Chapter 7
A Mathematical Model of a Linguistic Database

Abstract In this chapter a broadly applicable mathematical model of linguistic database is constructed, that is, a model of a collection of semantic-syntactic data associated with primary lexical units and used by the algorithms of semantic–syntactic analysis for building semantic representations of natural language texts.

7.1 The Principles of Designing Semantics-Oriented Linguistic Processors

Most often, semantics-oriented natural language processing systems, or *linguistic processors (LPs)*, are complex computer systems, their design requires a considerable time, and their cost is rather high. Usually, it is necessary to construct a series of LPs, step by step expanding the input sublanguage of NL and satisfying the requirements of the end users.

On the other hand, the same regularities of NL are manifested in the texts pertaining to various thematic domains.

That is why, in order to diminish the total expenses of designing a family of LPs by one research center or group during a certain several-year time interval and in order to minimize the duration of designing each particular system from this family of LPs, it seems reasonable to pay more attention to (a) the search for best typical design solutions concerning the key subsystems of LPs with the aim to use these solutions in different domains of employing LPs; (b) the elaboration of formal means for describing the main data structures and principal procedures of algorithms implemented in semantic-syntactic analyzers of NL-texts or in the synthesizers of NL-texts.

That is why it appears that the adherence to the following two principles in the design of semantics-oriented LPs by one research center or a group will contribute, in the long-term perspective, to reducing the total cost of designing a family of LPs and to minimizing the duration of constructing each particular system from this family:

V.A. Fomichov, *Semantics-Oriented Natural Language Processing*, IFSR International Series on Systems Science and Engineering 27, DOI 10.1007/978-0-387-72926-8_7,

- the *Principle of Stability* of the used language of semantic representations (LSR) in the context of various tasks, various domains, and various software environments (stability is understood as the employment of a unified collection of rules for building the structures of LSR as well as domain- and task-specific variable set of primitive informational units);
- the *Principle of Succession* of the algorithms of LP based on using one or more compatible formal models of a linguistic database and unified formal means for representing the intermediate and final results of semantic-syntactic analysis of natural-language texts in the context of various tasks, various domains and various software environments (the succession means that the algorithms implemented in basic subsystems of LP are repeatedly used by different linguistic processors).

The theoretical results stated in Chaps. 2, 3, and 4 of this monograph provide a basis for following up the principle of stability of the used language of semantic representations. Chapter 4 defines a class of SK-languages that enable us to build semantic representations of natural language texts in arbitrary application domains.

This chapter is based on the results stated in previous chapters and is aimed at creating the necessary preconditions for implementing the succession principle in the design of LP algorithms.

In this and next chapters, we introduce a new method of transforming natural language texts into their semantic representations for the sublanguages of English, Russian, and German languages being of practical interest. This involves solving the following problems:

- Formalizing the structure of a linguistic database allowing for finding various conceptual relations, e.g., in the combinations “Verb + Preposition + Noun,” “Verb + Noun,” “Noun1 + Preposition + Noun2,” “Numeral + Noun,” “Adjective + Noun,” “Noun1 + Noun2,” “Participle + Noun,” “Participle + Preposition + Noun,” “Interrogative pronoun + Verb,” “Preposition + Interrogative pronoun + Verb,” “Interrogative Adverb + Verb,” “Verb + Numerical Value Representation” (a number representation + a unit of measurement representation).
- Formalizing the structure of data used as an intermediate pattern of the input natural language text semantic structure to provide a basis for building later a semantic representation of the input text.
- Using the solutions to Problems 1 and 2 for developing a domain-independent method of transforming an input NL-text (question/command/statement) from the sublanguages of English, Russian, and German languages into its semantic representation.

In this chapter, we apply the theory of SK-languages to building a broadly applicable formal model of a linguistic database (LDB). This model describes the logical structure of LDB being the components of natural-language interfaces to intelligent databases as well as to other applied computer systems. The expressions of SK-languages enable us to associate with the lexical units the appropriate simple or compound semantic units.

7.2 Morphological Bases

Let's formally represent the information about the elements being the primary components of natural-language texts.

Morphology is a branch of linguistics studying the regularities of the alteration of words and word combinations (depending on grammatical number, case, tense, etc). A linguistic database (LDB) must include a morphological database (MDB) with the content depending on the considered language. In contrast to the English language, the Russian language (RL) and the German language (GL) are very flexible, that is, the words in these languages can be changed in many ways. That is why, though an MDB is rather simple for English, the situation is different for RL and GL.

There are many publications devoted to the formalization of morphology of Russian, German, and many other languages. However, in order to develop a structured algorithm of semantic-syntactic analysis of NL-texts (the input texts may be from Russian, English, and German languages), it was necessary to propose a new, more general look at morphology of Russian, English, German, and many other languages in comparison with the available approaches.

The goal was to indicate the role of morphological analysis as a part of semantic-syntactic analysis, avoiding too detailed treatment of the morphological problems. For achieving this goal, the notions of morphological determinant, morphological space, and morphological basis are introduced in this section.

Definition 7.1. *Morphological determinant (M-determinant)* is an arbitrary ordered triple of the form

$$(m, n, maxv), \tag{7.1}$$

where m, n are the positive integers; $maxv$ is a mapping from the set $\{1, 2, \ldots, m\}$ into the set of non-negative integers N^+.

Let Det be an M-determinant of the form (7.1), then m will be interpreted as the quantity of different properties (which are called morphological) of the words from the considered language; n be the maximal amount of different sets of the values of morphological properties associated with one word. If $1 \leq i \leq m$, then $maxv(i)$ is interpreted as the maximal numerical code of the value of the property with the ordered number i (see Fig. 7.1).

For example, three sequences of the values of morphological properties can be connected with the Russian word "knigi" ("book" or "books"): if "knigi" is a word in the singular form, then this word is in the genitive case; if "knigi" is a word in the plural form, then it can be both in nominative case and in accusative case. That's why $n \geq 3$.

Let us suppose that the morphological properties with the numerical codes 1 and 2 are the properties "a part of speech" and "a subclass of a part of speech." That is why every integer k, where $1 \leq k \leq maxv(1)$, will be interpreted as a value of a part of speech, and every r, such that $1 \leq r \leq maxv(2)$, will be interpreted as a value of a subclass of a part of speech.

Figure 7.2 illustrates the structure of one sequence of the values of morphological properties associated with one word.

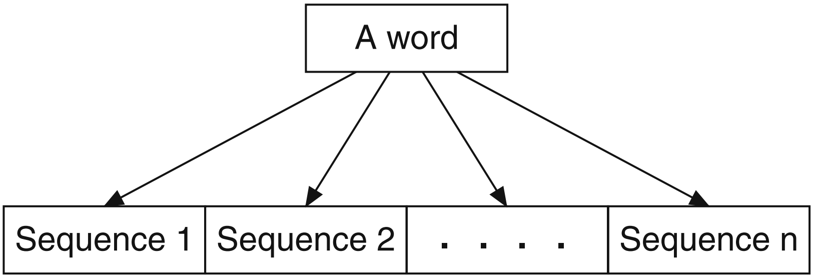

Fig. 7.1 The structure of the array of the values of morphological properties associated with one word

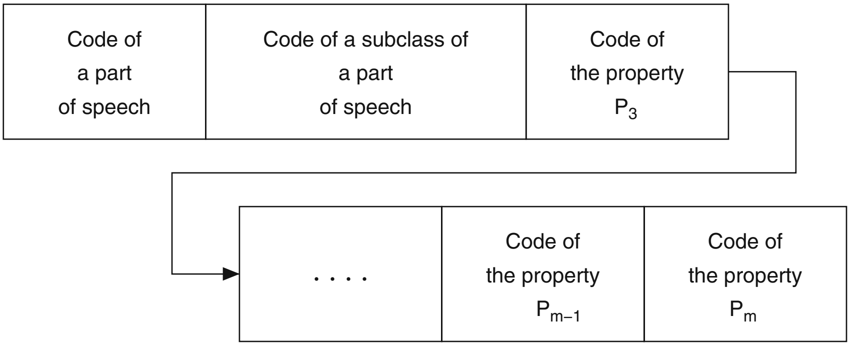
Code of a part of speech
Code of a subclass of a part of speech
Code of the property P_3
. . . .
Code of the property P_{m-1}
Code of the property P_m

Fig. 7.2 The structure of one sequence of the values of morphological properties associated with one word

We will suppose that every word from the considered language can be associated with only one part of speech and with one subclass of a part of speech. On one hand, this assumption is true for a very large subset of Russian language and, for example, of German language. On the other hand, such assumption will allow for diminishing the complexity of the elaborated formal model of LDB (without any real harm for applications).

Definition 7.2. Let *Det* be a M-determinant of the form (7.1). Then the *morphological space defined by the M-determinant Det* is the set *Spmorph* consisting of all finite sequences of the form

$$(x_1, \ldots, x_m, x_{m+1}, \ldots, x_{2m}, x_{2m+1}, \ldots, x_{nm}), \tag{7.2}$$

where (a) for every $k = 1, \ldots, n-1$, $x_{km+1} = x_1$, $x_{km+2} = x_2$; (b) for every $k = 1, \ldots, n$ and every such q that $(k-1)m+1 \leq q \leq km$, the following inequality is true: $0 \leq x_q \leq maxv(q-(k-1)m)$.

The conditions (a), (b) from this definition are interpreted in the following way: In the element of the form (7.2) from a morphological space, x_1 is the code of the part of speech. This code is located in every position separated by the distance m, $2m$, $\ldots$, $(n-1)m$ from the position 1; x_2 is the code of the subclass of the part of speech, this code is located in all positions separated by the distance m, $2m$, $\ldots$, $(n-1)m$ from the position 2.

Let q be the numerical code of the letter designation of a morphological characteristic (or property). Then the mapping $maxv$ determines the diapason of the values of this characteristic $[1, maxv(q)]$. That is why for each position q, where $1 \leq q \leq m$, the inequality $0 \leq x_q \leq maxv(q)$ takes place for the sequence of the form (7.2).

If $1 \leq q \leq m\, x_q$ is a component of the sequence of the form (7.2), and $x_q = 0$, this means that a word associated with this element of the morphological space doesn't possess a morphological characteristic (or property) with the numerical code q. For instance, the nouns have no characteristic "time."

Every component x_s of an element of the form (7.2). of a morphological space, where $s = q+m, q+2m, \ldots, q+(n-1)m$, is interpreted as a possible value of the same morphological property as in case of the element x_q. That's why the inequality

$$0 \leq x_s \leq maxv(s-(k-1)m)$$

indicates the diapason of possible values of the element x_s, where the integer k in the borders from 1 to n is unambiguously defined by the condition

$$(k-1)m + 1 \leq s \leq km.$$

The definition of a morphological basis introduced below gives a new mathematical interpretation of the notion "a morphological database." Temporarily abstracting ourselves from mathematical details, we describe a morphological basis as an arbitrary system $Morphbs$ of the form

$$(Det, A, W, Lecs, lcs, fmorph, propname, valname), \qquad (7.3)$$

where Det is a morphological determinant, and the other components are interpreted as follows: A is arbitrary alphabet (a finite set of symbols); the symbols from A are used for forming the words of the considered sublanguage of natural language (English, Russian, etc.). Let A^+ be the set of all non empty (or non void) strings in the alphabet A (in other terms, over the alphabet A). Then W is a finite subset of A^+, the elements of this set are considered as words and fixed word combinations (for example, "has been received") used for constructing natural language texts. The elements of the set W will be called words.

The component $Lecs$ is a finite subset of the set W, the elements of $Lecs$ are called *the lexemes* and are interpreted as basic forms of the words and fixed word combinations (a noun in singular form and nominative case, a verb in the infinitive form, etc.).

The component lcs is a mapping from the set W into the set $Lecs$, associating a certain basic lexical unit with a word; the component $fmorph$ is a mapping associating an element of the morphological space $Spmorph(Det)$ with a word wd from W.

The component *propname* (it is an abbreviation from $property-name$) is a mapping linking a numerical code of a morphological characteristic (or property) with the letter designation of this morphological characteristic. For example, the following relationship can take place:

$$propname(1) = part\_of\_speech.$$

More exactly, it is a mapping

$$propname: \{1, 2, ..., m\} \longrightarrow A^{+} \setminus W,$$

where $\setminus$ is the sign of set-theoretical difference.

The component *valname* is a partial mapping with two arguments. The first argument is the numerical code k of a morphological characteristic (or property). The second argument is the numerical code p of a certain possible value of this characteristic (property). The value of the mapping $valname(k, p)$ is the letter designation of p. For instance, the relationship $valname(1, 1) = verb$ can take place.

Definition 7.3. Let A, B be arbitrary non empty sets, and $f : A \longrightarrow B$ is a mapping from A into B. Then $Range(f)$ is the set of all such y that there is such element x from A that $f(x) = y$.

Definition 7.4. *Morphological basis* is an arbitrary 8-tuple *Morphbs* of the form (7.3), where *Det* is an M-determinant of the form (7.1), A is an arbitrary alphabet, W is a finite subset of the set A^{+} (the set of all non empty strings in the alphabet A), *Lecs* is a finite subset of the set W, $lcs : W \longrightarrow Lecs$ is a mapping from W to *Lecs*, $fmorph : W \longrightarrow Spmorph(Det)$ is a mapping from W to the morphological space defined by the M-determinant *Det*, *propname* is a mapping from the set $\{1, 2, \ldots, m\}$ to the set $A^{+} \setminus W$, *valname* is a partial mapping from the Cartesian product $N^{+} \times N^{+}$ to the set $A^{+} \setminus (W \cup Range(propname))$ defined for the pair (i, j) from $N^{+} \times N^{+} \Longleftrightarrow 1 \leq i \leq m,\ 1 \leq j \leq maxv(i)$.

Definition 7.5. Let *Morphbs* be a morphological basis of the form (7.3). Then

$$Parts(Morphbs) = \{valname(1, 1), \ldots, valname(1, maxv(1))\},$$

$$Subparts(Morphbs) = \{valname(2, 1), \ldots, valname(2, maxv(2))\}.$$

Thus, $Parts(Morphbs)$ is the set of the letter designations of the parts of speech, $Subparts(Morphbs)$ is the set of the letter designations of the subclasses of the parts of speech for the morphological basis *Morphbs*.

For instance, it is possible to define an English-oriented morphological basis *Morphbs* in such a way that the following relationships will take place:

$$Parts(Morphbs) \supseteq \{verb, noun, adjective, preposition, pronoun, participle,$$
$$adverb, cardinal\_numeral, ordinal\_numeral, conjunctive\},$$

$$Subparts(Morphbs) \supseteq \{common\_noun, proper\_noun\} .$$

Definition 7.6. Let *Morphbs* be a morphological basis of the form (7.3), $z \in Spmorph(Det)$ be an arbitrary element of morphological space, and $1 \leq i \leq mn$, then $z[i]$ is the i-th component of the sequence z (obviously, z has $m \cdot n$ components).

Definition 7.7. Let *Morphbs* be a morphological basis of the form (7.3). Then the mapping *prt* from the set of words W to the set $Parts(Morphbs)$ and the mapping *subprt* from the set W to the set $Subparts(Morphbs) \cup \{nil\}$ are determined as follows: for arbitrary word $d \in W$,

$$prt(d) = valname(1, fmorph(d)[1]) ,$$

if $fmorph(d)[2] > 0$,

$$subprt(d) = valname(2, fmorph(d)[2])$$

else

$$subprt(d) = nil .$$

Thus, the strings $prt(d)$ and $subprt(d)$ are respectively the letter designations of the part of speech and the subclass of the part of speech associated with the word d.

Example. A morphological basis can be defined in such a way that

$$W \ni cup, France;\ prt(cup) = noun,\ subprt(cup) = common\_noun ,$$

$$prt(France) = noun,\ subprt(France) = proper\_noun .$$

7.3 Text-Forming Systems

Natural language texts include not only words but also the expressions being the numerical values of different parameters, for example, the strings 90 km/h, 120 km, 350 USD. Let us call such expressions the *constructors* and suppose that these expressions belong to the class of elementary meaningful lexical units. It means that if we are building a formal model of linguistic database, we consider, for example, the expression 120 km as a symbol.

Of course, while developing computer programs, we are to take into account that there is a blank between the elements “120” and “km,” so “120 km” is a word combination consisting of two elementary expressions. However, the construction of every formal model includes the idealization of some entities from the studied domains, that is why we consider the constructors as symbols, i.e. as indivisible expressions.

Except the words and constructors, NL-texts can include the *markers*, for instance, the point, comma, semi-colon, dash, etc, and also the expressions in inverted commas or in apostrophes being the names of various objects.

Definition 7.8. Let Cb be a marked-up conceptual basis. Then *a text-forming system (t.f.s.) coordinated with the basis Cb* is an arbitrary system $Tform$ of the form

$$(Morphbs, Constr, infconstr, Markers), \tag{7.4}$$

where

- $Morphbs$ is a morphological basis of the form (7.3),
- $Constr$ is a countable set of symbols not intersecting with the set of words W,
- $infconstr$ is a mapping from the set $Constr$ to the primary informational universe $X(B(Cb))$,
- $Markers$ is a finite set of symbols not intersecting with the sets W and $Constr$,

and the following requirements are satisfied:

- for every d from the set $Constr$, the element $tp(infconstr(d))$ is a sort from the set $St(B)$;
- the sets W, $Constr$, $Markers$ don't include the inverted commas and apostrophes.

The elements of the sets W, $Constr$, and $Markers$ are called *the word forms (or words), constructors, and markers* of the system $Tform$, respectively.

Obviously, every morphological basis $Morphbs$ determines, in particular, an alphabet A and the set of words W.

Definition 7.9. Let $Tform$ be a text-forming system of the form (7.4), $Morphbs$ be a morphological basis of the form (7.3). Then

$$Names(Tform) = Names1 \cup Names2\,,$$

where $Names1$ is the set of all expressions of the form "x," where x is an arbitrary string in the alphabet A, and $Names2$ is the set of all expressions of the form "y," where y is an arbitrary string in the alphabet A;

$$Textunits(Tform) = W \cup Constr \cup Names(Tform) \cup Markers;$$

$Texts(Tform)$ is a set of all finite sequences of the form $d_1, \ldots, d_n$, where $n \geq 1$, for $k = 1, \ldots, n$, $d_k \in Textunits(Tform)$.

Definition 7.10. Let Cb be a marked conceptual basis, $Tform$ be a text-forming system of the form (7.4) coordinated with the basis Cb. Then the mapping $tclass$ from $Textunits(Tform)$ to the set $Parts(Morphbs) \cup \{constr, name\}$ and the mapping $subclass$ from $Textunits(Tform)$ to $Subparts(Morphbs) \cup \{nil\}$, where nil is an empty element, are determined by the following conditions:

- if $u \in W(Tform)$, then $tclass(u) = prt(u)$;
- if $u \in Constr$, then $tclass(u) = constr$;
- if $u \in Names(Tform)$, then $tclass(u) = name$;
- if $u \in Markers$, then $tclass(u) = marker$;
- if $u \in W(Tform)$, then $subclass(u) = subprt(u)$;

- if $u \in Constr$, then $subclass(u) = tp(infconstr(u))$, where $infconstr$ and tp are the mappings being the components of the text-forming system $Tform$ and of the primary informational universe $X(B(Cb))$ respectively;
- if $u \in Names(Tform) \cup Markers$, then $subclass(u) = nil$.

7.4 Lexico-semantic Dictionaries

Let us consider a model of a dictionary that establishes a correspondence between the elementary meaningful text units ("containers," "have prepared," etc.) and the units of semantic (or, in other words, informational) level. Lexico-semantic dictionary is one of the main components of a linguistic database. One part of the informational units corresponding to the words will be regarded as symbols; they are the elements of the primary informational universe $X(B(Cb))$, where Cb is a marked-up conceptual basis (m.c.b.) built for the considered domain, B is a conceptual basis (c.b.) being the first component of Cb.

The examples of such units are, in particular,

$$publication,\ entering1,\ entering2,\ station1,\ station2,$$

etc. Other informational units are compound. For example, the adjective "green" from W can be connected with the expression $Color(z1, green)$.

Definition 7.11. Let S be a sort system (s.s.) of the form

$$(\ St, P, Gen, Tol\ ).$$

Then *the semantic dimension* of the system S is such maximal number $k > 1$ that one can find such sorts $u_1, \ldots, u_n \in St$ that for arbitrary $i, j = 1, \ldots, k$, where $i \neq j$, u_i and u_j are comparable for the compatibility (or tolerance) relation Tol, i.e. $(u_i, u_j \in Tol)$. This number k is denoted $dim(S)$.

Thus, $dim(S)$ is the maximal number of the different "semantic axes" used to describe one entity in the considered application domain.

Example 1. Let us consider the concepts *"a firm"* and *"a university"*. We can distinguish three *semantic contexts* of word usage associated with these concepts. First, a firm or a university can develop a tool, a technology etc., so the sentences with these words can realize *the semantic coordinate "intelligent system."* Second, we can say, "This firm is situated near the metro station 'Taganskaya,'" and then this phrase realizes *the semantic coordinate "spatial object."* Finally, the firms and institutes have the directors. We can say, for example, "The director of this firm is Alexander Semenov." This phrase realizes *the semantic coordinate "organization."*

In the considered examples, we'll presume that semantic dimension of the considered sort systems is equal to four or three.

A *lexico-semantic dictionary* is a finite set *Lsdic* consisting of the $k+5$-tuples of the form

$$(i, lec, pt, sem, st_1, \ldots, st_k, comment)\,, \tag{7.5}$$

where k is the semantic dimension of the considered sort system, $i \geq 1$ is the ordered number of the $k+5$-tuple (we need it to organize the loops in the algorithms of processing NL-texts), and the rest of the components are interpreted in the following way:

- *lec* is an element of the set of basic lexical units *Lecs* for the considered morphological basis;
- *pt* is a designation of the part of speech for the basic lexical unit *lec*;
- the component *sem* is a string that denotes one of the possible meanings of the basic lexical unit *lec*.

The component *sem* for verbs, participles, gerunds is an informational unit connected with the corresponding verbal noun. For example, the verb *"enter"* has, in particular, the following two meanings: (1) entering a learning institution (in the sense "becoming a student of this learning institution"); (2) entering a space object ("John has entered the room," etc.).

So, for example, one system from a possible lexico-semantic dictionary will have, as the beginning, the sequence

$$i_1,\ enter,\ verb,\ entering1\,,$$

and the other will have, as the beginning, the sequence

$$i_2,\ enter,\ verb,\ entering2\,.$$

Number k is the semantic dimension of the considered sort system, i.e. $k = dim(S(B(Cb)))$, where Cb is the considered marked-up conceptual basis; $st_1, \ldots, st_k$ are the different *semantic coordinates* of the entities characterized by the concept *sem*. For example, if $sem = firm$, then $st_1 = ints$, $st_2 = space.ob.$, $st_3 = org$, $k = 3$.

If an entity characterized by the concept *sem* has the various semantic coordinates $st_1, \ldots, st_p$, where $p < k$, then $st_{p+1}, \ldots, st_k$ is a special empty element *nil*. The component *comment* is either a natural language description of a meaning associated with the concept sem or an empty element nil.

Definition 7.12. Let *Cb* be a marked-up conceptual basis of the form

$$(B, Qmk, Setmk, Cmk)\,,$$

Morphbs be a morphological basis of the form (7.3), *Qmk* be a questions marking-up of the form (5.1), and let the primary informational universe $X(B(Cb))$ and the set of variables $V(B(Cb))$ not to include the symbol *nil* (empty element).

Then *a lexico-semantic dictionary coordinated with the marked-up conceptual basis Cb and with the morphological basis Morphbs* is an arbitrary finite set *Lsdic* consisting of the systems of the form (7.5), where

- $i \geq 1$, for each $lec \in Lecs$;

- $pt = prt(lec)$, $sem \in Ls(B(Cb)) \cup \{nil\}$;
- $k = dim(S(B(Cb)))$;
- for each $p = 1, \ldots, k$, $st_p \in St(B(Cb)) \cup \{nil\}$;
- $comment \in A^+ \cup \{nil\}$;

and the following conditions are satisfied:

- no two systems from *Lsdic* may have the same first component *i*;
- if two systems from *Lsdic* have different values of the *sem* component, then these two systems have different values of the *comment* component.

Example 2. A set *Lsdic* can be defined in such a way that *Lsdic* includes the following 8-tuples:

$$(112, container, noun, container1, dyn.phys.object, nil, nil, \text{“}reservoir\text{”}),$$

$$(208, enter, verb, entering1, sit, nil, nil, \text{“}enter\,a\,college\text{”}),$$

$$(209, enter, verb, entering2, sit, nil, nil, \text{“}enter\,a\,room\text{”}),$$

$$(311, aluminum, adj, Material(z1, aluminum), phys.ob., nil, nil, nil),$$

$$(358, green, adj, Color(z1, green), phys.ob, nil, nil, nil),$$

$$(411, Italy, noun, certn\,country * (Name1, 'Italy'), space.ob, nil, nil, \text{“}country\text{”}),$$

$$(450, passenger, adj, sem1, dyn.phys.ob., nil, nil, nil),$$

where

$$sem1 = Purpose(z1, movement1 * (Object1,$$
$$certn\,set * (Qual - compos, person))).$$

7.5 Dictionaries of Verbal – Prepositional Semantic-Syntactic Frames

Verbs, participles, gerunds, and verbal nouns play the key role in forming sentences due to expressing the various relations between the entities from the considered application domain.

Thematic role is a conceptual relation between a meaning of a verbal form (a form with time, an infinitive, a participle, a gerund) or a verbal noun and a meaning of a word group depending on it in the sentence.

Thematic roles are also known as *conceptual cases, semantic cases, deep cases*, and *semantic roles.*

The concept of deep case was proposed by the world-known American linguist C. Fillmore in 1968. This concept very soon became broadly popular in computer linguistics and theoretical linguistics, because it underlies the basic procedures that find conceptual relationships between a meaning of a verbal form and a meaning of a word group dependending on it in a phrase.

Example 1. Let T1 = "The bulk carrier 'Mikhail Glinka' has arrived from Marseilles to Novorossiysk on the 27th of March." The compound verbal form "has arrived" denotes a certain event of the type "arrival" that can be connected with the label $e1$(event1). In the text T1, the following objects are mentioned: a certain ship $x1$; a certain city $x2$ named "Marseilles"; a certain city $x3$ named "Novorossiysk."

In the event $e1$, the object $x1$ plays the role "Agent of action" (Agent1), $x2$ plays the role "Initial place of movement" ($Place1$), $x3$ plays the role "Place of destination" ($Place2$). Then we can say that the text T1 realizes the thematic roles $Agent1$, $Place1$, $Place2$ as well as the thematic role $Time$.

Example 2. Let T2 = "The bulk carrier 'Mikhail Glinka' has arrived from Marseilles." The text T2 explicitly realizes only the thematic roles $Agent1$ and $Place1$, whereas the thematic roles $Time$ and $Place2$ only are implied because of the semantics of the verb "arrive." Thus, the phrases with the same verb in the same meaning can explicitly realize the different subsets of thematic roles.

Formally, we will interpret thematic roles as the names of binary relations with the first attribute being a situation and second one being a real or abstract object playing a specific role in this situation. In this case, if an element $rel \in R_2(B)$, where B is a conceptual basis, and rel is interpreted as a thematic role, then its type $tp(rel)$ is a string of the form $\{(s,u)\}$, where s is a specification of the distinguished sort sit (situation), and u is a sort from the set $St(B)$.

The dictionaries of verbal – prepositional frames contain such templates (in other terms, frames) that enable us to represent the necessary conditions of realizing a specific thematic role in the combination

$$Verbal form + Preposition + Dependent\,word\,group,$$

where $Preposition$ can be void (let nil be the sign of void preposition), and

$Dependent\,word\,group$ is either a noun with dependent words or without them, or a construct, that is, a numeric value of a parameter.

For example, such expressions include the combinations "has arrived to the port," "left the city," "prepare 4 articles," "has bought the Italian shoes," "arrived before 16:30."

Definition 7.13. Let Cb be a marked-up conceptual basis of the form (5.4), $Tform$ be a text-forming system of the form (7.4) coordinated with Cb, $Morphbs$ be a morphological basis of the form (7.3), $Lsdic$ be a lexico-semantic dictionary coordinated with Cb and $Tform$.

Then *a dictionary of verbal – prepositional semantic-syntactic frames (d.v.p.f.) coordinated with* Cb, $Tform$, and $Lsdic$ is an arbitrary finite set Vfr consisting of the ten-tuples of the form

$$(k, semsit, form, refl, vc, sprep, grcase, str, trole, expl), \tag{7.6}$$

where

- $k \geq 1$, $semsit \in X(B)$, $form \in \{infin, ftm, nil\}$, $refl \in \{rf, nrf, nil\}$, $vc \in \{actv, passv, nil\}$,
- $sprep \in W \cup \{nil\}$, where nil is an empty element, W is the set of words from the text-forming system $Tform$; if $sprep \in W$, then $prt(sprep) = preposition$;
- $0 \leq grcase \leq 10$, $str \in St(B)$,
- $trole$ is a binary relational symbol from the primary informational universe $X(B)$, $tp(trole) = \{(s, u)\}$, where $s, u \in St(B)$, s being a concretization of the distinguished sort sit ("situation") (i.e. $sit \rightarrow s$);
- $expl \in A^+ \cup \{nil\}$.

The components of an arbitrary 10-tuple of the form (7.6) from Vfr are interpreted in the following way:

- k is the ordered number of the collection;
- $semsit$ is a semantic unit identifying the type of situation (arrival, departure, receipt, etc.);
- $form$ is a verb form property;
- $infin$ is the indicator of the infinitive verb form;
- ftm is the indicator of a verb form with time, i.e., of the verb in indicative or subjunctive mood;
- $refl$ is the property of reflexivity of the verbs and participles, rf is the indicator of reflexive form, nrf is the indicator of non reflexive form;
- $actv$, $passv$ are the indicators of active and passive voices.

The components $semsit$, $form$, $refl$, vc define the requirements to a verbal form, and the components $sprep$, $grcase$, str formulate the requirements to a word or word group being dependent on the verbal form and used in a sentence for expressing a thematic role $trole$.

The string $sprep$ is a simple or compound preposition (for example, the preposition "during" is translated into Russian as "v techenie") or the sign of the void preposition nil; $grcase$ is the code of a grammatical case (that is why $1 \leq grc \leq 10$) or 0 (it is the sign of the lack of such information); str is a semantic restriction for the meaning of a dependent word group or word; $trole$ is such thematic role that the necessary conditions of its realization are represented by this collection (frame); $expl$ is an example in NL that explains the meaning of the thematic role or it is the empty example nil.

The maximal value 10 for the numerical code of a grammatical case is chosen with respect to the fact that the quantity of grammatical cases is 4 for the German language and 6 for the Russian language.

Example 3. Let us construct a certain dictionary $Vfr1$ helping us to find the conceptual relations in the sentences with the verb "to prepare." This verb has, in particular, the meanings $preparation1$ (the preparation of a report, article, etc.) and $preparation2$ (the preparation of the sportsmen of highest qualification, etc.).

In particular, the dictionary $Vfr1$ can be useful for the semantic analysis of the texts like T1 = "Professor Semenov prepared in June a report for the firm 'Sunrise' ";

T2 = “A report for the firm ‘Sunrise’ was prepared in June by Professor Semenov;”
T3 = “Professor Semenov prepared three Ph.D. scholars in chemistry during 2003–2008.”

One can create such marked-up conceptual basis *Cb* that its first component is a conceptual basis *B* and the following relationships take place:

$$St(B) \supset (Sorts1 \cup Sorts2)\,,$$

where

$$Sorts1 = \{org, ints, mom, inf.ob, dyn.phys.ob\}\,,$$

$$Sorts2 = \{sit, event, qualif, space.ob, string\}\,;$$

$$X(B) \supset (Units1 \cup Units2 \cup Units3 \cup Units4 \cup Units5)\,,$$

where

$$Units1 = \{\#now\#, firm, university, professor, phd-scholar, learn.inst\},$$

$$Units2 = \{person, {}'Semenov', preparation1, preparation2, report1\},$$

$$Units3 = \{Name, Surname, June, 3, 2003, 2008, Qual\},$$

$$Units4 = \{Agent1, Object1, Object2, Product1, Time\},$$

$$Units5 = \{Place1, Place2, Recipient1, Educ\_inst\};$$

$$sit \rightarrow event\,,$$

because the events are special cases of situations;

$$tp(preparation1) = tp(preparation2) = \uparrow event;$$

$$tp(firm) = tp(university) = \uparrow org * space.ob * ints;$$

$$tp(person) = \uparrow ints * dyn.phys.ob;$$

$$tp(professor) = tp(phd-scholar) = qualif;$$

$$tp(Agent1) = \{(event, ints)\};\ tp(Recipient1) = \{(event, org)\},$$

$$tp(Time) = \{(event, mom)\};$$

$$tp(Place1) = tp(Place2) = \{(event, space.ob)\};$$

$$tp(Surname) = \{(ints, string)\},\ tp({}'Semenov') = string;$$

$$tp(Object1) = \{(event, dyn.phys.ob)\}.$$

In that case, let $Vfr1$ be the set consisting of the following sequences:

$$(1, preparation1, ftm, nrf, actv, nil, 1, ints, Agent1, Expl1),$$

where $Expl1 = {}'P.Somov\,prepared\,(a\,textbook)'$,

$$(2, preparation1, ftm, nrf, passv, by, 5, ints, Agent1, Expl2),$$

where $Expl2 = ''(This\,book)\,was\,prepared\,by\,Professor\,Semenov'$,

$$(3, preparation1, ftm, nrf, actv, nil, 4, inf.ob, Product1, Expl3),$$

where $Expl3 = '(P.Somov)\,prepared\,a\,book'$,

$$(4, preparation1, ftm, nrf, passv, nil, 1, inf.ob, Product1, Expl4),$$

$Expl4 = 'This\,article\,was\,prepared\,(during\,three\,weeks)'$,

$$(5, preparation2, ftm, nrf, passv, nil, 4, qualif, Object2, Expl5),$$

where $Expl5 = 'Many\,masters\,of\,sport\,were\,prepared\,(by\,this\,school)'$,

$$(6, preparation1, nil, nil, in, 0, mom, Time, Expl6),$$

where $Expl6 = 'prepared\,in\,2007'$,

$$(7, preparation2, nil, nil, in, 0, mom, Time, Expl7),$$

where $Expl7 = 'prepared\,in\,2007'$.

In this example, the numerical codes of the grammatical cases are indicated for the Russian language, where six grammatical cases are distinguished. The reference to the Russian language helps to become aware of the significance of the component *grcase* of the elements of the dictionaries of verbal – prepositional frames for highly flexible languages.

7.6 The Dictionaries of Prepositional Frames

Let's consider in this section the following problem: how it would be possible to find one or several conceptual relationships realized in the word combinations of the form

$$Noun1 + Preposition + Noun2$$

or of the form

$$Noun1 + Noun2.$$

Example 1. Let us assume that *Expr*1 is the expression "an article by Professor Novikov," and a linguistic database includes a template of the form

$$(k1, 'by', inf.ob, ints, 1, Authors, 'a\,poem\,by\,Pushkin'),$$

where *ints* is the sort "intelligent system," 1 is the code of common case in English. We may connect the sorts *ints* and *dyn.phys.ob* (dynamic physical object) with the basic lexical unit “professor.” We see that the expression $Expr1$ is compatible with this template having the number $k1$.

Definition 7.14. Let Cb be a marked-up conceptual basis of the form (5.4), $B = B(Cb)$, $Morphbs$ be a morphological basis of the form (7.3), $Tform$ be a text-forming system of the form (7.4) coordinated with m.c.b Cb; $Lsdic$ be a lexico-semantic dictionary consisting of the finite sequences of the form (7.5) coordinated with Cb and $Tform$.

Then *a dictionary of prepositional semantic-syntactic frames coordinated with* Cb, $Tform$, and $Lsdic$ is an arbitrary finite set Frp consisting of the ordered 7-tuples of the form

$$(i, prep, sr1, sr2, grc, rel, ex), \tag{7.7}$$

where

- $i \geq 1$; $prep \in Lecs \cup \{nil\}$, where nil is the string denoting the void (empty) preposition; if $prep \in Lecs$, then $prt(prep) = preposition$;
- $sr1, sr2 \in St(B)$; $1 \leq grc \leq 10$;
- $rel \in R_2(B)$, where $R_2(B)$ is the set of binary relational symbols (thefore, it is a subset of the primary informational universe $X(B(Cb)))$; $ex \in A^+$.

The components of the 7-tuples of the form (7.7) from the set Frp are interpreted as follows: The natural number $i \geq 1$ is the ordered number of the 7-tuple (it is used for organizing the loops while analyzing the data from the dictionary Frp), $prep$ is a preposition from the set of basic lexical units $Lecs$ or the void (or empty) preposition nil.

The elements $sr1$ and $sr2$ are interpreted as the sorts that may be associated respectively with the first and second nouns in the linguistically correct combination of the form “Noun1 + Preposition + Noun2”; grc (grammatic case) is the code of such grammatical case that the second noun must be in this grammatical case when it is a part of the correct combinations of the kind.

The component rel is a designation of such conceptual relation that this relation can be realized in such combinations when the specified conditions are satisfied; ex is an example being an expression where the same relation rel is realized.

Example 2. It is possible to build such marked-up conceptual basis Cb, a morphological basis $Morphbs$, a text-forming system $Tform$, a lexico-semantic dictionary $Lsdic$, and a dictionary of prepositional semantic-syntactic frames Frp that Frp includes the semantic-syntactic template (frame) with the number $k1$ considered in Example 1 and also the templates

$$(k2, 'for', substance, illness, 1, Against1, Expr1),$$

$$(k3, 'for', phys.ob, ints, 1, Addressee, Expr2),$$

where 1 is the code of common case in English, $ints$ is the sort “intelligent system,” $Expr1$ = “pills for flu,” $Expr2$ = “a letter for Mary.”

Suppose that a dictionary of prepositional semantic-syntactic frames Frp contains no such 7-tuples where the components $prep \neq nil$, $sr1$, $sr2$, grc coincide but the components *rel* or *ex* don't coincide. In such cases the 4-tuple of the form $(prep, sr1, sr2, grc)$ unambiguously determines the relation *rel*.

7.7 Linguistic Bases

Let's take two final steps for constructing a formal model of a linguistic database.

7.7.1 Semantic Information Associated with the Role Interrogative Words

Let's define the notion of *a dictionary of the role interrogative word combinations.* Consider the following pairs of the form $(prepqw, qwd)$, where *prepqw* is a preposition or the void (or empty) preposition *nil*; *qwd* is an interrogative word being either a pronoun or adverb:

$$(nil, who), (by, whom), (for, whom),$$

$$(from, whom), (from, where), (nil, when)\,.$$

Our language competence enables us to associate a certain rather general conceptual relation with each of such pairs:

$$(nil, who) \Rightarrow Agent;$$

$$(by, whom) \Rightarrow Agent$$

$$(for, whom) \Rightarrow Addressee;$$

$$(from, whom) \Rightarrow Source1$$

$$(from, where) \Rightarrow Place1$$

$$(nil, when) \Rightarrow Time\,.$$

We'll say that such pairs of the form $(prepqw, qwd)$ are the role interrogative word combinations.

Taking this into account, we'll include one more dictionary into a linguistic database.

Definition 7.15. Let *Cb* be a marked-up conceptual basis of the form (5.4), $B = B(Cb)$, *Morphbs* be a morphological basis of the form (7.3), *Tform* be a text-forming system of the form (7.4) coordinated with m.c.b *Cb*; *Lsdic* be a lexico-semantic dictionary consisting of the finite sequences of the form (7.5) coordinated with *Cb* and *Tform*.

Then *a dictionary of the role interrogative word combinations* coordinated with the marked-up conceptual basis *Cb*, the morphological basis *Morphbs*, and the lexico-semantic dictionary *Lsdic* is an arbitrary finite set *Rqs* consisting of the ordered 4-tuples of the form

$$(i, prepqw, qwd, relq), \tag{7.8}$$

where

- $i \geq 1$, $prepqw \in Lecs \cup \{nil\}$,
- $qwd \in W$, $prt(qwd) \in \{pronoun, adverb\}$,
- $relq \in R_2(B(Cb))$;
- if $prepqw \neq nil$, $prt(prepqw) = preposition$.

Example. It is possible to define B, Cb, $Morphbs$, $Lsdic$, Rqs in such a way that *Rqs* includes the 4-tuples

$$(1, nil, who, Agent),\ (2, nil, whom, Addressee),$$

$$(3, for, whom, Addressee),\ (4, from, whom, Source1),$$

$$(5, with, what, Tool)\ (6, nil, when, Time),$$

$$(7, from, where, Place1),\ (8, nil, where, Place2).$$

7.7.2 The Notion of a Linguistic Basis

The linguistic bases are formal models of linguistic databases (LDB).

Definition 7.16. The ordered 6-tuple *Lingb* of the form

$$(Cb, Tform, Lsdic, Vfr, Frp, Rqs) \tag{7.9}$$

is called *a linguistic basis (l.b.)* ⇔ when *Cb* is a marked-up conceptual basis (m.c.b) of the form (5.4), *Tform* is a text-forming system (t.f.s) of the form (7.4) coordinated with m.c.b *Cb*; *Lsdic* is a lexico-semantic dictionary coordinated with m.c.b *Cb* and with t.f.s *Tform*, *Vfr* is a dictionary of verbal – prepositional semantic-syntactic frames coordinated with m.c.b *Cb*, t.f.s *Tform*, and lexico-semantic dictionary *Lsdic*; *Rqs* is a dictionary of the role interrogative word combinations coordinated with *Cb*, *Tform*, and *Lsdic*.

The structure of a linguistic basis is illustrated by Fig. 7.3. The introduced formal notion of a linguistic basis reflects the most significant features of broadly applicable logical structure of a linguistic database. This notion is constructive in the sense that it really can help to design LDB of practically useful linguistic processors, it is shown in the next chapters of this book.

The formal model of a linguistic database constructed above generalizes the author's ideas published, in particular, in [51, 54, 81, 85].

Problems

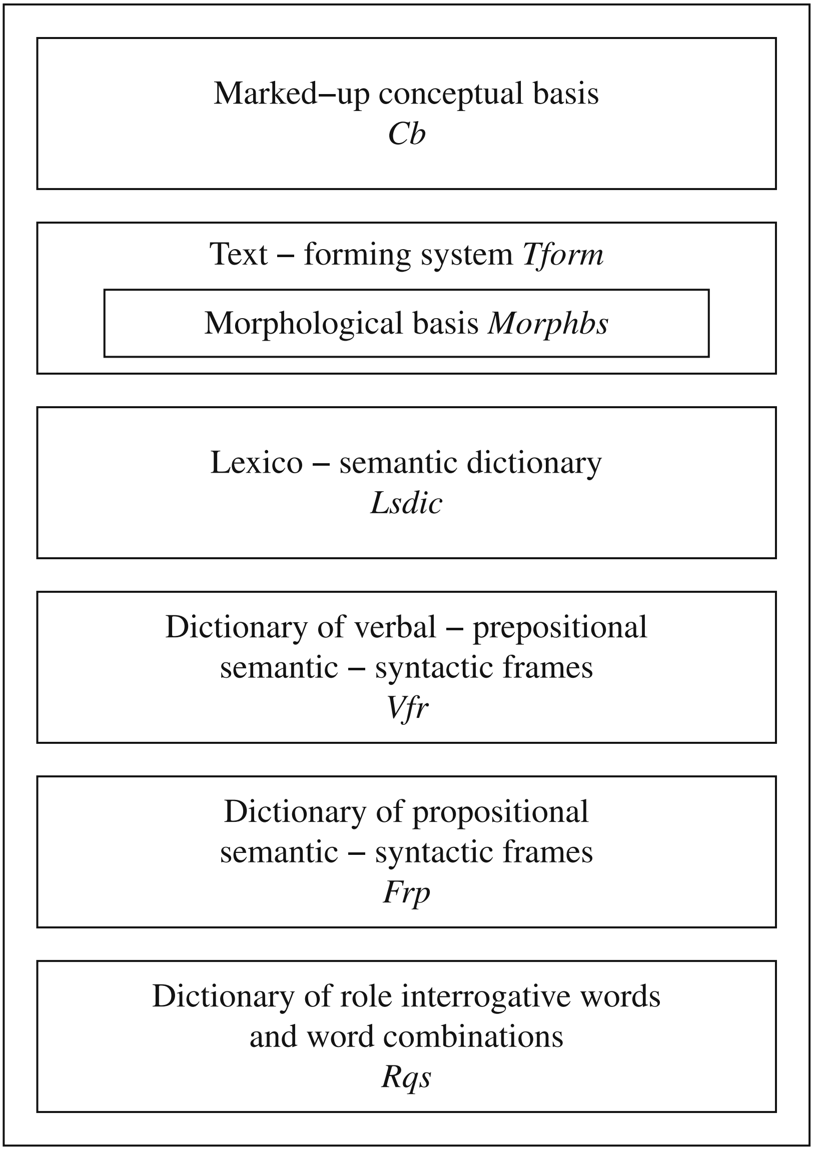
Marked–up conceptual basis
Cb
Text – forming system *Tform*
Morphological basis *Morphbs*
Lexico – semantic dictionary
Lsdic
Dictionary of verbal – prepositional
semantic – syntactic frames
Vfr
Dictionary of propositional
semantic – syntactic frames
Frp
Dictionary of role interrogative words
and word combinations
Rqs

Fig. 7.3 The structure of a linguistic basis

1. How are formally interpreted thematic roles (conceptual cases, deep cases)?
2. What is the semantic dimension of a sort system?
3. What is a morphological determinant (M- determinant)?
4. What is a morphological space?
5. What are the components of a morphological basis?
6. What are the components of a text-forming system?

7. What are constructs?
8. How is the subclass of constructs defined?
9. What is the role of the tolerance relation on the set of sorts in the definition of a lexico-semantic dictionary?
10. What is the structure of the finite sequences being the elements of a dictionary of semantic-syntactic verbal–prepositional frames?
11. What are the components of a dictionary of semantic-syntactic prepositional frames?
12. What is the structure of a linguistic basis?