The Central Asian Languages Corpora project (CALC). I: Modern Uzbek

Marc Vandamme & Hansje Braam

Vandamme, Marc & Braam, Hansje 1997. The Central Asian Languages Corpora project (CALC). I: Modern Uzbek. *Turkic Languages* 1, 254-261.

The CALC project aims to produce medium-sized (500,000-1,000,000 tokens) computer-readable text corpora of modern Central Asian Turkic languages. The texts are meant to be used as linguistic data as well as language teaching material. Care is taken to collect texts from a broad range of text types and usage domains. The Uzbek part has been finished recently and the Kazakh, Kyrghyz and other components are *in statu nascendi.*

The Uzbek corpus contains some 1,100,000 words, distributed over 243 corpus texts, covering about 45 text types. The larger part of the texts dates from after 1990. The corpus is available for analysis at Utrecht. In the future it will be possible to consult the data collection over the Internet.

Marc Vandamme & Hansje Braam, Department of Oriental Studies, Drift 15, NL-3512 BR Utrecht, The Netherlands.

1. Context

The development of descriptive and comparative linguistics of the modern Central Asian Turkic languages has been hampered as a result of the political constellations of the Cold War era on the one hand, and the relatively late introduction of digital information processing to this field on the other. These factors explain why computer readable language corpora of Turkic languages are almost nonexistent, and why we have so few adequate bilingual dictionaries of Turkic apart from Russian. With the disappearance of many if not most of the political obstructions since 1991, initiatives contributing to a solid empirical foundation of descriptive Turkic linguistics would seem apposite. Such initiatives are also necessary in consequence of the growing demand for educational materials and research tools such as adequate dictionaries.

In 1994 a plan was drawn up to meet this goal. This resulted in the CALC (Central Asian Languages Corpora) project, an international cooperation of institutes and scholars from Europe and Central Asia. The Utrecht group is responsible for the coordination and concrete realization of the project activities. At the Department of Oriental Languages and Cultures of Utrecht University the project functions within a broader research frame concerning the philology of the Central Asian Turkic languages, from the older stages (Chagatay) to the modern varieties. CALC focuses on the modern Turkic languages, albeit with a view to link the digital materials to older stages of the relevant languages in a later phase. Of course this only makes sense after lemmatization of the textual materials has taken place. In this process we work from the older towards the more recent stages (Vandamme, Boeschoten & Braam 1989). The lexicon of the text of Rabghūzī's *Stories of the Prophets*, completed in 1310 AD (Boeschoten, Vandamme & Tezcan 1995), serves as point of departure for the lemmatization work.

2. CALC goals

In the first place, the CALC project strives to construct high quality computer-readable medium-sized (0.5 to 1 million tokens) text corpora of modern Central Asian Turkic languages (Uzbek, Kazakh, Kyrghyz, Turkmen, Uyghur and others) and make them accessible to all interested scholars.

It is secondly our intention to function as a clearing house for other electronic language corpora and derived sets of data of the Central Asian languages concerned, yet not originally created by the Utrecht CALC group.

CALC also wishes to provide for the publication of research materials (data, bibliographical information etc.) and results (frequency lists, analytical studies of the data collections etc.), especially in electronic form.

3. Corpus aspects

3.1. Text selection

The data collection must support amongst others lexicographical, area-linguistic, textlinguistic and general comparative studies. The data bases should be flexible in use and of a sufficient size to enable the realization of related projects such as the compilation of dictionaries, language manuals tailored to various professions and activities (banking, law, agriculture, education etc.) and language courses. For reasons of efficiency we have excluded non-printed sources in this phase of CALC. We have been able to collect some manuscript materials for modern Uzbek, although they are not included in the corpus. Future inclusion of oral materials (from tape, existing transcriptions etc.) is necessary, especially since many dialects are in danger of extinction.

Texts are selected for inclusion in the corpora according to a predefined division into text types and domains. At the first CALC meeting in December 1994 the issue of selecting text types and domains was discussed. This resulted in a preliminary list of proposed text sorts of a more or less ad hoc character. Finally a list was established based on a more systematic approach. This list covered many of the pre-theoretic text sorts mentioned in Gülich & Raible (1975) and Heinemann & Viehweger (1991). Some extra text sorts proposed by participants of the meeting were also included. The problem of systematic classification still remains, as a single generally accepted text-linguistic framework which can be used to derive a clear-cut text type list does not seem to exist.

We had to settle for a trade-off: On the one hand, we wanted linguistically relevant variation (in text construction), on the other, we also had to choose according to lexical demands. We used the cognitive communication theory of Heinemann and Viehweger to produce, by parametric variance, a smallest set of text sorts and language usages. This ensured that texts taken from these categories would (most probably) contain the linguistic forms which are of interest to us. We tried to meet the lexical demands by selecting the texts from specific domains.

Heinemann & Viehweger (1991: 147-149) state that the systematic classification of texts should take into account four different levels: Function, situation, procedure and structuration. The *function* of the text is what is realized by using the text in interaction: To express oneself, to (re)present oneself, to contact another, to inform someone, to guide someone or to act aesthetically. These functions generally occur in combination. The *situation* of the text is not easily classified, the social structure of the communicative action is described using a sociolinguistic model of interaction. Text *procedures* are goal-oriented cognitive procedures applied in the process of text production and interpretation. Three main classes can be distinguished: Text unfolding procedures (to explain using an example, to make an issue more specific by giving extra information, to give a reason for something etc.), strategic procedures (choosing a narrative, a descriptive or an argumentative set-up of the text) and tactical procedures (additional specification or strengthening of the main procedure, for example, emotional strengthening). Text *structures* concern the way in which the different text parts are combined in order to produce a certain text (for example a request has as structure: Letterhead; Letter-nucleus: K, because of G; End-of-letter. G can have a complex argumentative substructure).

This means that text sorts can be distinguished using a four-dimensional structure, with values which can be combined (as a text can have several functions), and not necessarily holding only binary feature values. The sets of possible values for the categories of text situation, procedure and structure are large sets. Although we can draw up a very large systematic matrix of text types, the CALC project cannot cover all these values at the same time. This is due to a lack of money, time, workpower and available texts. For instance, it became clear that it was impossible to find instructive texts on housekeeping products, such as washing instructions and the like. And of course utterances like curses, obscenities etc. are also hard to find in printed sources. Consequently, we had to limit the extent of text selection.

In the end we came up with the following criteria for drawing up the text selection list, in decreasing order of importance:

1. Gather the same collection of text types (in the same domain, if possible with the same subject) for every language concerned.
2. Choose text types to cover the text functions mentioned above.
3. Choose texts which allow maximalizing the expected number of different linguistic phenomena. This means, of course, that one must vary between communicative situation, procedure and structure.
4. Choose texts according to a priority list of domains. This list reflects the main lexical domains we are interested in. For the lists of text types and domains which were actually selected for Uzbek, see section 4 below.

3.2. Representativity

The resulting data collections are intended to be exemplary, not representative (Bungarten 1979: 42). This means that the frequencies of linguistic phenomena occurring in the corpus for language A are not intended to be statistically good approximations of such frequencies in all A language utterances. As mentioned above, we have tried to increase the odds of finding special phenomena by previous selection (not random choice) of text type and domain.

Care has been taken that the corpus texts, the basic items of the corpus collection, generally contain about 5,000 tokens each. This ensures that the linguistic utterances can be studied in their larger context. However, in the case of texts of types which are generally smaller in size (for example, poetry or advertisements) a corpus text is made up of several of such smaller texts. The corpus texts are each homogeneous in text type and usage domain, although a few corpus texts contain material of several text types, for instance, periodical articles on football with some inserted laudatory verses.

3.3. Representation standards (transliteration, text headers)

The CALC project will eventually use in its electronic data collections the conventions of text encoding of TEI / CES (cf. Dunlop 1995 and Ide 1996) and for further planned explicit language description the conventions as published by the *Eurotyp* group (Bakker et al. 1993) as far as possible. The data will be made available in two formats: A platform independent representation that supports online data retrieval using Internet / WWW connections and also in the current Macintosh form. We are currently investigating the possibilities of a full SGML version of the corpus texts, using TEI / CES. This would ensure 100% transparency also with respect to future developments in hardware and software. At present only the Macintosh version is available for pattern searching at Utrecht.

As complex taggings are for the time being not part of the project, the only problem concerns the alphabets used to write the relevant languages. In order to save space and to be able to work easily with the texts we use a simple transliteration scheme (CATL: Central Asian TransLiteration), which follows in almost all cases a straightforward method. This scheme is strictly one-to-one inside one language or orthography. A few interlingual many-to-one and one-to-many symbol pairs could not be circumvented. In choosing the symbols we have striven to comply with the proposed national standards as far as possible, although no opinion concerning the current discussion on Romanizing the Cyrillic Central Asian writing systems is to be inferred. Example: The poetry line Ҳорғиним, ипакдай қош-кўзи чангларим, becomes in CATL *Horġinim, ipakday qoş-közi çanglarim*.

Textual structures have not yet been covered in full detail, due to the enormous amount of work which must be invested in such a process. For example, in a collection of short articles from a newspaper the bibliographical information is marked (using a simple SGML compliant markup coding system), however the internal structure of the text itself is not (headlines etc.). Information about the texts (contents, date, length, domain, text type etc.) is available through a database and through the TEI-textheaders of the corpus texts.

4. The Uzbek corpus

The Uzbek corpus was located, selected, and converted between early 1995 and autumn 1996. Some material was selected from the library collections at Utrecht, Mainz and Frankfurt, but the major part was collected in Uzbekistan by H. Ykema. The bulk of the material was keyboarded in Tashkent on PC's, a smaller amount was scanned using OCR methods. Because of the generally bad quality of print and paper of the Uzbek books and periodicals, the available OCR techniques did not always produce high quality scans; in some cases 99% correctness was obtained, but in most cases the quality was considerably lower.

The Uzbek corpus comprises 1,164,851 tokens, in 243 corpus texts, taken from 388 sources. Poetry and proverb collections are counted as one source each, the sources of slogans are not taken into account here. All corpus texts, except one, date from after 1963. We have 200 corpus texts dating later than 1980, of which 140 corpus texts date later than 1992.

The following text types are covered:

Academic text book	92,976
Advertisement	4,789
Announcement / TVguide-text	10,197
Autobiography	21,000
Congratulations	730
Cooking recipe	20,086
Decree, permission	28,670
Drama	52,513
Fairy tale	54,899
Guide	22,128
Holy text	30,211
Instruction manual	7,096
Newspaper report	24,278
Novel	52,299
Offical application form	4,606
Periodical article	73,465
Plan	10,426
Poetry	40,262
Proverb	9,925
Report	20,472
Rules of conduct	4,302
School book	3,189
School text	57,716
Scientific book	55,980

Joke	21,698	Short story	55,903
Jurisprudence	35,198	Slogan	2,383
Law text	113,346	Speech	6,867
Letter	5,045	Story	34,791
Manual	47,390	Survey	21,323
Newspaper article	79,252	Tales & Proverbs	3,730
Newspaper article, interview	26,272	Texts on soap boxes etc.	738

Grand total of tokens 1,156,889

The following domains are covered:

Arts & Culture	130,753	Politics&Economics	91,643
Daily life	76,872	Public administration	4,606
Education	110,856	Religion	71,062
Health	70,216	Society	28,401
Law	177,214	Sports	45,402
Literature	296,007	Technology	57,583

Grand total of tokens 1,164,851

The grand total by text type is slightly smaller than the grand total by domain because a few corpus texts contain parts belonging to different text types. These corpus texts were not taken into account in the first grand total.

5. Accessibility

As mentioned above, the Uzbek corpus is for the time being only accessible on location at Utrecht University. However, we shall try to answer requests for data extracts within reasonable limits. It is our intention to make the Uzbek corpus available worldwide via the Internet.

6. Status of other languages concerned

Corpora of other languages are still being put together: Kazakh (selected 80%, converted 20%), Kyrghyz (selected 70%, converted 5%), Uyghur (selected 30%). In addition to these planned corpora, CALC offers to function as an archive or clearing house for language materials from other Central Asian languages. CALC was already so fortunate to receive into its care samples of Yellow Uyghur, Tuvinian and some others.

Acknowledgement

Our special thanks to T. Atabaki, R. Dor, E. Gürsoy-Naskali, L. Johanson, M. Kirchner, A. Nauta, M. Ploeger, M. Roos, C. Schönig, U. Shamiloglu, H. Ykema and others for cooperation, advice and supplying textual materials.

References

Bakker, D. & Dahl, Ö. & Haspelmath, M. & Koptjevskaja-Tamm, M. & Lehmann, C. & Siewierska, A. 1993. *Eurotyp guidelines.* (Eurotyp Working Papers.) Berlin, Strasbourg: European Science Foundation.

Boeschoten H. E. & Vandamme, M. & Tezcan, S. 1995 (eds.) *Al-Rabghūzī. The stories of the prophets: Qiṣaṣ al-Anbiyā', an Eastern Turkish version.* Volume I. Leiden, etc.: Brill.

Bungarten, T. 1979. Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. In: Bergenholtz, H. & Schaeder, B. (eds.) *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora.* Königstein / Ts.: Scriptor.

Clear, J. H. 1993. The British national corpus. In: Landow, G. P. & Delany, P. (eds.) *The digital word: Text-based computing in the humanities* 1. Cambridge, Mass., etc.: MIT press. 163-187.

Dunlop, D. 1995. Practical considerations in the use of TEI headers in a large corpus. *Computers and the humanities* 29, 85-98.

Gülich, E. & Raible, W. (eds.) 1975. *Textsorten. Differenzierungskriterien aus linguistischer Sicht.* Frankfurt a. M.: Athenaion.

Heinemann, W. & Viehweger, D. 1991. *Textlinguistik.* Tübingen: Niemeyer.

Ide, N. 1996. *Corpus encoding standard.* http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html.

Vandamme, M. & Boeschoten, H. & Braam, H. 1989. Editing and linguistic analysis of a medieval Turkic text with the aid of computer facilities. In: Sagaster, K. (ed.) *Religious and lay symbolism in the Altaic world and other papers.* Wiesbaden: Harrassowitz. 77-99.