Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics

Serge Sharoff *

Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld,
Postfach 10 01 31, D-33501 Bielefeld, Germany,
serge.sharoff@uni-bielefeld.de, s_sharoff@yahoo.com

Abstract

The paper discusses the use of corpora for experimental studies in contrastive lexical semantics, in particular, for comparing how a state of affairs is expressed in different languages and by different translators. Three topics are addressed: (1) a lexicographic database, which is aimed at storing and maintaining contrastive descriptions of a class of lexical items in several languages; (2) an aligned parallel English-Russian corpus, including several literary texts and software manuals (the total size is about one million words), together with tools for querying the corpus by means of Perl-based regular expressions; and (3) an example of development of a lexicographical database of the most frequent size adjectives in English, German and Russian.

1. Introduction

If studies in lexical semantics are based on experimental evidence, they should study purposes, with which words are used in real texts, and conditions, under which they are used. If the same aim is required for studies in contrastive lexical semantics, we should start with comparison of uses of respective words in parallel texts. Dictionaries (mono- and multilingual) are a good repository of information on meanings that can be expressed by a word and on its possible translations. However, dictionaries cannot list all possible contexts, in which a word can be used and which can influence its translation into another language. Moreover, dictionaries are designed for helping a human reader (with some expertise) to understand or translate a word in a specific context, and not for providing linguistic analysis of lexical semantics or helping in machine translation or generation. For example, the English verb *leave* is quite polysemous, it is described by 17 senses in WordNet, 29 senses in the Collins-COBUILD dictionary, and 31 senses in the Oxford English Dictionary (the senses for the noun, phrasal verbs and idioms are not counted). When such huge repositories of senses are compared against real uses of words, some uses fit into several senses, while some fairly innocuous, i.e. non-metaphorical, uses violate necessary conditions for membership in any category defined by senses. Analogously, various English-Russian dictionaries list 9 to 16 Russian verbs that can be used for translating *leave*, yet it is easy to find several dozens of its contextually-dependent translations.

Here we adopt a communication-centered view, in which words are treated not as references to concepts, but as systematic hints that enable communication between the speaker and the hearer. This view can be defined as the "meaning is use" paradigm, widely accepted by various linguists and philosophers of language, e.g. (Wittgenstein, 1953), (Halliday, 1978), however, it is not supported by computational mechanisms for representing meanings as uses and by tools for facilitating the analysis of words in terms of meaning intentions corresponding to their uses.

In Section 2, the format of a lexicographic database is discussed. The database is aimed at storing and maintaining contrastive descriptions of behavior of a class of lexical items in several languages. The emphasis of the database design is on representing meanings intended by uses of words and on comparing uses across languages. Section 3 presents an aligned parallel corpus and tools used for querying uses in the corpus on the basis of regular expressions. An example of development of a description of uses of size adjectives in English, German and Russian is provided in Section 4.

2. The format of the lexical database

Unlike logic-based theories of meanings, which assume that a word denotes a concept (a word *has* a meaning as an entry in a dictionary), functional theories of meaning assume that the meaning of a word is the function of its purposeful use in communication. The computational model is based in terms of Halliday's systemic-functional linguistics (Halliday, Matthiessen, 1999). In this model, the lexicogrammar specifies the *potential* of meaning intensions for expressing speaker's goals. The goals are *realized* by various lexical and grammatical means available in language. Finally, the potential of the lexicogrammar is *instantiated* in the context of the exchange between the speaker and the hearer. In this theory the lexicogrammar is represented by a network of interrelated choices and realization statements following the choices. For instance, classification of the English mood starts with features 'indicative' vs. 'imperative'. Semantically, it corresponds to the opposition of speech acts referring to exchange of information vs. issuing commands. More delicate features in the network of mood are 'declarative' and 'interrogative', which are realized syntagmatically by the respective order of Subject and Finite attributes; more information on representing lexicogrammatical resources in these terms is in (Bateman, et al, 2000).

The same principles guide the design of the lexicographical database, which represents three types of information: the potential of communicative intentions, realization of meaning intentions by lexical items and instantiation of the potential, when lexical items are used in the discourse.

* The research was supported by the Alexander von Humboldt Foundation, Germany.

Standards for encoding large lexical resources are under quite active development now. The most established standard at the moment is given by TEI (Sperberg-McQueen, Burnard, 2001), which includes standard means for encoding printed mono- and multilingual dictionaries and terminological databases; the latter means are inherited in MARTIF (MAchine-Readable Terminology Interchange Format). In addition, current activity on the International Standards for Language Engineering (ISLE) includes, in particular, development of a standard for representing Multilingual ISLE Lexical Entry (MILE), (Calzolari, et al, 2001). The goal of Lexicograph, yet another lexicographical database, is to encode basic facts concerning syntactic and semantic behavior for a significant slice of the lexicon (Paducheva, 1997).

However, the aim of the presented research required the development of a specific format. The purpose of the TEI guidelines is to provide standard means for encoding of existing dictionaries, which are significantly different from means for encoding of lexicographical databases aimed at development of new descriptions of behavior of lexical items. Even though the aims of MILE are similar to aims of the database described here, no definite specification of the dictionary content of MILE is currently available. Also, existing descriptions of MILE are aimed at encoding lexical items with clearly defined semantics, like *lancet* (Calzolari, et al, 2001), but they do not address the issues of polysemy and language-specific ways for using respective lexical items. For instance, the structure of the lexical entry in MILE does not provide sufficient tools to encode and relate all the 22 senses of *high* as detected in Collins-COBUILD and 24 words used as its translation equivalents in Russian. TEI can easily encode the senses and translations, but cannot explicitly express their relationships and conditions in which they are used. The format of Lexicograph is based on tables in a relational database. This restricts both its structure (many facts about the behavior of lexical items do not fit into the strict relational model) and implementation (any change in data structures requires costly changes of the database relational model, anyway dBase used in Lexicograph is now archaic). Finally, the goal of the reported project was to encode not only lexical items, but also basic purposes for which they can be used, so that one can study senses and translation equivalence between words in different languages in context of their use. None format discussed above was designed for representing communicative intentions.

The database is represented in the XML format and inherits the TEI guidelines for encoding printed dictionaries, since TEI provides a well-established format, which is suitable for encoding any information available in dictionaries. Extensions can be easily introduced as modifications of respective DTDs (document type definitions) for XML files. Another advantage in using XML is that it allows to separate encoding of the structure, content and presentation of resources.

The extensions over the set of XML elements and attributes from the TEI guidelines include options required for representing communicative goals and for developing lexicographic descriptions. For this purpose, the database contains divisions: the first one ( ) consists of lexical items; it may either include all lexical items in all languages or be restricted to a particular language. This division considers the lexical resources "from below": from lexical items occurring in texts to purposes they are used for. The second division ( ) allows a view in the opposite direction: from communicative goals to their possible realization by lexical items.

2.1. The anatomy of lexical entries

A lexical item (<entry>) is composed of elements from the TEI guidelines, for instance, <gramGrp> (morphosyntactic properties), <sense>, <trans> (translations), <eg> (examples). <note> is used for storing unstructured comments on the behavior of lexical items. Possible values of attributes of some elements were extended, e.g. resp indicates a source of information (a dictionary or a researcher), types of examples include "imposs", "quest", corresponding to linguistic examples marked in publications with an asterisk or a double question mark. New elements were also introduced

- <fts> lists features from the network used for annotating a sense, example of use or translation;
- <collocateGrp> a list of collocations;
- <frequency>, which is measured according to ipm (instances per million), rank in the word list, band (COBUILD); for instance, for *slight*

```
<entry key="slight" lang="en">
  <frequency resp="cobuild" type="band" value="3"/>
  <frequency resp="bnc" type="rank" value="2271"/>
  <frequency resp="bnc" type="ipm" value="39"/>
  <collocateGrp resp="coubuild" type="t-score">
     <colloc value="5.32" type="magn">even</colloc>
     <colloc value="4.65" type="obj">doubt</colloc>
     <colloc value="4.64" type="obj">increase</colloc>
  </collocateGrp>
</entry>
```

The TEI guidelines suggest to use <usg> (notes with usage information) for coding various semantic cues like synonyms or collocations, because the function of such elements in printed dictionaries is similar to usage remarks, e.g. specifications of the domain, style or preference level. However, when a lexicographical database is developed, such cues constitute the backbone for establishing relationships between lexical items. Out of this reason, a new element class <semref> was introduced. A semantic reference has a type and a target, i.e. the identifier of another element in the database, this can be another lexical item or a feature in the network of communicative purposes. The following types of references are defined:

- syn synonym, ant antonym;
- hyper superordinate term, hypo subordinate term;
- para a word in the same paradigmatic classification: it is closely related to the headword, but differs sufficiently in its lexical behavior, so that it cannot be treated as a synonym, e.g. *high* and *tall*;
- mero meronym, a part-whole relation;
- trans a translation (in the current context);
- subj/comp a typical subject/complement (for verbs)
- obj a typical object (for adjectives)
- colloc a collocate of another type

Finally, there is a group of types of semantic references for lexical functions in the tradition of Meaning-Text Theory, as listed in (Mel'chuk, 1996). Lexical functions are functions in the mathematical sense, which map words to other words with respect to generalized goals of

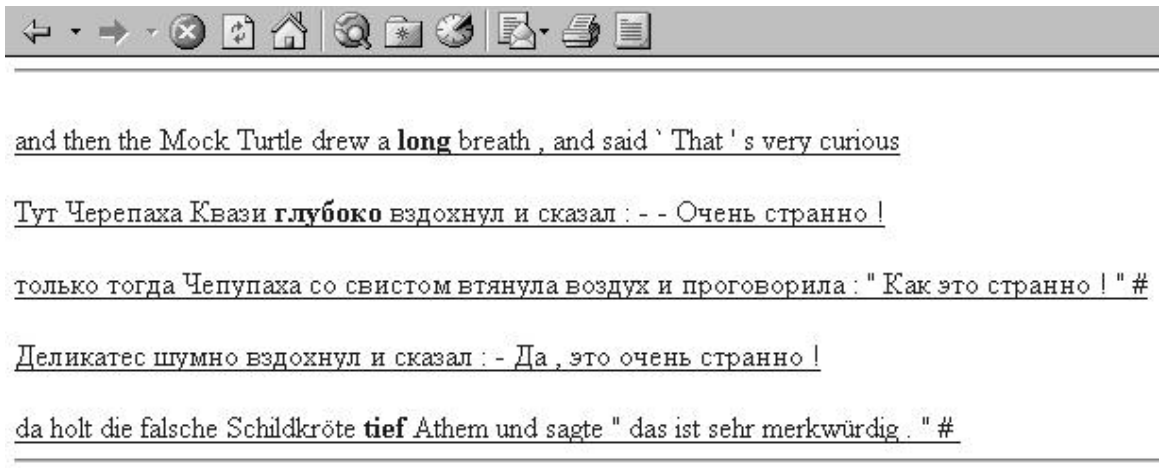

Figure 1. An output from processing the query: /<w id="(\S*)" lemma="long" pos="adj"/

the speaker. One example is bon, the standard praise for a concept expressed by a lexical item:

```
<entry key="advice" lang="en">
      <semref type="bon" target="sound"/>…</entry>
<entry key="analysis" lang="en">
      <semref type="bon" target="thorough"/>…</entry>
```

2.2. The anatomy of lexical choices

The second division ( ) consists of the following elements:

- <system> a paradigmatic class with a set of features;
- <chooser> semantic grounds for choosing a feature;
- <inquiry> interface to a procedure in the knowledge base, which makes a decision for a chooser.

<system> consists of a set of features for a class (<feature>) and their relationship with other features (<inputs>). A feature has a name and a set of realization statements constraining the lexicogrammatical structure. For instance, a system representing the indicative mood includes two features and an entry condition, which relates it to the entry mood system:

```
<system chooser="indicative-chooser">
      <inputs>indicative</inputs>
      <feature name="declarative">
          <rln>(order subject finite)</rln></feature>
      <feature name="interrogative">
          <rln>(order finite subject)</rln></feature>
</system>
```

<chooser> consists of an unstructured description, which relates the properties of concepts and objects to choosing features, and a formal definition (if any) expressed in LISP syntax. For instance, the simplified chooser for the indicative type:

```
<chooser name="indicative-chooser">
<note>possible communicative intentions for information
exchange are: demanding or providing information</note>
<def>(ask (information-exchange-q speech-act)
          (demanding (choose interrogative))
          (providing (choose declarative)))
</def></chooser>
```

Formal definitions of choosers and inquiries are not necessary for standalone databases, but it is helpful, when a database is used as a resource in an application, for instance, information retrieval, understanding or generation, cf. (Matthiessen, Bateman, 1991).

3. The aligned parallel corpus

Given multiple translation equivalents of most common words (such as examples with leave and high above suggest), research, which is aimed at contrastive study of, say, semantics of 'motion away from a place' or 'size of objects', requires access to a corpus of aligned parallel texts and the possibility to search for lexical items, their translations and corresponding contextual conditions. The availability of several translations of the same text in the corpus also allows for empirical analysis of paraphrastic possibilities.

Modern text alignment methods on the basis of cue words and character-length comparison (Gale, Church, 1993) are quite efficient for semi-automatic alignment of large volumes of text, yet a lot of manual work is required for ensuring the alignment quality. Out of this reason, in comparison to huge amount of monolingual corpora[1], relatively few aligned corpora are publicly available. This mostly concerns English-French, e.g. Hansard (Simard, Plamondon, 1998), English-German, e.g. (Schmied, Schäffler, 1996). None English-Russian aligned corpora were publicly available by 2000, so a corpus of aligned parallel texts was developed; its total size is about one million words (MW).

The corpus consists of several technical texts with descriptions of software, e.g. Microsoft Word Help, and literary texts, e.g. "Alice's Adventures in Wonderland" by Lewis Carroll. For the latter text, a German translation was also included (in addition to its three Russian translations). The texts have been aligned at the sentence level by means of Marc Alister (Paskaleva, Mihov, 1997). The corpus stores texts, sentences, words and alignments as XML elements; morphosyntactic and lexical-semantic properties of words are expressed as attributes of word elements (English, German and Russian texts were processed by respective POS-taggers). Cf. (Sharoff, 2001) for a more elaborate description of the corpus content and encoding format.

[1] They are available simply because of availability of electronic documents. Note the relative scarcity of spoken language corpora.

```
<entry key="deep" lang="en">
  <sense n="1" resp="cobuild"><def>If something is deep, it extends a long way down from the ground
                                or from the top surface of something.</def>
    <semref type="ant" target="shallow"/>
    <eg resp="alice"><q id="alice.11">she found herself falling down a very deep well. </q>
        <trans><tr id="alice-d.17" >она начала падать, словно в <semref type="trans">глубокий</semref>
            колодец.</tr>
            <tr id="alice-g.13">sie fiel, wie es schien , in einen <semref type="trans">tiefen</semref>
                Brunnen.</tr></trans></eg>
    <eg resp="cobuild"><q>Den had dug a deep hole in the centre of the garden.</q></eg>
    <fts>spatiotemporal big neutral-interpersonal vertical non-emphasized spatial-lex depth-size</fts>
  </sense>

  <sense n="11" resp="cobuild"><def>A deep sound is low in pitch.</def>
    <semref type="ant" target="high"/>
    <eg><q id="alice.1387">said the Mock Turtle in a deep, hollow tone. </q>
      <trans><q id="alice-d.1913">проговорил Квази <semref type="trans">глухим</semref> голосом</q>
          <q id="alice-n.1502">ответила Чепупаха <semref type="trans">глубоким</semref> голосом</q>
          <q id="alice-g.1224">sprach die falsche Schildkröte mit <semref type="trans">tiefer
            </semref>Stimme</q></trans></eg>
  </sense>
```

Figure 2. The structure of a lexical entry for deep

In addition, another corpus of about 20 MW has been developed: it is stored in the same format (without references to aligned sentences) and consists of modern Russian fiction. As no comprehensive Russian corpus with complex search facilities was available (comparable to the BNC in English or COSMAS in German), the present corpus served for in-depth corpus-based analysis of lexical semantics for Russian words.

An important feature of the parallel corpus is the possibility to consult uses and translation of words. The tools for querying the corpus are based on Perl regular expressions and allow to check co-occurrence of words or groups of words, specific morphological or lexical features of words. The result of query processing (Figure 1) is output as an HTML file, which is hyperlinked to sentence identifiers in the corpus, so that the wider context can be also explored. In addition to the source text, fragments aligned to the source can be also output. The keywords of both the source and the target texts are highlighted in the output. Translations are highlighted on the basis of a simple heuristics: translations typically belong to the same group as source words, thus, if we study uses of size adjectives, the list of Russian size adjectives is a good approximation for possible translations of size adjectives from English, though not always, as the two examples in Figure 1 suggest (of course, the list of candidates can be extended).

The mechanism of regular expressions in Perl is used for shallow parsing of the XML format of the concordance. The query language based on regular expressions is not always user-friendly, but it is powerful, so that it can extend the abilities of the encoding by shallow syntactic parsing, when the corpus lacks syntactic annotations. For instance, the most of uses of verbs of motion with a direct complement can be found by:

```
&lemma('run|come|go')&lemma('\w+','pos="(noun|adj|det)')
```

This means that the pattern matches, if one of the specified motion verbs is followed by an arbitrary noun, adjective or determiner (&lemma is a short-cut, which extends into a full-fledged pattern). The mechanism is also useful for detection of German verbs with separable prefixes, when the finite form of the verb is at the second position of the clause, while its prefix, on which the meaning crucially depends, is at the end:

Nach Angaben Seidlers nahm die Zahl der Arbeitslosen vor allem durch den weiteren Anstieg der Berufsanfänger und Aussiedler um knapp 15 000 Personen zu

(lit: according to Seidler's data the amount of unemployed rose primarily due to continuing increase of new applicants and emigrants by approximately 15000),

Taken as separate word forms, neither *nehmen* nor *zu* mean *to rise*. The pattern that catches the most of such uses is:[2]

```
&lemma('nehmen', 'feats=".*?finite').*?&lemma('zu')<punct
```

4. An example of database development: the case of size adjectives

The reported research established a methodology for describing lexical items in a specific domain (like size adjectives). It involves the following steps:

1. compilation of a list of lexical items in the domain, using existing frequency lists and/or corpora;
2. detecting basic choices for expressing events and their properties in the domain by lexical items from the list;
3. checking how the basic choices cover the existing descriptions in available sources, i.e. mono- and multilingual dictionaries and lexicographical descriptions, and detecting most important contexts of their use;
4. extending the network of choices to cover all possible uses of the lexical items under consideration.

The frequency lists of size adjectives for English, German and Russian were taken from, respectively, the BNC word lists (Kilgariff, 1996), (COSMAS, 2000), and (Zasorina, 1977). The final list consists of 66 words (23 for English, 21 for German, and 22 for Russian). Their basic uses can be quite simply arranged according to di-

[2] If *zu* occurs before a punctuation mark and a finite form of *nehmen* precedes it, most probably, it is a stray prefix.

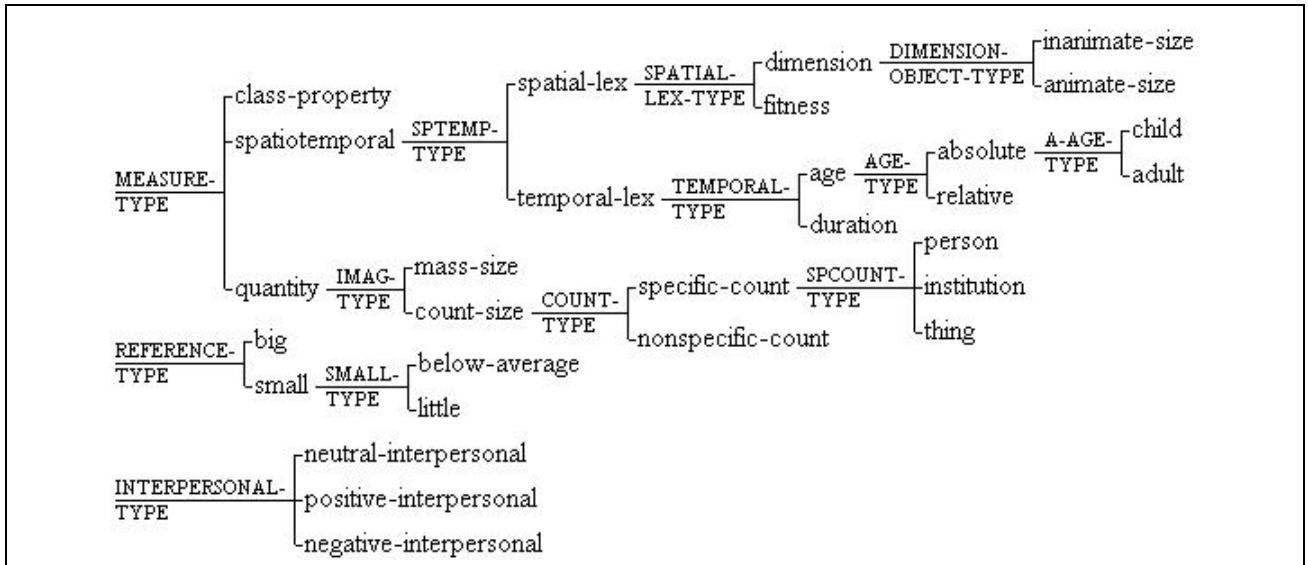

Figure 3. The basic options for size adjectives

mensions (*large, high, wide, thick*) and size (*tiny, small, average, big, huge*).

Even though the scheme is based on the intuitive understanding of what size adjectives are, it fails to take into account the complete range of possible uses of size adjectives and does not make a clear picture of the communicative purposes they are used for. As it is often the case with frequent words, the size adjectives in the list have a large number of polysemous senses in dictionaries. For instance, great has 12 senses in COBUILD and 9 in WordNet; high has respectively 22 and 12 senses. In total, the lexical division of the database lists 365 senses of 66 size adjectives.

The next step consisted in further development of the lexical division by introducing semantic references between the senses (synonyms, antonyms, collocates, etc), and examples of uses and their translations from corpora. The resulted structure can be searched with respect to elements, attributes and their values, for instance, looking for all synonyms referring to brief or all possible translations of *deep* as detected in the German and Russian divisions; this is done by means of options available in the XML mode of Emacs (for simple searches) or in an XML query language, e.g. XQL, XPath, XQuery.

One example of data exploration: searches in the database for typical translations of English size adjectives into German and Russian confirm that the degree of one-to-one correspondence between words is relatively high, i.e. size adjectives, like *large, high, long, wide,* regularly correspond to *groß, hoch, lang, breit* and *boljshoj, vysokij, dlinnyj, shirokij,* respectively. This also concerns many non-spatial senses, like *high quality, hohe Qualität, vysokoe kachestvo, long list, lange Liste, dlinnyj spisok*. However, the cases of mismatches warrant for language-specific options in systems and choosers. For instance, several English word pairs, like *large* vs. *big*, *high* vs. *tall* or multiple words for referring to sizes below average in German and Russian, e.g. *klein, wenig, gering, mäßig, nebolshoj, malenjkij, malyj, melkij,* require more delicate features in the respective networks, while some translation options require language-specific choosers for features available in all languages: *deep delight* is rendered in German and Russian in terms of non-directional properties: *große Freude, bolshoe naslazhdenie*; compare *long breath* to *tiefer Atem* and *gluboko vzdohnul* from Figure 1.

A simplified lexical entry for *deep* is given in Figure 2. Since all relationships between lexical items are explicitly encoded, it is easy to collect all synonyms or translation equivalents of *deep* and check contexts of their uses.

Thus, the elaborated lexical division of the database helps in development of the network division, which encodes communicative intentions and relates them to specific lexical items in the three languages. The main part of the network of size adjectives is presented in Figure 3. It distinguishes three basic classes of uses: spatiotemporal (*a large room*), quantity (*a large amount of cash and jewelry*), and class property (the size of an object is not measured, but it is classified according to specific criteria, compare *a big coward* to *a big country*). Among other options missed in the original intuitive model, there is a difference in lexical means for referring to the dimension proper (this is the most frequent choice) and to the degree of fitness, when an object (primarily the human body) fits into the space in another object (primarily clothing). This often leads to specific word choices, e.g. *weit* vs. *knapp* in German, *velik* vs. *mal* in Russian, which correspond to *loose* vs. *tight* in English (less frequent English size adjectives are *spacious, cramped*). Finally, there may be a lexical difference in referring to the size of animate vs. inanimate objects (*a tall man* vs. *a high house*). Similar important differences in the lexical options are possible for the temporal dimension and for the small-average domain.

Another option that is missed in the intuitive model concerns the interpersonal attitude. The reference to the size of an object/person can be used to justify the need in taking care (*small*, sympathy) or being cautious (*big*, antipathy). On the other hand, importance is typically described in the opposite way (*big* - important, *small* - unimportant). The interpersonal attitude influences the pattern of uses of roughly synonymous lexical items. *Little* correlate much stronger with the positive attitude, while *small* is typically used in less favorable contexts (the same pattern is in Russian: *malenkij* vs. *nebolshoj*).

5. Conclusions

The research reported in the paper led to several important contributions. The first contribution is a parallel aligned corpus for the English-Russian (partly, German) language combination (about 1 MW), as well as a corpus of modern Russian fiction (more than 20 MW). The corpora are stored in an XML-based format and are furnished with query tools based on regular expressions. The aligned corpus and the tools are available from http://purl.org/net/concordance or from TELRI.

The second contribution is the format of a lexicographical database, which is aimed at elaborate description of ways of using significant slices of lexicon in several languages. The XML-based storage format of the database encodes communicative intentions and lexical items, which can be used to realize intentions in the discourse. The database also helps in contrastive study of languages, because it adds the possibility to compare uses across languages.

The third contribution concerns several slices of lexis stored in the database. The database concerning the size adjectives in English, German and Russian is completed (it comprises 66 size adjectives with 365 analyzed senses and 52 features in the network of communicative intentions); the databases for verbs of motion (about 200 lexical items) and words used for expressing emotions (about 750 lexical items) are under construction. The database can be used as a resource for studies in contrastive semantics or machine translation, as it encodes multiple relationship between words and contexts of their use in the three languages. The network of size adjective also allows an extension to other languages, this involves changes in the lexical part, but most probably no significant alterations in the network of communicative intentions.

6. References

Bateman, J., E. Teich I. Kruijff-Korbayova, G.-J. Kruijff, S. Sharoff, H. Skoumalova, 2000. Resources for multilingual text generation in three Slavic languages. In *Proc. Language Resources and Evaluation Conference, LREC'2000,* Athens. 1763-1768.

Calzolari, N., A. Zampolli, A. Lenci, 2001. Towards a Standard for a Multilingual Lexical Entry: the EAGLES/ISLE Initiative. In *Proc. CICLING'2001,* Mexico.

COSMAS 2000. http://corpora.ids-mannheim.de/~cosmas/

Gale, W., K. Church, 1993. A Program of Aligning Sentences in Bilingual Corpora. *Computational Linguistics,* 19:75-102.

Halliday, M.A.K., 1978. *Language as social semiotic. The social interpretation of language and meaning.* London: Edward Arnold.

Halliday, M.A.K., C.M.I.M. Matthiessen 1999. *Construing experience through meaning: a language-based approach to cognition.* London: Cassell.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10: 135-155. http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

Matthiessen, C.M.I.M., J.A. Bateman 1991. *Text generation and systemic functional linguistics: experiences from English and Japanese.* London: Pinter Publishers.

Mel'chuk, I., 1996. Lexical Functions: a tool for the description of lexical relations in a lexicon. In L. Wanner (ed.) *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins: Amsterdam. 37-102.

Paducheva, E.V. 1997. Verb categorization and the format of a lexicographic definition. In L.Wanner (ed.) *Recent Trends in Meaning-Text Theory*. Amsterdam: John Benjamins.

Paskaleva, E., S. Mihov 1997. Second Language Acquisition from Aligned Corpora. In *Proc. of the International Conference "Language Technology and Language Teaching"*, Groningen, April 1997.

Schmied, J., H. Schäffler H., 1996. Approaching Translationese through Parallel and Translation Corpora. In C.E.Percy, C.F.Meyer, I.Lancaster (eds), *Synchronic Corpus Linguistics*. Amsterdam: Rodopi.

Sharoff, S. 2001. Through the looking glass of parallel texts. In *Proc. of the Corpus Linguistics Conference.* Lancaster, March-April, 2001. 543-552.

Simard, M., P. Plamondon 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13:59-80.

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

Wittgenstein, L. 1953. *The Philosophical Investigations.* Oxford: Blackwell.

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