The Compreno Semantic Model as an Integral Framework for a Multilingual Lexical Database

Ekaterina MANICHEVA Maria PETROVA Elena KOZLOVA Tatiana POPOVA

ABBYY SOFTWARE HOUSE, Otradnaya str. 2b/6, 127273, Moscow, Russia

Ekaterina_M@abbyy.com, Maria_Pet@abbyy.com, Helen_Koz@abbyy.com, Tatiana_P@abbyy.com

ABSTRACT:

The paper presents an integral framework for multilingual lexical databases (henceforth MLLD) based on Compreno technology. It differs from the existing approaches to MLLD in the following aspects: 1) it is based on a universal semantic hierarchy (SH) of thesaurus type filled with language-specific lexicon; 2) the position in the SH generally determines semantic and syntactic model of a word; 3) this model proposes a suite of elaborate tools to determine universal and language-specific semantic and syntactic properties and deals efficiently with problems of cross-lingual lexical, semantic and syntactic asymmetry. Currently, it includes English, Russian, German, French and Chinese and proves to be a compatible MLLD for typologically different languages that can be used as a comprehensive lexical-semantic database for various NLP applications.

KEYWORDS: multilingual lexical database, semantic and syntactic model, cross-lingual asymmetry

1. Introduction: Integral Framework for the MLLD

Over the past decade, NLP has witnessed a surge in the development of multilingual lexical databases and tools for cross-lingual tasks such as information retrieval, machine translation and foreign language acquisition.

Most of the large-scale lexical databases that lately evolved into multilingual frameworks for language-specific lexicons have been initially designed as monolingual databases and developed independently without referring to any particular processor or potential NLP applications. In order to integrate typologically different languages into these frameworks, adjust them to certain processors and guarantee their cross-platform applicability communities of developers have carried out a great amount of work to develop tools for cross-platform integration and universal standards for semantic representations. Still these projects encounter a lot of problems of uniformity and consistency across languages, categories and applications.

By contrast, the Comreno semantic model developed by ABBYY was initially designed for multilingual purposes and aimed at machine translation, without being limited to it. The system consists of a language database and includes interrelated modules: morphological, syntactic, semantic and statistical ones. The semantic module is based on a universal semantic hierarchy of thesaurus type which is filled with lexical information. The morphological and the syntactical

Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), pages 215–230,
COLING 2012, Mumbai, December 2012.

modules, in turn, are language-specific. This approach proved to be efficient to provide high-quality machine translation for English<->Russian pair (refer to Anisimovich et al., 2012).

At present, we continue working on German, French and Chinese languages. Currently, we have described more than 96000 English, 85000 Russian, 12 000 German, 11 000 French and 8500 Chinese lexical classes. The choice of the languages is mostly determined by the applied tasks of machine translation within corresponding language pairs, though as we have languages here that are typologically different such a choice allows testing the universality of the Compreno model as well.

The format in which the lexical data is implemented has been worked out for this particular system by ABBYY developers. Compreno Parser is available on a fee-for-service basis.

In the following, we briefly present existing multilingual lexical databases (2) and linguistic problems they have to encounter (3); give an overview of Compreno semantic framework (4) and, finally, present in more detail how Compreno MLLD deals with cross-lingual asymmetry and serves as a basis for machine translation (5).

2. Snapshot of the Existing Multilingual Lexical Databases

In this section we provide an overview of the most representative wide-scale projects aimed at constructing multilingual lexical resources in terms of their theoretical approaches and potential NLP applications leaving aside other less known MLLDs for the reason of space limits.

2.1 EuroWordNet project

The mainstream approach to the construction of wide-scale multilingual resources has been demonstrated by the EuroWordNet (Vossen, 2004) and the following Global WordNet Grid initiative. In these projects the goal is to build a worldwide grid of wordnets by means of an interlingual platform.

EuroWordNet consists of individual databases for seven European languages (Dutch, English, Italian, Spanish, German, French, Czech and Estonian) and is analogous to the original Princeton WordNet for English. EuroWordNet provides a fine-grained formal concept analysis for nouns. However, it comes with a poor database of illustrating examples and lacks information about the syntactic behavior of verbs and nouns.

Besides, in EuroWordNet, each language-specific WordNet is an autonomous language-specific ontology where each language has its own set of concepts and lexical-semantic relations based on the lexicalization patterns of that language. EuroWordNet differentiates between language-specific and language-independent modules. The language-independent modules consist of a top concept ontology and an unstructured Inter-lingua-Index (ILI) that provides mapping across individual WordNet structures and meanings.

2.2. PAROLE/SIMPLE lexicons

The initial goal of the LE PAROLE project conducted by the Council of Europe was to produce a head of the harmonized corpora and lexicons for 12 European languages: Catalan, Danish, Dutch,

English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish. These efforts resulted in monolingual morphological and syntactic lexicons for these languages, the volume of each lexicon amounting to 20000 entries.

The next step towards cross-lingual usage of these resources was the SIMPLE project, when existing morphological and syntactic data were provided with semantic representations. The SIMPLE lexicons were developed in line with the EAGLES (Experts Advisory Group on Language Engineering Standards) requirements on lexical-semantic representations for NLP tasks. Thus developers tried to bear in mind potential NLP applications; still they did not refer to any particular applications that would use this information. The SIMPLE lexicons cover 10000 word meanings for the above mentioned languages; they are built around the same head ontology and the same set of semantic templates.

Just as EuroWordNet SIMPLE is constructed not as a property-rich ontology but as a hierarchical net of the lexical items that imposes constrains to its NLP applicability: it lacks disambiguating power and the relations between entities are insufficient (Nirenburg, 2004). To ensure an overlap of lexical senses certain EuroWordNet's Base concepts were converted into each language providing linking of the lexical stock.

The theoretical foundations of the semantic description in SIMPLE are based on the extensions of Generative Lexicon theory (Pustejovsky, 1995) that makes it different from EuroWordNet. A SIMPLE lexical entry includes the following semantic information: 1) semantic type, corresponding with the SemU (semantic unit); 2) domain information 3) lexicographic gloss 4) argument structure for predicate 5) selectional restrictions on the arguments 6) event type to characterize the aspectual properties of verbal predicates 7) link of the arguments to the syntactic subcategorization frames, as represented in PAROLE lexicons 8) Qualia Structure 9) information about regular polysemous alternation in which a word sense may enter 10) cross-part-of-speech relation (derivation) 11) synonymy (McShane et al., 2004).

2.3. FrameNet and FrameNet-like lexicons

Another large-scale multilingual project is FrameNet (Baker et al, 1998). FrameNet is based on Fillmore's Frame Semantics (Fillmore, 1976). Frame Semantics models the lexical meaning of predicates in terms of *frames*; frames describe a conceptual structure or prototypical situation together with a set of semantic roles, or *frame elements (FEs)* involved in the situation. FrameNet currently contains about 600 frames. FrameNet projects employ the deep syntactic representations provided by large-scale lexical functional grammars as syntactic basis for frame-based meaning assignment. As an additional knowledge source FrameNet uses the public SUMO/MILO ontology whose classes are also aligned with WordNet.

By employing semantic frames as interlingual representations, FrameNet, as opposed to other MLLDs, focuses on organizational units larger than words. Besides, each FrameNet entry contains exhaustive information about its semantic and syntactic combinatorial potential and semantically annotated example from large parallel corpora. Thus FrameNet's database deals effectively with paraphrase patterns across languages.
Currently, there are several autonomous FrameNet and FrameNet-like lexicons for English, German, Danish, French, Swedish, Spanish, Japanese and Chinese languages, all on different stages of completion.

3. Challenges for Multilingual Lexical Databases

Construction of MLLDs faces even more complicated problems than those encountered in the creation of monolingual lexical databases (Boas, 2005). Among the main issues developers of the MLLDs have to face are: 1) cross-linguistic polysemy; 2) asymmetry of source and target semantic and syntactic structures; 3) cross-language asymmetry in the delimitation of semantic fields.

- cross-linguistic polysemy

Dictionaries often vary in their organization of word meanings, which makes it difficult to compare definitions across different dictionaries. Besides, most polysemous words are usually the most frequent ones and their meanings are often domain-independent which may make disambiguation impossible. In the case of MLLDs for NLP tasks granularity of sense distinction is a key and controversial both to professional lexicographers and applications (Palmer et. al, 2006).

Cross-linguistic polysemy is even more problematic. It may vary from a complete overlapping of word senses through diverging polysemy to the absence of correspondences among senses across languages (Altenberg and Granger, 2002). Thus, consistent criteria for sense distinction and strategies for cross-lingual sense mappings are crucial for the successful implementation of a MLLD.

- semantic and syntactic asymmetry

In addition to providing information about different meanings of a word, any MLLD should accurately describe deep semantic model of each sense and all its possible surface realizations to ensure correct cross-language mapping.

- cross-language asymmetry in the delimitation of semantic fields

As Talmy (2000) points out, languages differ in the kinds of semantic components they lexicalize. This has a number of important implications for the overall architecture of a MLLD. Some languages might make semantic distinctions that are irrelevant in others. For example, English verbs use particles to show the path of motion (*"run into"*, "*go out", "fall down"*), whereas in Russian and German the path is encoded by affixation, in French – usually by the verb itself and in Chinese by directional modifiers.

Another challenge is posited by culture-specific vocabulary, lexical gaps and their translation equivalents across languages. In this sense, the conception of MLLD development should stem from the Principle of Practical Effability (Nirenburg and Raskin, 2004), which states that what can be expressed in one language can be *somehow* expressed in all languages, be it by a word, phrase, etc. It should also take into account fixed multiword expressions (idioms, terms and collocations) and include a description of how to map such multiword expressions across languages.

Below, we present in more detail the theoretical approaches that Compreno semantic model employs and demonstrate how it treats the problems mentioned above.

4. Key features of Compreno Semantic Model

The Compreno linguistic technology has been originally developed for machine translation, but now it is applied for a wider range of NLP applications aimed at semantic analysis.

In the following we will focus on the universal semantic module of the system and show how its mechanisms can be applied to describe a group of typologically different languages (English, Russian, German, French and Chinese).

4.1. Semantic Hierarchy

All words in our system are organized in the form of a thesaurus-like hierarchical tree which we call the **semantic hierarchy** (henceforth SH). The tree consists of language-independent branches called **semantic classes (SC)**, which are filled with lexical items of natural languages – **lexical classes (LC)**. Higher semantic classes denote general notions like entities, characteristics or actions, while their children have more specific meanings, so the deeper the class is the more particular notion it expresses:

ENTITY_LIKE_CLASSES > ENTITY > FOOD > SOUP > KHARCHO > kharcho

ENTITY_LIKE_CLASSES > ENTITY > FOOD > food

Each semantic class can have both semantic and lexical classes as its descendants (fig. 1).

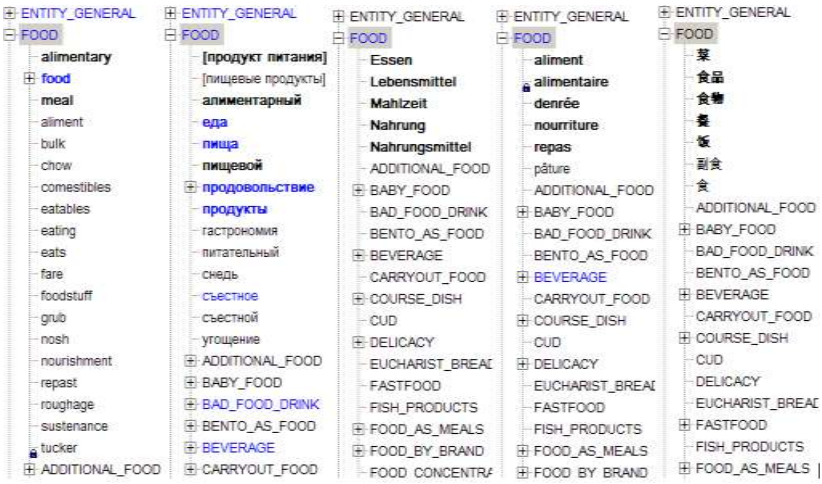

FIGURE 1 - Fragment of the Semantic Hierarchy.

Lexical classes, in turn, contain lexemes with morphological paradigms. Each lexical class can have several lexemes that are **grammatical derivates (GD)**: typical instances are verbs and verbal nouns (like "*translate – translation*") or adjectives and adverbs (like "*beautiful – beautifully*") that differ only in their part of speech type.

The **lexicographic description** of the classes includes the following information: 1) a gloss drawn from a dictionary; 2) compatibility examples; 3) semantic and grammatical restrictions for different surface realizations of the actant valencies; 4) examples of voice transformation (for verbs) and additional restrictions imposed by them, if any; 5) relevant grammatical information; 6) examples of nontrivial translations, set expressions and any other relevant information. For Chinese, we also indicate the transcription, the spelling in Traditional characters, variant spellings and give glosses for all examples. It is essential to provide exhaustive information for the core vocabulary as it serves as basis for the syntactic descriptions and parser. Later on, the work becomes more labor-saving as syntactic and semantic models of LCs are inherited from their ancestors and only local mismatches should be marked.

All words in the hierarchy are attributed with grammatical and semantic values, called **grammemes** and **semantemes** respectively. The usage of grammemes has been minutely examined in Anisimovich et al. 2012, some illustrations will be given below as well. **Semantemes** help to distinguish different lexical items within one semantic class (for other their functions see Anisimovich et al. 2012): i.e., *"beautiful, pretty, handsome"* have a <<PolarityPlus>> semanteme while *"ugly"* takes <<PolarityMinus>>. Semantemes are universal for all languages. We use more than 1100 semantemes in SH. On the contrary, **grammatical** system is unique for every language. So, the number of grammatical categories varies depending on the language. For example, in Russian we set up about 460 categories and 2500 grammemes, 420 / 2400 in English, 240 / 940 in French, 260 / 1300 in German and 60 / 160 in Chinese.

The LC-descendants of one semantic class that have a similar set of semantemes are synonyms. During translation, lexical choice at the synthesis stage usually favors the lexical class with the most similar set of semantemes. Such a choice gets a better evaluation than mismatches between input and output classes.

Words with the same root that differ not only morphologically but also semantically are introduced as **semantic derivates (SD)**: SDs are the descendants of one lexical class that differ in semantemes, for example – "*handsome – unhandsome"*.

The possibility to store multiple SDs under one lexical class is especially helpful for words with a big number of SDs. For instance, the verb "*go*" has about 30 SDs like "*go away*, *go back*, *go in*", etc., corresponding to such verbs as "*leave*, *return*" and "*enter*", so we can place all these verbs in one SC, where "*go*, *leave*, *return*" and "*enter*" will be different LCs while "*go away*, *go back*, *go in*"– the SDs of the LC "*go*". Both LCs "*leave*, *return*, *enter*" and the SDs "*go away*, *go back*, *go in*" acquire the semantemes <<From>>, <<Back>> and <<To>>, respectively. This ensures their distinction from the neutral "*go*".

Semantic derivates are formed by regular morphological models and express semantic relations which are typical for the derivates formed by these models: "*go away*, *fly away*, *swim away*" are all formed with *'away'* particle and express the semantics of leaving the place, or "*go in*, *come in*, *fly in*" are formed with the help of *"in"* particle and express the semantics of moving inside.

Such derivates can also differ in the semantic valencies they attach: for instance, valency indicating initial point ("*come [from school]*") is typical for neutral „*come*" but is rather marginal for the "*come in*" derivate.

The derivates are marked with **derivatemes** – fixed combinations of corresponding grammemes and semantemes, which describe both their syntactic and semantic features. For example, the German verb "*laufen*" ("*to run*") has 40 SDs such as "*durchlaufen*" ("*to run through*"), "*zurücklaufen*" ("*to run back*") or "*fortlaufen*" ("*to run away*") with the derivatemes <Durch_EnRouteLandmark>, <ZurückRück_Back> and <Fort_Depart> respectively. These derivatemes, in turn, contain semantemes <<En_Route>>, <<Back>> and <<From>>. At the current stage of the project the system numbers about 120 English derivatemes, 150 Russian derivatemes, 120 German derivatemes and 10 French derivatemes.

The following table provides data on language-specific descendents of the SC TO_RUN with a few illustrating examples:

	English	Russian	German	French	Chinese
number of LCs	9 (run, scatter, jog, lope, etc.)	3 (бежать, труситься, пробежка)	2 (laufen, rennen)	1 (courir)	2 pǎo (跑, bēnpǎo 奔跑)
number of SDs	37	42	42	2	N/A
number of GDs	46	52	44	3	N/A
<<Back>>	*run_back*	-	*zurücklaufen*	-	pǎohuí 跑回
<<To>>	-	*прибежать*	-	*accourir*	pǎodào 跑到
<<From>>	*run_away whip_off*	*убежать*	*davonlaufen*	-	pǎoqù 跑去

TABLE 1 - Language-specific descendents of the SC TO_RUN

N/A in some fields of the table means 'not applicable'. In Chinese a verb with a directional and resultative complement can insert potential marker between a main verb and a complement and a lot of disyllabic verbs can be used nominally; thus we decided to treat Chinese verbs differently. We do not add them to the SH as grammatical derivates, but describe their derivation paradigm as high as possible on ancestor SC. Examples on the derivates are provided in the LC commentary and nominal syntactic usage is marked with grammeme <VerbNoun>.

- cross-language asymmetry in the delimitation of semantic fields

The asymmetry between different languages is neutralized by marking semantic classes with a representativity feature: this feature defines the relation between a given class and its parent.

There are 3 types of representativity: a SC can be **non-representative**, **semi-representative**, or **fully representative**. A non-representative SC is completely cut off from its parent, so the

translation equivalent for a source concept will be chosen among the LCs of this semantic class only (that is actually a normal situation, where no language asymmetry occurs). A semi-representative SC allows choosing translation equivalents from the parent SC as well (an option for cases where no direct correspondence in a target language can be found and the optimal equivalent is a hyperonym for the word). Finally, a fully representative semantic class is "transparent", i.e., it allows choosing translation equivalents both in the parent and child semantic classes. For instance, English "*go*" and French "*aller*" mean both "*go on foot or by vehicle"* while in Russian or German different verbs must be used here: correspondingly „*идти*" and "*gehen*" for motion on foot and "*examь*", "*fahren*" for motion by vehicle. When translating "*go*" and "*aller*" into Russian or German we normally have to choose between these verbs. So we put "go" and "*aller*" in a parent class that has two representative SC-descendents: MOTION_WITHOUT_DEVICES with "*идти, gehen"* and MOTION_ON_DEVICES with „*examь, fahren'*. The choice between them depends on the semantic valencies expressed at a given sentence, their filling and statistics as well.

We claim that the tree of semantic classes is universal for the classification of all languages. It may certainly still look a bit contrastive. The fact is that we cannot simultaneously fill the hierarchy with a correct representative sample of meanings for both typologically similar and typologically different languages. But our successive description of Russian, English, Chinese, French and German has clearly showed that the structure of semantic classes underwent practically no important changes: cases of language-unique lexicalization lead us to adding low-level semantic classes.

Another problem concerning cross-language asymmetry is a phenomenon of semantic incorporation, so to say: under semantic incorporation we mean here cases like an English verb "*fish*" – "*to catch fish"*. Such incorporation is not universal and occurs within words with different meanings in different languages. Thus Russian lacks a verb like "fish", and intransitive usage of "fish" must be translated with two words – "*ловить-catch рыбу-fish*".

To solve this problem we create a SC TO_FISH with English LC "*fish*" and put the whole expression – "*ловить рыбу*" in the Russian part of the class. This verb can attach an [Object] slot as well – "*to fish [for trout]*", but its usage without the [Object] slot is also possible - in "*he is fishing*" the semantic valency of [Object] is not expressed explicitly and is incorporated in the semantic structure of the verb.

- lexical gaps and multiword expressions

SH is a dynamic database that can be revised (mainly on its lower SCs) and supplemented when we add new languages and have to describe culture-specific realities. For example, when describing the Chinese word "旗袍 qípáo" which denotes traditional Chinese body-hugging one-piece dress we create a new SC and fill it with corresponding loan-words in other languages – "*ципао*" in Russian, "*qipao*" and "*cheongsam*" in English.

If a language lacks the necessary loan-word and the translation requires the use of several words, we put the whole necessary expression in the SC. For instance, we created SC S_BAHN_RAILWAY for German-specific entity "urban railway". This SC is filled with LC S-Bahn in German, loan-word S-Bahn in English and a multiword expression "*городская*

железная дорога" ("*urban railway"*) in Russian as it is the only way to translate this word into Russian.

- language-specific challenges: some examples

Each language can have some peculiarities that require special attention in formal descriptions. Thus, we have elaborated consistent methodological guidelines for each language that take into account language-specific features to guarantee effective semantic and syntactic parsing.

For instance, upon adding German compounds to the SH, we consider whether their translation can be derived from their internal structure. If not, we add them to the SH into existing SCs or create new ones. For instance, the analysis of the compounds "*Geldautomat*" (*"ATM"*) and "*Straßenbahn*" (*"tram"*) is technically possible as there are lexical classes "*Geld*" *("money")*, "*Automat*" *("automat")*, "*Straße*" (*"street")* and "*Bahn*" ("*train"*) in the SH and there are semantic slots that can describe semantic relations between them. However, possible interpretations, e.g. "*der Automat mit Geld*" *("an automat with money"*) and *"die Bahn auf der Straße"* ("*a train in the street*") do not make any sense since they are not equal to the notions these compounds represent. So we add them to the SH into existing SCs or create new ones.

Possible disadvantage is that adding new languages, like German here, may demand the adding of new SCs to the SH as well, so the number of the universal SCs may grow to provide the necessary translation correlations. But the necessity of adding new SCs doesn't seem to cause any inconvenience for the model in general.

Another example: Chinese has a relatively strict word order and limited freedom to attach dependent constituents to the left or to the right of the head-verb. This often leads to asymmetry in the semantic model of Chinese and the semantic model of the target/source language. Thus, in order to translate a sentence with several dependent constituents attached to the head verb into Chinese we have to resort to one of the following transformations:

- to reduplicate a verb,
- to move a child constituent to another head, usually downwards a syntactic tree,
- to add another coordinated or dependent predication,
- to move a dependent constituent into a topic position.

Thus, it is essential for Chinese to provide 'negative' information in the verb LCs indicating which of semantic slots typical for the SC cannot be attached to the head and what type of transformation will be needed. For more details concerning Chinese-specific challenges and solutions refer to Manicheva et al., (2012).

4.2 Compatibility, semantic and syntactic model

Semantic relations between words are described in terms of **semantic slots** that partially correlate with the notions of Tesnière"s valencies (Tesnière, 1976), Fillmore"s cases (Fillmore, 1968), as well as with semantic and thematic roles in later theories. The key difference in the Compreno system is that most theories usually focus on verbal arguments only, underlining the difference between complements and modifiers, while in Compreno project we introduce the semantic slots for all possible semantic dependencies, more than 300 slots in total.

This means there are semantic slots for verbal actants (such as [Agent] in "*[the man] came in*" or [Possessor] in "*[I] have a pen*"), adjectival and adverbial modifiers (such as [Ch_Parameter_Dimensions] in "*[large] drops*" or [Ch_Evaluation] in "*[good] idea*"), circumstantial adjuncts (spatial or temporal, for instance, as [Time] and [Locative] in "*[yesterday] I saw him [in the street]*") and plenty of others.

Semantic slots are language-independent and get surface syntactic realizations (we call them **surface** or **syntactic slots**) in every language ([Agent] usually corresponds to the subject surface slot in an active mood and characteristical slots like the above-mentioned adjectival and adverbial modifiers are often expressed by attributive modifiers).

The semantic hierarchy is organized according to **inheritance** principle: many slots, especially the circumstantial ones like adjuncts or characteristics, are introduced on the upper levels and the child classes inherit them, as such constituents can be governed by almost any heads ("*an [important] person, book, meeting, work*" or "*[last year] she worked there/had this opportunity/was very rich*").

Other constituents, especially the arguments, are introduced on lower levels. For instance, verbs like "*have*" or "*possess*" need a [Possessor] slot while verbs like "*work*" or "*run*" do not have this valency as they have an [Agent]-subject. So the [Possessor] slot is introduced in the necessary semantic class only.

The inheritance principle means that most part of manual work is done on the initial stage of the description, when the core vocabulary is added to the SH, as words placed to the SH later inherit the most part of their semantic and syntactic model.

In different branches semantic slot can have different **status**: usually the **allowed** one, **normal** or **preferred**. For instance, the [Possessor] slot in "*[I] have a pen"* has the preferred status, while the [Possessor] slot in *"[my] pen"* has the normal status.
Each semantic slot can be **filled** with a fixed set of the semantic classes. I.e., [Possessor] is filled with beings, organizations and some territorial units: "*[my/our school's/Russia's] property"*, while slots for characteristics are filled with classes containing, for instance, adjectives and adverbs with corresponding semantics ([Ch_Evaluation] is filled with LCs like "*good, bad, excellent*", etc.).

The instantiation of semantic slots can be restricted to semantic classes. For instance, the [Object] slot can be filled with a wide range of vocabulary (*"to have [a cat/good health/an advantage]"*), but some verbs require additional constraints on filling: "*to read [a book]*", but * "*to read [a chair]*". Still, one can find marginal examples like "*I'll eat [my hat] if Kim ate [a motor-bike]"* (Soehn, 2005). For such cases, we define two sets of fillers: the allowed one and the preferred one. Thus, additional restrictions are normally imposed by further constraining the preferred fillers.

There are as well special cases of nontrivial compatibility, when a lexeme in some meaning can be combined with only one or several words. For example, we can say "*broad difference*" in the sense of "*big difference*" but can hardly say "*broad love*" in the same meaning. To describe this type of restricted compatibility A.K. Žolkovskij and I.A. Mel'čuk introduced a mechanism of

lexical functions (LF) in their "Meaning-Text Theory" (Žolkovskij, Mel'čuk, 1967 and later papers of the authors).

We have adopted the idea for Compreno system. If the descendants of some semantic class have such narrow compatibility, we declare this class to be a lexical function, mark the semantic slot where the narrowing is necessary, and indicate the fillers of this slot (the LF-arguments) for each LC-descendant of the semantic class. The arguments can be both the dependent or parent constituents. I.e., the SC GROUP_OF_ANIMALS is a LF and includes LCs like "*swarm*" or "*shoal*", the former usually combines with insects, the latter – with fish. Here "*swarm*" and "*shoal*" syntactically govern their LF-arguments ("*swarm [of insects]*", "*shoal [of fish]*") while in the example with "*[broad] difference*" "*broad*" is a dependent constituent.

The mechanism of LF proved to be an indispensable tool to describe classifiers and measure words in Chinese. Classifiers and measure words are used together with numbers to define the quantity of a given object. Different groups of nouns collocate with different classifiers:
yī bǎ yǐzi
一/把/椅子 - One chair (one + m. w. for objects with a handle + chair)

liǎng zhāng zhuōzǐ
两 / 张 / 桌 子 - Two tables (two+ m. w. for objects with flat surface + table)

4.3 Sense distinction and disambiguation problem

Sorting out meanings and positioning them in the SH is a controversial issue. On the one hand, we should describe them thoroughly and consistently in terms of the source language. On the other hand we need to correlate meanings with the material in other languages to ensure appropriate translation.

It often happens that dictionaries define several meanings of a word that can be actually added to the same SC in the SH or at least to the neighboring SCs. However, having homonyms that have no clear distinction expressed in mutually exclusive formal terms in closely-related classes is highly problematic. The choice of the necessary homonym becomes a problem and the number of hypotheses at the analysis stage grows. So the general principle of our lexicographic description is to merge homonyms with similar models and use other mechanisms to define the differences in translation (such as collocations, for instance).

Another key NLP problem is disambiguation. In most cases proper description of the semantic model of the word helps to distinguish its different meanings. For instance, we can understand that
(1) *I took to London,*
(2) *I took a book,*
(3) *I took a shower*
have different instances of "take" (from different semantic classes) as in the first sentence "take" has no [Object] slot which is obligatory for its usage in two other meanings, and we know that the example (2) can't have "take" in the meaning we have it in the example (3) as "take" from the third example evidently has rather narrow compatibility, so it is located in a LF-class and has narrow arguments thus.

Still, nothing in the semantic description prevents us from understanding *"take"* in sentence (3) as equal to "take" in sentence (2): indeed, sentences like *"I took the shower in my left hand"* are also possible. Here the statistical mechanism comes into play.

To describe a semantic model of a word and to differentiate its meanings we also use grammemes as well – for example, reflexivity or transitivity grammemes. Consider some French examples: "POSITION_IN_SPACE: trouver" ("*to be situated*") is used in a reflexive form only and thus has a grammeme <OnlyReflexive> ("*La maison se trouve à Paris*"– "*The house is situated in Paris*"), while "TO_SEEK_FIND : trouver" ("*to find*") is non-reflexive ("*J'ai trouvé un emploi*" – "*I have found a job*") or self-reflexive ("*Je me suis trouvé un emploi*". – "*I have found a job for myself*").

.

5. Compreno MLLD as basis for machine translation

Compreno MLLD serves as a lexical-semantic database for a rule-based MT system. Currently it provides a high quality machine translation for the EN<->RU language pair. It was also tested on a limited text material for GE<->RU and FR<-> RU language pairs. Below we briefly describe the translation process with a special focus on the processing of the semantic model.

When the program translates "*food*" from English to Russian, for example, the following operation is being done: we see the lexeme "*food*" which is in the corresponding English lexical class in the semantic class FOOD, go to the universal level – SC FOOD, and descend back to the necessary lexical class in the Russian language – "*еда*":
food => FOOD => еда

Important convenience is that generally when adding some new language (French, for instance) we do not have to describe French-Russian and French-English translation separately. We just add a necessary lexical class "*nourriture*" in French and thus get all the desired translation pairs (that"s an ideal situation though).

Of course, there is a lot of asymmetry between languages when such a straightforward translation is impossible. Let's consider some examples and illustrate briefly different mechanisms that can help (here we will just show different possibilities of the description without going into details and arguing where each of these mechanisms shall be chosen).

To treat cross-lingual asymmetry effectively, we have elaborated a wide range of universal instruments. Important tools related to the semantic module are 1) **collocations** and 2) **transformational rules**. Basically, both 1 and 2 represent a formalized description with conditions expressed in terms of SCs, LCs, semantemes, grammemes, semantic and/or syntactic slots and are aimed at setting exact correspondences between languages.

Collocations are used in more trivial cases, where the transformation of the structure is not very hard (usually to ensure the correct lexical choice or to set correspondences between different semantic models). Some collocations are written manually, other are gathered automatically. For instance:

(1) English construction *"Y-sized X"* must be translated in Russian like *"X размером с Y"* (the Russian variant roughly corresponds to the English *"X like Y in size"*), so we need a transformation of the structure here and add a collocation specifying all the necessary semantic, syntactic and grammar conditions for both languages. The collocation is written on a relatively high level of the SH as at least any entity can correspond to the X. Hence, we get proper translations for *"egg-sized hail"* <-> "*град размером с яйцо"* and etc.

(2) German prepositions like "*angesichts*" (*"in the face of*"), "*gegenüber*" ("*towards*") can correspond to noun phrases in other languages:
German: *Grausamkeit [gegenüber Object_Relation: Tieren],*
English: *cruelty [towards Object_Relation: animals] / cruelty [with Ch_Relation: respect [to Relation_Correlative:animals]],*
Russian: *жестокость [по Ch_Relation: отношению [к Relation_Correlative: животным]]* (a structure equal to the English one *"cruelty [with respect [to animals]"*).
Some collocations are gathered automatically, some are written by linguists.

Transformational rules are applied when the transformation is rather complicated, especially when the head of the constituent must be changed, or when dealing with regular cross-lingual asymmetry. Consider some examples:

(1) French expression "*l'ensemble [de x]*" means *"[all] x's*", i.e. "*l'ensemble [de messages]*" – "*[all] messages*". In French the variable [x] depends on "*ensemble*", while in English "*message*" becomes the head.

(2) In European languages numerals that go between thousand and million are counted by thousands, while in the numeral system of Chinese there is a special word for ten thousands – "万 wàn", and all the following numerals are derived from it. I.e., "百 万 bǎiwàn" (100 wans) stands for million, "二 十 五 万 èrshíwǔwàn" (25 wans) stands for 250,000.

Thus we have to add a new SC WAN_NUMBER to SH with a semanteme <<Rank_Wan>>. WAN_NUMBER is a descendent of the SC NUMBER along with other numeral units - TENS, THOUSANDS, etc. As we see, a direct translation through semantic classes is impossible, so we make the transformation with the help of a transformational rule that translates numerals over 9999 from/into Chinese through converting the numerals from one language into another.

Conclusion

The Compreno technology combines both multilingual lexical database and parser technology. It includes several levels of language description: the morphological, semantic, and syntactic ones, and possesses a wide range of powerful tools to describe lexicon and grammar of typologically different languages and establish correlations between them as well.

The universal and full description of the semantic models of the lexicon together with additional mechanisms like collocations, transformational rules and statistics allows to cope with the problems typical for NLP applications, i.e. the problems of language asymmetry and language polysemy.

The existing description shows that Compreno semantic model can serve as a universal integral framework for multilingual lexical databases and be successfully applied for different NLP tasks such as machine translation, text mining, information retrieval, fact extraction and other problems concerned with semantic analysis.

Furthermore, the English, Russian, German, French and Chinese lexical-semantic dictionaries can be studied from a cognitive perspective, as filling universal semantic hierarchy with language-specific vocabulary gives a vivid representation of the structure of language-specific vocabulary, lexicalization patterns and different conceptualizations of the world.

References

Altenberg, B. and Granger, S. (2002). Recent trends in cross-linguistic lexical studies. In B. Altenberg and S. Granger (ed.), *Lexis in Contrast*. Amsterdam/Philadelphia: Benjamins, pages 3-50.

Anisimovich K. V., Druzhkin K. Y., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K.A.(2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies //Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii. 'Dialog' 2012 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"], Bekasovo.

Baker, C. F., Fillmore C. J. and Lowe J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98*, Montréal, Canada.

Boas, H C. (2005). Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, Volume 18(4), pages 445–478.

Fillmore , C.J. (1968). The case for case. In *Universals in linguistic theory*, Bach, E. and Harms, R. (ed.), pages 1-90, New York.

Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, Vol. 280, pages. 20-32.

Manicheva, E. S., Dreyzis, Y. D., Selegey, V.P. Razrabotka leksiko-semanticheskogo slovar'a kitaiskogo yazika dlya mnogoyazichnoy sistemi analyza teksta [Development of Chinese language lexical-semantic dictionary for multilingual NLP system] 2012. Komp'juternaja lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoj konferencii. 'Dialog' 2012 [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"], Bekasovo.

McShane, M., Zabludowski, M., Nirenburg, S. and Beale, S. (2004). OntoSem and SIMPLE: Two multi-lingual world views. In *Proceedings of ACL-2004 Workshop on Text Meaning and*

Interpretation, Barcelona, Spain.

Nirenburg, S., McShane, M., Beale, S. (2004). The rationale for building resources expressly for NLP. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Nirenburg, S. and Raskin V. (2004). Ontological Semantics. Cambridge, MA: MIT Press.

Palmer, M., Dang, H. T., Fellbaum C. (2006). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. In *Natural Language Engineering*, Volume 13(2), pages 137-163.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge, pages 11-21.

Soehn, J.-P. (2005). Selectional Restrictions in HPSG: I'll eat my hat! In Stefan Müller (ed.), In *Proceedings of the HPSG-2005 Conference*, University of Lisbon, Portugal, Stanford: CSLI Publications.

Talmy, L. (2000). Toward a cognitive semantics. Cambridge, MA: MIT Press.

Tesnière L. (1976). Éleménts de syntaxe structural, Paris: Klincksieck, 1976.

Vossen P. (2004). EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. In *International Journal of Lexicography*, Volume 17 (2), OUP, pages 161-173.

Žolkovskij, A. I., Mel'čuk, I. A. (1967). O semantičeskom sinteze. Problemy kibernetiki, Volume 19, pages 177–238.