G.I. Kustova
Moscow State Pedagogical University
E.V.Paducheva
VINITI/Russian Academy of Sciences, Moscow

Semantic Dictionary as a Lexical Database*

Abstract

We investigate general principles implemented in a lexical Database named LEXICOGRAF. In our Database lexical definition is divided into a sequence of syntactically independent semantic components. The priority is given to those components that can be used to predict and to explain non–trivial peculiarities of surface behaviour shared by a class of words – such as co–occurrence restrictions, systematic polysemy etc. The Database is associated with what may be called a Knowledge Base – a repository of general rules and regularities pertaining to theoretical lexicography.

The paper presents a description and a linguistic substantiation of an expert system named LEXICOGRAF.[1] The project is being carried out at VINITI, Moscow, by an interinstitutional research group supervised by Elena V. Padutcheva.The basic idea is to present lexicographic information in the form of a lexical data base of a relational type: all information about a lexeme (= a word taken in one of its senses) is distributed between several domains, each having some definite range of values. As a result, the user is enabled to conduct a search of any depth within the frame of the given set of parameters. The system includes two components: (1) the Database proper and (2) the Knowledge base. The main attention is paid to semantic information about words. So, in fact, the system in question is a semantic dictionary presented in the form of a relational database.

1. Database

The Database is partitioned into several segments (see Krasil'shchik, Rakhilina 1992 about the segment 'Names of Objects'). This paper is concerned with verbs.

The lexical *entry* of a verbal lexeme is constituted by the following domains:

(1) Lexeme with illustrations of its use;
(2) Morphology (e.g. Aspect–form);
(3) Actants (i.e. arguments);
(4) Taxonomic category;
(5) Lexicographic definition;

(6) Aspectual characteristics;
(7) Derived meanings.

Domains (3) Actants and (5) Definition are, in their own turn, divided into subdomains.

A crucial semantic characteristic of a lexeme is its *taxonomic category* (T–category). The role of the taxonomic category in lexical semantics is similar to that of the part of speech in grammar. In particular, the T–category of a given verb determines the format of its lexicographic definition. The basic T–categories of the verb are State, Process, Action and Happening (approximately, after Z.Vendler).

The central domain of the lexical entry is the *definition*. The definition has a certain *format*. First of all, our definitions are divided into separate syntactically independent components (*features*), – approximately as in Wierzbicka (1987) and in contrast to syntactically coherent definitions in a "Meaning – Text" model. The components of a definition all have a predicative form, e.g.:

> 'The Subject is doing smth'
> 'The process in the Object takes place'
> 'The Object moves', etc.

The formatting of definitions becomes possible due to the fact that each component is considered to be a value of a certain *parameter* (parametrized definitions are used in Wierzbicka (1987) where the components of definitions of speech act verbs are identified as assumption, dictum and illocutionary purpose (cf. also partially formatted lexicographic definitions in Iordanskaja 1972; Zaliznjak 1983). Parameters serve as names for the subdomains of the definition. There are such parameters as 'activity', 'causation', 'the initial state', 'the final (new) state', 'process', 'result', 'limit' and the like. *A set of semantic components (presented as values of these parameters) with a partial syntactic ordering constitutes a definition.* The syntactic ordering is needed, e.g. because there is such a parameter as 'causation': for causation its arguments must be pointed out, otherwise it is pointless. This is why a set of semantic features, no matter how sophisticated, is not sufficient).

The format of the definition is the same for all the verbs of the same T–category and it is determined by the presence of a certain group of parameters. E.g. verbs of action are characterized by a set of parameters necessarily including the parameters 'activity', 'causation' and 'result (corresponding to the purpose of the Agent)'; the definition of a verb of happening invariably includes the parameters 'initial state' and 'resulting state'.

The system of T–categories is hierarchically organized; e.g. different values of the parameters 'limit' and 'causation' divide the T–category

Process into the subcategories Bound process (*tajat'*, 'melt') and Non–bound process (*kipet'*, 'boil'); the T–category Happening is divided into subcategories Ordinary Happening concerning the Subject (*upast'*, 'fall down': *The stone fell down*) and Happening concerning the Object (*zagorodit'* 'block': *The stone blocked the entrance to the cave* = 'The stone moved; as a result the entrance became blocked').

As far as the definition is concerned, it is not its exhaustiveness that is at stake. Only those meaning components are of interest that are common to a certain class of lexemes; thus, only recurrent semantic oppositions should be mentioned in the definition. Individual semantic peculiarities of a certain lexeme may not be taken into consideration, e.g. the verbs *napolnit'* i *zapolnit'* got identical definitions though they have perceptibly different meanings.

The information contained in other domaines of a lexical entry is to a certain extent predetermined by the definition. Thus, the aspect of the verb mentioned in domain (2) Morphology is deducible from the definition: the definition is different for two verbs differing only in aspect. The information about the aspectual form of a verb (domain 6) – the existence of an aspectual counterpart and the set of possible context– dependent aspectual meanings – also can often be deduced from the definition. For example, states (such as *znat'*, 'know'), as is known, are not used in the Progressive in English; and in Russian they do not allow the Progressive interpretation either, cf. *schitat'* 'believe', *razdrazhat'* 'bother'. Verbs denoting a constant feature or relation (*vesit'* 'weigh', *stoit'* 'cost') do not have standard aspectual counterparts; cf. the pair *vmeshat'* – *vmestit'* 'be able to contain' with a non–standard semantic relation where Ipfv implies the observer who watched the experiment, Glovinskaja (1982).

However, there are grammatical restrictions which are specific to a given word and cannot be derived from its meaning; e.g. according to Maslow (1948), the absence of the Imperfective (with iterative meaning) corresponding to *ochnut'sia* 'to come to oneself', must be treated as a lexical gap. Not the same thing with *dunut'* 'to make a blow': the absence of the iterative here is semantically motivated.

The information in domain (3) Arguments is only partially deducible from the definition. Each argument is characterized along the following three parameters:

(1) syntactic characteristics, i.e. surface case; we draw a tripartite distinction viz. the Subject and the (direct) Object and the Peripheral arguments. The prepositional and inflexional forms characterizing indirect objects and modifiers are not presented in the Database as they are considered to have no immediate semantic relevance;

(2) semantic characteristics, i.e. deep case; we distinguish, e.g. Agent, Patient, Experiencer, Goal, Place;

(3) taxonomic characteristics ('person', 'Physical object', 'substance', 'measurable parameter', 'situation' etc.).

Among the characteristics of the argument the deep case is often superfluous; indeed, the component 'Subject is doing smth' – in the definition of a verb of action – makes it clear that the Subject of this verb is an Agent). On the other hand, neither the syntactic nor the taxonomic characteristics of an argument can be deduced from the definition, cf. different syntactic roles of the Experiencer in *Ja ljublju* 'I love' i *Mne nravitsja* 'I like'; the T–category of the object of *stirat'* is only 'linen', not 'body' or 'floor' as for English *wash*.

The combination of an actant's surface case with its deep case characterizes the verb from the point of view of the *communicative perspective* in Jakobson–Fillmore's terms. For example, a group of verbs with completely affected Object, including, e.g. *napolniat'* 'fill', *zavalivat'* 'block' (see Paducheva & Rozina 1993) have a shifted perspective: their Patient is denoted by a peripheral argument, in the Instrumental case (*napolniat' vodoj*), whereas for the Object, which is a central argument, included in the perspective, its deep case is nothing but Place (though it is much more common for the Object to denote the Patient). The shift of the perspective gives a peculiar semantic effect of complete affectedness of the Object, cf. as an example (following Apresjan and Fillmore): *to load the sacks on the truck* is not the same as *to load the truck with sacks*: in the latter phrase, when Place becomes an Object and thus enters the perspective, the truck is understood as *filled* with sacks.

The Database is set up to be used in several different ways. First of all, the user is provided with semantic information about every particular lexeme. On the other hand, the system can supply the user with lists of verbs constituting various semantic classes – in principle, a class may be formed by any characteristics or by any set (or syntactic configuration) of characteristics contained in the Database.

Examples of queries: verbs with inanimate Subject, such as *paxnut'* 'smell'; physical actions that are not consistent with the use of an instrument, such as *sxvatit'* 'grasp'; actions that imply the use of an instrument; e.g. one can *polot'* 'weed' or *rvat'* 'tear' with one's own hands – in contrast to *paxat'* 'plough' or *rezat'* 'cut', which demand an instrument. Traditional semantic classes can also be described in terms of semantic components contained in the Database – mental acts (such as *choose*) and mental states (such as *know*); verbs of motion; verbs of possession; speech act verbs; emotional states etc.; e.g. verbs of movement are verbs whose definition includes either component 'X moves towards the Place' or a pair of components <'at t_i X is not in the Place', 'at t_j X is in the Place'>.

The user may also get information about some parameter as a set of components. Thus the following queries are possible: a complete list of T–categories of verbs; of arguments; all different values of the parameter 'causation'; T–categories of the argument Agent, etc.

Yet another mode of use of the Database is to test various hypotheses concerning the correlation between the meaning of a word and its surface behaviour. E.g. the Database makes it possible to check if it is true that the meaning of a Pfv verb always includes the component 'the beginning of a new state' as suggested by Wierzbicka in (1967). The answer is negative, cf. *posvetit'* 'to give light [for some time]', *zashchitit'* 'protect' (as in *The coat protected me from the rain*) which are oriented towards the end of the state and not the beginning.

To describe all possible correlations between the semantics of a word and its surface behaviour we use the notion of a *relevant* (semantic) class, which is cognate to the notion of a non-trivial (syntactic) class in Apresjan (1980). A relevant class is a class

(1) defined by a set or a configuration of semantic characteristics accounted for by the Database and
(2) constituted by lexemes that share some peculiarity of surface behaviour i.e. by some co-occurrence restriction, morphological combinability, a disposition to semantic derivation etc.

For example, a T-category is a relevant class: words of the same T-category (i.e. having the same characteristics in domain (4) combine with modifiers of time, purpose etc. in the same way; traditionally this type of combinability is considered to be free, while, in fact, it depends on the T-category.

It may be the case that a configuration of characteristics not only defines a relevant class, but also serves as an explanation of some peculiarity of surface behaviour; thus, the presupposition of activity having taken place that is inherent in the meaning of the negated Pfvs of verbs like *ugovorit'*, *ubedit'*, 'to persuade' (from *ne ugovoril* follows *ugovarival*, see Apresjan 1980) is explained by the semantic component 'good luck' which plays a part in the lexical decomposition of these verbs as opposed to such verbs as *wash* or *cook*.

2. Knowledge base

The other component of the system, its expert part, might be called the *grammar of the lexicon* – it is a fragment of theoretical lexicology, cf. the distinction between lexicography and lexicology as it is drawn in Nunberg & Zaenen (1992). The Knowledge base contains semantic rules and generalizations which make it possible to manipulate semantic and grammatical information contained in the Database. Our basic conviction is (cf. Wierzbicka 1988) that very many peculiarities of a word's behavior may be related to its meaning. The grammar of the lexicon should contain all possible generalizations of this kind.

At present the Knowledge Base includes the following parts:

(1) *T–categories of verbs.* The suggestion that tense–aspect meanings of verbs correlate with their lexical meaning has been made by a number of linguists. Vendler's classification gained the highest acclaim (cf. also the semantic classification of verbs suggested by Ju.S. Maslov in 1948). However, even Vendler's classification does not encompass all verbs and needs elaboration. There is an urge for compiling a complete list of Russian verb T–categories, and, consequently, a task to describe in the utmost detail the peculiarities of lexeme surface behaviour related to its T–category. It is clear, at present, that the T–category of a verb determines: the format of its lexicographic definition; the set of its basic actants (e.g. actions always have Agent and Patient); possibilities to have an aspectual counterpart and semantic categories of the latter; restrictions imposed on the set of aspectual meanings; the ability to motivate marked actionalities; e.g. verbs belonging to the T–category many–act Activity/Process, such as *stuchat'* 'knock', without restrictions motivate delimitatives (*postuchat'* [lit.] 'to knock repeatedly for some time') and inchoatives (*zastuchat'* 'to start knocking'); and, last but not least, combinability with modifiers of place and time.

Non–monotonic formalism is in order here: cancellable predictions – i.e. conclusions subject to invalidation – are better than none: we begin with the supposition of universal combinability and then make it more precise if and when new information is brought in. In other words, we make use of default reasoning drawing plausible conclusions from incomplete information in the absence of evidence to the contrary. Thus, when we describe co–occurrence possibility we first start a generalisation and then look for those factors that serve as 'brakes' (Russ.*tormoz*) to the generalization postulated. Cf. co–occurrence of the verb *ubit'* 'kill' with modifiers of place and time described in Paducheva (1991).

(2) *Semantic combinability.* Different kinds of restrictions on combinability are determined not by the T–category, but by some component of a lexicographic definition. E.g. abstract verbs of physical action, such as *rasshirit'* 'to widen' (with a semantic component 'the mode of action is not specified') do not easily combine with Instrument; thus **rasshirit' jamu lopatoj'* 'to widen the pit with a spade' is unacceptable, even though the action is almost surely conducted with an instrument; e.g. *kopat' jamu lopatoj* 'to dig a pit with a spade' is acceptable.. We suppose that many generalizations of this type are possible.

(3) *Context–dependent aspectual meanings.* The restrictions on the range of aspectual meanings of a verb are determined mainly by the T–category. However, sometimes they are related to some components of the definition; e.g. unacceptability of the use of an Ipfv verb in the meaning of Progressive (or restrictions on such use) may be determined by the component 'the process in the Object is non–synchronous with the Subject's activity', as in *streliat'* 'shoot', *vzryvat'* 'blow up', *otravliat'* 'poison', *ubivat'* 'kill'. The

component 'the process in the Object is very short', as in *udariat'* 'knock', is less definite in this respect.

(4) *Semantic derivation.* Systematic polysemy (according to Ju. Apresjan), or semantic derivation, is characteristic of the vast majority of verbs. The grammar of the lexicon must answer the question: what makes a certain semantic class of words *predisposed* to a certain type of semantic derivation. A principal source of regular polysemy is a metonymic transfer; cf. diathetic shift, such as *Storozh napolniajet bassejn vodoj – Voda napolniajet bassejn* 'The guard is filling the bathing–pool with water – Water is filling the bathing–pool'; *V svojej komedii on vysmeivajet intelligentsiju – Jego komedija vysmeivajet intelligentsiju* 'He mocks the intellectuals in his comedy – His comedy mocks the intellectuals'. Another example of a diathetic shift –*polot' sorniaki – polot' griadki* 'to weed ([lit. 'to weed weeds') – to weed the beds'. While metaphorical transfer is usually considered unpredictable, we would argue that this is not quite true. E.g. the T–category of a potential metaphorically derived meaning can sometimes be predicted. For example, it is highly likely that a verb of action will have a derived meaning of the T–category Happening (*porezal khleb – porezal palec* 'cut the bread – cut the finger').

Such are the problems approached by LEXICOGRAF. If these problems are even partially illuminated it will be a step towards a new level in the systematic description of the lexicon (Apresjan 1992).

Notes

* In 1993 the project was supported by the Russian Foundation for Fundamental Research.

References

Apresjan Ju.D. 1992. "Systemic lexicography" in *Euralex'92. Papers submitted to the 5th EURALEX International Congress of Lexicography.* Tampere .

Glovinskaia M.Ja. 1982. *Semanticheskie tipy vidovyx protivopostavlenij.* Moscow: Nauka.

Iordanskaja L.N. 1970. "Popytka leksikographicheskogo tolkovanija gruppy russkix slov so znacheniem chuvstva" in *Mashinnyj perevod i prikladnaja lingvistika.* 12. Moscow.

Nunberg G., Zaenen A. 1992. "Systemic Polysemy in Lexicology and Lexicography" in *EURALEX'92. Papers submitted to the 5th EURALEX International Congress of Lexicography.* Tampere.

Paducheva E.V., Rakhilina E.V. 1990. "Predicting co–occurrence restrictions by using semantic classifications in the lexicon" in *COLING–90. Papers presented to the 13–th International Conference on Computational Linguistics.* Vol.2. Helsinki.

Paducheva E.V., Rozina R.J. 1993. "Semanticheskij klass glagolov polnogo okhvata: tolkovanije i leksiko–semanticheskije svojstva". *Voprosy jazykoznanija* 6. Moscow.

Wierzbicka A. 1987. *English Speech Act Words.* Sydney, etc.:Academic press.

Wierzbicka A. 1980. *Lingua Mentalis.* Sydney, etc.: Academic press.

Zalizniak Anna A. 1983. "Semantika glagola 'bojat'sja' v russkom jazyke" *Izv. AN SSSR. Ser. lit. i jazyka* 1. Moscow.