RussNet as a Semantic Component of the Text Analyser for Russian

Irina V. Azarova
Department of Applied Linguistics
St-Petersburg State University
Universitetskaya nab. 11
199034 St-Petersburg, Russia
azic@bsr.spb.ru

Vadim Ivanov
IdeoGraph Company
Stavropolsky str., 10, 312
191124 St-Petersburg, Russia
artifex.i@gmail.com

Ekaterina A. Ovchinnikova
IdeoGraph Company
Stavropolsky str., 10, 312
191124 St-Petersburg, Russia
e.ovchinnikova@gmail.com

Anna A. Sinopalnikova
Brno University of Technology
Bozetechova 2
61266, Brno, Czech Republic
sino@fit.vutbr.cz

Abstract

In this paper we present a text analysis system developed on the basis of the AGFL grammar and RussNet – a wordnet-like lexicon for the Russian language. We describe a basic architecture of the system, in particular the characteristics of its semantic components – lexico-semantic, morpho-semantic, and syntactic-semantic ones. Principally, the effectiveness of the system benefits from the fact that its semantic modules are extended with syntactic-semantic descriptions – that of valency frames and predicate proposition formalism. The output structures may be used for various NLP tasks including text mining, fact extraction etc.

1 Introduction

Numerous NLP tasks ranging from machine translation to fact extraction require accurate text analysis as a preprocessing step, thus need reliable natural language parsers to be applied.

The key problem of parsing concerns the inherent ambiguity of natural languages, which covers all levels of representation: morphologic, lexical, syntactic, pragmatic. In practice, high ambiguity of the input (texts) results in a high number of possible outputs of a parser. E.g. a grammar extracted from Penn Treebank and tested on a set of sentences randomly generated from a probabilistic version of the grammar has an average 7.2×10^{27} parses per sentence (Moore 2000). In languages with free word order and reach morphology, such as Slavonic languages, the number of ambiguous phenomena multiplies enormously. A traditional solution of this problem is presented by probabilistic parsing techniques aiming at finding the most probable parse of a given input sentence. These methods are based on the relative frequency of occurrences of the possible relations in a representative corpus (Blunt, Nijholt 2000; Horák, Kadlec & Smrž 2002).

It's really unattainable to mention all valuable NLP approaches to parsing, e.g. hundreds of them are presented in the Digital Archive of Research Papers in Computational Linguistics (2005). Equally, it is hardly allegeable and necessary to distinguish some of them as outstanding – it's a matter of personal preferences. Roughly, these approaches may be classified into: (1) systems with modules reflecting traditional linguistic levels and aiming at the comprehensive text analysis and (2) heuristic systems with "distributed" architecture and complex linguistic description. The shortcomings of the former are: the ambiguity on each language level is treated isolately, thus the number of ambiguous parses grows as the analysis runs, and, there appear difficulties with compatibility between modules. The latter, though work effectively, don't give an exhaustive notion of underlying mechanisms for text analysis and appear to be an "ad hoc" realisation applicable to one particular case and not extendable to others.

In our approach, we attempted to combine advantages of both types: use linguistically based modules, which are embedded into each other. The outer module being semantic, and innermost modules being morphologic and derivational ones. This "nesting" mode of data aggregation affords to reduce ambiguity on each level and facilitates data flow between modules.

Another anchor point of our system is a wordnet. The retrieval experiments with WordNet (Voorhees 1998) were a little bit discouraging, manifesting that a text cohesion in non-terminological texts should have more sophisticated projection on thesaurus structure than "lexical chains" within the wordnet trees (Hirst & St-Onge 1998).

2 System design

Our system was developed within an Ideograph project, which is a joint initiative of researchers from the Saint-Petersburg State University and the IdeoGraph company (IdeoGraph 2004). The system gains from the results of our earlier projects – RussNet (Azarova et al 2002; Azarova, Sinopalnikova 2004) and AGFL for Russian (Azarova 2002; Rus4IR 2004). The specific features of the IdeoGraph technology concerns the usage of:

- AGFL formalisms (Koster 1991) for grammatical description;
- several features of HPSG grammar formalism (Pollard & Sag 1994) for syntax and semantic analysis;

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 19–27.

- typed feature structures (Carpenter 1992) for internal representation of linguistic objects;
- RussNet – a wordnet-like lexicon extended by valency frames;
- a semantic module for resolving the grammatical and lexical ambiguity, and for producing output propositions;
- IdeoLog platform for logic inference.

The overall architecture of the system is presented on Fig. 1.

The IdeoLog platform is an efficient implementation of an abstract machine (Takaki et al. 1997) supporting the unification procedure defined on typed feature structures (TFS). The platform supports parsers for Prolog, AGFL and TFS formal syntax, that provides Prolog predicate extensions to AGFL grammar transduction, with TFS representing linguistic objects and obtaining data from the wordnet thesauri.

Main linguistic components of the system are:

- grammatical modules supporting morphological, derivational and syntactic text analysis;
- semantic modules for morpho-semantic and syntactic-semantic analysis, and word sense disambiguation.

The characteristics of the latter component is a topic of the current paper, particularly, the structure of the semantic representation based on the wordnet, though we will briefly describe the grammatical component too: especially, the dataflow between modules and interaction of grammatical and semantic pieces of information.

3 Grammatical component: RUS4IR

The grammatical analyser was developed on the basis of a generative Affix Grammar over a Finite Lattice (AGFL) (Koster 1991), adapted for effective natural language processing. AGFL belongs to the family of two level grammars, along with attribute grammars: a first, context-free level is augmented with set-valued features for expressing coordination between word forms in syntactic constructions. Using AGFL parser generation system, a Rus4IR module (Russian parser for Information Retrieval) was developed – a powerful tool aimed to generate parses for texts written in Russian (Rus4IR 2004). Rus4IR was implemented into the Ideograph system as its grammatical component, supporting manifold morphological, derivational and syntactic analyses of texts.

The main advantage of the grammatical component is that it selects those parses of items which are pertinent to the respective outer units. Thus high ambiguity of the morphological structures is solved (or significantly reduced) by the syntactic information supplied, e.g. in Russian when a preposition allows to solve the ambiguity of the noun form, or when agreeing adjective helps to choose the correct form of the noun.

Moreover, combining syntactic and morphologic analysis, we could effectively identify and process complex and compound word forms: analytic forms of verbs (e.g. future simple: будет слушать ‘(he) will listen’; conditionals: пришел бы ‘(he) would have come’), complex numerals (три тысячи триста тридцать три ‘3333’), complex conjunctions and prepositions (в течение ‘during’, так как ‘as’, cf. English *in order to*, *of course* etc.). This is particularly important in case of splitting the word form constituents, e.g. in preposition-pronoun phrases like ни у кого ‘by nobody’(literally ‘*no- **by** body’)*, ни с кем ‘with nobody’(literally ‘*no- **with** -body’)*.

Developers of generative formalisms are as a rule antagonists of a statistical approach, however, we follow the idea of a “hybrid parsing” (Beinema, Koster 2004). It implies utilising statistical characteristics of syntactic constructions in a corpus of modern texts to improve the calculation time and to ensure the grammar robustness.

AGFL morphologic modules are based on the **stem vocabulary**, which covers all words included into RussNet. The basic vocabulary is expanded by a **derivational** submodule added to the grammar formalism. It generates new stems from a given one, appending to them productive prefixes and suffixes, and linking them by a semantic relation to the particular synset incorporated into RussNet (Azarova et al. 2002). For example, a prefix *anti-* attached to a noun or an adjective stem forms a new stem, which is linked to the basic one by a semantic relation DER_ANTONYM_OPPOSITE[1]. Some derivatives with this prefix occur regularly in the text corpus: e.g. антисоветский ‘anti Soviet’ (10.2 ipm), антивоенный ‘unmilitary’ (1.48 ipm), антитело ‘antibody’ (2.14 ipm), so they are listed in RussNet and the stem vocabulary. But others may appear in the processed text occasionally: e.g. антиастматический ‘anti asthmatic’, антигерой ‘antihero’. These words are considered to be potential – that’s why we don’t list them in RussNet, but in case they are encountered in a text, we have a hypothesis about their grammatical characteristics (POS, case, number, etc.) and semantic status determined through the semantic relation and the related RussNet synset, e.g. {герой}[2].

The derivation procedure is time-consuming (decelerating analysis for 10 %), so stems from vocabulary are checked first, before generated ones. Proper names and abbreviations are processed using regular expressions by the tokeniser.

An output of the grammatical modules includes:

- lemma and its POS tag for each wordform in the text;
- for a generated lemma – its basic lemma with a pointer to a particular semantic relation in RussNet;
- a list of grammatical tags for each wordform (case, number, animate, etc.);
- a dependency tree for each sentence (a set of syntactically linked word pairs with established relations “head-daughter”);
- a phrase structure in terms of semantic-syntactic functions, such as “proposition”, “subject”, “object”, etc.

[1] It is a subtype of antonym link, particularizing that two senses have no intermediate. Cf. with oppositions between gradual attributes: *hot – (warm) – cold*.

[2] Compare this approach with that of the Princeton WordNet, where all opposites with productive prefixes *un-, ir-, im-, non-* are not differentiated.

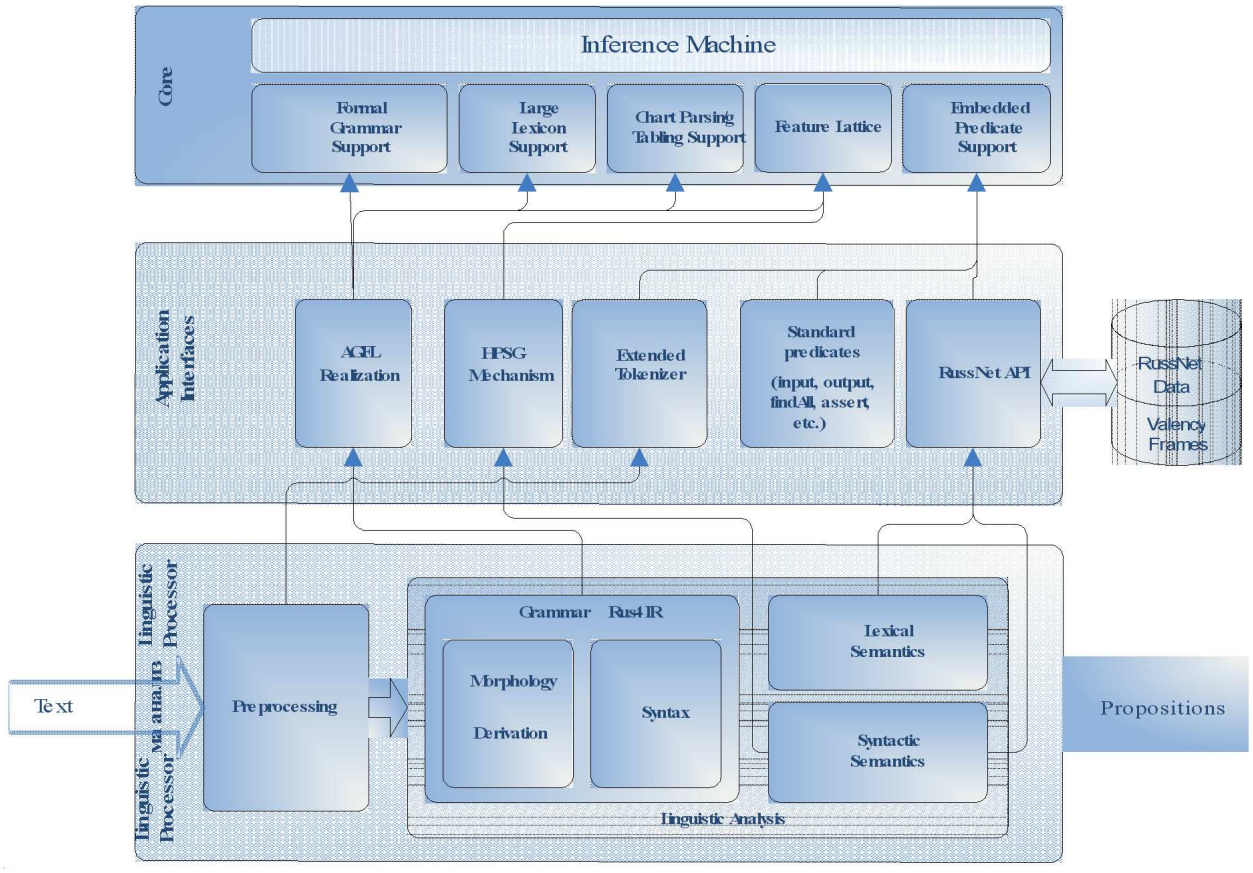

Figure 1: The overall architecture of the Ideograph system.

4 Semantic component

The semantic component of the IdeoGraph system includes three interacting procedures:

- a lexico-semantic module supports the access to RussNet synsets;
- a disambiguation module integrates the matching syntactic structures and synset representations for structural nodes;
- a syntactic-semantic module extracts elementary propositions from the text interpreting verified structures.

Below we present the detailed description of these information resources.

4.1 Lexico-semantic subcomponent: RussNet

Lexical information used by the IdeoGraph system is encoded by means of RussNet, sharing its essential characteristics with other wordnets.

- It present core vocabulary of the Russian language, thus including primarily basic, **non-terminological** word senses.
- The top structure of RussNet consists of about 2000 Basic Concepts – the most **frequent words** (nouns, verbs, adjectives, adverbs) with a frequency more than 100 ipm.
- RussNet is based on **a balanced corpus** of modern texts. It includes texts from 1985–2004 numbering 21 million words, mainly (60%) from newspapers and magazines presenting common vocabulary, as well as political, economic, and scientific terms combined with a small portions of fiction (15%), legislation (10%), and scholarly papers (15%).

Specific features were added to the RussNet structure (Azarova, Sinopalnikova 2004), which were motivated by a specific nature of Russian (its high derivational index) or they were by-products of the data pre-processing.

- The list of semantic relations is extended with **semantic-derivational** ones, they are similar to INVOLVED/ROLE relations in EuroWordNet project stressing the fact that an agent (DER_AGENT: сеятель – сеять ‘sower < to sow’), or an instrument (DER_INSTRUMENT: сеялка – сеять ‘seeder < to sow’), or other situation participant are named after the corresponding action.
- Different senses of an ambiguous word are distinguished not along the system of traditional definitions

in the explanatory dictionaries, but according the context marker distribution. The latter is specified in terms of particular grammatical forms, which accompany the word usage in a particular sense, and semantic constrains of neighbouring words in terms of RussNet semantic trees, or both. To be considered statistically significant, these markers must appear in the corpus regularly: in more than 35% contexts for a meaning in question. The context markers form **valency frames** of particular senses, and are presented within RussNet.

For lexical analysis, another complex problem concerns the so-called **Multiword expressions** (MWEs) (Calzolari et al 2002, Sag et al 2002). These "words with spaces", although equivalent in their meanings to 'true' words, behave in a sentence like phrases, each part of the MWE is changed (conjugates or declines) according to its own POS, not the POS of the MWE as a whole. The borderline between MWEs and free word combinations is quite vague. In conventional dictionaries it is drawn with a shade of subjectivity. More objective way of differentiation relies upon frequency of co-occurrence of MWE parts in a contact position without intrusion of other words, and a corresponding value of the MI-score (Church and Hanks 1991; Azarova et al. 2005) marking the non-random character of the combination.

In RussNet synsets we included mainly MWEs that are 'intrusion sensitive', i.e. do not allow other words to be inserted into their structure, otherwise they lose (or change) their meaning. This principle is rather strict for MWEs with a head noun (e.g. время года 'season', головной убор 'head-dress', микроволновая печь 'microwave oven'). If some adjective is incorporated into MWE большой палец 'thumb', it changes its meaning to 'large finger': большой толстый палец ('large thick finger'), большие кривые пальцы ('large crooked fingers'). However, verb-based MWEs (e.g. иметь в виду 'to have in mind', literally 'to have in view', иметь значение 'to mean', literally 'to have meaning', набрать номер 'to dial', literally 'to pick the number') are more inconsistent allowing for adverbs and particles to put in between the parts of the MWE. In those cases the statistical characteristics – frequency of occurrences and high value of the MI-score – are crucial factors.

Each identified synset of a particular POS activates the so-called "**vicinity**" of a thesaurus node: pointing to hyperonyms, hyponyms, meronyms, etc. Generated lemmas are linked to RussNet synsets with a specified semantic-derivational relation. Selected synsets are localised tracing the hyperonymy links to the top nodes of semantic trees (synsets from top level of the RussNet taxonomies) as domain labels. In general, the amount of synsets and roots may differ, because several synsets may belong to the same tree (e.g. рука occurs in two synsets 'upper extremity' and 'hand' in the tree with the node *body part*), and one synset may have several hyperonyms (e.g. {стакан$_1$} 'glass' is a part of two trees: *artifact* and *container*). Another problem concerns the treatment of words not included into the wordnet, but crucial for the text analysis (e.g. pronouns). As other wordnets, RussNet contains lexical meanings only for four main POS. However, words of other POS may receive a **projection** on the **semantic** tree structure. It's usual for pronouns, e. g. pronouns of the 1st and 2nd persons (я 'I', ты 'you', мы 'we', вы 'you') refer to the semantic tree *human*, meanwhile, masculine and feminine forms (он 'he', она 'she') of the 3d person pronoun refer to a number of trees: *human* as well as *animal*, *object*, *food*, *plant*, etc.

Thus, lexico-semantic module of the system has at its input a set of lemmas identified in the text, together with respective lists of grammatical tags (POS, and categories). An **output** of the **lexical** analysis of a text includes

- a list of **synsets** for identified lemmas from corresponding POS;
- a list of synsets of basic lemmas with derivational links for generated lemmas;
- a list of respective **top nodes** (roots), specifying the trees which comprises active synsets;
- **valency frames** according to RussNet.

Thus, lexico-semantic module affords to represent the variety of word meanings in the text, projecting them to RussNet structure.

4.2 Syntactic-semantic subcomponent: valency frames

This subcomponent provides the interaction of grammatical and lexical outputs for the sake of data disambiguation on both levels. The information core of this procedure is supported by the so-called **valency frames.** This term refers to a set of local context markers (valencies) assigned to a particular sense of a word in RussNet. A number of valencies in frames is variable, however, it was observed that the abstract words usually have less specific environment than concrete ones.

In our approach, valency frames are identified and fixed at the stage of the RussNet data pre-processing – that of word sense identification within particular lexical groups. In order to distinguish different senses of a word, a list of contexts is marked up according the set of context markers, which accompany the occurrence of the word in a particular sense (additionally, a list of senses defined in the explanatory dictionary is used as a zero hypothesis). The percentage of contexts per sense is used for their thesaurus ordering (being a standard wordnet procedure). The percentage of particular context markers for a sense is calculated. Regular valencies, which appear in word's contexts consistently, are enumerated. Thus, in case of RussNet, valency frames have rather statistical nature than speculative. The threshold of regularity is taken to be 35%. Valencies that occur with utmost stability – in 88–95% contexts – are considered to be **obligatory**. Valencies, which appear in less than 60% contexts are treated as **optional**.

We experimented with a number of contexts required for sense identification and observed that 100 random contexts gave the same valency distribution as sets with 1000 and

more contexts[3]. The problem of statistical approach to valency frame identification is that it does not allow to assign valency frames to rare words reliably, as the minimal sufficient number of contexts is assessed to be 25.

Valencies are classified according to several categories (see Fig. 2). The facet concerning the regularity was mentioned above. Next category refers to the function of a word in the phrase, whether it is grammatical head or daughter. From this point of view valencies are divided into **active** and **passive.** Active valencies are characteristic for predicate words (usually verbs, adjectives, and their derivatives). The active valency frame specifies grammatical and semantic features of daughter words, which regularly accompany the head word in texts. The passive valency specifies the grammatical form of a noun, which is attached as a daughter to the top node of some POS semantic tree. E.g. when attached to verbs of speech, a Russian wordform в лицо means "without ceremonies" (literally 'in face'). This does not happen when в лицо occurs with verbs of perception or manipulation with objects, in such cases it keeps the sense "into the face".

Valency frames are identified within two types of **main segments**: (1) non-referential predicative units – a proposition domain; (2) referential units (usually noun phrases denoting objects of propositions and their equivalents). The former are embracing constructions for predicative valencies, the latter – for attributive valencies.

Grammatical specification (*morph_data*) of a valency position includes POS and relevant grammatical tags (e.g. case and preposition description for nouns). For example, the verb направиться in the sense 'to move in some direction' has an active valency frame with two obligatory positions. The first position is expressed by the nominative, the second has two regular grammatical forms: preposition "в" (*in*) with the accusative and preposition "к" (*to*) with the dative. These forms cover 71% occurrences of this valency – other ways of grammatical expression of this situation aspect (denoting the movement destination point) are occasional. Thus, from the variety of all possible grammatical forms we choose the kernel group of the valency occurrences.

POS reference to a "noun" value may involve transposition equivalents: noun phrases, pronouns, abbreviations, citations, etc., which are defined by general rules.

Some grammatical forms may have standard alternations. For example, in case of a sentence negation the direct object may appear in genitive, e.g. создавать помехи (Acc) – не создавать помех (Gen). There is no need to add the genitive form to each frame for transitive verbs, because such altering is regular. In this case the general procedure for checking type identity should account for case alternations on condition that specified grammatical label (*negative direct object*) is presented in the sentence structure marker.

In Russian, **word order** is relatively free, but that does not actually mean "scrambling": the direct (or objective) word order dominates statistically (about 80%). However, in some constructions the indirect word order turns out to be regular, thus if context information evidences for this, special grammatical parameter – an irregular object position (*place*) is inserted into the valency description.

Semantic characteristics of valencies includes two levels. The general semantic property of a valency type refers to some attribute – a role feature, which corresponds to a formal structure of a semantic proposition (discussed in the next section): *subject, object1, attribute*. Semantic characteristics of a valency are specified in terms of references to RussNet trees. For example, the first valency position of the verb направиться mentioned above points to the hyponymy tree *human*. Otherwise, semantic description may point to some subtree. For example, a valency position of an adjective большой in the sense 'possessing the high intensity of an attribute' (большой друг 'great friend', большой артист 'great actor') has semantic reference to that part of the tree *human*, in which people are designated by their qualifying feature. Synsets from another part of the tree *human* mentioning age, e.g. ребенок 'child', or sex, e.g. юноша 'youth', женщина 'woman', and others don't conform with this meaning, cf. большой мальчик 'big boy'. Occasionally, semantic characteristics refer to a particular synset (e.g. agent valency of the verb ржать 'to neigh' could be occupied only by the word *horse*). Regularly, semantic description of a valency comprises a grouping of semantic trees with its own title: *animate* (*human & animal*), *object* (*natural object & substance & artefact &*...), *entity* (*animate & object)*.

The reference to RussNet synsets may support **anaphor** resolution. Another complex problem is the valency frames **ellipsis** under modal influence – negation or modal assessment, for example: нельзя выпить море literally 'it's impossible to drink a sea'; не сердись*!* literally'don't be angry'; не надо сердиться literally'no need to be angry'. In these circumstances even obligatory valencies are regularly omitted or expressed by nouns of "inappropriate" semantic types. In our approach, calculating valency occurrences we exclude such contexts from the context set, as they are obviously occasional, and don't exceed the 5% contexts.

The **disambiguation procedure** uses as its input data the full output of grammatical and lexico-semantic components described above: a number of dependency trees with marked-up levels (each specifying a particular lemma set) and a collection of RussNet synsets with tree specification and valency frames. Dependency trees with corresponding RussNet data are compared with grammar and semantic descriptions of the active valency frames. If some syntactic structure meets the requirements of a certain valency frame, it is considered to be **verified**. If several frames are verified, the text passage is not disambiguated in full. Ambiguity is not cleared up, if there are no valency frames in the RussNet description of respective words or context markers are not sufficient for disambiguation.

Let's consider an example with Russian sentence "Я был знаком с тобой" ('I was acquainted to you/ knew you' or 'With you I was a sign'). On the grammatical level we

[3]Cf. the similar comparison of the training sets of 25 and 200 sentences by (Leacock & Chodorow, 1998).

```
<VALENCY_FRAME active="yes" main_segment="proposition">
  <VALENCY obligatory="yes" role="object1" >
    <morph_data POS="noun" CASE="acc" place="preposition" />
    <sem_data TYPE="top" ID="RUS-nObject" />
  </VALENCY>
  <VALENCY obligatory="no" role="subject" >
    <morph_data POS="noun" CASE="nom" place="postposition"/>
    <sem_data TYPE="top" ID="RUS-nHuman" />
  </VALENCY>
</VALENCY_FRAME>
```

Figure 2: An example of the XML representation of valency frames.

interpret it twofold: a wordform знаком may be lemmatised as (1) a predicative form (singular, masculine) of an adjective знакомый 'acquainted' or (2) an ablative singular of a noun знак ('a sign'). For the adjective знакомый ('acquainted') RussNet fixes an optional valency, grammatically expressed by the preposition с 'with' and the ablative of a noun (**c+N5**), which is accompanied by the semantic tree *human* (in the sense 'personally acquainted with somebody') or *object* (in the sense 'having knowledge of something; encountered before'). A personal pronoun тобой (2nd person, singular, ablative) is equal to the occurrence of a noun from the tree *human.* That affords us to choose the first sense (synset) of знакомый ('acquainted'). The optional valency of the noun знак (*sign*) is a noun in genitive (знак остановки, знак приоритета, 'a sign of smth'), however the context gives us no evidences for this choice. Thus, the verification of the RussNet valency frames against the context markers observed in texts provides the device for disambiguation of grammatical structures and alternative synsets.

4.3 Semantic-syntactic subcomponent: propositional structures

4.3.1 Proposition structure

The syntactical semantic module produces as its output a set of propositions – elementary informative structures inside a simple sentence, with several relations defined on them (such as causation, taxis etc.). A proposition includes a kernel structure and peripheral quantifying and qualifying attributes.

A simple sentence usually has a core propositional structure presented by its predicate, though "hidden" (or abbreviated) propositions may be inserted into various grammatical positions. In this sense, a simple sentence may have complicated propositional structure (the same as complex or compound sentences). Each proposition may have a modality component. It is more usual for a core predicative proposition; however, other propositions may be qualified too – in a more indirect manner.

The propositional structure is as follows. Its **kernel part** includes: (a) a link to a synset in RussNet, which is the predicate representing the given proposition; (b) links to subject and object semantic structures, including the head-noun synset reference and its various attributes.

The periphery of a proposition includes time, place, quality and other specifications. The semantic structure of referential objects consists of a reference to the synset of the phrase head (usually a noun) and quantifying and qualifying attributes. For example, a phrase *two red balls* will be written as follows:

```
proposition
   [ ID id.arrive
     SUBJECT X = object [ID id.man]
     OBJECT3 object [ID id.Vienna]
     TIME T1 ],

proposition
   [ ID id.come
     SUBJECT Y = object [ID id.man]
     PLACE Z = object [ID id.place]
     TIME T2 ],

proposition [ ID id.meet
              PLACE Z ],

proposition [ ID id.agree
              OBJECT1 Z
              TIME T3 ],

before(T1, T2), before(T3, T2).
```

Figure 3: An example of the propositional output.

```
object [ID id.ball,
    QUALITY attribute[id id.red],
    QUANTITY: number [value: 2]].
```

A typical predicate is a verb, though some attributes (adjectives and nouns) may occupy this position as well. A subject role doesn't imply that this position refers to some active component of the situation (such as *causal agent*), but merely that this semantic function is potential for a proposition. The **subject** position, as well as an object one, may be occupied by a **list**, members of which are grammatically equal in the surface text structure (*In semi-final R and M. fought*), or differentiated by grammatical positions (*In semi-final R. fought with M.*).

Various **object** functions are assigned to valency positions in order to differentiate them rather than to express some hidden semantic role (like *agent, patience*, etc.) E.g., for a verb бросать in the sense 'by arm swinging, to cause to fly something hold in hands' the valency frame comprises the

only position ***object1.*** Nonetheless, in the surface structure it may be expressed twofold: бросать камни and камнями (accusative or ablative forms) meaning the same 'to through stones'.

Several object positions are inserted in those cases, when they are **opposed** in the proposition structure. For example, a verb забить with the meaning 'to push something inside something with strokes' has two object positions: the proper object, which refers to the point of strength application, and other object, which is affected by the first. In the former case, conventional role assignment is quite apparent, however this does not hold for the latter. In our approach, they are differentiated as ***object1*** and ***object2***. Corpus evidences – context distribution – shows that besides these objects, there often appears another one (забить уши ватой 'to cram the ears with cotton wool'), which will be designated as ***object3***. Thus, an order of object positions reflects the frequency of their occurrence in the texts for synsets from a particular RussNet tree, some object position prevailing for all or most of all synsets of the tree.

Lets take as an example Russian sentence После нашего приезда в Вену я отправился на заранее определенное место встречи 'After arriving to Vienna, I came to the meeting place agreed before', its propositional structure is shown on Fig. 3.

4.3.2 Interpretation of syntactic rules

The interpretation module includes a set of semantic rules defined on syntactic structures. These rules are verified against the corpus that provides an evidence for frequency ordering and cutting down rare instances. For example, a Russian noun phrase "noun-head + noun-genitive daughter" may express a possession (дом отца – 'my father's house'), an action (создание традиции – 'creation of a tradition'), quantity (тысяча человек – 'thousand of people'), etc. The most frequent case is a phrase, in which head word is a verb-derived noun like 'creation of the tradition'. This noun may be included into RussNet, linking by a relation DER_TRANSPOSITION_ACTION to the corresponding verb synset {создать$_1$} 'create'. Otherwise, it will be processed in a derivational module as a suffix production from a verb, connected with verbal synsets by a link DER_TRANSPOSITION_ACTION.

A daughter noun in a pair may have the most "broad" semantics included into the *entity* set of RussNet trees. So the first rule of semantic interpretation of a genitive construction shows that a phrase comprises a hidden proposition, the predicate of which is the corresponding verbal synset and the object is a dependant noun synset, the proposition subject is unspecified without other attributes.

```
(1) N1 der_transposition_action [ID] +
N2 gen ents-id => P* [object1: N2, id: ID]
```

The second rule interprets phrases like создатель традиции ('a creator of a tradition'). The semantic rule is just the same as concerns the hidden proposition and object handling, though differs in the type of a RussNet link DER_AGENT (*creator* <= *create*) from the head noun to the verb synsets, and an additional subject position in the predicate structure, which is to be filled with a phrase coreferential with a given one (for example, the subject position of a sentence).

```
(2) N1 der_agent[ID] + N2 gen ents-id =>
P* [subject: ref, object1: N2, id: ID]
```

The third rule describes phrases with a subject of the abbreviated proposition specified by a genitive noun: e. g. падение профессора ('professor's falling').

```
(3) N1 der_transposition_action [ID] + N2
gen animate-id => P* [subject: N2, id: ID]
```

The syntactic-semantic module is tightly connected with a lexical one: semantic rules give a generalized interpretation of a construction, while a particular valency structure may put additional restrictions on some elements of the propositional structure or even introduce some rare interpretations, which were not described in common interpretation rules.

For example, a phrase убийство профессора ('killing of a professor' or 'killing by a professor') will receive two interpretations: the first according to rule (1), and the second – rule (3). However, the most frequent interpretation would be inappropriate for a phrase страдание профессора ('suffering of a professor') due to the "void" value of the object positions (see the typed feature structures for valency frame of a Russian verb страдать 'to suffer' on Fig. 4).

The order of semantic interpretations for the syntactic construction is defined by their recurrence in the text corpus, though contradicting information from valency frames can rearrange them. The obligatory valency parameter in a frame is used to filter and rank different interpretations. If an interpretation lacks an obligatory argument, it won't be verified by the system. If a phrase is ambiguous, and two interpretations are verified, then an alternative with a factual optional position will be output as preferred.

5 Conclusion and Future Work

To sum up we may say that presenting a text in terms of propositional structures afford us to exclude from the output semantic structures those variants of the surface presentation, which are inessential for sense identification. Matching propositional structures against RussNet semantic trees makes it possible to generalise the sense, if necessary, or to narrow it, including the utmost semantic details. The generalised form of propositions is to be exploited later in grasping the communicative perspective – a gradual growth of details connected with some microtopics, which are the very essence of the information content of a text document.

The described semantic procedure is still under construction. We are looking for appropriate methods to facilitate the preliminary stage of text processing, especially conceptualised context classification, the promising direction was described in (Leacock, Chodorow 1998) and (Pantel, Lin 2005). Besides, the applicability of this method to Russian, a free word-order language with rich morphology, shows that the described semantic procedures may be applied with different wordnet-based systems, moreover, may serve for wordnet structure evaluation. To further directions of our work there also belong the generation of Word Sketches for

```
proposition[
            id: ID.suffer,
            subject: object[id: RUS-nHuman, morph: noun[case: nom]]
            object1: void
            object2: void
            object3: void
           ]
```

Figure 4: Propositional structure for a valency frame of a verb истрадать ‘to suffer’.

Russian (Rychlý, Smrž 2004) – a structured set of sophisticated semantic patterns (relations and propositions) automatically extracted from annotated corpora.

The general applicability of the described system allows it further usage for the text analysis in other languages. Semantic-derivational module proved to be very productive in the solving many problems concerned to parsing with unrestricted lexicon. Derivational phenomena grasped by the system are characteristic not only for Slavonic languages, they belong to the core issues of the text analysis in agglutinative languages, some traces could be found even in English (e.g. productive prefixes *un-*, *in-*, *ir-*, *non-)* – a language with relatively poor morphology.

The thorough investigation is necessary for valency frames inheritance in wordnet verbal troponymic trees. The sophisticated anaphor and ellipsis resolution procedures are essential in order to describe in full contextual features.

Acknowledgements

Our thanks go to the Russian Federation Foundation of assistance to small & medium enterprises in technological & scientific area, which provided the financial support for our project “Development of Software Technology for Information Extraction from Electronic Documents”.

References

Advances in Probabilistic and Other Parsing Technologies. (2000) Blunt H., Nijholt A. (eds.) Kluwer Academic Publishers.

Azarova I. (2002) *The matching of AGFL subcategories to Russian lexical and grammatical groupings* In “Proceedings of the Second AGFL Workshop on Syntactic Description and Processing of Natural Language”. Radboud University Nijmegen, the Netherlands, `http://www.cs.ru.nl/agfl/papers/`

Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. (2002) *RussNet: Building a Lexical Database for the Russian Language*. In “Workshop on WordNet Structures and Standardisation, and how these affect Wordnet Application and Evaluation. 28th May 2002”. Las Palmas de Gran Canaria, pp. 60–64.

Azarova I., Sinopalnikova A. (2004) *Adjectives in Russnet.* In “Proceedings of the Second International WordNet Conference”, GWC 2004, Brno, Czech Republic, January 20–23, pp. 251–259.

Azarova I., Sinopalnikova A., Smrž P. (2005) *Distinguishing between Multiword Expressions and Collocations in RussNet.* In “Proceedings of the Dialog-2005”. Moscow, pp. 5–12.

Beinema P., Koster C.H.A. (2004) *AGFL Grammar Work Lab: Manual for the AGFL system.* At `http://www.cs.ru.nl/agfl/papers/manual.pdf`. 62 p.

Calzolari N., Fillmore C., Grishman R., Ide N., Lenci A., McLeod C., Zampolli A. (2002) *Towards Best Practice for Multiword Expressions in Computational Lexicons.* In “Proceedings of LREC 2002”. Las Palmas, Spain.

Carpenter B. (1992) *The Logic of Typed Feature Structures.* Cambridge University Press, Cambridge, England. 270 p.

Church K, Hanks P. (1990) *Word Association Norms, Mutual Information and Lexicography.* Computational Linguistics, 16 (1), pp. 22–29.

Hirst G. & St-Onge D. (1998) *Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms.* In “WordNet: An electronic lexical database”, Ch. Fellbaum (ed.), The MIT Press, pp. 307–332.

Horák A., Kadlec V., Smrž P. (2002) *Enhancing Best Analysis Selection and Parser Comparison.* In “Proceedings of the TSD”. Brno, Czech Republic, pp. 463–466.

Kay M. (1986) Parsing in Functional Unification Grammar. In “Readings in Natural Language Processing”, B. J. Grosz, K. Spark Jones & B. L. Webber, ed., Morgan Kaufmann Publishers, Inc., Los Altos, California, pp. 125–138.

Koster C.H.A. (1991) *Affix Grammars for natural languages.* In “Attribute Grammars, Applications and Systems”, International Summer School SAGA, Prague, Czechoslovakia, June 1991.

Leacock C., Chodorow M. (1998) *Combining Local Context and WordNet Similarity for Word Sense Identification.* In “WordNet: An Electronic Lexical Database”. C. Fellbaum (ed.) MIT Press. pp. 265–283.

Makino T., Torisawa K. and Tsujii J. (1997) *LiLFeS – Practical Programming Language For Typed Feature Structures*. In “Proceedings of Natural Language Pacific Rim Symposium ‘97”.

Pantel P., Lin D. (2003) *Word-for-Word Glossing with Contextually Similar Words.* In “Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, May 27 – June 1, Edmonton, Canada”.

Pollard C. & Sag I. (1994) *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press. 440 p.

Rychlý P., Smrž P. *Manatee, Bonito and Word Sketches for Czech.* In “Proceedings of the International Conference ”Corpus Linguistics-2004“, St.-Petersburg, Russia, pp. 324–334.

Sag I., Baldwin T., Bond F., Copestake A., Flickinger D. (2002) *Multiword Expressions: A Pain in the Neck of NLP.* In ”Proceedings of the CICLING 2002“. Mexico City, Mexico.

Voorhees E. M. (1998) *Using WordNet for Text Retrieval.* In ”WordNet: an Electronic Lexical Database“, Ch. Fellbaum, ed., MIT Press, pp. 285–303.

Internet resources

Digital Archive of Research Papers in Computational Linguistics (2005) `http://acl.ldc.upenn.edu/`

IdeoGraph (2004) `http://www.ideograph.ru`

Rus4IR (2004) – the Russian parser for Information Retrieval `http://www.phil.pu.ru/depts/12/AGFL`