transferred to UTF-8 format. All texts in the database are now automatically converted from the ACT XML format into TEI-compliant XML, in order to be compatible with future developments. Fig. 1 shows the Java annotation tool: each first line contains a coding of the original graphic form, холюбче, with c͡ above хо, the second line contains the spelled-out «rendered form», христолюбче, the third contains the POS tag (so far only the main word class, N,), and the fourth the lemma, in this case in its Old Church Slavonic form, хрьстолюбьць.

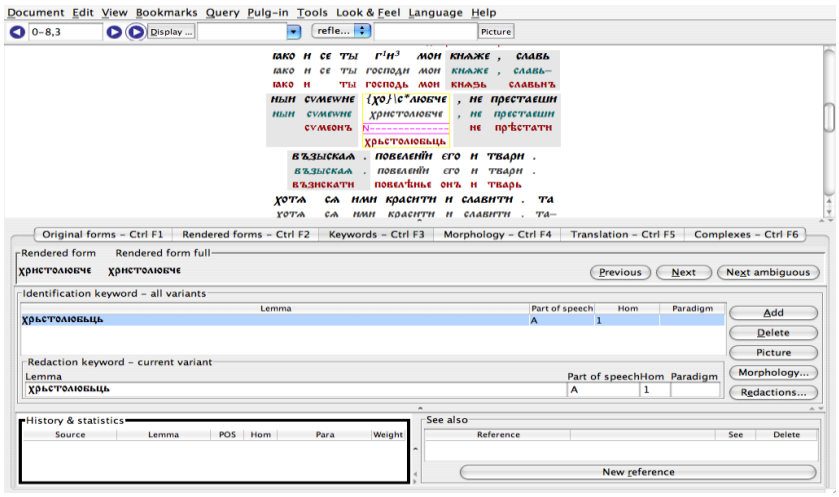

Fig. 1. ACT Java interface for annotation

The ACT PHP web interface has been adapted and is being developed further. We simplified the concordance view, added exact source descriptions to the hits and put in a paragraph view. Besides

[1] *Bubník J.* Automatizované značkování (středověkých) textů – heslová slova, morfologie, komplexy, korelace. Diploma thesis, UFAL, Charles University, Prague, 2003.

the usual SQL, a regular expression search can now be used. This facilitates queries for specific parts of words, morphemes, and stems. In order to guarantee an optical impression close to the original, we worked out a mechanism in html for properly setting letters above each other, without the need for a special font apart from a full cyrillic unicode one (see fig. 2). The original form (encoded as in fig. 1), is output as unicode characters set above each other where appropriate. For searching, the normalized «rendered form» has to be used. ACT allows for the definition of a set of regular character equivalences, so that e.g. search words containing *o* hit also on *w*, those containing *ı* also hit on *ї* etc. For the texts with lemma information, all the instances of a lemma can be retrieved at once. Fig. 2 illustrates part of the result of a lemma search for богъ, rendered in a way graphically close to the original. As expected, all the inflected forms, shortened or in full, are found. The original encoded form is rendered properly.

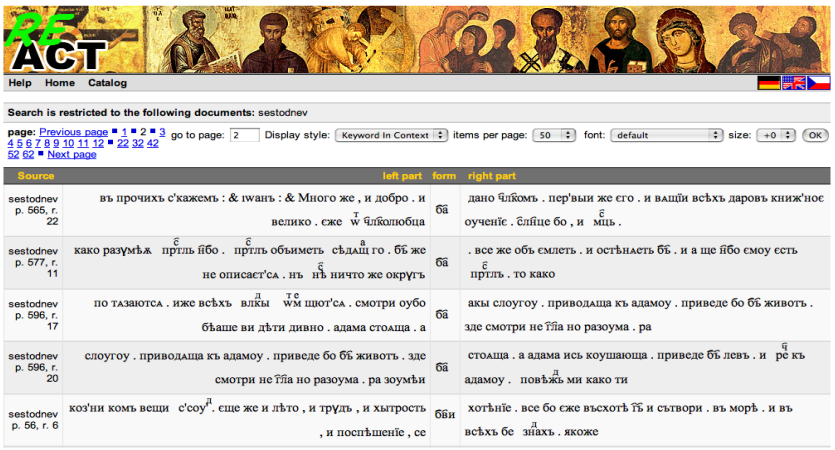

Fig. 2. PHP interface: Partial output of a lemma search for богъ

4. Summary and Outlook

As is obvious from sections 3 and 4, the Regensburg Diachronic Corpus has already fulfilled part of the requirements we argued for, but several others still have to be worked upon. We avail of a PHP

interface with a rendering component, regular expression and lemma search, the annotated texts are saved in TEI-conformant XML, and lemmatization is under way. The most important goals for the future are (i) to gain a properly balanced corpus of text samples, (ii) to realize a frequency and collocation module for the web interface, and (iii) to add morphosyntactic information for each token. Since several similar projects are under development, one of the most important hopes for the future is a common XML standard for document exchange. There are definitely too many valuable and important texts around for a single research group to manage, so close coordination and cooperation will be useful for everybody in the end.