**Meta**

# ROSS: Semantic Dictionary for Text Understanding and Summarization

Nina N. Leontyeva

Volume 40, numéro 1, mars 1995

URI : id.erudit.org/iderudit/003464ar
DOI : 10.7202/003464ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN 0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Leontyeva, N. (1995). ROSS: Semantic Dictionary for Text Understanding and Summarization. *Meta*, *40*(1), 178–185. doi:10.7202/003464ar

Résumé de l'article

Le dictionnaire sémantique russe à usage général (ROSS) est un outil pour I'analyse sémantique et informationnelle de tout texte russe cohérent. La structure de ROSS reflète la philosophie et les niveaux de représentation adoptés par le système de comprehension de texte POLIText actuellement en cours d'implementation à l'Institut des États-Unis et du Canada.

## ROSS: SEMANTIC DICTIONARY FOR TEXT UNDERSTANDING AND SUMMARIZATION*

***RÉSUMÉ***

*Le dictionnaire sémantique russe à usage général (ROSS) est un outil pour l'analyse sémantique et informationnelle de tout texte russe cohérent. La structure de ROSS reflète la philosophie et les niveaux de représentation adoptés par le système de compréhension de texte POLIText actuellement en cours d'implémentation à l'Institut des États-Unis et du Canada.*

### THE PURPOSE OF ROSS

The Russian general-purpose semantic dictionary, ROSS, is a tool for the semantic (both linguistic and informational) analysis of texts. The rich semantic information contained in the dictionary makes possible local (within one phrase) semantic interpretation as well as the semantic analysis of coherent texts. Some zones of the dictionary provide the possibility of logico-situational analysis of a text and a link to different domains.

The present version of ROSS is designed, firstly, to construct a base of textual facts (BTF) for a given collection of Russian texts. The vocabulary of the the present version is based on communications about political events. The procedure proposed for BTF construction (Leontyeva 1992) enables any user to form an individual BTF ordering the desired degree of compression of the initial content. Secondly, the dictionary is to ensure multi-language, above all, Russian-to-English, knowledge based machine translation (KBMT) (see Johnson, King and Des Tombe 1985; Nirenburg 1989; Papagaaij 1986).

Though similar in many aspects to dictionaries with rich semantic data belonging to some MT

systems (*e.g.* Papagaaij *et al.* 1986, Шаляпина 1974; Апресян *et al.* 1992), ROSS differs from them essentially. ROSS as well as the main semantic dictionary (Леонтьева Н Н *et al* 1979) of the FRAP system (French-to-Russian automatic translation implemented by 1985, (see *Machine Translation and Applied Linguistics* 1987 and Hutchins 1986) implies a clear-cut distinction between the syntactic and semantic levels of text representation. Moreover, semantics itself is split into several sublevels, making possible very important information processes in a system of text understanding, such as tuning to the required domain, shifting the focus of interest, compressing and varying of the initial text content, etc.

The nature of semantic information in ROSS will be clearer if we outline the stages in the understanding of natural language texts (NLT), that is, the linguistic model underlying both text analysis and ROSS as its instrument, which are being implemented in the system POLIText.

## STAGES OF THE NLT UNDERSTANDING PROCESS

For us, text understanding is very closely related to information analysis. The text understanding process is a sequence of operations that generate information with the desired degree of completeness from any given NLT for any given reader. We distinguish three major steps in this process (Leontyeva 1987):

1) a complete linguistic analysis of all NLT sentences, which includes all traditional levels of parsing and produces a sequence of representations for a given NLT — graphematic, morphological, syntactic and local semantic: GraphR, MorphR, SynR, local SemR, the last one having the form of initial semantic space (SemSpace). Similarly, any intermediate, not final syntactic structure may be called "syntactic space" (SynSpace); we apply this notion also for the whole text when it is partially processed by the syntactic analyser (*e.g.* only noun and verb phrases).

This step (or stage) may be characterised as "local understanding" of a text;

2) construction of an internal textual representation in terms of textual fact (TF) constituents. The construction of TF units is a generation process which begins within analysis. This second step of text analysis is "global understanding": one TF may be gathered from the lexical material of the whole text and represented in the global TF structure with the required degree of compression;

3) developping the TF structure into information structure in terms of the given domain and / or according to the user's request or information interest. The global TF structure must be matched against the dictionary of the chosen domain objects, *i.e.* "rewritten" in terms familiar to the given addressee of information (that is, "matched" against and tuned to the lexicon of the user). It may be focused around the required notions as well. This third stage produces units valid, or informative, for the reader or any other perceiving intellectual system. We therefore call it "relative understanding".

We do not consider the process of translation into another NLT as one more level of text understanding. The final representation of every above level could be translated into some another NLT or even into units of another semiotic system (*e.g.* the "language of actions"), — the results of translation may rather show what kind of understanding has occured: local, global or specialized (relative).

In planning translation component for our POLIText system we have chosen the second and third kinds of understanding. We proceed from the assumption that they involve less computation than every-detail-of-the-text understanding, not to mention the fact that the user needs not the finest information from the text, but only essence (*cf.* Shank's approach). So, we aim at building the TF structure or the preceding structure — situational representation (SitR) — and at translating them into another natural language. In this context, the task of translation is similar to text generation system tasks, though more complex.

## THE STRUCTURE OF ROSS

The structure of ROSS is hierarchical: the lower level comprises fields that assume concrete values. The higher level is represented by zones, *i.e.* names of groups of fields. The present version of the dictionary contains 10 zones, comprising over 50 fields. The list of zones is given below.

(Note that all the names of zones and fields, as well as the examples below, are in Russian in the original version of the ROSS dictionary. We have converted most of them into English to make the paper accessible to a wider scientific audience.)

GEN: General information on the word (C) to be described;

MORPH: Morphological characteristics, including the word-forming potential;

SYN: Syntactic properties of the word C (syntactic class, specific syntactic structures, complements);

LEX: Lexical combinability, cliches, idioms, lexical functions;

SEM: Semantic characteristics of C, semantic valencies, hypothetic realisation of valencies within one phrase and through the whole text. Actant lexical functions and non-standard questions to actants. Necessary corrections of and other semantic operations on the C-containing primary semantic representation;

THES: Thesauric links of C as a concept within a certain domain, C-containing terms, explicit definitions of concepts, encyclopedic functions;

SIT: Structure of situations related to the semantics of C and relevant to the given subject area, relationships among situations, with particular emphasis on relations of temporal precedence ("before") and consequence ("after");

PRAGM: Pragmatics of the situation in the domain: events, inferences, presuppositions, evaluation of the event or its parts;

EQUIV: Equivalents of the initial lexeme (and terms within the entry) into other languages (with selection criteria);

COMM: Comments of the lexicologist (+ name).

Some zones and their fields need more detailed consideration especially when they are non-trivial or different from standard patterns.

**Zone GEN. Fields: TITLE, TYPE, CAT, NO, REF, OTHER, ILL**

TITLE — name / title of the entry, *i.e.* any lexical unit which can be described semantically. In the present version of ROSS, the entries are represented by one-word lexical units (TYPE = LEX).

● CAT — semantic category of the lexeme related to the way it is represented in the semantic notation. This field can acquire the following values:

● LBL — word-label, takes up position A or B in the semantic relation R(A,B). The words of this category have the maximal initial informational weight. This is the largest, open and mobile class that must cover the lexical nucleus of the chosen domain. With ROSS tuned to different domains, the greatest changes in the composition, system of meanings, system of references to the vocabularies of the subject area occur in the LBL categorie. There are two subcategories in category LBL: SIT and OBJ (SITuation, *e.g. war, discussion*, and OBJect, *e.g. President, book, bill*). Those words-labels which can not be specified as SIT or OBJ yield the LBL symbol (*e.g. problem*). The field CAT provides the highest level of classification; further semantic differentiation of words-labels is derived from fields SEMF (semantic features) and VAL (valency) of the zone SEM.

● REL — relations (or semantic relations, SemRel) occupy position R in the formula R(A,B). These words have lower, as compared to LBL, initial informational weight since they denote links (relations) between units of other categories. In the majority of cases, REL is ascribed to auxiliary parts of speech (prepositions, conjunctions, punctuation marks). Nouns that coincide with the name of some semantic relation (*cause, part*) fall into the category of aspect words; they become relations only at the next step of analysis.

● ASP — aspect words. They have a common structural feature, *i.e.* the way they are represented in the semantic notation: the lexeme defines the name of the semantic relation and fills in its first place in the formula R(A,B): PARAMETER(size,B); PART OF(member,B).

Aspect words are non-homogeneous both grammatically (nouns, adjectives, verbs, adverbs) and semantically, and they can be specified in SEMF and VAL fields of SEM zone (see below) by more specific semantic relations: STAGE(A,B), MODALITY(A,B), NAME(A,B), FUNCTION(A,B), INDENTIFIER(A,B), and others. ASP words have lower informational weight than the above.

OPER – words-operators whose individual behaviour is described by algorithms of transformation of a part of semantic representation (SemR): each operator has its own sphere of action and leads to a fairly complicated transformation of already constructed part of SemR. This category comprises pronouns (*he, they*, etc.), particles (*even, only*), adverbs (*respectively, particularly, contrary*), parenthetical words (*in particular, incidentally*), quantors (*every, all*), etc. These units operate on SemSpace structure (at the stage of semantic analysis proper) when there already exist units that can become arguments of the semantic relations that are introduced.

The informational weight of these lexemes given in the dictionary is minimal, but the transformations they start may change the informational weight of term B in the semantic formula.

● ELSE – a quasi-category for doubtful and unclear cases.

NO – the field to indicate the number of meanings (in brackets) and the ordinal number of the given one: *key1*(2).

REF – reference to a word or its individual zones and fields (in ROSS) with similar information; REF is used for unification of information and effort-saving input, *e.g.* TITLE – *key*; NO – 2(2); REF — *key1* (MORPH, SYN), which means that *key2* (*e.g. This is the key to the solution of our problem*) has the same MORPH and SYN zones as key1 (*e.g. Open the door with a key*).

OTHER – other meanings. The field indicates (in a free form) other meanings (not yet described in the dictionary) to provide a contrasting example or to prepare the ground for further description in the dictionary.

ILL – illustrations: an example of the most typical context for the given meaning of the word. The word itself is substituted by ~. Illustrations can be also given in the fields ILL1-7 of the SEM zone with the emphasis on each of the satisfied valencies.

**Zone MORPH. Fields: MCL, CHANGE, DERIV**

MCL – morphological class; each class has its own morphological characteristics. For example, for verbs: GASP – grammatical aspect of the verb: (perfect, imperfect, perf / imperf, and so on).
TRANSIT = transitivity ("+" and "-").
REFLEX = reflexivity ("+" and "-").

CHANGE – availability of the paradigm ("+", "-").

DERIV – morphological derivation. The field denotes the ways to form other parts of speech from the given TITLE, as well as reflexive forms, comparative and superlative degrees of comparison, negative forms. For verbs, the field indicates:

■ the way to form the aspect pair, *e.g.*
TITLE = решать; DERIV = PERF (-з + нть);

■ the way to derive the passive form, *e.g.*
TITLE = решать; DERIV = PASSIV (anal, ~ся), *i.e.* analytically and with the help of the postfix.

■ potential word formation through prefixes, *e.g.* TITLE = делать; DERIV = PREF (на~), (за~).

More complex forms of derivation are described in zone LEX and field ALF of zone SEM.

**Zone SYN. Fields: SynCl, ROLE, COMPL, CONSTR**

SynCl — syntactic class: NOUN (noun), VERB (verb)...

ROLE — role in the phrase: NP (noun phrase), VP (verb phrase), ATTR — attribute, etc.

COMPL — complement, "strong government", in the form of a) a pair "preposition and the symbol of the noun's case", *e.g.* "В + Acc"; b) the symbol of the noun's case (6 cases in Russian); c) the name of syntactic class.

CONSTR — prediction of non-trivial syntactic constructions containing the title word as the key one.

**Zone LEX. Fields: P-left, P-right, COLLOC, LF**

P-left, P-right — left and right parts of fixed phrases.

TITLE = *power*; P-left = *under* ~;

COLLOC — different non-free collocations:

TITLE = peace; COLLOC = *the cause of peace; peace champion*;

LF = lexical functions (according to the "meaning-text" theory, see Mel'čuk 1988).

TITLE = *deputy*; LF = $Oper_1$: *have a deputy (of their own)*; $Oper_2$: *be a deputy*; $Caus_1Oper_2$: *elect a* ~; S1: *elector*; Fact: *justify the trust of electorate*.

**Zone SEM. Fields: SEMF, VAL1-7, SEMF1-7, ALF1-7, QUEST1-7, SYN1-7, MORPH1-7, ILL1-7, POS, INCOMP, COMPAT, ADD, CORR, RESTOR, OBLIG, MODIF, TRANSF(SEMF), TRANSF(VAL)**

SEMF — semantic feature, or characteristic (see the list below p. 183).

VAL — a set of valencies of the word (up to 7); candidates for filling in the slots of valencies are introduced by symbol Ai (i=1,2...7). The form of notation in VAL field: R,Ai,C or R,C,Ai. TITLE = *message*; VAL = agent,A1,C; addressee,A2,C; topic,A3,C; content,A4,C.

SEMFi — semantic feature of each possible actant of each valency.

ALFi — actant lexical functions:

TITLE = *elect*; ALF1 = *elector, electorate*.

QUESTi — non-trivial question to each actant.

SYNi — syntactic characteristics required of actants; syntactic relations between the actant and C.

MORPHi — morphological characteristics required of each actant.

ILLi — illustrations for each valency.

POS — interrelated positions of C and its actants in the phrase:

TITLE = *delegate*; MORPH1 = 1. from + N; MORPH3 = 1. of + N; POS = < MORPH3.1 < MORPH1.1 (*delegate to the Congress from the state of Michigan*).

INCOMP — incompatibility of morphological characteristics of actants; written in the form of conjunction of MORPH's.

COMPAT — compatibility of MORPH's and/or SEMF's; written in the form: inference, slash, condition, *e.g.*

TITLE = *compensation*; VAL = agent,A1,C; ob,A2,C; addressee,A3,C;

MORPH1 = 1. poss.pronoun; 2. adj; 3. of + N; 4. by + N;

MORPH2 = 1. of + N; 2. for + N;

INCOMP = MORPH1.3 and MORPH2.1

COMPAT = MORPH2.1/MORPH1.4 (If there is MORPH1.4, then there is also MORPH2.1), *e.g. compensation of damages; compensation of damages by an insurance company; compensation of the insurance company was too little*.

ADD — additional semantic relations among the actants:

TITLE = *compensation*; ADD = 1. passive actant,A2,A3; *e.g. compensation* to NN(A2) *for the damage*(A3).

CORR — the rules of correction of the valency structure written in the form: initial SemRel, —> (symbol of transition), resulting SemRel, /condition; *e.g.*,

TITLE = *ruin*; VAL = agent,A1,C; passive actant,A2,C;

CORR = agent,A1,C, —>, cause,A1,C/SEMF1 \= animated; *e.g. The flood*(A1) *has ruined the village*(A2).

RESTOR — the rules for reconstructing a member of the valent structure, in particular, the subject of action expressed by infinitive:

TITLE = *order*; VAL = agent,A1,C; content,A2,C; addressee,A3,C.

MORPH2 = 1. INF; RESTOR = agent,A3,A2/SEMF2 = action; *e.g., The commander*(A1) *ordered his soldiers*(A3) *to remain*(A2) *seated*.

OBLIG — obligatory realisation of a valency:

TITLE = *order*; OBLIG = 2.

TRANSF(SEMF) — transformation of semantic characteristic, *i.e.* regular change of semantic characteristic of C under certain conditions (presence of certain grammatical features, or presence of a certain context), *e.g.*

TITLE = *stress*; SEMF = phenomenon; TRANSF (SEMF) = parameter/value,A1,C; (*cf. Stress is a physical phenomenon* vs *The stress was equal to 1,5 j/kg*).

TITLE = *audience*; SEMF = receptacle; TRANSF (SEMF) = anim, assemblage.

TRANSF(VAL) = transformation of valencies, written in the form: Ri, Rj, which means that a strong syntactical valency which serves to express Ri becomes a means of expression for Rj, while Ri becomes completely devoid of a strong syntactic expression, *e.g.*

TITLE = *publish*; VAL = agent,A1,C; ob,A2,C; loc,A3,C; MODIF = agent, loc (*The newspaper*(A1)

*has published an article*(A2) —> *Somebody*(A1) *published an article*(A2) *in the newspaper*(A3));

**Zone THES. Fields: DOM, TERM, HIGH, ENCYC, VAR, ASSOC, EXPLIC**

DOM — domain, or subject area, one or several of the following: economics, politics, jurisprudence, military problems, general vocabulary. In brackets, we indicate the informational weight of the word (5 — for the most informative words, 1 — for terms indifferent to given DOMain), *e.g.*
TITLE = *election*; DOM = politics(5).

TERM — C-containing terms which must be then incorporated to ISCRAN databases, *e.g.*
TITLE = *convention*; TERM = *Geneva Humanitarian Convention*.

HIGH — generic concept for C.
TITLE = *component*; HIGH = *part*;
TITLE = *bill*; HIGH = *document*;
TITLE = *deputy*; HIGH = *representative*; *human and socio*.

ENCYC — encyclopaedic functions (Anti, Mult, Pars, Param, Sing):
TITLE = *senator*; ENCYC = Mult: *Senate*;
TITLE = *deputy*; ENCYC = Mult: *Parliament*, *Congress*, *Supreme Soviet*;
TITLE = *profit*; ENCYC = Anti: *damage*, *loss*;
TITLE = *meeting*; ENCYC = Sing: *participant*, *delegate*.

VAR — variants of C (of different degrees of closeness):
TITLE = *compensation*; VAR = *contribution*, *insurance*, *reimbursement*, *indemnification*.

ASSOC = other associated concepts:
TITLE = *deputy*; ASSOC = *election campaign*, *legislative bodies*, *campaigning centre*, *canvassing*, *ballot*, *ballot paper*, *polling station*, *constituency*, *ballot-box*, *suffrage*, *electoral franchise*, *eligibility*, *polling district*, *glasnost*, *perestroika*

EXPLIC — definition or explicit description for words with informational weight 4 and 5; can be formulated in terms of SIT or taken from an encyclopaedic dictionary.

**Zone SIT. Fields: MSit, ESit (ES1-n), PRECED, POST**

MSit — main situation, designated as comb,C,[A], where [A] means the set of actants, and comb indicates its compatibility with C; this notation signifies that the main situation is represented by multi-term predicate C.

ESit — description of elementary situations in the form of a set of semantic relations R,A,B, *e.g.*
TITLE = *export*; VAL = agent,A1,C; ob,A2,C; end-point,A3,C.

ESit = 1. belongs-to,A2,A1 / SEMF1 = organisation; 2. loc,A3,A2 / SEMF1 = space, state; 3. belongs-to,A2,A3; 4. loc,A3,A2.

Further on, elementary situations are referred to as ES1, ES2, etc.

PRECED — elementary situation preceding MSit.

POST — elementary situation following MSit, *e.g.*
TITLE = *export*; PRECED = ES1&ES2; POST = ES3&ES4.

**Zone PRAGM. Fields: EVENT, CONCL, PRESUP, EVAL, LOG**

EVENT — event (main situation denoted by the word and / or one of its actants with the greatest informational weight which may be a nucleus of some event in the indicated domain), *e.g.* EVENT = A3.

CONCL — standard inference in the form of a production rule: If (SIT1), then (SIT2); if (SIT2), then NON (SIT3).

PRESUP — presupposition (names of situations already introduced in some field, or being formulated by the linguist, which are indispensible for C to be true).

EVAL — evaluation ("+", "-", "0", or "?", the latter signifies that the evaluation depends upon certain conditions), *e.g.*
TITLE = *export*; EVENT = MSit; EVAL = ? / (2) (is inherited from A2).

LOG — more complicated situations that characterise the logics of the events, formulated in terms of SIT and production rules.

**Zone EQUIV. Fields: ENG, FR**

ENG — English equivalents with selection conditions.

FR — French equivalents (not indicated in the present version).

**Zone COMM. Fields: COM, NAME**

COM — commentary in a free form.

NAME — name or any other designation of the lexicologist.

## SOME REMARKS ON THE SEMANTIC COMPONENT USING ROSS

The semantic component (SemComp) is responsible, first of all, for the semantic interpretation of all the units of the initial text (or of its syntactic structure, even incomplete). In our system, SemComp receives syntactic space (SynSpace) as a source and makes possible the interpretation of those nodes and relations of SynSpace on which information is available in one of the semantic dictionaries (ROSS1, ROSS2 ... ROSSn). Secondly, SemComp must ensure the matching of textual units to and their compatibility with the units of the chosen domain and / or knowledge representation (KnowR). Thirdly, it does the same in regard of the units of information representation: it is SemComp that is responsible for correspondence between two languages — the KnowR of the system and the KnowR of the user.

The initial semantic structure, semantic space (SemSpace), is a result of a local (within the sentence boundaries) semantic interpretation of the initial text in terms of binary relations only, relations not associated with a particular notional sphere or domain. It may be accompanied by the structural decomposition of some lexemes which have valency structure, or by combining some simple nodes of SynSpace into one complex term, all operations being guided by ROSS and the thesaurus of political terms.

From the structural point of view, SemSpace is a sequence of formulas meeting definite rules of semantic grammar, —these are the now well-known semantic triads of the form R(A,B), where "A" is a dependent notion, "B" is a chief notion, "R" is a semantic relation between them. We have used this form of semantic representation since 1965, not only for verbs as is in frame-based languages, but for all the lexical material of any text, be it Russian, French or any other.

Here are some examples of semantic features and semantic relations.

SEMANTIC FEATURES (fragment)
ARTEFACT — man-made object: *table*, *weapon*, *sputnik*;
ABSTRACT CONC. — *theory*, *system*;
SUBSTANCE — *sand*, *clay*, *uranium*;
PERCEPTION — *hear*, *see*, *observe*;
HARM — anything connected with life hazards: *war*, *threat*, *damage*;
STATE — *country*, *USSR*;
MOTION — *go*, *carry*, *float*, *move*, *boat*, *train*, *car*;
CHANGE — *increase*, *augmentation*;
INTELLIGENCE — *opinion*, *lecture*;
SPACE — *river*, *platform*, *region*;

SEMANTIC RELATIONS (fragment)
AUTHOR, A,B — *poem*(B) *by Kay*(A);
ADDRESSEE,A,B — *to send*(B) *smth to the press*(A);
ASPECT,A,B — *sort*(B) *by size*(A);
TIME,A,B — *went*(B) *in 1991*(A);
VALUE,A,B — *velocity*(B) *of 100km per hour*(A);
IDENTIFICATOR,A,B — *President*(B) *Gorbachev*(A);
NAME,A,B — IBM(A) *Company*(B);
SOURCE,A,B — *laser*(A) *ray*(B);
QUANTOR,A,B — *each*(A) *party*(B);
MODALITY,A,B — *must*(A) *come*(B);

The Semantic Grammar is defined by two axes:

- syntagmatics, that is rules of combining words into formulas. Ex.: REASON,A,B , where A (and also B) = SIT or a whole semantic formula; NAME,A,B , where A = NAME(A, ?), B = any;
- paradigmatics, that is rules of substitution for terms and relations. Ex.: IDENTIFICATOR,A,B can be specified as (that is, substituted in the course of semantic analysis by) NAME,A,B or SYMBOL,A,B. TIME,A,B can be specified by two formulas: STARTING POINT(?,B) and FINAL POINT(?,B), etc. There exist more complex relations between semantic units (relations, formulas and other units).

As for features, they are also partially hierarchizied, ex.: SITUATION > PREDICATE > PROCESS > ACTION. This Grammar is used for compression and other transformations over the SemSpace.

SemSpace is a homogeneous structure. All transformations within it aim at the construction of more "intelligent" units, called situations (Sits) and "textual facts" (TFs). Both are multiplace predicates. The former are built on the basis of dictionary descriptions of words and the lexical material of the text, that is they are "local" semantic units. The latter must be "generated" on the basis of the quality of Sits constructed in conformity with rules of global semantic analysis, and the quality of coherent text structure. A TF and its actants are calculated as notions with maximum informational weight (maximally meaningful) for the given text and/or the given thesaurus, domain and pragmatic orientation. (See in Leontyeva 1992 the first attempt to formulate some rules of similar pragmatic analysis.)

Example of a textual fact:

TF1 = COUP D'ETAT (1,2,3,4,5)
Lexical variant: *seizure of power*
1. Agent = the USSR State Emergency Committee (SEC)
   Variant = the Soviet leadership
   Identification = G.Yanayev, V.Pavlov, O.Baklanov, B.Pugo, V.Starodubtsev, A.Tizyakov, V.Kruchkov, D.Yazov
2. Counter-agent = [former power, President Gorbachev]
3. Cause = destabilisation of political and economic situation in the USSR: Sit
4. Goal = to overcome economic and political crisis in the USSR: Sit
5. Time: Starting point = August 19, 1991: Sit

The base of textual facts BTF for a given corpus of texts will be a condensed, secondary structure, a result of comparison, generalization and generation of new units which can be translated into superficial forms other than the initial ones, as in the case of knowledge based machine translation.

Text generation (TG) from BTF can yield new texts, such as English or French summaries of initial Russian texts. Some of the existing multi-language TG projects and systems (see McKeown 1985; Kittredge *et al.* 1991) allows us to look forward to valid translations from BTFs.

## IMPLEMENTATION AND PERSPECTIVES

Some components of local linguistic analysis in the POLIText system, particularly graphematical, morphological and, in part, syntactic (noun phrases, verb phrases) ones, have already been implemented. At present, domain-oriented analysis is being created, matching every noun phrase against the thesaurus of political terms and thus constructing nodes which

obtain the denotation status in a future semantic structure.

The next step is the semantic interpretation of all other nodes and relations based upon the ROSS dictionary. The prototype of this module and of the domain-oriented module were programmed by N. Lucashevich.

The creation of the Russian semantic dictionary is the most labour-consuming part of the project. This work is done by a team of five lexicologists of the Moscow State Linguistic University (M. Shatalova *et al.*). ROSS has been entered into IBM PC by means of a special scenario in an online mode (man-machine dialogue). A special data base management system, derived from the Clarion Database, has been evolved by B. Dobroff, for more flexible storing, updating and development of ROSS, including access from applied programs.

## CONCLUSION

When computer systems faced the severe necessity to provide the description of the semantic component, the traditional linguistics fell short of suitable semantic theories. Meanwhile no serious systematic text processing would be possible without a full employment of the lexicon as a whole. The most vulnerable area proved to be that of interaction with data and knowledge bases. Usually this problem would be solved by computer scientists, albeit scholars in the field of detailed linguistic analysis made a successful attempt to "crush through" to the KnowR (Апресян *et al.* 1992; Boguslavskij and Tsinman 1991).

A full picture of computational semantics was proposed in Pustejovsky 1992, the central idea being that "word meaning is highly structured, and not simply a set of semantic features." Two tasks — "a fully compositional semantics for natural language and its interpretation into a knowledge representation model", as well as "the mapping from the lexicon to syntax" — are considered to be the most important.

We see a certain correlation between these ideas and our approach to levels of semantic description.

NINA N. LEONTYEVA
*Institute of the USA and Canada,*
*Russian Academy of Sciences, Moscow, Russia*

### Notes

* The work on the ROSS dictionary was supported by the MacArthur Foundation (a 1994 grant) and the SOROS (Cultural Initiative; a 1994-1995 grant). I would like to express my heartfelt gratitude to both bodies for their support.

## REFERENCES

BOGUSLAVSKIJ, I. M. and L. L. TSINMAN (1991): "Semantics in a Linguistic Processor", *Computers and Artificial Intelligence*, n° 3, pp. 3-20.

HOVY, E. F. (1988): "On the Study of Text Planning and Realization", *AAAI Workshop on Text Planning*, St. Paul.

HUTCHINS, W. J. (1986): *Machine Translation: Past, Present, Future*, New York, 382 p.

Шаляпина З.М. (1974) Англо русский многоаспектный автоматический словарь (АРМАС) // Машинный перевод и прикладная лингвистика Вып. 17, pp. 7-67

Апресян Ю., Богуславский И.М., Иомдин Л.Л. и др. (1992) Лингвистический процессор для сложных информационных систем. М.: Наука, 256 p.

Леонтьева Н.Н., Кудяшова И.М., Соколова Е.Г. (1979) Семантическая словарная статья Iв системе // ФРАП Ин-т русского языка АН СССР, Вып. 64 p.

JOHNSON, R., KING, M. and L. DES TOMBE (1985): "EUROTRA: a Multilingual System Under Development", *Computational Linguistics*, Vol. 11, n° 2-3.

KITTREDGE, Richard, KORELSKY, Tanya and Owen RAMBOW (1991): "On the Need for Domain Communication Knowledge", *Computational Intelligence*, Vol. 7, n° 4, December.

LEONTYEVA, Nina N. (1987): "Stages of Information Analysis of Natural Language Texts", *Int. Forum Inf. and Docum.*, Vol. 12, n° 4.

LEONTYEVA, Nina N. (1992): "Textual Facts as Units of Coherent Text Semantic Analysis", *International Workshop on the Meaning-Text Theory*, Karen Haenelt and Leo Wanner (Eds.), Darmstadt.

*Machine Translation and Applied Linguistics. Problems Related to the Development of Automatic Translation Systems* (1987): Issue 271, Moscow's Maurice Thorez State Institute of Foreign Languages.

MCKEOWN, Kathleen R. (1985): "Discourse Strategies for Generating Natural-Language Text", *Artificial Intelligence*, 27.

MEL'ČUK, Igor A. (1988): "Paraphrase et lexique dans la théorie linguistique Sens-Texte", *Cahiers de lexicologie*, 52-1, pp. 5-50, 53-2, pp. 5-53.

NIRENBURG, S. (1989): "Knowlege-based Machine Translation", *Machine Translation*, 4-1.

PAPAGAAIJ, B. C. (1986): "Word Expert Semantics. An Interlingual Knowledge-based Approach", *Distributed Language Translation*, Toon Witkam (Ed.), Dortrecht, Reverton, 254 p.

PAPAGAAIJ, B. C., SADLER, V. and A. P. M. WITKAM (1986): "Experiments with an MT-directed Lexical Knowledge Bank", *COLING-86*.

PUSTEJOVSKY, James (1992): "The Generative Lexicon", *Computational Linguistics*, Vol. 17, n° 4.

ROESNER, D. (1987): "The Generation System of the SEMSYN Project. Towards a Task-independent Generator for German", *First European Workshop on Language Generation*, Paris.