Baltic International Yearbook of Cognition, Logic and Communication

Volume 10 *Perspectives on Spatial Cognition* Article 5

2015

A Description Of Space Relations In An NLP Model: The ABBYY Compreno Approach

Aleksey Leontyev
ABBYY Moscow

Maria Petrova
ABBYY Moscow, Institute of Linguistics, RAS

Recommended Citation

Leontyev, Aleksey and Petrova, Maria (2015) "A Description Of Space Relations In An NLP Model: The ABBYY Compreno Approach," *Baltic International Yearbook of Cognition, Logic and Communication*: Vol. 10. https://doi.org/10.4148/1944-3676.1096

The Baltic International Yearbook of
Cognition, Logic and Communication

December 2015 Volume 10: *Perspectives on Spatial Cognition*
pages 1-25 DOI: http://dx.doi.org/10.4148/1944-3676.1096

ALEKSEY LEONTYEV
ABBYY Moscow

MARIA PETROVA
ABBYY Moscow, Institute of Linguistics, RAS

A DESCRIPTION OF SPACE RELATIONS IN AN NLP MODEL: THE ABBYY COMPRENO APPROACH

ABSTRACT: The current paper is devoted to a formal analysis of the space category and, especially, to questions bound with the presentation of space relations in a formal NLP model. The aim is to demonstrate how linguistic and cognitive problems relating to spatial categorization, definition of spatial entities, and the expression of different locative senses in natural languages can be solved in an artificial intelligence system. We offer a description of the locative groups in the ABBYY Compreno formalism – an integral NLP framework applied for machine translation, semantic search, fact extraction, and other tasks based on the semantic analysis of texts. The model is based on a universal semantic hierarchy of the thesaurus type and includes a description of all possible semantic and syntactic links every word can attach. In this work we define the set of semantic locative relations between words, suggest different tools for their syntactic presentation, give formal restrictions for the word classes that can denote spaces, and show different strategies of dealing with locative prepositions, especially as far as the problem of their machine translation is concerned.

1. INTRODUCTION

Space category and the expression of locative meanings in different languages have been widely discussed in linguistics, especially in cognitive studies (for example, Aurnague et al. 2007; Bloom et al. 1996; Hickmann & Robert 2006; Levinson & Wilkins 2006; Levinson 2003; Shay & Seibert 2003; Svorou 1994; Van der Zee & Slack 2003). In addition, there are several works that focus on the description of the locative dependencies for various NLP applications (such as Creary et al. 1989; Jørgensen & Lønning 2009; Oliver & Gapp 1998) or suggest machine translation models for different language pairs (for instance, Trujillo (1995) for the English-Spanish pair, Jørgensen (2004) for English-Norwegian, or Japkowicz & Wiebe (1991) for English-French).

As many studies have shown, the description of spatial domain requires an interdisciplinary approach and involves studies from different fields, including linguistics, cognitive psychology, artificial intelligence, and others. Key issues for the description of the space category in language, cognition, and artificial intelligence systems include both categorization of spatial relations and spatial entities as well as the elaboration of formal tools for the precise analysis of the data.

In the present paper we analyze the problems bound with the formal description of locative dependencies, which concern both semantics and syntax, and show how the spatial domain can be presented in a functional NLP model. Specifically, we demonstrate how an artificial intelligence system can deal with the problems of spatial categorization, of the definition of spatial entities, and of the expression of different locative senses in natural languages.

The current paper offers a description of the space relations in the ABBYY Compreno model, which is an integral NLP system aimed at solving a wide range of problems bound with the semantic analysis of texts, such as semantic search, fact extraction, or machine translation (for details, see Anisimovich et al. 2012). At present, the system functions for English and Russian and, in addition, partial models for German, French, and Chinese are available.

Compreno is based on the thesaurus-like universal semantic hierarchy, where the description of each word includes its positioning in the hierarchy and all semantic and syntactic links of the word (Manicheva et al. 2012; Petrova 2014). For example, the verb *walk* can attach valen-

cies such as [Agent] in *[the boy] walks*; locative adjuncts with various semantics such as *[in the street, around the room, from the park, to the station, along the road, two kilometers]* and so on; different modifiers such as *[fast, barefoot]*, and many other dependencies (*[on crutches, with a stick, with large steps, yesterday]*).

The model defines the necessary semantic relations for each dependency, which are called deep, or semantic, slots (DSs), about 300 slots in total. The notion of the DS is close to the concept of valency in L. Tesnière's dependency grammar theory (Tesnière 1976) or the deep case in Ch. Fillmore's case grammar theory (Fillmore 1968); unlike valencies, however, which are usually associated with actant dependencies only, DSs cover all possible semantic dependencies a word can have, as in the examples above.

An additional restriction is that each slot can be filled with words of the appropriate semantics only, namely, [Agent] can be filled with words denoting beings or organizations and locative adjuncts can be filled with words denoting places, such as physical or spatial objects.

The description of each language contains all possible syntactic realizations for each DS. For instance, [Agent] can correspond to the subject in the active voice (*[The boy] took the box*) or to the *by*-group in the passive voice (*The box was taken [by the boy]*). These syntactic realizations are expressed in the form of syntactic, or surface, slots (SSs), for example *$Subject* or *$Object_Indirect_By*.

DSs are supposed to be universal for all the languages included in the model. SSs, in turn, are special to each natural language.

The hierarchy is organized in accordance with the inheritance principle: most DSs (for example, the [Locative] slot and slots for various modifiers such as *good, beautiful, equal*) are introduced high in the hierarchy, and words of lower levels inherit them. Therefore, one does not have to describe hundreds of possible dependencies for each word – most of the work is done on the upper branches of the hierarchy.

The description of locative adjuncts is of special interest for the model, as locatives differ significantly from the majority of other dependencies. We have already mentioned some of the problems bound with a formal description of the space category in Leontiev & Petrova (2014), where we briefly characterized the formal presentation of several locative adjuncts and showed the descriptive opportunities that the given NLP model suggests.

The current paper is aimed at giving a more detailed description of the space domain, covering most of the locative adjuncts, and at analyzing how problems bound with the cognitive presentation of the space category can be solved in an NLP model. Such analysis includes the formal structuring of the space semantic field, a description of locative syntactic realizations in different languages, and a comparison of the locative and non-locative adjuncts (especially the temporal ones), as far as their semantic and syntactic parallelism is concerned.

The semantic field of space is rather complicated. First, the domain of locative relations includes groups with different semantics, such as where-groups (*lie [under the table]/live [in England]*); groups with the meaning of the initial and final point (*look [from the window]/come [from abroad]* and, correspondingly, *put [into the box], go [to school]*); route-groups (*walk [along the street/across the road]*); and distance-groups (*walk [two miles]*). We thus have to introduce different locative DSs for different semantic relations, such as [Locative], [Locative_InitialPoint], or [Locative_Route].

Second, each locative adjunct includes numerous syntactic realizations through prepositions with different semantics (*under/on/in/at the box*), and the variety of such prepositions distinguishes locative DSs from most other DSs (excluding, primarily, temporals). Usually, the semantics of each DS is rather narrow, and all possible surface correspondences of a DS are more or less synonymous (for example, *[wooden] furniture* vs. *furniture [of wood]* or *[the boy's] pen* vs. *the pen [of the boy]*). Therefore, a formal model needs to elaborate additional mechanisms for distinguishing different senses within each locative slot as well.

Third, locative adjuncts also include groups such as *in the country* and *on the island*, that is, groups with different prepositions but similar semantics, as *in* and *on* in examples such as *in the country* and *on the island* do not express the same difference as they do in the examples *in the box* and *on the box*. The combination of *in* and *on* with *country* and *island*, respectively, seems to be idiomatic. Here the choice of the preposition is determined not through the preposition's semantics only; rather, it is the core noun that determines the choice of the preposition. For a formal model this means that such prepositions should be treated

differently, as we will argue in section 3.

The fourth problem is to define the set of words which can be the fillers of locative DSs, as the border between the locative and non-locative groups is not always easy to determine. Usually, locative adjuncts correspond to physical objects and spaces, e.g., *country, forest, house*, or *table* can fill the locative DSs. However, there are also nominal groups with similar semantics and surface realizations but which do not denote spaces in a literal sense, for instance, *on the Internet, on TV, at the demonstration*, or *in a meeting*.

On the one hand, it seems reasonable to regard such groups as locative adjuncts because of their semantic proximity and syntactic properties. On the other hand, the distinctions between these groups and the locative groups are significant. In terms of semantics, we are not speaking of a place but are rather using metaphor or metonymy: for instance, in the sentence *He is in a meeting now* the event functions as the place where it occurs. In terms of syntax, *on the Internet* and *in a meeting* cannot be used with locative prepositions such as *under, above, near* (in the locative sense), and so on, and these restrictions should be indicated in a formal model.

In the next section we give a detailed semantic description of the space category in the current NLP model. Namely, we define the set of locative DSs and consider what groups of words can be used as their fillers. In the third section we focus on the syntactic part of the locative model and suggest two different approaches to the description of so-called "semantic" prepositions (as in groups such as *in/on the box*) and "default" prepositions (as in groups such as *in the country/on the island*). The fourth section is devoted to the borders between the locative and non-locative groups and considers different cases of the overlap between the locative, temporal, and sphere adjuncts in both their semantics and syntax. The conclusion offers a short summary and determines further perspectives.

2. THE SEMANTICS OF SPACE RELATIONS IN THE COMPRENO NLP MODEL

The semantic description of the space dependencies faces two main issues. The first is to define different locative relations, as there are locative relations with different meanings. In the given formalism, it means to define the set of the necessary DSs as well as to introduce various semantic slots for various locative semantic relations. Our semantic locative model consists of several "basic" DSs:

- [Locative] slot for where-groups, such as the examples in square brackets in (1) (most of the examples we give here and below are taken from the real text corpora, which contain a wide range of texts, such as fiction, different documents, scientific texts and many others; however, sometimes we use the shortened sentences for the simplification of the description):

(1) She was awakened the next afternoon by a clatter [in the street].
Sticker [on the box].
The coach lurched and accelerated again; they were [in France] now.
The boy, keeping his right hand [under his coat], looked down.

- [Locative_InitialPoint] slot for groups denoting initial point:

(2) I came [out of my house].
I drew a breath and a spider fell [from the ceiling].
Remittances [from abroad].

- [Locative_FinalPoint] slot for groups denoting final point:

(3) They don't go [to school] on the weekend.
He took his cash card out of his wallet, and put it [into the cash machine].
I came [home] from the beach.

- [Locative_Route] slot for groups with the route semantics:

(4) He walked [along the road] in the darkness.
Now I can't run [across the street].
I looked [through the window].

- [Locative_Distance] slot for groups denoting distance:

(5) The trio stopped [three to four hundred meters away from the main traffic route].
He walked [quite a long distance].
We had to run [about six miles].

As in the case for all DSs in the current model, each of the loca-

tive DSs is filled with a strict set of words, namely, [Locative], [Locative_InitialPoint], [Locative_FinalPoint], and [Locative_Route] can be filled with words denoting spaces and physical objects, as in the examples above, and the [Locative_Distance] slot can be filled with the necessary units of measure and words such as *distance*.

When analyzing the sentence *The boy walks home*, for instance, the system, roughly speaking, checks whether the verb walk has a [Locative_FinalPoint] slot in its model, and whether home is included in the filling of the [Locative_FinalPoint] slot (in fact, the analysis process is more complicated, and we will consider it in more detail in section 3).

The restriction on how the DSs can be filled is an important feature of the model, as it reduces significantly the number of possible hypotheses at the analysis stage. At the same time, it helps to deal with homonymy and to differentiate between various homonyms as well as between different homonymic constructions. For instance, let us take sentences (6) and (7):

(6) He was in the car.

(7) He was in anger.

The *in*-group in (6) has a locative meaning, whereas the *in*-group in (7) has nothing to do with the locatives. The parser can easily differentiate between these cases, as *car* is included in the filling of the [Locative] slot and *anger*is not.

Thus the second question arises: which words should be referred to as fillers of the locative DSs and which should not? (Similar problems on the nature of spatial entities in language and cognition are discussed in Aurnague et al. (2007a) and in other papers in Aurnague et al. (2007).)

There are some evident cases, for instance, *live [in England]* or *put [into the box]* are definitely locative adjuncts, and *be [in anger]* or *achievements [in medicine]* are definitely not. But there are also groups such as *read [on the Internet]*, *see [in one's head/imagination]*, or *be present [in the meeting/at the demonstration/at the rehearsal]*. Consider the examples in (8) and (9):

(8) At least [in your imagination], complete it!
I have read [on the Internet] that it has been sold in the UK for 3 years already, but haven't seen it yet.

(9) If you were [at the presentation] and have further questions, don't hesitate to get in contact with us!
A chair was set for her on the stage [at the rehearsals].

Examples like (8) do not mention physical places in a literal sense, but rather refer to some kind of metaphoric spaces, and in examples like (9) different events function as places where they occur. Both cases are close to the locative groups, as all of them describe some kind of space, but they are not the same as the "usual" locatives: at least, these groups generally cannot be used with prepositions such as *under* or *above* in the locative sense, for example, *be present in a meeting*, but not **be present under a meeting*. Only usage with the so-called "default" locative prepositions is possible for such nouns.

This means that we have to differentiate between "usual" or "core" locative adjuncts and "peripheral" cases like those illustrated in (8) and (9), because in a formal model one has to set the necessary restrictions for locative prepositions in such "peripheral" cases.

For this reason we have introduced two additional groups of locative DSs which are close to the locative slots in their semantics but differ through their filling and possible syntactic realizations. Specifically, we have introduced the [Metaphoric_Locative] slot for examples like (8) and [Locative_Event] slot for examples like (9).

[Metaphoric_Locative] is filled with words such as *imagination, memory, dream, Internet, book, document*, and so on, that is, with words denoting various types of "informational storage", for example, a person's internal world or some sort of printed matter. (Calling this type of the locative DSs "metaphoric", we use "metaphoric" to denote such "informational storages" only, not the metaphor in a wide sense.)

As indicated in Leontiev & Petrova (2014), [Metaphoric_Locative] includes fillers which are absent among the fillers of the [Locative] slot, such as *imagination* or *Internet*, and fillers which are present in both sets of fillers, [Locative] and [Metaphoric_Locative], such as *book* or *head*. These groups can be analyzed through both DSs, although the sense would be different: for instance, in (10) *book* functions as a kind of informational storage, therefore, *[in the book]* is a [Metaphoric_Locative] here, whereas (11) is an example of [Locative]:

(10) It is written [in the book].

(11) A dollar bill [in the book] will increase its value.

The fillers of the [Locative_Event] slot are words denoting different events, for example *meeting, conference, presentation, exhibition, rehearsal, demonstration, lesson*, and so on, which can often be used to indicate places of their occurrence. The fillers of [Locative_Event] do not overlap with those of [Locative] and [Metaphoric_Locative].

Words that fill the [Metaphoric_Locative] and [Locative_Event] slots can also be used in the groups with initial and final point semantics, as in examples (12)-(13) and (14)-(15), respectively:

(12) It was as if that knowledge had been erased [from my memory].

(13) They were returning [from the demonstration] on 10 June.

(14) The function generates a random password for the user and puts it [into the database].

(15) Once, by chance, I came [to the exhibition of rare cars].

By analogy with the [Locative_IntialPoint] and [Locative_FinalPoint] slot, we have introduced the [Metaphoric_InitialPoint] slot for cases like (12), [LocativeEvent_InitialPoint] for (13), [Metaphoric_FinalPoint] for (14), and [LocativeEvent_FinalPoint] for (15).

All these slots are grouped into several classes according to:
a) the semantics of the locative relation a DS expresses, and
b) the filling of a DS.
This is shown in Figure 1:

[Locative_Class]
- [Locative]: *book [on the shelf]*
- [Metaphoric_Locative]: *read [in the book]*
- [Locative_Event]: *be [in a meeting]*

[Locative_Route_Class]
- [Locative_Route]: *walk [along the street]*

[Locative_InitialPoint_Class]
- [Locative_InitialPoint]: *come [from abroad]*
- [Metaphoric_InitialPoint]: *erase [from memory]*
- [LocativeEvent_InitialPoint]: *come [from the meeting]*

[Locative_Distance_Class]
- [Locative_Distance]: *walk [a long distance]*

[Locative_FinalPoint_Class]
- [Locative_FinalPoint]: *return [home]*
- [Metaphoric_FinalPoint]: *put [into the database]*
- [LocativeEvent_FinalPoint]: *go [to the meeting]*

Figure 1: Classes of the locative deep slots.

In fact, these classes include several other DSs as well, but they are organized on the same principles and concern rather marginal cases, so we will exclude them from our consideration here.

To facilitate the description of a DS's filling, we sometimes use *distributional semantemes* (for more information on semantemes see Anisimovich et al. 2012); the term *semanteme* here denotes a semantic informational unit. We have used the term by analogy with K. Pike's "grammeme" (Pike 1957), which refers to units of grammatical information. For example, *food* or *medicine* have the semanteme «Eatable», whereas *mile* or *kilogram* are marked with a «Unit_Of_Measure» semanteme (so "semanteme" is not used here as widely as in Mel'čuk (2012: 37), where it relates to the meanings of lexical units as well).

Usually, to set the filling of a DS, we just enumerate the necessary branches of the semantic hierarchy, for instance, the [Experiencer] slot is filled with beings, organizations, and countries (or other administrative units). As all beings are gathered in one branch, and the same is true for organizations and administrative units such as countries or cities, we indicate that the [Experiencer] slot can be filled with the following branches (the names of the branches are printed in capitals): BEING, ORGANIZATION and ADMINISTRATIVE_AND_TERRITORIAL_UNIT.

But there are also DSs, including locatives, that can be filled with branches from different parts of the hierarchy; thus *Internet*, *document*, and *imagination*, which fill [Metaphoric_Locative], are positioned in various branches. In such cases it is more convenient to mark the necessary words with special semantemes, which indicate that a word (or a branch of the hierarchy) can be the filler of some (locative) DS. It allows one to avoid enumerating a large number of small branches but rather to indicate one large branch and restrict it with the necessary semanteme. Thus we have marked the possible fillers of the [Locative] slot with a «Place» semanteme, the fillers of [Metaphoric_Locative] with «MetaphoricPlace», and the fillers of [Locative_Event] with «EventPlace».

Therefore, the filling of the [Locative] slot includes the branch ENTITY with the «Place» semanteme. This branch, in turn, consists of different descendants: physical objects (such as *table* or *bag*), mental objects (such as *idea*, *thought*, or *opinion*), abstract and scientific ob-

jects (such as *formula* or *logarithm*), countries, organizations, and so on. Since not all of these branches can be used as locatives, we use the «Place» semanteme to mark only the branches that can be used as locative identifiers, for instance, we assign «Place» to all physical objects but do not use it to mark mental or abstract objects.

Now let us discuss the syntactic part of the locative description, and then analyze some cases of the correlation between the locative and non-locative adjuncts, including both semantic and syntactic aspects.

3. THE SYNTAX OF SPACE RELATIONS IN THE COMPRENO NLP MODEL

The syntactic description in our system is similar to the semantic description. The notion of SS is close to a syntactic valency in modern linguistic theories, therefore, an SS is an approximate analogue of terms such as complement, specifier, or adjunct. Unlike DSs, SSs are closer to the input text, as they appear in earlier stages of analysis. The key restriction for a DS is its semantic filling, whereas SSs are restricted grammatically, that is, the definition of an SS is based on the following features:
- *Government*, or the grammeme restrictions that a constituent must satisfy in order for one to analyze it through the SS. For example, grammemes specify case forms and prepositions, that is, we use the grammeme of null preposition in the government of the English $Object_Direct slot (as in *I see [the picture]*) or indicate the *through* preposition in the government of the $Object_Indirect_Through slot (as in *He spoke [through an interpreter]*);
- *Linear order* that describes the linear positions, where the SS is allowed. For instance, the linear order for the $Object_Dative slot in English does not allow the leftmost position in a sentence (*She gave [him] a book*, but not **[Him] gave she a book*), whereas for the $Adjunct_Concession_Clause slot this position is quite normal (*He came [although it was difficult]* vs. *[Although it was difficult], he came*);
- *Punctuation* that describes the punctuators (comma, bracket, semicolon, and so on), which are allowed in the SS.

Therefore, the output of the parser is a dependency tree, where each arc is marked with a label corresponding to the DS/SS pair, as shown in Figure 2.

When analyzing the sentence *The man came to Africa by sea*, for instance, the parser does the following:
- it defines what SSs the verb *come* can attach ($Subject for the [man]-dependency, $Adjunct$_F$inalPoint for the [Africa]-dependency, and $Adjunct_Route for the [sea]-dependency);
- it determines what DSs can correspond to these SSs in the model of *come* ([Agent] slot for the $Subject SS, [Locative_FinalPoint] slot for the $Adjunct_FinalPoint SS, and [Locative_Route] for the $Adjunct_Route SS);
- it checks whether *man* is included in the set of fillers of the [Agent] slot, *Africa* in the set of fillers of the [Locative_FinalPoint] slot, and *sea* in the set of the [Locative_Route] slot.

Figure 2 is an illustration of the analysis process for the sentence *The man came to Africa by sea*:

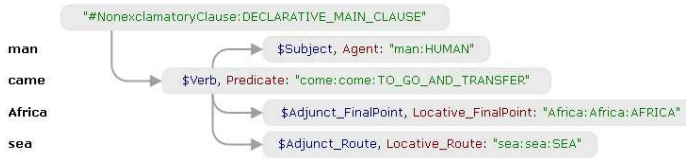

Figure 2: Semantic and syntactic analysis in the ABBYY Compreno model.

In an ideal case, each constituent should be analyzed through one SS only. (If there are fewer SSs available for a single constituent, the parser builds fewer hypotheses at the analysis stage.) However, in practice it is not always possible to avoid overlap between different syntactic positions, as there can be, for instance, SSs with similar government but different word order. Locative adjuncts are a good example here. The absolute majority of the locative prepositions can be used in non-locative contexts as well, for example, the preposition on marks the locative group in (16) and the non-locative group in (17):

(16) He stood [on the hill] alone.

(17) Research [on the locative dependencies].

However, there are some syntactic differences between the locative and non-locative usage of the *on*-preposition. First, locatives can occupy the leftmost position, and this usage does not seem emphatic, as

in example (18). For the majority of non-locative groups, this position is either emphatic or scarcely admitted, as in (19):

(18) [In Hertford, Hereford, and Hampshire], hurricanes hardly ever happen.

(19) *[In locative adjuncts] this problem consists.

The second difference concerns the usage of relative pronouns, specifically, non-locative groups, unlike locative groups, never use *where*-relativizers. For example, *where*-relativization is impossible in (20), where *on* marks the theme valency, yet it is fine in (21), where *on* is a part of the locative group:

(20) The issue you advised me on. vs. *The issue where you advised me.

(21) He lived on the roof. vs. The roof where he lived.

Therefore, there is evidence that locative dependencies require separate SSs. Introducing such SSs is an extra burden for the parser, as it increases the number of hypotheses at the analysis stage, but at the same time, it allows for the solution of other problems of locative description, which we will focus on below.

The current model uses the following locative SSs for each language: $Adjunct_Locative for the [Locative_Class], $Adjunct_FinalPoint for the [Locative_FinalPoint_Class], $Adjunct_InitialPoint for the [Locative_InitialPoint_Class], $Adjunct_Route for the [Locative_Route_Class], and $Adjunct_Distance for the [Locative_Distance_Class]. However, [Locative_Route] and [Locative_Distance] seem to be less homogeneous than the former classes, and their syntactic description has to include several other SSs as well.

Let us now consider first the syntactic description for the [Locative], [Locative_InitialPoint], and [Locative_FinalPoint] classes and, after this, discuss the more complicated cases of the [Locative_Route] and [Locative_Distance] classes.

3.1. A syntactic description of the [Locative], [Locative_InitialPoint], and [Locative_FinalPoint] classes

As we have stated above, locative adjuncts include locative adverbs (like *below*), adverbial pronouns (like *here* or *whence*), and a large

variety of prepositions (like *on*, *in*, *at*, *under*, and so on). These instances are heterogeneous and, furthermore, there are many lexicalization cases (for example, *in the country* and *on the island*). To take all the necessary expressions and restrictions into account, we have introduced a new grammatical category named *FormOfLocativeCircumstance*, which consists of the following values:

Grammemes allowed in $AdjunctLocative:
- DefaultLocativeLikeForm,
- SemanticLocativeLikeForm.

Grammemes allowed in $AdjunctInitialPoint:
- DefaultFromForm,
- SemanticFromForm.

Grammemes allowed in $AdjunctFinalPoint:
- DefaultToForm,
- SemanticToForm.

Every possible filler of the locative slot (in other words, each word that can fill a locative slot) is provided with the necessary grammemes from the category. For instance, to indicate that the word *table* can fill the $Adjunct_Locative slot with the preposition *on*, one has to indicate the preposition *on* in the DefaultLocativeLikeForm pattern of the word *table*.

3.2. "Default" and "semantic" prepositions

As shown above, each adjunct includes two grammeme patterns: the "default" form and the "semantic" form (which have already been discussed in Leontiev & Petrova (2014)). The semantic pattern covers prepositions such as *under*, *behind*, or *near*: their locative semantic interpretation is not determined by the noun they modify, and these prepositions have exact counterparts in other languages, which can serve as translation analogues in all the contexts (for example, English *near* and Russian *okolo*). Usually these prepositions denote one of the peripheral spatial localizations like AD or APUD (in terms used in Plungian 2000).

Unlike the semantic prepositions, default prepositions form collocations with the nouns they modify. This means that different nouns demand different locative prepositions, and the choice of the preposition here is highly lexicalized and language-specific. Such prepositions correspond to IN-Localization in Plungian's terms.

For example, the English prepositions *in*, *on*, and *at* in groups like *in the country*, *on the island*, and *at the pole* denote the same localization. It should be pointed out that "default" vs. "semantic" opposition is not an intrinsic trait of the preposition itself. One and the same preposition can be "default" in collocation with one word (for instance, *in the city*) and "semantic" in some other context (for instance, *in the cupboards*).

In a formal model it is convenient to describe these prepositions differently.

As there is no significant difference in the usage of the semantic prepositions with different nouns, it is better to introduce the necessary pattern for them at the highest levels of the hierarchy, since one usually does not have to modify it lower in the hierarchy.

The default prepositions, on the contrary, should be indicated for specific lexical units. Therefore, one has to define the default pattern for the lower levels. Furthermore, the default prepositions, unlike the semantic ones, can correspond to DSs such as [Metaphoric_Locative] and [Event_Locative], as these positions denote a virtual, not a real, space, where the localizations like APUD are irrelevant.

This two-pattern approach allows us to avoid problems in the description of these restrictions. Moreover, the two-pattern model proved to be efficient for the correct translation of locative prepositions as well.

The sense of semantic prepositions is saved through the system of special semantemes and transfer rules (which are discussed in Anisimovich et al. 2012; Bogdanov & Leontyev 2013; Leontiev & Petrova 2014): each preposition corresponds to the necessary semanteme (for example, the preposition *under* corresponds to the semanteme «Under»), and the calculated semanteme demands the necessary preposition at the synthesis stage. Consider example (22):

(22) The tea was in the cup.
Chaj byl v chashke.

Here the preposition *in* is semantic. Its sense is rendered by a special semanteme, «Inside», which evokes the corresponding preposition at the synthesis stage directly, without reference to the locative patterns.

The mechanism for translating the default prepositions is different. A preposition of this type is ignored, and the input for the transformational rules is the DefaultLocativeLikeForm grammeme. According to this form, a special semanteme, «Default_Location», is computed, which, in turn, demands the necessary default preposition in the target language at the synthesis stage.

Using this approach, we can get the correct Russian translations for the English sentences (23) and (24):

(23) *The Statue of Liberty is in New York.*
Statuja Svobody stoit v Nju-Jorke.

(24) *The sun rises in the East.*
Solnce vstajet na Vostoke.

In the English sentences (23) and (24) one preposition, *in*, is used, whereas the Russian translations demand different prepositions: *v* and *na*. In these examples the preposition in is the default, so it corresponds to the special semanteme «DefaultLocation». At the stage of building the output Russian structure, this semanteme will not evoke any concrete preposition but rather a link to the default locative pattern, DefaultLocativeLikeForm, where the preposition *v* is indicated for the noun *Nju-Jork* "New York" and the preposition *na* for the noun *Vostok* "East". This means that the resulting preposition depends on the default locative pattern only and does not correspond directly to any preposition in the input structure.

3.3. A syntactic description of the [Locative_Distance] and [Locative_Route] classes

The [Locative_Distance] and [Locative_Route] DSs are less homogeneous, which influences their syntactic description as well. The [Locative_Distance] slot can be divided into two cases: locative-like adverbs (25) and nominal adjuncts (26):

(25) He is [far away].

(26) He ran [60 miles].

Locative adverbs denoting distance do not differ syntactically from

other locative adverbs, such as *up* or *down*, and, therefore, they can be analyzed through the $Adjunct_Locative SS, which is used for the ordinary locative dependencies. As for cases like (26), it would be better to describe them through an SS, where all the conditions are indicated directly in the government of the SS, since the range of nouns allowed in examples like (26) is not wide, and this group of nouns is rather homogeneous, as most of them denote units of measure and distance.

The [Locative_Route] position is a more complicated case. This slot allows different prepositions, some of which are highly lexicalized whereas others are not. Examples (27) and (28) demonstrate the lexicalized usage of the *by*-preposition (27) and the free compatibility of the preposition *through* (28):

(27) They travelled [by air] / * [by hole].

(28) They flew [through the air] / [through the hole].

Furthermore, the *through*-dependency can be attached to many cores for which the by-dependency in the route meaning is semantically impossible. For example, the verb look in sentence (29) allows only *through*-dependencies in the [Locative_Route] slot, whereas verbs of motion such as *walk* in (30) usually can combine with a full set of route prepositions:

(29) He looked [through the window]. / * He looked [by the window].

(30) He walked [through the building]. / He walked [by the building].

For these reasons, we have split the $Adjunct_Route SS into several SSs: the $Object_Indirect_Through slot is set as a syntactic correspondence for the [Locative_Route] DS in all the necessary branches without any restrictions on the filler itself, whereas the $Adjunct_Locative SS, which includes the by-preposition as well, is organized according to the two-pattern principle.

As one can see, the syntactic model includes different descriptive opportunities, permitting us to choose and combine different instruments in order to make the description of different adjuncts most convenient.

4. CORRELATION AND OVERLAP BETWEEN THE LOCATIVE AND NON-LOCATIVE GROUPS

As we stated in section 2, it is not always clear where to make the border between the locative and non-locative groups, as semantically this border seems to be rather vague. In addition, there are also groups which are not locative according to their semantics, but their syntactic behavior is very close to that of the locative domain. In the first part of the current section, we analyze some cases demonstrating the overlap between the locative and non-locative DSs, and in the second part, we focus on non-locative DSs that have the same syntactic description (and the same SSs) as locative DSs.

4.1. Semantic borders between the locative and non-locative groups

We have widened the locative domain by including in it the so-called metaphoric and event locatives. Such a description, however, has both advantages and disadvantages. On the one hand, it is convenient to draw parallels between "usual" locatives and these peripheral cases, but on the other hand, it causes several problems, especially as far as the event locatives are concerned.

The first problem deals with the universality of spatial concepts, as it is not always clear whether the concepts underlying spatial entities in language and cognition are universal or whether they depend on individual and cultural factors (Aurnague et al. 2007a, 5).

Event locatives are a contentious issue in this respect, as the expression of the event locatives in different languages is highly lexicalized. Namely, there can be a locative-like group in one language and no locative group for the same sense in another. For instance, in English and Russian there are groups like *in a meeting – na soveshchanii* or *at the demonstration – na demonstracii*, which are close to locatives in both languages. But there are as well cases like those illustrated in examples (31) and (32):

(31) Dazhe na ohote on nikogda ne ispytyval takogo straha.
even at hunt he never not experienced such fear
Even while hunting he had never been so afraid.

(32) On byl na rybalke.
he was at fishing

He was fishing.

In (31) and (32) locative-like adjuncts with the verbal nouns *ohota* ”hunt” and *rybalka* “fishing” are possible (and frequently used) in Russian but would scarcely be used in English, where the same sense is more likely to be expressed through other means, such as the temporal group in (31) and the verb “to fish” instead of the combination “be + nominal group” in (32). So a correction of the structures is needed here when translating (such transformational rules are briefly characterized in Manicheva et al. 2012).

Second, such groups are semantically close to the temporal adjuncts as well, as example (31) shows, that is, in most cases the transformation of the event locative group into the temporal group is possible, for example, “*The film was demonstrated at the presentation*” ⇒ “during the presentation”. Therefore, semantically, it is problematic to make a strict border between the locative meaning of such groups and their temporal usage.

Actually, temporal analysis would be sufficient in most cases. The necessity of turning to locative description occurs, first of all, for verbs which have a locative valency as an obligatory slot, and it concerns not only where-locative adjuncts, but the locatives of the initial and final point as well. For example:

(33) He is [at the meeting].

The model of *be* in the “position” meaning (when the verb denotes the location of an object in a particular place) in (33) demands an obligatory locative valency, such as [Locative] (*He is [here]*), [Metaphoric_Locative] (*It is still [in my memory]*), or [Locative_Distance] (*He was [two meters behind me]*). If no locative-like valency is available for cases such as *in a meeting/at the exhibition*, the interpretation of such examples becomes problematic. For the formal model this means that it would be difficult to distinguish be with the position meaning from other *be*-homonyms.

A similar situation is shown in examples (34) and (35):

(34) He left [the building]. / He went away [from the building].

(35) He left [the meeting]. / He went away [from the meeting].

Both examples in (34) have the same semantic model, and the DS for *building* is [Locative_InitialPoint] in both cases (the verbs *leave* and *go away* differ through the SSs, which correspond to the [Locative_InitialPoint] slot only).

The semantic structure of the sentences in (35) seems equal to the sentences in (34), as the semantic relation between the core verb and *meeting* can also be determined as the locative of the initial point, so the introduction of the [LocativeEvent_InitialPoint] slot seems to be an appropriate decision here.

Examples (36) and (37) demonstrate the same relations for the locatives of the final point:

(36) He visited [the building]. / He came [to the building].

(37) He visited [the meeting]. / He came [to the meeting].

It seems reasonable to provide these sentences with the same semantic models and describe both *visit [the building]/come [to the building]* and *visit [the meeting]/come [to the meeting]* as the locatives of the final point, namely, as [Locative_FinalPoint] in (36) and [LocativeEvent_FinalPoint] in (37).

There are also parallels between the locative groups and some other groups. For instance, DSs with sphere meaning (for example, *achievements [in medicine], he works [in the field of cultural policy]*) are also sometimes expressed through groups that look like the initial and final point groups, as happens in cases of metaphorical shifts of motion verbs:

(38) The management has already confirmed that they will exit [from the real estate business].

(39) Fifty years ago when I came [into science], we rarely talked about ethical issues.

Nevertheless, such cases do not have locative semantics, and the words that fill these groups usually do not refer to spaces. Therefore, the models of such verbs require other DSs, not the locative ones. However, syntactically, non-locative groups like these and locative groups can share the same SSs, as will be shown in the following section.

4.2. Application of the two-step locative syntactic model to non-locative cases

The two-step model used for the locative domain proved to be an effective solution for cases of prepositional lexicalization, especially as far as their collocational usage is concerned. Therefore, this approach has been transferred to some cases outside the locative domain.

For example, we were able to apply the same syntactic model to describe the [Sphere] DS (as in the groups like *achievements [in medicine]* mentioned above). First, the surface realization of the [Sphere] slot includes lexicalized prepositions, as occurs with locative groups. For instance, see the Russian sentences (40) and (41):

(40) On rabotal v sfere obrazovanija.
he worked in sphere education
He worked in the sphere of education.

(41) On rabotal na nive prosveshchenija.
he worked on field education
He worked in (literally: on) the field of education.

The Russian noun *sfera* "sphere" demands the v-preposition ("in") in the [Sphere] context in sentence (40), whereas the noun *niva* "field" demands the *na*-preposition ("on") in sentence (41).

Second, [Sphere] groups can occur in the leftmost position, like the locative groups, as shown in example (42):

(42) [In this area], you have a few choices.

Third, [Sphere] groups allow where-relativization as well, as in (43):

(43) The area where he worked.

Therefore, it turned out to be rational to use locative SSs for these examples. It did not require much effort: we just had to add a *DefaultSphereForm* pattern to the list of the locative patterns and a special semanteme which is computed for it.

Another area in which one might apply the two-step strategy is the temporal domain, since in the temporal adjuncts the choice of the preposition can also be highly lexicalized. Groups in square brackets in (44) and (45) are temporal adjuncts, and the choice of the preposition here is determined by the core of the temporal group, *2000* in (44) and *Eve* in (45):

(44) He was born [in 2000].

(45) He was born [on Christmas Eve].

Moreover, there are certain semantic restrictions on the compatibility of different temporal prepositions with different nouns. For example, prepositions such as *before* or *after* are easily combined even with animate beings, as in (46):

(46) [Before Marx] no one thought about this problem.

Prepositions denoting time point, such as *in* and *on* in (44) and (45), respectively, do not allow such compatibility in the temporal meaning; their temporal usage is possible with nouns denoting time periods or temporal points only.

The two-pattern model provides an elegant solution to this problem. We have created different temporal patterns for different groups of prepositions. Therefore, the pattern for *before* and *after* is introduced high in the hierarchy, whereas the pattern for time point is introduced only in the branches of the temporal and situational nouns.

5. CONCLUSION

In the current paper we have analyzed different aspects of the space domain and problems bound with its cognitive presentation in a formal NLP model. We have defined different locative semantic relations and presented them in the form of locative deep slots, which can be filled with a strict set of words with appropriate semantics. DSs have surface syntactic correspondences in natural languages – SSs, which include grammatical information of the constituent.

We have also widened the boundaries of the locative semantic field by including metaphoric and metonymic cases, for example, *read [on the Internet]* and *be present [in a meeting]*, as such cases are close to the locatives both in their semantics and their syntactic realizations. The important distinction is that nouns like meeting or Internet have some restrictions on the locative prepositions they combine with, namely, in the locative groups they allow only the "default" locative prepositions, not the "semantic" ones *(read [on the Internet]* and *be present [in a*

meeting], but not **under the Internet* or **above a meeting*).

To take this into account, we have introduced additional locative DSs in the model (for example, [Metaphoric_Locative] or [Locative_Event]), which have the same semantics as [Locative] but differ through their filling and syntactic realization.

Furthermore, we have proposed different mechanisms for dealing with "default" and "semantic" prepositions, which helps not only to restrict the prepositional compatibility of slots such as [Metaphoric_Locative], but also to provide a synthesis of the proper prepositions for the machine translation. This description is based on a two-step strategy, which implies the use of special grammeme patterns instead of giving direct links to the prepositions. This strategy proved to be efficient for complicated cases of prepositional lexicalization in the locative domain and in some cases beyond it (namely, for the description of temporal and sphere adjuncts). However, it lays an additional burden on the parser.

The Compreno model is efficient in terms of adding new languages in the formalism. When adding the German verb *gehen* or the French *aller*, which mean "walk", we do not have to describe their semantic model (namely, the DSs which these verbs can attach and the filling of the slots) – we simply position the verbs in the same place in the hierarchy as the already-described verbs with equivalent meaning (such as English *walk* and Russian *idti*), and they thus acquire the same model. That is, the semantic part of the formalism is supposed to be universal for different languages (cross-language asymmetry cases are analyzed in Manicheva et al. 2012; Petrova 2014). The addition of new language vocabulary in the system demands mainly its syntactic description, namely, the description of SSs that correspond to the universal DSs, as SSs are language-specific, and the necessary transformational rules and collocations in asymmetry cases.

For the description of the locative domain, adding new languages in the model requires the description of the necessary sets of locative prepositions, locative SSs, and transformational rules.

As far as the evaluation of the description is concerned, we have estimated the efficiency of our model on the English and Russian text corpora consisting of different texts – fiction, news, various terminological fields, such as medicine, sport, law, economics, computer science,

and many others.

Our basic text collections consist of the English and Russian sentences with the manually done text mark-up, which sets the correct analysis of each fragment. The essential corpus for every day testing consists of more than 20000 examples, and there is a large number of additional text collections, which we use for different purposes as well. During the testing process, each example is analyzed both with and without the mark-up, then the two analyses are compared; in cases of discrepancy or bad analysis, the debugging process starts.

Estimating the quality of the locative description is a part of the general testing. As the analysis of the locative groups in the given corpora shows, the current approach allows one to solve most of the problems the locative adjuncts evoke. Therefore, the borders between the locative and non-locative domain (namely, between the [Locative], [Metaphoric_Locative] and [Sphere] adjuncts) can remain rather vague, which leads to the homonymy in the interpretation of the cases such as *read [in the book]*.

References

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'. In 'Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"', vol. 11(18), 91–103. Moscow: RGGU.

Aurnague, M., Hickmann, M. & Vieu, L. (eds.). 2007. *The categorization of spatial entities in language and cognition*. Amsterdam: John Benjamins.

Aurnague, M., Hickmann, M. & Vieu, L. 2007a. 'Introduction: Searching for the categorization of spatial entities in language and cognition'. In M. Aurnague, M. Hickmann & L. Vieu (eds.) 'Categorization of Spatial Entities in Language and Cognition', 1–32. Amsterdam: John Benjamins.

Bloom, P., Peterson, M. A., L., Nadel & Garrett, Merrill F. 1996. *Language and space*. Cambridge, Mass., MIT Press.

Bogdanov, A. V. & Leontyev, A. P. 2013. 'Description of the Russian External Possessor Construction in a Natural Language Processing System'. In 'Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"', vol. 11(19). Moscow: RGGU.

Creary, L. G., Gawron, J. M. & Nerbonne, J. 1989. 'Reference to locations'. In 'Proceedings of ACL-89', 42–50. Vancouver: University of British Columbia.

Fillmore, Ch. 1968. 'The case for case'. In E. Bach & R. Harms (eds.) 'Universals in linguistic theory', 1–90. New York: Holt, Rinehart and Winston.

Hickmann, M. & Robert, S. (eds.). 2006. *Space in Languages: Linguistic Systems and Cognitive Categories*. Amsterdam and Philadelphia: John Benjamins.

Japkowicz, N. & Wiebe, J. M. 1991. 'A System for Translating Locative Prepositions from English into French'. In 'Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics', 153–160. Berkeley, CA.

Jørgensen, F. 2004. *The Semantic Representation of Locatives in Machine Translation*. Ph.D. thesis, Leiden University.

Jørgensen, F. & Lønning, J. T. 2009. 'A minimal recursion semantic analysis of locatives'. *Computational Linguistics* 35: 229–270.

Leontiev, A. P. & Petrova, M. A. 2014. 'The Description of Locative Dependencies in a Natural Language Processing Model'. In 'Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog"', vol. 13(20), 318–329. Moscow: RGGU.

Levinson, S. C. 2003. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press.

Levinson, S. C. & Wilkins, D. P. (eds.). 2006. *Grammars of space*. Cambridge, Cambridge University Press.

Manicheva, E., Petrova, M., Kozlova, E. & Popova, T. 2012. 'The Compreno Semantic Model as Integral Framework for Multilingual Lexical Database'. In M. Zock & R. Rapp (eds.) 'Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING', 215–229. Mumbai: The COLING 2012 Organizing Committee.

Mel'čuk, I. A. 2012. *Jazyk: ot smysla k tekstu*. Moscow: Jazyki slavjanskoj kul'tury.

Oliver, P. & Gapp, K. P. 1998. *Representation and Processing Of Spatial Expressions*. Mahwah, NJ: Laurence Elbraum Associates Inc.

Petrova, M. A. 2014. 'The Compreno Semantic Model: The Universality Problem'. *International Journal of Lexicography* 27, no. 2: 105–129.

Pike, K. 1957. 'Grammemic Theory'. *General Linguistics* 2, no. 2: 35–41.

Plungian, V. A. 2000. *General morphology. Introduction to the problems*. Moscow: Editorial URSS.

Shay, E. & Seibert, U. 2003. *Motion, direction and location in languages*. Amsterdam and Philadelphia: John Benjamins.

Svorou, S. 1994. *The grammar of space*. Amsterdam and Philadelphia: John Benjamins.

Tesnière, L. 1976. *Éleménts de syntaxe structurale, 2nd edition*. Paris: Klincksieck.

Trujillo, I. A. 1995. *Lexicalist Machine Translation of Spatial Prepositions*. Ph.D. thesis, University of Cambridge.

Van der Zee, E. & Slack, J. (eds.). 2003. *Representing direction in language and space*. New York, Oxford University Press.