Methods and tools for development of the Russian Reference Corpus*

Serge Sharoff

Centre for Translation Studies, University of Leeds

Abstract

The paper discusses the history of development of Russian corpora and presents methods and tools that are used in the ongoing development of the Russian Reference Corpus. Development of the corpus follows the key design principles of the BNC and extends them further by introducing an elaborate model of text typology and by adding lemmatisation and morphosyntactic annotations to POS tagging. The paper also discusses problems in development of the corpus that are related to the Russian language and culture.

1. The history of development of Russian corpora

It is not too big a generalisation to say that development of Russian computer corpora followed the pattern established by English corpora. The Brown Corpus (Kucera, Francis, 1967) set up the standard for the design, size and coverage of general-purpose corpora in other languages, including Russian. In 1970s a corpus of 1 mln words was developed by Zasorina and her colleagues; it consisted of 500 samples of 2000 words each and covered four types of genres: mass media, fiction, science (including humanities) and drama (as an attempt to cover the spoken language). The study resulted in a frequency dictionary (Zasorina, 1977), but not in a publicly available resource. The best known comprehensive Russian corpus was developed in the 1980s in Uppsala, Sweden; it also resulted in a frequency dictionary (Lönngren, 1993). The Uppsala Corpus (UC) consists of 1 mln words in 600 samples equally divided between fiction and non-fiction texts. UC is popular for various reasons, partly because it can be freely accessed via the Internet, but for modern standards it is too small and restricted in the genre coverage. It also lacks morphosyntactic annotations and lemmatisation. The lack of lemmatisation hinders the search of multiple word forms, which often cannot be found using regular expresions, e.g. the verb *vyjti* (to leave) in Russian has about 40 forms, including many dissimilar forms like *vyjdu*, *vyshla*, *vyshedshij*. The lack of morphosyntactic annotations hinders even simple searches of grammatical relations, for example, searching for uses of the partitive case or for complements of a particular verb in the dative case.

Another attempt to develop a comprehensive corpus was made in the Soviet Union in the mid 1980s. It is known as the Computer Fund of Russian Language (CFRL). Its aims were similar to those of the British National Corpus (BNC), which was to be developed few years later. The main goal was to create a very large corpus of general language and subcorpora for various genres that would help in the development of NLP applications. The set of corpora would also

provide resources for studying and teaching the Russian language, including development of dictionaries, grammars, textbooks, etc (Andryuschenko, 1989). It was also expected that the corpus would include a historical component to cover the development of the Russian language from the earliest available sources (10th century AD). However, the project did not produce the expected outcome: no representative corpus has been collected. Resources available from the CFRL now include Russian literature of the 19th century and samples of newspapers from 1997. The progress in development of OCR software resulted in multiple ad hoc collections of Russian fiction and reference texts, for instance, Moshkow's Library (ML), but such collections are not balanced and representative. The same applies to collections of newspapers available online.

Currently corpus studies of Russian are based mostly on the Internet. The Internet can be considered as the largest Russian corpus, because the amount of Internet documents available for Russian search engines can be estimated at about 250 billion words (1,5 TB of unique texts indexed by Yandex), much larger than any conceivable corpus. However, there are three types of problems that hinder its use for corpus studies.

First, it cannot be claimed that the material is representative and that there is a balance of text types. Texts presented on the Russian Internet are chaotic: their set depends on preferences and interests of a very specific group of Russian language speakers that are active on the Internet. The recall of search results also cannot be evaluated, because it depends on unknown parameters: which texts are available or not available on the Internet; which texts available on the Internet were not found by the search engine used for the query, etc.

Second, search engines address the needs of information retrieval, rather than linguistic search. Even though search engines provide lemmatisation, so that one can search for all forms of a word, a query cannot be formulated in terms of grammatical features, including tenses, cases or word classes. As for lemmatisation performed by search engines, it is not designed to handle the queries of (corpus) linguists. For example, normal users, who are interested in information retrieval, pay no attention to the aspect of verbs used in their queries and want to get pages corresponding to the verb irrespective of its form. Search engines anticipate the need and index verbs of the perfective and imperfective aspect under one lemma. However, this technique drastically decreases the precision of linguistic searches and leads to some funny results, when *pomni* used in a query leads to pages with *myatyj*, because *pomyat'* and *myat'* form an aspectual pair.

Third, search engines present search results in a way that also does not correspond to the needs of a linguist. The pages are ordered in terms of their information rank that has nothing to do with linguistic criteria. The output also does not form a concordance, because pages in the output are separated by documents, rather than by contexts of their uses. Finally, search results are based on words occurring in titles of pages or keywords or even in other pages that refer to the link being displayed as relevant.

2. The content of the Russian Reference Corpus

From the viewpoint of corpus linguistics, Russian is one of few major world languages that lack a comprehensive corpus of modern language use. However, the need for constructing such a corpus is growing in the corpus linguistics community both in Russia and in the rest of the world. The objective of the project presented in the paper is to develop the Russian equivalent of the BNC, namely the Russian Reference Corpus (BOKR, BOljshoj Korpus Russkogo yazyka). It is designed as a corpus of 100 mln words with the proportional coverage of major varieties of texts in modern Russian, with POS annotation and lemmatisation. The annotation scheme (which is based on TEI) also marks noun phrases and prepositional phrases, because they are important for the resolution of the ambiguity and can be reliably detected. The corpus consists of texts originally written or uttered in Russian by native speakers[1] in recent years (the exact diachronic sample depends on the text type and is discussed below).

Table 1. Corpus composition

	Russian Standard	**BOKR**
quantity	10 mln words (500 texts)	100 mln words (10,000 texts)
quality	a representative sample of Russian fiction written between 1960 and 2002	a representative corpus of modern Russian, balanced according to a text typology
annotation	POS tags, morphological and partial syntactic properties with manual disambiguation	POS tags, morphological and partial syntactic properties with automatic disambiguation
access	public Internet access with a query interface shared between the two corpora (Russian Standard is a subcorpus of BOKR)	

BOKR will include the Russian Standard, a subcorpus of 10 million words of modern fiction representative for the standard literary language. The relationship between the two corpora is described in Table 1. The two corpora differ mostly in their foci: on the large size, wide coverage and the balance of genres in BOKR and on selection of culturally salient modern literary works and manual disambiguation of morphosyntactic annotations in the Russian Standard. The latter aspect is similar to the design intentions of the hand-corrected core BNC subcorpus (Leech, 1997). The Russian Standard is aimed to be the basic source of information for the development of corpus-based Russian grammars for academic and teaching purposes, while BOKR will provide a complementary source of grammatical information and will be the basic source of lexical information.

In one aspect, the design of the Russian Standard is remarkably different from the design of the core BNC subcorpus. The core BNC is based on a proportional selection of texts from the whole set of the BNC files, while the Russian Standard

is based on literary texts. This reflects the difference in the cultural status of the language of imaginative writing in the British and Russian cultures: in Russian the literary language is treated as the authoritative source, which effectively defines the language used by native speakers. This fact is also the reason for the higher proportion of fiction in the Uppsala Corpus and the corpus used by Zasorina (1977): fiction texts covered about the half of their content, much higher than the proportion of fiction in the Brown Corpus (25%) and the BNC (17%), cf. also the balance of genres proposed for BOKR in the discussion below.

2.1 The typology of texts

The balance of genres in BOKR is based on a text typology that is more sophisticated than that of the BNC. The basic principles for describing texts in BOKR follow the EAGLES guidelines (Sinclair, 1996), which distinguish between text-external (E) and text-internal (I) parameters in text classification:

1. E1 (origin) - parameters concerning the origin of the text, i.e. the creation date, the author's age and sex, the place of his/her origin, other circumstances of text creation that can affect linguistic parameters;
2. E2 (state) - the appearance of the text, in particular, the distinction between written and spoken text modes (including written-to-be-spoken and electronic communication as the two border cases), and between published sources (books, magazines and newspapers), ephemera and correspondence within the written mode;
3. E3 (aims) - matters concerning the reason for making the text and the intended effect it is expected to have, including (1) the size of the audience (and subclasses for private and public speech) and (2) the communicative function of the text, i.e. discussion, information, recommendation, instruction or recreation.
4. I1 (topic) - the main topic of the text, following a shallow classification of knowledge domains similar to classes used in the BNC, e.g. natural sciences, applied sciences, life or politics;
5. I2 (style) - "the patterns of language that are thought to correlate with external parameters" (Sinclair, 1996), such as formal or informal, one-way or interactive, etc.

The changes in the finer classification of parameters in comparison to Sinclair (1996) are based on the experience in development of other representative corpora, such as the Brown Corpus, BNC, and the TEI guidelines (Sperberg-McQueen, Burnard, 2001), as well as considerations from Russian texts. This concerns, for example, the use of an additional mode (written-to-be-spoken), which is borrowed from the BNC (E2), the intended audience age (E3.1), a classification of fiction genres (E3.2) and styles (I2). It was considered helpful to extend the classification of text styles with separate subclasses for fiction and non-fiction texts. The patterns of language detected for fiction include the following styles (some better known writers that often use the style are also indicated):

1. neutral, — the style characteristic for standard literary texts in Russian,

2. regional, *derevenskaja proza* — an imitation of regional, mostly rural, language varieties, e.g. Astafiev, Rasputin,
3. lowly, *snizhennyj* — an imitation of the spoken language used by a "lesser educated" population, often slang, e.g. Ju. Aleshkovskij, Limonov,
4. official, *socrealism* — the official style of the Soviet literature, e.g. Dangulov, Markov,
5. individual, — a marked way of language use with significant deviations from the neutral style, this style is typically the result of linguistic or stylistic experiments, e.g. S. Sokolov.

Each style in the list instantiates a specific set of implications on lexicogrammatical properties (with the exception of the individual style, which is often author-specific, but this is exactly the reason to classify a text in this way). Nonfiction is classified according to the following styles: neutral, formal, informal, and academic writing.

Since the project is aimed at a representative sample of modern Russian, all meaningful combinations of parameters should be represented in the corpus by at least a handful of texts, though the number of texts in each group depends on the estimated number of respective texts in the Russian discourse and the availability of their electronic copies. The text length is another important technical parameter. It is easier to develop a large corpus using longer texts. However, this means that the corpus contains fewer texts, so an idiosyncratic use of language in each text significantly influences lexicogrammatical properties that can be described using the corpus. This is the reason for the balance of texts of various sizes in the two corpora, i.e. both shorter and longer texts should be included in each category with a greater number of shorter texts to alleviate the influence of longer ones.

The intended coverage of knowledge domains (I1) roughly follows the proportion used in the BNC. The comparison is shown in Table 2 (the data are from the BNC Index by David Lee). Since the typology of texts in the BNC is based on other principles, the comparison presents the content of texts in BOKR, as if they were described in terms used by the BNC. For instance, spoken language is treated as a domain in the BNC, so the figures in Table 2 also include it, even though a spoken discourse can be devoted to any other topic in the list of domains, so it is described as the mode of speech in BOKR (E2).

It would be desirable to increase the proportion of spoken language in BOKR at least to the coverage of the BNC, if not to 50% of the total corpus, but the small amount of available transcribed recordings make the ideal target impractical. The major departure from the BNC is the already discussed higher proportion of fiction texts, which are not considered in our scheme as a knowledge domain of its own (similar to the spoken domain), but as the most important component of the knowledge domain "Life" (cf. respective sections in newspapers, which in the Russian context often include short fiction stories). Note that the corpus is currently under development, so the figures in the third column in Table 2 are approximations for the expected coverage.

Table 2. The proportion of knowledge domains

Domains as in the BNC	**BNC**	**BOKR**
Spoken (not a domain in BOKR)	10,7 %	5 %
Imaginative Texts (Life in BOKR)	16,7 %	30 %
Natural Sciences	3,8 %	5 %
Applied Sciences	7,2 %	10 %
Social Sciences	14,2 %	12 %
World Affairs (Politics in BOKR)	18,9 %	15 %
Commerce	7,6 %	5 %
Arts	6,8 %	5 %
Belief/Thought (Religion and philosophy in BOKR)	3,1 %	3 %
Leisure	11,2 %	10 %

Currently tools and techniques for working with BOKR and the Russian Standard are tested using a corpus of 40 mln words. Its subcorpus of about 1 mln words of fiction texts (corresponding to the Russian Standard) has POS annotations that have been automatically assigned and manually inspected. It is also used for correcting the POS tagger used for processing the larger corpus. It is expected that the final release of the corpus will be available by the end of 2004.

2.2 The methodology for achieving the proportional coverage

The costs of compiling a representative corpus now are smaller than 10 years ago, when the BNC was collected. Many types of source texts are readily available in electronic form, in particular, fiction and news texts are widely accessible via the Internet and can be legally available for the corpus. Other types of the discourse, like business or private correspondence, are harder to obtain and deposit in a corpus because of legal obstacles. Yet other types of sources, like samples of spontaneous speech, are rare for technical reasons. The proposed solution is to increase the amount of ephemera (including leaflets, junk mail and typed material), correspondence (business and private) and spoken language samples whenever possible, because they reflect everyday language produced and reproduced regularly in the discourse. Anyway, various types of published texts will take the rest of the share. In this respect, the situation is similar to the early time of the BNC: the amount of texts from unpublished sources in the written part of the BNC is about 4.5%. It is unlikely that in BOKR we will have significantly more: even though the majority of source texts are available in electronic form now, their holders are unwilling to share them.

For the reasons of protection of privacy, personal and business letters are subjected to an anonymisation procedure with respect to names of persons and companies. Person names are replaced with MX, FX or CX tags (for male, female or child participants respectively) and names of companies with CoX (X is the identification number of a participant in the text; the same practice is also used in the Bank of English). In some cases, text providers manually replace names with codes. In other cases, they provide original texts, but when texts are stored in the corpus, names are replaced automatically using the lists of known given names and surnames of persons and names of companies. Care has been taken so that names of prominent figures and characters from popular books and films have not been replaced, for instance, even though *Karamazoff* and *Putin* are valid Russian names, it is much more likely they are *not* participants in the exchange, so their names are left as they are in texts (given that the corpus lacks private letters from or to prominent figures).

We understand that the anonymisation procedure is not completely satisfactory, cf. analysis in (Rock, 2001). First, it does not lead to complete anonymity: contextual clues are left in texts and allow the detection of participants. Second, a text in which names are replaced with codes looks less natural. Third, errors in setting identification numbers of participants are possible; they can lead to problems in discourse studies based on such texts. Finally, the anonymisation removes the possibility to study the frequency of personal names, discourse patterns of their uses, as well as phonological patterns. However, according to our views, this is the best possible practice for storing private and business letters without violating the privacy of their authors and addressees.

The text description framework is much more elaborate than a list of domains, so the balance of texts should be achieved on the basis of the text typology described above. The typology can be represented in terms of the systemic network of interrelated choices (Martin, 1987). For instance, when a text is described as fiction, it can be described in terms of the style of fiction, such as, stylistically neutral, low or regional, and in terms of the genre of fiction, such as general, historical, science fiction, etc, but not in terms of the interaction between the author and the audience, because it is not produced spontaneously in the presence of the audience.

The network of options is traversed using the Systemic Coder. A person that encodes metatextual information about a document has access to its record, including the author, the title, the year of creation, the location of the file and its size in words. Encoding options are selected from a list of categories, for instance, the age of the intended audience is selected from adult, child, teen or x-age, and the age of the author at the time of text creation is selected from child, teen, young, mid, senior (mid-aged authors is the broadest category that covers ages from 22 to 55). Even though the typology is elaborate, experience shows that most texts can be described in few seconds.

Some combinations of features are logically impossible, for instance, a personal letter aimed at a very large audience or a private discussion on TV. Some other combinations are very unlikely, for instance, books written in the domain of

natural sciences in formal style, aimed at a very large female audience for entertainment: the combination of formal style and entertainment or natural sciences and a sex-targeted audience is unlikely. However, if a combination of parameters is meaningful, an effort should be taken to cover it in the corpus by, at least, several texts. As an example, we can consider the set of texts within the knowledge domain "Politics", subdomain "home affairs" (the parameter I1 in our typology). Variation over other parameters involves selection of texts written in neutral, formal, academic and informal styles (I2), texts created by male or female authors or texts with corporate authorship, texts written within the period of 1990-2000 in different regions of Russia (E1), texts printed in newspapers, magazines, or books, as well as letters and reports, or spoken discussions, on site, on TV and radio (E2), texts aimed at different audiences (general vs. informed vs. professional, public vs. private, etc), and aimed at various communicative functions, e.g. discussion, recommendation, instruction or entertainment (E3). Each text in the corpus should be described by this set of options, for instance, the Russian Constitution is a text written in formal style (I2) in 1993 in Moscow, the authorship is corporate (E1), it is a written material printed as a book of 9500 words (E2), aimed at a very large audience (even if it is rarely read by the majority of population), with the intended function of recommendation, as a legal document.

The typology ensures that every text to be included in BOKR can be described in terms of the parameters listed above. Since texts aimed at more public audiences are easier to obtain, extra efforts are taken to cover texts aimed at more private audiences. The text collection activity could lead to a corpus significantly larger than 100 mln words. The next step is to balance the collection. The balance takes into account the proportion of basic genres according to Table 2, as well as the proportion of texts within each parameter. For instance, the classes of the intended outcome of a text include discussion, recommendation, instruction, recreation and information. The exact proportion of intended outcomes in the corpus can hardly be determined, e.g. the allocation of 25% for discussions or 10% for instructions looks fairly arbitrary. However, if texts classified as a recommendation take 90% of the total corpus, this is a clear sign of the disproportional coverage, which should be corrected. The balance will be monitored using statistical tools available in the Systemic Coder, such as Cell Analysis or Significance Tests.

Another problem with sources concerns the choice of diachronic sampling, because the turbulent history of Russia in the 20th century radically affected the language. For instance, according to the frequency list (Zasorina, 1977), which was compiled on the basis of texts from 1930-1960, such words as *sovetskij* (Soviet) and *tovarishch* (comrade) belonged to the first hundred of Russian words on a par with function words, but this is no longer valid in modern texts. The language of fiction has not been so radically affected. The decisions on the chronological limits of the study are different for different text types, for instance, fiction texts are taken starting from 1960, scientific texts from 1980, political

texts and ephemera from 1990 (earlier ephemera texts are also hard to obtain), while news texts from 1997.

3. The principles of morphosyntactic annotation

English corpora, including the BNC, are annotated with complex tags, like NNS for plural common nouns, a technique, which is impossible in a highly inflective language, such as Russian. For instance, an adjective inflects for case, gender and number, giving 36 basic adjectival categories in total, while a verb in addition to its own 14 basic categories has up to 4 participles, each of which declines for adjectival categories. This leads to thousands of separate tags that cannot be effectively searched. For this reason, each word is annotated with a list of features, which names provide unique identification of their type according to the morphosyntactic codes from EAGLES (Calzolari, McNaught, 1996), for instance, *bylo* (was) is annotated as "verb,ifve,int,act: n,sg,past", which stands for verb, imperfective, intransitive, active voice (the features describe the lexical item *byt'*), followed after a colon with features that describe the form: neutral gender, singular number, past tense. Separate features from a feature bundle associated with each word can be selected in a window of the query interface (Figure 1).

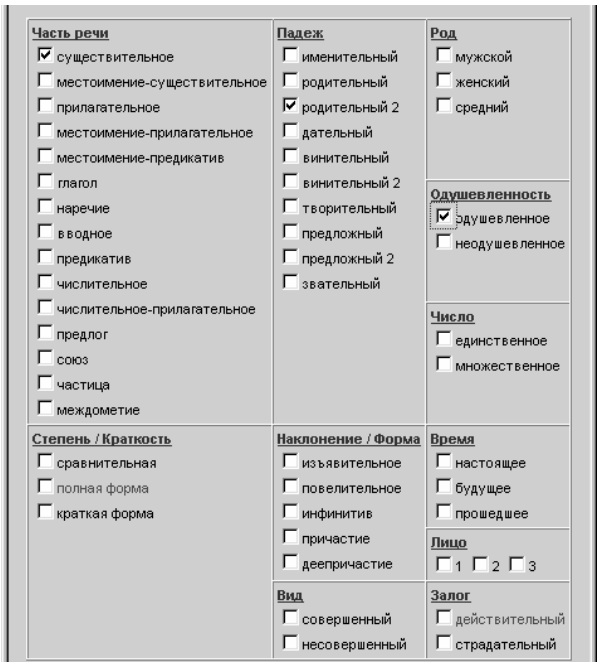

Figure 1. The query interface

3.1 How to resolve the ambiguity

The POS class and the morphological properties of a word are reflected in the flexion and the probability of deciding the lemma and the POS class from a word form is higher than, for instance, in English. As the result, there are many morphological analyzers for Russian, which make decisions on the basis of word forms, but virtually no Russian taggers which take into account the local context.

However, if real Russian texts are to be tagged with the intention of making a corpus that includes lemmatisation and morphosyntactic annotation, the level of ambiguity is high: very frequent word forms like *stali*, *shli*, or *ego* can correspond to several lemmas, i.e. *stali* – *stat'* (verb, to become) vs. *stal'* (noun, steel), *shli* – *idti* (to go) vs. *slat'* (to send), *ego* – *ego* (possessive) vs. *on* vs. *ono* (both are personal pronouns).

The ambiguity between POS classes is relatively rare, but many noun forms have multiple readings, for instance, the word form *pole* is an instance of four different nouns *pol* (floor), *pole* (field) and *pola* (lap) and *Polya* (a person name, when it is capitalised). Also, in many cases the ambiguity concerns the set of morphological features of the same lemma, e.g. *knigi* is the singular form in the genitive case or the plural form in the nominative or accusative case, while *znakomoj* is the singular form in the genitive, dative or prepositional case of either a noun or an adjective. According to initial experiments, the frequency of ambiguous detection of lemmas and POS classes in running text is about 25%, while the frequency of ambiguous morphological properties is about 55%. The values are too high for a corpus with morphosyntactic annotations and the ambiguity should be reduced.

Currently there are no tools available that allow reliable parsing in a corpus of this size for Russian (or any other language). Use of language-independent POS taggers based on statistical models cannot improve the quality of the output, because they are typically based on considerations of the word order, which is flexible in Russian, and because the genuine ambiguity between POS classes is relatively rare, and the most frequent type of the ambiguity concerns different readings of a word form (cf. the example with *pole*) and morphological features (cf. the example with *knigi*). If word forms or sets of morphological features (e.g. plural+dative+feminine) are treated as POS classes, then their number increases and the quality of language-independent POS tagging declines. However, partial parsing that detects nominal and prepositional phrases is reliable enough and can be used for deciding the reading of ambiguous forms. In BOKR and Russian Standard we use Dialing, a morphological analyzer with simple mechanisms for syntactic and semantic analysis. Since two analyses of the same word form have distinct morphological properties (the case, number and gender), the agreement between the noun and the adjective in noun phrases removes some types of ambiguity, e.g. the word combination *znakomoj knigi* from *Otkroem stranitsy etoj horosho znakomoj knigi* (Let's open pages of this well-known book) can be parsed only as the genitive singular form of both words, and the first word is an adjective. Another simple mechanism that requires only partial parsing is the

agreement between the subject and the predicate, it can remove, for instance, the plural reading of a noun in the nominative case (*knigi*), when the predicate is singular, as in *Knigi na polke ne bylo*.

Some types of the ambiguity of lemmas and POS classes are left after partial parsing: 12% of forms are ambiguous with respect to lemmas, 22% with respect to morphological properties. Since the BOKR corpus is annotated without human intervention, ambiguous analyses are subjected to further filtering according to statistical heuristics. For instance, two nouns *spina* (back) and *spin* (spin) have several identical word forms, which cannot be separated by means of parsing. However, *spin* is a term in theoretical physics, so the reading can be ignored in normal texts. Few other cases resist even complete parsing and statistical considerations, for instance, the ambiguity between the two readings of the word form *banke* (*bank* vs. *banka*) in *Xranite svoi denjgi v banke* (keep your money in a bank/in a jar) can be resolved only on the basis of semantic and pragmatic constraints. Such cases of ambiguity are retained in BOKR. The same applies to the ambiguity in morphological properties left after the syntactic filter. Currently, 3.6% of forms remain with ambiguous lemmas. The ambiguity in the Russian Standard is corrected manually.

3.2 How to store annotations

The design of the annotation format of the two corpora follows the best practices in corpus development established in 1990s, namely EAGLES (European Advisory Group on Language Engineering Standards) and TEI (Text Encoding for Interchange). Even though XCES (XML Corpus Encoding Standard) is expected to become the international standard for language resources (Ide, Romary, 2002), the annotation scheme based on it is extremely verbose (the size of an annotated file is a hundred times larger than a plain text file and, for a corpus of 100 mln words, the size really matters). The XCES scheme is also not suited for querying word uses in the corpus, because information on similar properties is represented at different levels of the XML structure. Thus, we have postponed the use of XCES until the standard is established and there are publicly available software tools for working with the format.

The format of BOKR is based on the TEI scheme and uses standard tags, like <phr>, <s>, <w> for representing phrases, sentences and words. Morphological annotation is stored in <ana> tags that describe word properties in lemma and feats attributes. Ambiguity is represented using multiple <ana> tags. The following is an example from the beginning of a sentence:

Mne bylo ochen' zhalko svoih chasov, …

(I was very sorry about loosing my watch, …)

```
<s id="kozlotur.1476">
  <w n="1">Мне<ana lemma="я" feats="pron,sg,1: dat"/></w>
  <w n="2">было<ana lemma="быть" feats="verb,ifve,int,act: n,sg,past "/></w>
 <phr type="ADV+ADV"> <w n="3">очень<ana lemma="очень" feats="adv"/></w>
  <w n="4">жалко<ana lemma="жалко" feats="adv"/></w>
 </phr>
```

```
<phr type="ADJ+NOUN">
  <w n="5">своих<ana lemma="свой" feats="pron,poss: pl,gen"/></w>
  <w n="6">часов<ana lemma="час" feats="noun,m,in: pl,gen"/>
                <ana lemma="часы" feats="noun,pl: gen"/></w>
</phr>
</s>
```

The parser was able to resolve the ambiguity between two analyses of *mne* (the dative or prepositional case), *svoikh* (the genitive, accusative or prepositional case of either a personal pronoun or a possessive pronoun), and *zhalko* (an adverb or an adjective). However, the ambiguity between two readings of the word form <w n="6"> is left in the output in the two <ana> tags. It can be read as *chas* (an hour) and *chasy* (a watch), in the latter case it is *pluralia tantum*, which is reflected in the position of the number value before the colon.

Metainformation about a document is stored in the header and is also based on the TEI format. In some cases, TEI provides tags for encoding the contextual settings used in the text typology, for instance, <creation> or <textClass>. In other cases, the information on the expected outcome or the size of the audience is expressed using the general framework of taxonomy specifications by means of <catRef> (category reference) tags.

4. References

Andryuschenko, V.M. (1989). *Konzepziya i arhitectura Mashinnogo fonda russkogo jazyka* (The concept and design of the Computer Fund of Russian Language), Moskva: Nauka, 1989

Calzolari, N., McNaught, J. (eds.) (1996). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. EAGLES document EAG-CLWG-MORPHSYN/R http://www.ilc.cnr.it/EAGLES96/morphsyn/morphsyn.html

Ide, N., Romary, L. (2002). Standards for language resources. In *Proc. of Language Resources and Evaluation Conference (LREC02).* May, 2002, Las Palmas, Spain. 59-65.

Leech, G. (1997). A brief users' guide to the grammatical tagging of the British National Corpus, UCREL, Lancaster University. http://www.hcu.ox.ac.uk/BNC/what/gramtag.html

Lönngren, Lennart (ed.) (1993). *Chastotnyj slovar' sovremennogo russkogo jazyka.* (A Frequency Dictionary of Modern Russian. With a Summary in English.) Acta Universitatis Upsaliensis, Studia Slavica Upsaliensia 32. 188 pp. Uppsala.

Martin, J.R. (1987). The meaning of features in systemic linguistics. In M.A.K. Halliday, R.P. Fawcett (eds.) *New Developments in Systemic Linguistics.* Vol. 1. London: Pinter Publishers. 14-40.

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics,* 6(1).

Sinclair, J. (1996). Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P. http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html

Sperberg-McQueen, C. M., Burnard, L. (eds.) (2001). *Guidelines for Electronic Text Encoding and Interchange.* http://www.hcu.ox.ac.uk/TEI/P4X/index.html

Verbitskaya, L.A., Kazanskij, N.N., Kassevich, V.B., (forthcoming). Nekotorye problemy sozdanija natsional'nogo korpusa russkogo jazyka. *NTI, Series 2*. (in Russian)

Zasorina, L.N. (ed.) (1977). *Chastotnyj slovar' russkogo jazyka.* Moscow: Russkij Jazyk.

Internet links

BNC Index: http://www.comp.lancs.ac.uk/ucrel/bncindex/

BOKR: Boljshoj Korpus Russkogo yazyka (the Russian Reference Corpus, a description of the project), http://bokrcorpora.narod.ru/

CFRL: the Computer Fund of Russian Language, http://irlras-cfrl.rema.ru/

Coder, a markup and classification tool: http://www.wagsoft.com/Coder/

Dialing, the morphological analyser: http://www.aot.ru/download.htm

Moshkow's Library: http://lib.ru/

RS: the Russian Standard (online access), http://corpora.yandex.ru/

UC: the Uppsala Corpus, available from the University of Tübingen, http://www.sfb441.uni-tuebingen.de/b1/en/korpora.html

Yandex, the search engine: http://www.yandex.ru/

Notes

* The research presented in the paper has been supported by the Alexander von Humboldt Foundation, Germany, when the author was affiliated with the University of Bielefeld and the Russian Research Institute for Artificial Intelligence. I am grateful to my Russian colleagues, in particular, to Vladimir Plungian and Katia Rakhilina, who took the leadership in the ongoing development of the Russian Reference Corpus.

[1] Development of a translation corpus is considered to be a separate task.