Semantic dictionary viewed as a lexical database

Elena V.Paducheva
Ekaterina V.Rakhilina
Marina V.Filipenko
Institute of scientific and technical information (VINITI)
Academy of sciences of Russia
125219 Moscow, Usievicha 20a
e-mail psy-pub@comlab.vega.msk.su
Telefax: (7.095) 9430060
Telex: 411249

Abstract

In this paper an expert system is described which is called **Lexicographer** and which aims at supplying the user with diverse information about Russian words, including bibliographic information concerning individual lexical items. It is supposed that the system may be of use for a practical computational linguist and at the same time will serve as an instrument of linguistic research.

1 Lexical database and its advantages over traditional dictionaries

In this paper we investigate general principles implemented in an expert system (called **LEXICOGRAPHER**), designed to supply the user with diverse information about Russian words, cf. [2].

The system is conceived as an aid both in the area of natural language processing and in the traditional lexicography.

The system consists of two basic components:

- Lexicon (containing some 13.000 most common words);
- Bibliographical database.

It is the lexicon that is of primary concern in this paper.

The idea was to present the Lexicon in a form of a lexical database (LDB).

LDB is a vocabulary presented in a machine readable form and consisting of several domaines, as in a usual relational database. The user may get information about morphology, syntactic combinability and semantic features

of individual lexical items. It is semantics that we concentrate upon in this paper.

Many attempts have been made to use traditional dictionaries in order to assign word senses to general semantic categories, cf. [1].

Our LDB contains semantic information that cannot be elicited from the existing dictionaries. The priority is given to semantic features influencing lexical or grammatical co-occurrence. In this paper possibilities are discussed of predicting selectional restrictions, syntactic features and other formal characteristics of the utterance - such as the array of arguments and their semantic interpretation, the meaning of an aspectual form of a verb etc., - on the basis of semantic features of a word in the lexicon.

The main advantage of a lexical database as compared with a traditional dictionary consists in the fact that a database makes it possible to present semantic information in a format enabling the computer to locate efficiently various types of information specified for a given class of words. To put it differently, the main advantage of a database consists in the possibility of compiling lists of words possessing a common feature or a set of features.

There are three main principles that the system is based upon.

1. We are convinced that semantic features of words determine co-occurence to a much greater extent than it is usually acknowledged. In other words, we claim that many aspects of syntactic subcategorization of lexical items are predictable from their meaning.

2. A semantic feature of a word is essentially a semantic component (or components) in its lexicographic definition.

3. A great amount of information about the meaning of a lexical unit; about its combinatory possibilities; prosody; referential features; or about its regular ambiguity, need not be stored in the dictionary: this information belongs to what may be called **a grammar of lexicon** and should be formulated in a generalized form. In this form it can be stored in a **Lexical Knowledge-Base** of semantic and syntactic regularities. This Knowledge-Base has not yet been designed, but semantic features of words in LDB are conceived as an input for general rules that will be stored in this hypothetical Knowledge-Base.

2 Lexical Database for Concrete Nouns

There are different layers of lexicon that require specific formats of a database, and the choice of the format is one of the main problems of database formation.

In what follows we list domains in the Lexical Database for Concrete Nouns - one of the components of Lexicographer, now implemented in a working program. Each domain is interpreted as a feature that can take a definite set of values.

Domain I. Morphological and syntactico-

morphological information (taken from the grammatical dictionary [3]).

This domain is subdivided into three domains:

I.1. Gender (fem., masc., neuter., common).

I.2. Animate/Inanimate

I.3. Declension and accentuation.

All the other domains contain semantic information. We do not mean that the system of semantic features would provide a word with an exhaustive lexicographic definition - this is not the appropriate task for a lexical database. The purpose of a database is to highlight those semantic aspects of a word that unite semantically cognate words and differentiate many of semantically different words from one another. In other words, lexical database is an instrument of predicting and calculating all sorts of useful semantic classes of words.

Domains II.1 and II.2 specify Mereological status of a word (more precisely, of a lexeme - namely, of a word taken in one of its lexical meanings). The values of the feature II.1 may be: PART, SET or WHOLE. In the later case domain II.2 is empty while in the first two cases it specifies the WHOLE for the PART and the ELEMENT for the SET: PART (SET) of what? E.g.,

(1) krylo 'ail'

M-status | PART

Of what | body

(2) stado 'herd'

M-status | SET

Of what | animals

(3) chelovek 'man'

M-status | WHOLE

Of what | -

Domain II.3 provides a lexeme with a taxonomic supercategory, such as Person, Plant, Animal, Metal, Building, Sphere of activity etc. This domain is of primary importance and it is this domain that defines the most interesting classes of concrete lexemes. The system of taxonomic categories has a hierarchical structure. Thus, the possibility is provided to state implicative dependencies between categories, so that the lower category inherits all the information from the category of a higher level. E.g.,

T-category (osobnjak 'private residence') = dom 'house';

T-category (dom) = postrojka 'building';

T-category (postrojka) = sooruzenie 'construction'.

Thus, lexeme osobnjak will be assigned not only to the class of houses but also to the class of buildings and to the class of constructions.

Domain II.4 specifies a Predicate semantically connected with the noun in question. It turns out that such predicates occupy the most prominent place in lexicographic definitions of a great majority of concrete nouns. Usually these are predicates that determine a standard way in which the corresponding object is used

(functional predicates):

Predicate (house) = to live

Predicate (chair) = to sit

Predicate (goblet) = to drink. There are also nouns that imply a non-functional predicate in their lexicographic definition - a predicate that determines its characteristic property, cf.

Predicate (liquide) = to flow.

Some nouns require predicates of both types, cf.

Predicate (cellar) =

1) to store products;

2) digged under the floor of a house.

For some classes of nouns Domain I.4 Predicate is empty, e.g., for some (not all!) names of the so-called natural classes and for the names of parts of the corresponding objects, cf.

krab 'crab':

M-status | WHOLE

Of what | -

T-category | animal

Predicate | -

Inclusion of predicates into a lexicographic definition of concrete nouns may be considered an attempt to fertilize theoretical lexicography with the ideas of frame semantics.

Domain II.5: Predicate may have a Restriction as for the range of possible taxonomic classes of its arguments, e.g.

khishchnik 'beast of prey'

M-status | WHOLE

Of what | -

T-category | animal

Predicate | to eat

Restriction | animal

The Database for Concrete Nouns is ready for demonstration. The database for verbs and a small base for pronouns are in a stage of preparation.

3 Combinability predictions for concrete nouns

Here are some examples of how semantic information contained in the database can be used to predict syntactic regularities.

Example 1. As was stated earlier, domains II.1, II.2 define the following relations:

1) PART-WHOLE;

2) SET-ELEMENT.

There are propositions that differentiate these two relations; thus, combinations in (a), with PART-WHOLE relation are possible with a preposition U, while combinations in (b), with a SET-ELEMENT relation, are not:

a. nozka 'leg' U stula 'chair'

pugovica 'button' U paljto 'coat'

b. *chaschka 'cup' U serviza 'service'

*korova 'cow' U stada 'herd'

Note that Genitive Case can be used to express both relations.

Example 2 makes use of the domain Predicate: it is the predicate implied by a lexicographic definition of a noun that determine, in very many cases, the exact interpretation of the Genitive construction with a concrete noun as a head.

Thus, a noun gnezdo 'nest' has a possessive valency gnezdo orla 'nest of an eagle', chjo gnezdo? 'whose nest', because of the predicate 'to live' included in the lexicographic definition of gnezdo 'nest' has an unbounded variable: who lives? On the other hand, for such a noun as professor 'professor' Genitive construction realizes its object valency, cf. professor matematiki 'professor of mathematics', because of the Predicate 'to study', included in its lexicographic definition; an unbounded variable here corresponds to the object valency: studies what?

Examples of this kind are abundant.

To sum up, the following aspects of the proposed type of a semantic dictionary are of primary importance.

1. The fact that information is presented in the form of a database, which provides the facility of compiling all sorts of lexical lists.

2. Intensive use of T-categories (and other recurrent semantic features), which gives semantic explications for combinability restrictions.

3. Division of lexical information into two parts - Lexical Data Base and Lexical Knowledge Base, which widens the range of possible lexicographic generalizations.

References

[1] Gellerstam M. (ed.) Studies in computer-aided lexicology. Stockholm, 1988.

[2] Paducheva E.V., Rakhilina E.V. Predicting co-occurence restrictions by using semantic classifications in the lexicon. COLING-90, Helsinki, 1990.

[3] Zalizniak A.A. Grammaticheskij slovar' russkogo jazyka. 2-d ed. Moscow, 1980.