Through the looking glass of parallel texts

Serge Sharoff (sharoff@aha.ru)
Russian Research Institute for Artificial Intelligence /
Alexander von Humboldt Fellow at the University of Bielefeld, LiLi Faculty
Postfach 10 01 31, D-33501 Bielefeld, Germany

1. Introduction

The project reported in the paper was aimed at the comparative analysis of meanings of linguistic expressions taken in their context, in particular, how meanings intended by the speaker/writer are realized via lexical items. The goal of the project was to assess the influence of context, on the one hand, and different languages, on the other hand onto meanings delivered by the same lexical item. Since no access to enumeration of original meanings is possible, the method for analysis was based on the investigation of corpora. Several types of lexical items (verbs of motion, names of plants and animals and size adjectives) were chosen from the English or Russian corpora. Then their translations into respectively Russian or English were checked.

The corpora used in the project include two technical texts:

- Microsoft Word'97 User Manual and its translation into Russian;
- excerpts from the AutoCAD v.13 User's Manual (Chapter 2: Drawing objects) and its translation into Russian[1];

and three literary texts:

- Lewis Carroll's *Alice's Adventures in the Wonderland* and three of its translations into Russian by Demurova, Nabokov and Zakhoder;
- Vladimir Nabokov's *The Vane Sisters* and two of its Russian translations made by Ilyin and Barabtarlo;
- Vladimir Nabokov's *Lolita* and its Russian translation made by Nabokov himself.

All the texts were originally written in English and translated into Russian. The corpora were aligned at the sentence level in order to compare expressions of the same state of affairs both between languages and between translators. The corpus (it comprises about 250 thousand words in total) is relatively small in comparison to parallel corpora developed in other projects, for example, the Chemnitz English-German/German-English corpus (Schmied, Schäffler, 1996). However, its size allowed to check the most occurrences of lexical items under investigation manually within the reasonable amount of time (about 3200 instances of verbs of motion and 1100 size adjectives were consulted in total).

The paper starts with the description of the XML representation for multilingual concordances and the Perl-based tools developed for creating, maintaining and consulting them. Then, Section 3 discusses results of the annotation using lexical semantic features taken from the dictionaries. Finally, Section 4 presents a network of oppositions, which are important for modeling lexical semantics of verbs of motion and size adjectives in English and Russian. Here, lexical semantics is represented by means of the systemic network, which is used in multilingual generation applications within the KPML environment (Bateman, *et al,* 1999).

2. Software for working parallel concordance

A Perl-based software has been developed within the project to create and maintain parallel concordances, which are represented in XML. The concordance structure encodes:

1. the general information about each text;
2. the alignment between sentences in parallel texts;
3. morphosyntactic and lexical-semantic properties of words of each text.

A concordance consists of three types of entities: a concordance text, sentence and a word. The general description of a document to be concordanced includes the following attributes:

1. docid—the identifier of the document (typically its file name);
2. lang—the language, in which the document is written (actually, the lang attribute sets the *default* value for intepretation of all words in the concordance. The default value may be

[1] The texts were prepared within AGILE, the EU project aimed at multilingual generation of CAD/CAM manuals, cf. (Bateman, *et al,* 2000).

redefined for a separate sentence or a word, for example, when a foreign language citation is found);

3. author—the author of the document.

The sentence description includes the following attributes:

1. sentenceid—the identifier of a sentence (it is composed of the document identifier plus the sentence index within the document);
2. correspondencelist—the list of identifiers of respective sentences in other parallel texts (if a sentence corresponds to several sentences in one text, this is also reflected in the list).

The word description includes the following attributes:

1. wordid—the identifier of a word instance (it is composed of the sentence identifier plus the word index within the sentence);
2. lemma;
3. POS—the part of speech, e.g. POS="verb";
4. morphfeatures—morphological features of the word instance, e.g. morphfeatures="3pers, sing";
5. lexfeatures—lexical features of the word instance, e.g. lexfeatures="wordnet5, motion, physical, away, source-fg".

The concordance structure is quite flexible for describing parallel texts and is suitable to be the basis for research in contrastive semantics. Also, it allows to use several languages and several translations of the same text, in particular, the German translation of Alice is planned to be added to the concordance. The structure mostly conforms to the EAGLES guidelines (EAGLES, 1996), in particular, the corpus annotation is separated from the corpus itself. The most important difference from the EAGLES scheme concerns keeping the parts of speech (major categories in EAGLES terms) in one attribute (POS) and other values (like person and number) in another one (morphfeatures). This simplification is suitable for the current purposes of the concordance development, i.e. consultation of uses of lexical items. For instance, a concordance query, which consults uses of the verb *leave* in the corpus, can easily discard the noun *leaves* and the adjective *left*, which are irrelevant for the query, by means of specifying the following condition POS="verb" in the query.

As the result of XML conversion, the concordance is significantly larger than the original text (approximately by the factor of ten), however, the XML representation of attributes keeps the size of corpora smaller than the position-based encoding, which is used, for example, in the Susanne corpus (Sampson, 1995). Alignment of the text pairs has been performed by means of the Marc Alister software package (Paskaleva, Mihov, 1997), which implements the Gale-Church algorithm for language-independent alignment (Gale, Church, 1993). Since the language-independent alignment algorithm frequently fails to detect translation equivalents, the alignment results should be corrected manually. The alignment of “Alice in the Wonderland” and respective texts with its Russian translations has been corrected to the possible extent; the alignment of other texts has been mostly left unchanged. However, the quality of automatic alignment of other texts was much better, because the Gale-Church algorithm is noise-sensitive. The “noise” in Alice is related to the reported speech, which is coded in the original English and Russian texts quite differently. By this reason, sentence boundaries and their alignment cannot be determined automatically. Anyway, alignment errors that remain in the corpora do not cause significant problems when searching for translation equivalents, because the consultation function (see below) may present a wide context for each occurrence of a lexical item. A Perl script allows an incremental annotation of lexical items that conform to a condition. Initially, the lexical features were based on the set of senses of these lexical items under investigation from WordNet (Miller, 1990) for English and ECD and (Apresjan, 2000) for Russian. In the course of the project development, the set of features for annotation was extended to include oppositions discussed in Section 4.

A concordance helps in development of an instance-based vocabulary, which provides another layer of representation of semantic properties of lexical items. A lexical item is an XML entity, which has the following attributes:

1. lexid--the identifier of the lexical item (typically it is the same as lemma, but may have indices for representing homonyms);
2. lemma;
3. lang--language;
4. lexfeatures--the lexical features of the lexical item; they correspond to the set of features, which are pertinent to the lexical item in all or most typical cases of its usage;
5. comments.

Also, the XML entity for a lexical item has the following multiple-value attributes:

1. use--its values refer to the identifiers of sentences, in which the lexical item occurs;
2. synonym--in addition to the sentence identifier, it specifies the lexical item that is considered as synonymous in the context of the sentence;
3. translation-- in addition to the sentence identifier, it specifies the identifier and the language of the lexical item that is considered as a translational equivalent in the context of the sentence.

The lexid, lemma, lang, and use attributes are filled automatically, when the vocabulary is constructed from the concordance, while lexfeatures, synonym and translation are filled manually in the process of corpus analysis. The vocabulary is shared across all concordances and their languages.

The consultation function produces an HTML file with a KWIC (KeyWords In Context) list, which items are sentences that conform to the set of search criteria specified as a dedicated Perl function. The search criteria are applied to one sentence and may include all the information stored in the concordance about this sentence, e.g. co-occurrence of words within the sentence, specific morphological or lexical features of words, as well as their synonyms or translation equivalents. The search criteria can restrict the length of the context, which is extracted from found sentences and presented to the user. Also, the output can be sorted with respect to the left or right context of the found items or according to any other condition defined by a Perl function. As for translation equivalents, the respective parallel sentence or a hyperlink to it may be also included in the output. Each sentence in the output list can be explored with respect to its wider context, since it is hyperlinked to its position in the original text. The screen shot in Figure 1 shows the selection of English and Russian sentences, in which a size adjective (such as *little* or *small*) is accompanied with a noun denoting a child (*girl*, *boy*, *child* or *children*).

Figure 1 A concordance query result

The developed software allows an incremental research for lexical semantics of specific groups of lexical items. For example, the dictionary provides all lexical items, from which a list of, say, verbs of motion that appear in the corpus can be extracted. The concordance query uses the list in order to classify contexts in which such verbs can be used. As the result, specific contexts can be marked in the concordance (the information is stored in the lexfeatures attribute) and can be used in further concordance queries.

3. Two methodologies in lexical semantic representation

Approaches to lexical meanings fall roughly into two groups, which can be (superficially) labelled

as logic- and communication-centered paradigms[2]. The first paradigm assumes that lexical meanings are concepts that belong to an ontology, which represents real-world objects and their properties. Lexical items may refer to one or several concepts and by virtue of this reference they are endowed with a meaning. This relationship between words and meanings is primary with respect to communication, since the ontology exists and the mapping from words to concepts is defined independently from any possible act of communication. Within computational approaches to lexical semantics, good representatives of this paradigm are WordNet (Miller, 1990) and the Explanatory Combinatorial Dictionary, ECD, (Mel'chuk, 1988).

The second paradigm assumes the primacy of communication: human languages are not aimed at the correct representation of the world, but at the communication of experience. In this view, language is a tool for acting in the world, and words are hints, which refer to meanings intended by the speaker. This view can be defined as the meaning-as-use position. It is also shared by a wide community of philosophers of language and linguists, for example, Wittgenstein, Harman, Halliday. If the meaning of a word depends on its contribution to the ongoing exchange between the speaker (S) and the hearer (H), it should be analysed in terms of its occurrence in an utterance taken in its context. The two paradigms seems to be contradictory. At the same time, they are complementary both in terms of their purposes and their results. Thus, they can mutually benefit from their interaction. On the one hand, the proposed description utilizes lexicographical resources such as WordNet, ECD, and (Apresjan, 2000), as well as the analysis of verbs of motion from (Levin, 1993). On the other hand, it is based on the communication-centred systemic-functional linguistics, SFL (Halliday, Matthiessen, 1999).

The communication-oriented approach exercised in the project contrasts to the representation of meanings of lexical items as definitions in a dictionary. A dictionary entry lists senses as concepts that can be referred to by means of the respective lexical item. Each concept in the list of senses is considered as a separate item, which may be related to other concepts, which are other senses of the same word, but this is not specified explicitly in its definition. When a word is used in an utterance, its meaning is an element which is selected from the list of senses. A use of a word is considered ambiguous, when it refers to more than one element. The length of the list of senses of a word depends on the word and the lexicographer, however, the list is typically long. For example, for the verb *leave* it contains 17 senses in WordNet, 15 senses (without idioms) in the Random House Webster's, and 31 senses (also without idioms) in the Oxford English Dictionary[3]. Analogously, (Apresjan, 2000) analyzes 19 senses of *vyjti*, which is the most typical translation equivalent of *leave* in Russian.

When a formal lexicographic description (of the type found in WordNet or ECD) is applied to corpora, two problems arise. On the one hand, some examples of real and felicitous usage do not fit into the fixed list of senses in the definition of a lexical item. On the other hand, some examples fall simultaneously into several senses. This is not related to the disambiguation, since all the senses are relevant for a human judge, who also does not consider the usage as ambiguous. Another problem concerns translated texts: the piece of reality described in parallel texts (or parallel translations of the same text) is assumed to be essentially constant, but its lexicogrammatical realization varies significantly and is not constrained to choosing different synonyms or translation equivalents.

The concordance software and the parallel aligned corpus (reported above) provided the possibility to analyse English and Russian corpora with respect to:

1. the usage of verbs referring to motion, e.g. go, *leave, vyjti*;
2. the usage of nouns referring to natural-kind objects, such as names of trees and animals, e.g. *camomile, rabbit*;
3. the usage of adjectives referring to the size.

The annotation of lexical features was based on their senses from WordNet for English and ECD and (Apresjan, 2000) for Russian. The experiment shows that, in the case of verbs of motion, 6% uses of English verbs in the corpus do not fit into any sense from WordNet, while about 35% are ambiguous, i.e. more than one sense can be used for the annotation, in spite of the fact that their use is not ambiguous. For example, WordNet contains two subsets of senses of *leave*, which refer, respectively, to leaving a place and a person. The first case contains two senses: *go away* and *move*

[2]According to (Matthiessen, Bateman, 1992: 54-55) the origins of both logical- and communication-(rhetorical) centered paradigms dates back to the very beginning of thinking about language in the Western tradition. Nowadays, the two perspectives are partly mirrored in the opposition of formal and functional linguistics.

[3] The senses for the noun and phrasal verbs were not counted.

out of. In many utterances the distinction between them cannot be drawn, so the use should be considered as ambiguous. The second case requires the death or the divorce, so no sense from WordNet is applicable to:

(1) *But her sister sat still just as she left her*

The same situation is true for size adjective. Properties of objects with respect to their size are particularly important for the story of 'Alice', but the usage of such words as *large, little, small* does not always follow their typical definitions, for example, of *large* as 'greater than average size, quantity, or degree'. Each time, when such properties are mentioned, they refer to more than just physical characteristics of an object in comparison to its average size, because the size of an object is mentioned only if the reference to it is appropriate for some purpose of the author, often in comparison to the size of Alice:

(2) *a small passage, not much larger than a rat-hole* (so that Alice could not pass through it with the size she had at that moment)

(3) *[she went on crying] until there was a large pool all round her.*

As the result, some references to the size are not translated into Russian at all or are expressed in a different way. For example, in Nabokov's translation of 'Alice', (3) is rendered as:

(4) *… posredine zaly obrazovalos' glubokoe ozero*
in-center-of hall appeared-3sg deep lake.

Large vs. *glubokij* (*deep*, in this sense) and *pool* vs. *ozero* (*lake*) cannot be considered as translation equivalents in a normal dictionary. From the logical viewpoint, *"a large pool"* is in no way synonymous to *"a deep lake"*. Yet, basically the same meaning is delivered in the translation. As for annotated occurrences, 11% uses of English size adjectives are not covered in WordNet and about 65% are ambiguous. This is partly related to the unclear design of lexical entries for size adjectives, for example, WordNet has the following senses of *little*:

1. limited or below average in number or quantity or magnitude or extent;
5. of little importance or influence or power; of minor status;
6. (informal terms) small and of little importance;
8. contemptibly narrow in outlook, e.g. "a little mind consumed with trivia"; "petty little comments";
11. used of persons or behavior; characterized by or indicative of lack of generosity.

It is unclear which sense is implied in:

(5) *- Oh, you wicked little thing! - cried Alice, catching up the kitten, and giving it a little kiss to make it understand that it was in disgrace.*

Both occurrences of *little* in the example correspond simultaneously to several senses from the set: *little thing* means 1, 5, 6 and 8 (also, *little*$_4$--young), while *a little kiss* means 1, 6 and 11.

4. Words as Resources for Communication

The description reported in the previous section is mostly negative: the real usage of words does *not* necessarily conform to their definitions in a dictionary. However, the concept and the format of modern monolingual dictionaries depend on the history of their development[4] and on the function they serve in the society, namely, to be an authoritative source, which helps in understanding a specific use of a word or in checking/ensuring the correctness of its use. As the result, a dictionary entry is designed as a list of senses that denote to events, objects or their properties. In other words, it statically describes the result of the dynamic process, when words are used by S to refer to events, objects, and their properties[5]. Also, S uses words not only for the purposes of referring, but also for acting on H by means of referring.

The model, which is proposed in the paper, is aimed at the description of how words are used in purposeful communication. The model is based on Systemic-Functional Linguistics (SFL), cf. analysis of its traditions for representing lexical meanings in (Wanner, 1997). In spite of all the differences in

[4] Origins of the tradition of sense enumeration in a monolingual dictionary are related to the development of printed discourse, particularly the new periodicals, in England in the eighteenth century. This brought about a re-evaluation of the nature of meaning, cf. (Kilgariff, 1997).

[5] The Collins COBUILD English Dictionary is an exception in this respect; it presents senses according to the communicative potential of lexical items.

approaches, the computational formalism for describing lexicogrammatical meanings in SFL is the systemic network, which represents choices between interrelated oppositions. For instance, classification of the English mood starts with the features 'indicative' vs. 'imperative'. Semantically, it corresponds to the opposition of speech acts referring to exchange of information vs. issuing commands. Thus, the grammatical choice is controlled by an inquiry to relevant parameters, which are beyond the grammar, cf. (Matthiessen, Bateman, 1992).

The proposed communication-oriented description of lexical items separates the *potential* of possibilities, which defines possible usage of lexical items and is represented by oppositions in the systemic network, from *instantiation* of the potential, which is developed for their instances in the context of an utterance. The set of features, which are fired for a lexical item as the result of its use in the discourse, may correspond to its sense in a dictionary. However, specific communicative goals of S or S's idiolect may also result in a different set of features. The meaning of a lexical item in the context depends not only on the set of features chosen for itself, but (a) on its contribution to the lexicogrammatical structure of the utterance (the principle of compositionality) and (b) on inquiries, which are responsible for their choice, i.e. how features correspond to extralinguistic concepts, traditions in the register and to S's communicative goals. SFL also distinguishes between *paradigmatic* classifications, which are represented by the feature inheritance network, and *syntagmatic* realizations, which are implied by the choice of features. In the case of lexical semantics, realization statements constrain the choice of lexical items.

The proposed description also addresses the issue of multilingual differences between senses. The systemic network is based on the notion of (multiple) inheritance of features. The least delicate, more general, choices tend to be shared across languages, while more delicate choices tend to be language-specific (Bateman, *et al*, 1999). For example, many languages have the type of motion verbs and also distinguish between the motion towards or away from the reference object, as its subtypes, e.g. *enter* vs. *leave*. However, in Russian there are further subtypes, which are lexicalized using prefixes, e.g. the away-from motion subtype can be designated by such verbs as *vykhoditj, otkhoditj, uhoditj* (all of them are typically translated as *leave*), but they denote different configurations of source and destination properties: *vyhoditj* implies motion from within the source, *otkhoditj* implies motion from the vicinity of the source, *ukhoditj* implies motion towards a remote destination. Such subtypes are language-dependent choices. Thus, the systemic network represents *both* commonalities and differences between languages.

The following subsections contain systemic network fragments that are required for representing lexical semantics of verbs of motion and size adjectives. The description is oriented towards automatic generation of expressions for denoting respective events and properties.

4.1 A fragment of the systemic network for verbs of motions

The most designations of motion processes in English and Russian are based on the following oppositions[6]:

The Type of Motion: physical vs. imaginary

(6) He left the room vs. He left the job.

Most often, the reference to physical or imaginary motion is not an inherent feature of a lexical item, since either interpretation is possible depending on context: *to go through a lot of trouble; to advance the claim, to jump to the conclusions*. Within the physical motion there may be a difference in the manner of motion: *run, crawl, fly*, or the manner is left unspecified. The lexical item may also express the cause for the motion. Two clear cases are (a) external circumstances, e.g. *to quit the job* (compare to *to leave the job*); and (b) habits or legal requirements, e.g. *withdraw* (Apresjan, 2000).

[6] There are other oppositions that influence lexical meanings of verbs of motion, e.g. activity vs. achievement (Levin, 1993). However, they are not special for verbs of motion.

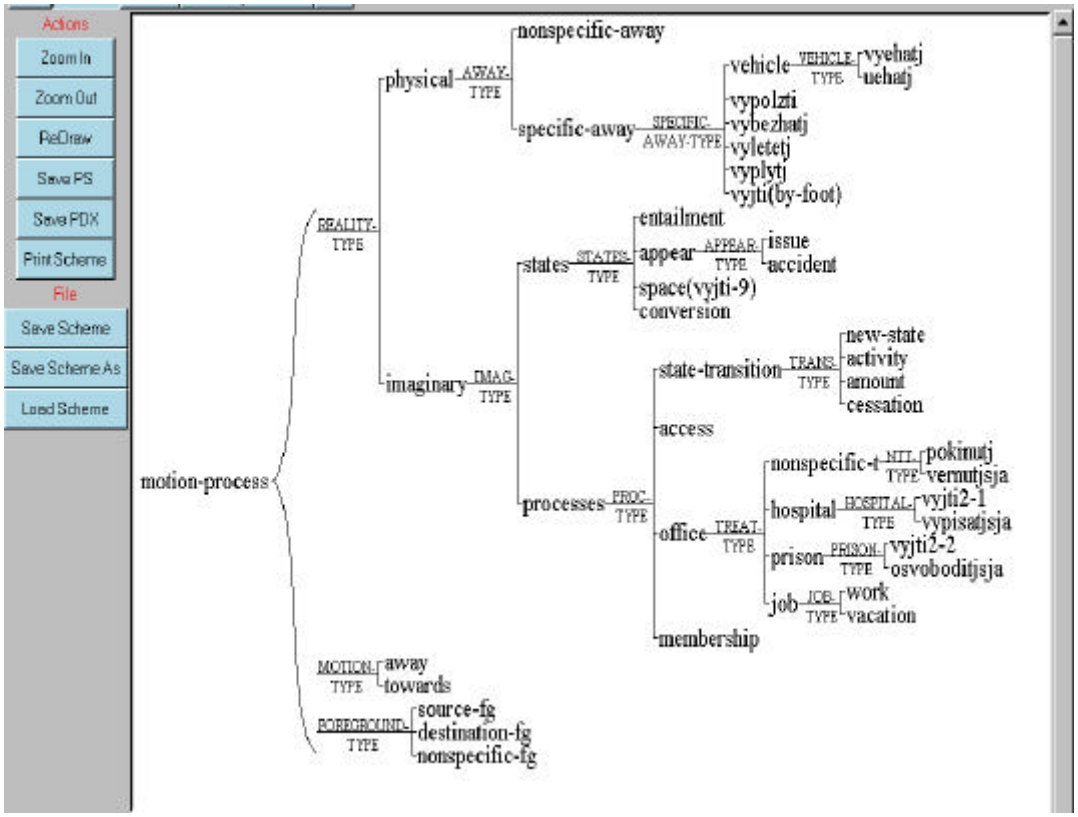

Figure 2: The systemic network for Russian verbs of motion (the "away" subtype).

Within the imaginary motion subtype, the opposition is drawn between those processes that retain some kind of metaphorical motion and those that use verbs of motion to designate an activity without an overt relation to spatial properties, e.g. *The boy left school in the middle of the year* (when an institution is considered as a space to be left) vs. *He had left off writing on the slate*. In Russian, the number of options for the imaginary motion includes, among others, processes for changing one's status with respect to a social institution (cf. Figure 2), in particular:

(7) vyjti, vypisat'sja iz bol'nicy (a patient leaves a hospital);

(8) vyjti, osvobodit'sja iz tjur'my (a prisoner leaves a prison);

(9) vyjti na rabotu, vyjti na pensiju, vyjti v otstavku
(to start one's duties, to become a pensioner, to resign);

(10) vyjti iz otpuska (to end vacation).

The Direction: unspecified vs. towards vs. away

(11) *She ran for two hours* vs. *She entered the room*.

A lexical item may either designate an inherently directed motion with respect to a reference point (*arrive, depart, descend, return*) or be indifferent to it (*roll, slide, walk*). In the case of inherently directed, it can be treated as away from a reference point or towards it (*to enter a room* vs. *to leave a room*). This opposition is also applicable to the imaginary motion.

The Focal Point: source- vs. destination-foregrounded

(12) *He came in with a teacup* vs.

(13) *A torrent of Italian music came from an open window.*

This opposition differs from the direction of motion with respect to the reference point, since the motion may be directed towards it, but the source of motion is foregrounded, cf. (12) and (13). The opposition does not necessarily concerns the semantics of the prepositional phrase. The meaning of

the focal point of an utterance can be incorporated into the verb as well, cf. *ausgehen* (to go out for an entertainment) in German or *vyjti v svet* (a publication is issued) in Russian.

Thus, the example (1) can be represented by the following sequence of features: motion-process, physical, nonspecific, directed-motion, away, destination-fg.

4.2 A fragment of the systemic network for size adjectives

As with verbs of motion, the investigation of size adjectives goes beyond the domain of pure size descriptions. The following oppositions are considered (they are shared in English and Russian):

The type of measurement: spatiotemporal vs. amount

Firstly, the majority of adjectives which refer to size are not specific for this purpose. They may also refer to the description of quantities, e.g. *little door* vs. *little hope* vs. *little pattering of feet.* From the cognitive viewpoint, a reference to size implies a mapping from a qualitative state into a one-dimensional space, which represents the measure for the state. So many words that designate the mapping are naturally applicable both to sizes and quantities. However, the opposite is not true, because some adjectives are applicable only to amount, e.g. *a lot of, insignificant*. Secondly, many words that refer to size can also be used for a reference to time. In language (at least, in Indo-European languages) temporal qualities are often expressed by the same means as spatial ones, e.g. *little stay, long duration* (as well as *from April to June*). In the case of spatiotemporal measurements, the properties denoted by an adjective may be either directional (*deep, early*) or non-directional (*small*). It is not easy to classify adjectives along measured directions, an example is the adjective *deep*, which has the set of senses referring to all possible directions: *a deep well* (vertical), *a deep shelf* (horizontal), *a deep border* (broad). However, all the senses of *deep* are related to the presence of a remote surface, cf. also metaphorical uses *deep space, deep thoughts, deep sleep*. The described set of senses of the English adjective *deep* corresponds to its Russian translation equivalent *glubokij*.

The reference of measurement: small vs. big

This opposition is simultaneous with the distinctions in the type of measurement. Both references to spatiotemporal and amount properties, as well as metaphorical expressions can be classified in this respect, cf. *deep sorrow, shallow thoughts*; *large breakthrough, small problem*. The options for realization of the opposition in Russian are often based on diminutive suffixes to render the smaller side of the scale. In the translation of *large pool* in (4), the contextually important features, such as the large physical measurement and the water surface (to swim in), are preserved, thus making the translation contextually appropriate.

The interpersonal dimension: sympathy vs. neutral vs. antipathy

A reference to a property of an object is often aimed at providing a rhetorical impact onto H. The interpersonal dimension of the way how properties are denoted is responsible for what Sinclair (1991) calls *semantic prosody*. The reference to the size of an object/person may be mentioned to justify the need in taking care of it or being afraid of it. Often, the small scale of size is associated with sympathy, cf. the example (5). However, the sympathy vs. antipathy opposition can be detected even in roughly synonymous lexical items. Let's consider examples from the corpus, in which the adjectives *little* and *small* are used to refer to a child, and check their translations[7]:

(14) *ignorant little girl*
(D) *strashnaja nevezhda* (awful ignorant),
(N, Z) *durochka* (fool-femin-diminut)

(15) *I'm a little girl*
(D) *Ja malen'kaja devochka* (I'm a small girl)
(Z) *Ja devochka* (I'm a girl)

(16) *very few little girls of her age*
(D) *nemnogo najdjotsja devochek ejo vozrasta* (few find-refl girls of her age),

(17) *A <u>small boy</u> and a begrimed, bowlegged toddler lurked behind them. Malen'kij mal'chik i zamyzgannyj, kolchenogij mladenec zanajachili gde-to za nimi.* (Lolita)

[7] Translations with insignificant differences are omitted. D stands for Demurova, N for Nabokov, Z for Zakhoder.

(18) *there came a blinding flash, and beaming Dr. Braddock, two orchid-ornamentalized matrons, the small girl in white, and presumably the bared teeth of Humbert Humbert were immortalized*

The selection provides two observations. Firstly, the reference to the age may be omitted from the English original without a loss in the representation of a person. Thus, the reference is used mostly for adding a specific semantic flavour. The reference to the age is omitted in some Russian translations and preserved in others. Secondly, in spite of the fact that the most vocabularies, including WordNet, Random House Webster's and OED treat *little* and *small* as synonymous for denoting young children, each time, when *small* is used, the context is unfavourable, (though, this observation may be related to the limited size of the corpus).

5. Conclusions

The project reported in the paper presents an attempt of the contrastive semantic study on an empirical basis of the corpus-driven description. The study of verbs of motion and size adjective in terms of their use in texts illustrates that even though the lexicographic research aimed at the enumeration of senses of a lexical item provides an invaluable resource for lexical semantic analysis of texts, a dictionary cannot account for all the possible felicitous uses of words. When the list of senses gets more elaborate, more uses become ambiguous, since often they start to refer to more senses from the dictionary. The communication-centred perspective onto lexical semantics shifts the attention from the meaning of a lexical item to possible uses of groups of lexical items. The question of the lexical semantic analysis in this case is not “What is the meaning of a lexical item?”, but “How things are meant by lexical items?” This also better corresponds to the social foundations of linguistic interaction: the end of using language is in acting on others. The future direction of the reported research consists in the development of the large-scale corpus-based comparative description which models the usage of motion verbs in English, German and Russian and its implementation for automatic generation within the KPML environment. Another outcome of the reported research consists in the development of a parallel English-Russian corpus, which is aligned at the sentence level. No corpus of this type existed at the beginning of the project. The project also led to the development of the special-purpose software to create multilingual concordances and interact with them. Fortunately, Perl as the programming language and XML as the underlying representation allowed for rapid prototyping.

Acknowledgments

The research reported in this paper has been funded by the research support grant RSS321/1999 from the Open Society Institute, and by the Fellowship awarded by the Alexander von Humboldt Foundation, Germany.

References

Apresjan, J.D. (2000) *Systematic lexicography*. Oxford: Oxford University Press.

Bateman J.A., Matthiessen C.M.I.M., & Zeng L.. (1999) Multilingual natural language generation for multilingual software: a functional linguistic approach. *Applied Artificial Intelligence, 13*, 607-639.

Bateman J.A., Teich E, Kruijff G-J, Kruijff-Korbayova I., Sharoff S., & Skoumalova H. (2000). Resources for multilingual text generation in three Slavic languages. In *Proc. Languages Resources and Evaluation Conference LREC2000,* Athens, Greece, May 30-June 2, 2000, pp. 1763-1767.

EAGLES (1996) *Recommendations for the morphosyntactic annotation of corpora.* EAG-TCWG-MAC/R. Available from ftp://ftp.ilc.pi.cnr.it/pub/eagles/corpora/annotate.ps.gz

Gale W, & Church K., (1993) A Program of Aligning Sentences in Bilingual Corpora. *Computational Linguistics,* 19(1):75-102.

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

Kilgariff, A. (1997) "I don't believe in word senses". *Computers and the Humanities 31* (2), 91-113.

Levin, B. (1993) *Towards a lexical organization of English verbs.* Chicago: University of Chicago Press.

Matthiessen C.M.I.M., & Bateman J.A. (1992). Text generation and systemic functional linguistics: experiences from English and Japanese. London: Pinter Publishers.

Mel'chuk I.A. (1988). Semantic description of lexical units in an Explanatory Combinatorial Dictionary: basic principles and heuristic criteria. *International Journal of Lexicography 1*, 165-188.

Miller G., ed. (1990). WordNet: an online lexical database. *International Journal of Lexicography 3* (the special issue).

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

Sampson G. (1995) *English for the computer: the SUSANNE corpus and analytic scheme*. Oxford: Clarendon Press 1995.

Sinclair J. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

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

Wanner L. (1997) *Exploring lexical resources for text generation in a systemic functional language model*. PhD Thesis, University of Saarbrücken.