Semantic Resources for Textual Content Compression

Nina N. Leontyeva
Research Computing Center
Moscow State University
Leninskije Gory, Moscow, 119899
Russian Federation
leont-nn@yandex.ru

Abstract

We discuss an information-linguistic model designed to extract the most essential parts of the content from any text, being at the same time a model of "soft" understanding of texts. The linguistic aspects of the model and the properties of the semantic dictionary ensuring the content analysis with compression of the text structure are considered. Our dictionary resources serve an instrument for building a multilevel textual semantic structure. The primary semantic representation of a whole text may be used as the starting point for generating complex linguistic units ('Situation' and 'Textual Fact') as candidates for Events.

1 Multiple Dimensions of Textual Structure

The shortest representation of a text within an Information System is a "Search Pattern" of this text in the form of a simple list of key-words and/or terms. Any short exposition of the initial text – an abstract, a summary, an annotation, a free rendering, a paraphrase, etc., be it human or computer-aided product, – gives some dimensions to textual representation. The well-known aspect of text understanding in Information Extraction (IE) systems is "partial understanding," with different degrees of compression of the initial text content. IE-systems construct different special frames: of persons (including VIPs), organizations, geographical names, parametric or referential information, etc. (see, e.g., Grishman, 1999 and materials of MUC, 1-7). IE-systems that create various TASK- or USER-oriented frame structures often deal with very simple linguistic data. (Mani, 2001) regards those systems as different means of text compression: "condensation of document information content for the benefit of the reader and task." These and similar sets of specific databases (DBs) form several "content dimensions" of a text.

On the other hand, we can obtain the most detailed representation of a text as a result of linguistic analysis. First of all, theoretical linguistics does not allow for any loss of information expressed by the text; the same is true for Semantic structures or representations (SemS, SemR), being elaborated in Meaning-Text Theory (MTT). Systems of NLP-analysis based on MTT try to retain even the smallest portion of meaning expressed in a sentence, which is very important, especially for Machine Translation (e.g., the ETAP system). The whole text semantic structure appears in ETAP as a sequence of sentence SemRs. It is a special kind of understanding – the purely linguistic one. Including a set of linguistic structures in the set of whole text understanding (WTU) structures we emphasize the principal role of linguistics in modeling intellectual systems that must rely on full linguistic analysis.

In an inverse task of Text Generation (TG), the source structure may be a table, a schema, a picture or a specific database (see McKeown, 1988 and many others issues). If some of them may serve as a basis for producing a natural text (as in TG-systems by Kittredge and Iordanskaja), then the source structures may be regarded as some compressed content of a future text, the real "textual" representations, thus enlarging the set of compressed products.

Any set of the SemRs mentioned above and of similar ones would present a kind of multidimensional semantic (or simply, **multisemantic**) structure modeling different modes of human text understanding.

The final multisemantic structure of a given text or of a corpus would serve first of all the basis for differentiated search processes. The information retrieved may be combined into some compressed semantic representation of a corpus (in our model it would be a BTF, or Base of Textual Facts, see below). This compressed product may be expanded in its turn into natural text: summary, etc.

"Text understanding" denotes a sequence of operations that extract information from any given text with the completeness degree desired by the user. It is a central point of the proposed **Information-Linguistic Model** (ILM), aimed at the "soft understanding" of texts, **mostly by linguistic methods of analysis**. The task of ILM is to bridge the gap between Information Systems and purely Linguistic Systems. Starting from the second one (= understanding any "textual" material) and using all available information resources, ILM aims at building the more meaningful units than those achievable in both approaches separately.

The principle task of semantic analysis (SemAn) in ILM is constructing a dynamic structure that ensures transition from linguistic units (mostly lexemes, the nodes of a syntactic or syntactic-semantic structure) to complex content units of different kinds, familiar to users. In our model it is a primary SemR (Semantic Space, SemSpace), built by using two instruments: the initial Grammar of Semantic Relations (SemRel), that is, syntagmatic and paradigmatic organization of the metalanguage of SemRels, and a Semantic Dictionary (SemDict), both having to be adapted to different information tasks. The SemSpace as the first level textual structure is to be followed by more than one level of SemRs: SitR, EventR, Textual-Fact-Representation (TF-R), all of them aimed at receiving the conceptual status.

2 Outline of Semantic Grammar

A language for the level of SemAn in our model has as its elementary unit a record of the form R(A,B), with R being a binary semantic relation, and A,B, its terms. The same semantic language (sometimes enriched by specific domain relations) can be used for the representation of any domain-knowledge text. The whole list has about 60 SemRels, but only half of them are used in the course of SemAn. A fragment of **semantic relations** list with textual examples follows:

Author (A,B): *novel* (B) *by Hemingway* (A); *Pushkins* (A) *poem* (B)
Addressee (A,B): *to send* (B) smth. *to the press* (A)
Actant (A,B): general name for any type of actant
Quantity (A,B): *two portions* (A) *of soup* (B)
Aspect (A,B): *classify things* (B) *by weight* (A)
Stage (A,B): *at the beginning* (A) *of the meeting* (B)
Sphere (A,B): *work* (B) *in agriculture* (A)
Condition (A,B): *welding* (B) *at high temperature* (A)
Goal (A,B): *NN arrived* (B) *in N.Y. to discuss* (A) *some linguistic problems*
Similar_to (A,B): *The girl* (A) *resembles her mother* (B); this SemRel appears at the second stage of SemAn by the rule of correction of 2 primary formulas: **1-actant** (A,*resemble*) & **2-actant** (B,*resemble*). Words of the category Rel (in the dictionary, see below) such as *resemble*, *similar*, *equal* etc. behave the same way.

Another part of semantic grammar consists of semantic features (SemF), examples:

ARTEFACT (man-made object): *table*, *weapon*, *sputnik*
SUBSTANCE: *sand*, *clay*, *uranium*
PERCEPTION: *hear*, *see*, *observe*
HARM (smth. connected with life hazards): *war*, *threat*, *damage*; *kill*
INTERVAL: (in time) *week*, *intermission*; (in space) *distance*
INFORMATION: *message*, *advertisement*, *novel*
SPHERE: *industry*, *electronics*, *agriculture*
QUANTUM (a unit of smth): *bit*, *lexeme*
CAUSATION (often used in combination with other characteristics): *launch*, *transform*

The Semantic Grammar of Rels is defined by two axes:

- syntagmatic, that is, rules of combining words with given SemF into formulas. Examples:

Reason (A,B), where SemF(A) = SIT or a whole semantic formula; the same for SemF(B). In the phrase *It happened a cause de Peter* two incomplete Rels will be built: Reason (Sit?, happen) & Actant (Peter, Sit?).

Name (A,B), where SemF(A) = Name (A, ?), SemF(B) = Person or Thing or Device;

- paradigmatic, that is, rules of substitution for terms and relations. Examples:

Identificator (A,B) can be specified as Name (A,B) or Symbol (A,B)

Time(A,B) can be specified by two formulas: Starting point(?,B) and Final point(?,B)

There exist more complex superrelations between semantic units (relations, formulas and other units). As for SemFeatures, they are also organized in a partial hierarchy; an example:

SITUATION > PREDICATE > PROCESS > ACTION

Thus, if a word has SemF = Action, it may be filled in some formula where SemF Sit is predicted.

The list of SemRels is rather stable, taking into account that many general Rels may be specified by index; for instance, SemRel Loc(A,B) may be specified as Loc-inside, Loc-outside, Loc-under, etc., others are to be approved and specified by the rules of semantic analysis. At the moment, Semantic Grammar is used when describing meanings of words in Russian Semantic Dictionary (especially the valency structure) and in building local interpretations for optional and separated groups or parceled propositions in the course of SemAn. Some subtle properties of the Grammar may be used for compression and other transformations of the SemR. For lack of space, I have limited myself to this short exposition of our experimental semantic metalanguage. The full description would require the formal definition of all ILM entities. My task is to outline the whole model of humanlike understanding of a text and to give the general idea of how it may be implemented in the system with developed semantic component.

3 SemDict RUSLAN as a Tool for Textual Semantic Analysis with Compression

The Russian general semantic dictionary (RUSLAN) is the main instrument for intra-, inter-sentence and whole text interpretation of the initial text (Leontyeva, 1995). RUSLAN was implemented in a DB form (Sokirko, 2001). It contains the information on semantic categories and taxonomic characteristics of each lexeme (Semënova, 2000), its semantic valencies, typical contexts, thesaurus relatives, derivational capabilities, collocations, English translations, domain of usage and some other data (about 50 fields organized in 10 zones). Lexicographic descriptions are entered into the DB RUSLAN using our slightly formalized metalanguage of SemRels.

RUSLAN was designed, first of all, for automatic semantic analysis accompanied by gathering informative semantic nodes (SITs and the like units and their constituents) for a given collection of Russian documents. Recognizing and extracting data on people and institutions (on the basis of SemSpace) and building the corresponding semantic nodes were the initial steps along this path. Units of SemSpace being matched successfully against the given DB of VIP-persons or of Organizations may be considered as Concepts. Many of them were restored from textual abbreviated forms to full names of persons, their functions and institutions taken from special DBs. This system was a kind of IE-system, but based on the syntactic and the local semantic analysis. The partial results looked as a naïve Knowbase.

Some data stored in RUSLAN (e.g., data about syntactic realization of the word's valencies) were used only at the pre-semantic stages of NLP. On the other hand, some fields (both linguistic and informational) of a dictionary entry would provide a sort of tuning to the user's request: you may indicate the degree of compression (of the complete SIT-unit predicted in the Dictionary) that you wish. This procedure proposed for RUSLAN-based semantic analysis permits a user to gather an individual BTF according to his/her interests. I will now dwell on some dictionary fields (traditional and new), open to discussion.

3.1 Some Dictionary Fields

• CAT stands for "category". We single out five categories of lexical units (LU) at the moment. Only three of them are described in RUSLAN and have been involved in primary SemAn, their categories depend on the role/place of a unit in semantic formula R(A,B) and further in the semantic graph. These three main classes are: units that may turn 1. into nodes (A or B), 2. into relations (R), or 3. into a node plus relation R(A,?). Lexemes with "empty" meaning form the forth class: (half)empty words occupying places A or B are accompanied by the relation "REF(?,A)", where the question mark requires to restore the omitted meaningful referent. The fifth class includes words whose meaning is an operation on the already existing part of the SemGraph (*accordingly, so, though, nevertheless,* etc.).

LUs of the first category are the most meaningful units (CAT = NODE). Many NODEs may be specified by subcategories: NODE-Object, NODE-Sit, NODE-Attr. Only these lexemes are supplied with semantic features (SemF) and with information WEIGHT (from 1 to 5): WEIGHT(Sit) = 4 or 5, WEIGHT(Attr) = 3, etc.

• SemF are semantic features (ex. "thing", "animate", "process", "intellectual", "bad", etc.; lexical functions – LF – are used in this field as well). Ex.: ENTRY = *bank*; SemF = ORG, FINance. ENTRY = *okazyvat'*; its own SemF = LF Oper.

• VAL is the set of semantic valencies of the word C; candidates for filling in the slots of valencies are introduced by symbols Ai (i = 1, 2, ..., 7). The notation in VAL field is as follows: R(Ai,C) or R(C,Ai). The notation R,Ai,C; R,C,Ai is also possible. Ex.: ENTRY = *message*; VAL = Agent,A1,C; Addressee,A2,C; Topic,A3,C; Content,A4,C.

Each term of a valency is described separately in an abbreviated form: "SemF1" means "SemF of A1", "SemF2" means "SemF of A2", etc.; the same is true for grammatical characteristics: Gram1, Gram2, etc.

• ADD refers to additional semantic relations among the actants. Ex.: ENTRY = *compensation*; VAL = Agent,A1,C; Addressee,A2,C; Cause,A3,C; Value,A4,C; ADD = 1. Patient,A2,A3; 2. Belongs-to,A4,A2. Ex.: *compensation to NN* (A2) *for the damage* (A3); the value (A4) of compensation belongs to NN (A2).

These important data are to be included in SemSpace; they occur usually in text continuation and may be used for disambiguation and for proving the coherence of the text if it will develop that idea (ex., about A2 and event A3 that happens with NN), for inferences.

• CORR is the set of rules of correctness of the valency structure written in the following form: initial SemRel, ⇒ (symbol of transition), resulting SemRel / if … Ex.: ENTRY = *ruin*; VAL = Agent, A1, C; Ob, A2,C. CORR = Agent, A1,C, =>, Cause, A1,C / SemF1 = non-animate; e.g. *the flood* (A1) *has ruined the village* (A2). The *flood* will be Cause instead of Agent of a situation 'ruin' by this rule: Cause,*flood,ruin*; Ob,*village,ruin*.

• RESTOR is the set of rules for reconstructing a member of the valency structure, in particular, the agent of an action being expressed by infinitive.

In the SIT zone, the most important fields are:

• SIT: the fullest linguistic description of a lexeme of the SIT category. Usually, it comprises all Ai from the VAL-field, ex. SIT = {A1–A5} if there were 5 actants. But one may add other members (from ADD or SIT, or PRECED fields).

• ESit, or ES: the description of elementary situations in the form of a set of semantic relations R,A,B; e.g. ENTRY = *export*; VAL = Agent,A1,C; Ob,A2,C; End-Point,A3,C; ESit = 1. Belongs-to,A2,A1 / SemF1 = "organization"; 2. Loc,A1,A2 / SemF1 = "space", "state"; 3. Belongs,A2,A3; 4. Loc,A3,A2. Further on, elementary situations are referred to as ES1, ES2, etc.

• PRECED: an elementary situation preceding the main Sit (C); PRECED = ES1 or ES2.

• POST: an elementary situation following main Sit, e.g. ENTRY = *export*; POST = ES3 or ES4.

In the PRAGM zone, the important fields are:

• DOMAIN: polit., or econ., or war, or usual.

• WEIGHT: initial semantic importance of the lexeme (specified by linguist). Ex.: *war* 5, *start* 4; *nice* 3, etc.;

• EVENT: event (main situation denoted by the word C and/or one of its actants with the greatest informational weight that may be a nucleus of some event in the indicated domain), e.g. EVENT = A3 (actant A3 of C is announced to be the center of TF to be built); Ex.: ENTRY = *to help*; VAL = Agent,A1,C; Addressee,A2,C; Content,A3,C; Reason,A4,C; EVENT = A3.

• INFER: a standard inference in the form of a production rule: If SIT1, then SIT2; If SIT2, then SIT5.

• PRESUP: presupposition (the name of a situation already introduced in a field or formulated by the linguist, that is indispensable for C to be true).

• EVAL: evaluation ("+", "-", "0", or "?", where "?" signifies that the evaluation depends on certain conditions); e.g. ENTRY = *export*; EVAL = ? / (A2) (evaluation of C is inherited from A2): EVAL (*to export drugs*) = "-"(bad). This SemF is used mostly when analyzing criminal texts.

• LOG: a more complex situation that characterizes the logic of the event and is to be formulated in terms of SITs and production rules. (For ex., to connect notions *crime – trial – sentence – punishment*).

I will dwell on two fields of RUSLAN which differ from the ECD approach. These are VAL and GRAM fields, which are filled in the "top-bottom" way. You write in VAL all SemRels expected to be

expressed in the text, not splitting them into strong and optional relations&fillers. The GRAM field consists of two parts (written as SYN: MORPH). The SYN part predicts by what syntactic class or syntactic role this actant in adopted system of SynR can be expressed. The MORPH part lists all morphological means used to express this role. Each VAL may have more than one superficial expression. This gives the rules of SemAn more flexibility.

Another big part of the RUSLAN dictionary is the dictionary of **words-relations** (punctuation marks, conjunctions, prepositions, etc.). The WEIGHT of Rels depends on the WEIGHT of their terms. Really, this second category dictionary is an extension of the SemGrammar. The words of the third category are **words-parameters**, which occupy the first place in R(A,B) formula, being semantically dependent on B and thus having the lesser weight; the name of R is a generic notion for them: Param(***height***, *man*); Part(***arm***, *human*). Many words-parameters repeat the R-name. Ex.: Time(***time***, SIT); Part(***part***, *body*). The WEIGHT of them is usually 3. For the lack of space I can't develop these themes.

3.2 Dictionary Implementations and the State of Affairs

The first version (ROSS) of our semantic dictionary has been an important component of semantic analysis in the FRAP MT-system (French-to-Russian Automatic Translation) in the National Translation Centre. Then the Russian-to-English version was developed and has been included in the text understanding system POLITEXT elaborated in the Academy Institute of USA and Canada (ISCRAN). Some modifications were made to include a part of that dictionary (Russian and English parts separately) into a Russian-to-English MT-system (Sokirko, 2001). The Russian part of the system was used as the main tool for semantic interpretation of texts in the Internet network version (www.aot.ru). The Bulgarian Academy of Sciences began working on a Bulgarian simplified version of the dictionary. RUSLAN-1 is being implemented and upgraded at Moscow State University; at the moment, it contains about 12, 000 semantic entries. It may be called by any application dealing with syntactic, semantic and pragmatic analysis of Russian texts; it has thus the status of reusable dictionary resource. (However, for the last two years, the work on development of the SemDict was practically stopped.)

4 Types/Stages of Understanding in ILM

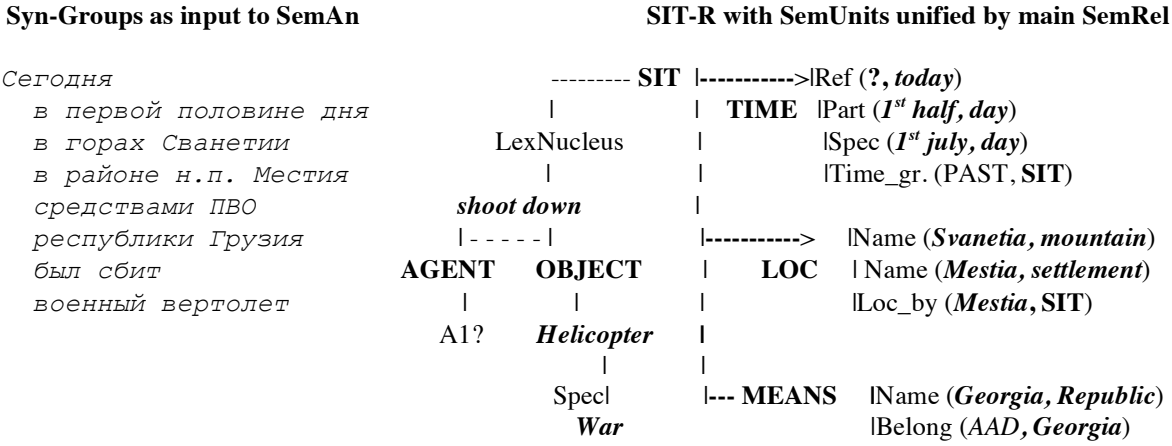
Stages of understanding, the corresponding representations and the instruments of analysis (roughly, dictionaries) adopted in ILM and partially realized in some systems (see Leontyeva, 1987 & later issues) are shown in the following table:

Types of understanding	Stages of analysis	Representations (R)	Dictionaries
Local understanding (within one sentence)	GraphAn, MorphAn, SynAn, primary SemAn	GraphR, LexR, MorphR, SynR, local SemRs of separate sentences	Grammatical Dicts, Special linguistic Dicts, ROSS/ RUSLAN
Global understanding (inter-sentence analysis)	Proper SemAn and SitAn (of utterance, text)	Semantic Space, SitR, Global Sem(SIT)R	RUSLAN, ling. frames, SIT-Schemata
Relative (user-oriented or domain-oriented) analysis/ understanding	Interpretation of SemR in terms of domain and user request / interest	Special R-s, Textual Facts (TF)-R, Event-R, TF-Base or BTF	Thesaurus, databases, Special Frames (of Orgs, Persons, Geogr., etc.)
Understanding in terms of another language	Translation, Text Generation Systems	English variants of Sit-Rs, Event-R, TF-R, BTF	Russ-Eng SemDicts

As for the fourth stage of understanding, it may be the translation of BTF in English or in another language; we hope that it is possible to use for this purpose an already existing TG-system.

Let us consider a simple illustration of three levels of Representation (SynR, SitR, TF-R) of a short newspaper text:

1 июля. Сегодня в первой половине дня в горах Сванетии в районе населенного пункта Местия средствами ПВО республики Грузия был сбит военный вертолет.
'1st july. | Today | in the first half of the day | in Svanetia mountains | near the settlement of Mestia | by means of AAD | of the Republic of Georgia | was shot down | a military helicopter.'



The TF-representation of the sentence under consideration is as follows:

EVENT = ***Сбили военный вертолет*** (1,2,3) – The node built according to the rule of meaningful SIT-name
1. **AGENT** = ***ПВО Грузии*** (textual SemNode deduced as the Agent due to absence of other candidates)
2. **TIME** = ***1989 г.- 1 июля*** (the node restored from the data of the text, and 1989 – from the publication date of the newspaper)
3. **LOC** = ***Грузия - Сванетия - Местия*** (the full name restored from the Geographical DB).

We have obtained a TF-representation of this message manually, by using Sem-Rules, the SIT-Structure, the external DB (Geography), the rule of inference (node A1 may be also generalized as Georgia), and Text attributes (see the full TIME-node).

5 Semantic Space

The local SemAn begins with sentence-for-sentence semantic interpretation (in terms of the adopted metalanguage) of the initial text. The basic mechanism of local SemAn has been implemented as a procedure of translating syntactic formulas r(a,b) of the SynR – adopted as input structure – into semantic formulas R(A,B):

1. attr (two,books) ⇒ Quantity (TWO,BOOK)
2. indirect obj (by Stern,books) ⇒ Author(STERN,BOOK) or more exactly:
   Author(STERN,1.), that is, Stern is the author of BOTH BOOKs (1. refers to the 1st formula).

Thus, at the beginning of SemAn we obtain rather a SynSemR, which corresponds to the first level of text understanding.

The local SemAn deals with the valencies of each lexeme. Those valencies that are not saturated within the limits of one sentence or do not satisfy some rules of semantic grammar and dictionaries are transmitted into SemR as invalid formulas: R(A,?), or R(?,B). These gaps in SemR become the important signs of text incoherence. Sometimes they may be filled in by reference to those portions of specific knowledge that are fixed in the SemDict or they have to be restored at the next stage – the global SemAn.

SemSpace may have conflicting formulas due to splitting of meanings of a single syntactic node; in this case, they are combined by the superrelation "INCOMPATIBLE(C,D)". Ex. *Сопротивление* 'resistance' *проводника* 'of conductor' will give two formulas: 1. Agent(*conductor* as "human", *resistance* as "action") and 2. Parameter(*resistance* as "parameter", *conductor* as "device"); they will be two members of the superrelation "INCOMPATIBLE(1.,2.)" in the same SemSpace – not to be split into different variants of SemR. Similar conflicts may be resolved by further steps of SemAn.

Now I will mention some other particularities of SemSpace that differ from standard linguistic SemRs in a principled way.

From the structural point of view, SemSpace is a sequence of formulas having simple syntax of binary relations R(A,B), where the second member is normally the main one and has therefore the greater

weight. (I will omit the procedure of semantic correction of some formulas if this main member has "empty" meaning in dictionary, as well as some other phenomena concerning our metalanguage formal demands.) As for members of SemRels, they are mostly lexemes, but can also be formulas, grammatical elements or even empty slots. Some SemRels may link not only lexical units, but nonterminal symbols as well. Ex. Repres (SIT5,Prop2), which means 'SIT5 is computed as the main representative of the Proposition 2'; Compos(25 Prop,Txt) 'Text consists of 25 propositions', etc. Those compositional relations are used in SemAn for references, when moving through the text, comparing formulas and making inferences, ex. Equal(Sit2 of Prop1,Sit7 of Prop8).

The divisions between sentences may be ignored. SemSpace without boundaries becomes a **continuous** textual structure being a basis for proper Semantic Analysis. As such, it has useful properties. SemSpace is a **flexible** structure: you may throw out or add some formulas, you may mix and change the order of sentences, utterances etc., you may analyze SemSpace from the last to the first sentence if it necessary for a particular task, etc. For example, the signature under a long order would permit you to reconstruct later the referent of the omitted name of the Author-valency of the first textual lexeme *PRIKAZYVAJU* 'I order': *1., 2.,* etc. As for the Content-valency of the same lexeme, it must be specified as the set of paragraphs introduced by numbers: *1., 2., 3.* etc. To collect such information you need to take into account even the visual form (GraphR) of the document under analysis.

As an experimental structure, SemSpace has a **dynamic character**: it allows the system to build meaningful complex units, to restore locally omitted units and to introduce lost "weak" groups in the proper slots throughout the text. Alike or comparable formulas may be found in the textual continuous structure due to homogeneous syntax and "natural" meanings of SemRel names. Moreover, the new compressed complex units of the next level (SitR, EventR) enter the same SemSpace along with its primary lexical units, etc.

SemSpace is an **adaptive structure**: its units may be compared with any external parts of the knowledge given by the user. If the user's request has been analyzed using the same metalanguage, it may be possible to replace step by step some text units by analogous terms of user's interest. The result will be an individual information representation, in the form of a frame corresponding to the ordered part of knowledge. As for the SemDict, the transition from one given Domain to another one does not lead to changing the dictionary. It will concern only few fields that are to be "tuned" to the Domain.

SemSpace as a kind of linguistic SemR adds one more dimension to the whole text understanding (WTU) – a **vague** one. SemSpace can include incomplete formulas, conflicting and redundant data, heterogeneous terms of SemRels, unknown words and other units without information.

6 Situational Analysis & Representation

The main problem of ILM is to develop mechanisms of global SemAn. The deficiencies of SemSpace have a **creative** character: they serve as indicators for restoring omitted parts of meaning, for compressing the redundant pieces of SemR, etc. All these processes aim at gathering greater units, thus compressing several empty formulas of the primary SemSpace into one complete unit. Global SemR is to be constructed according to the rules of semantic grammar of the following form: R1(A,?) & R2(?,B) ⇒ R3(A,B) // under certain conditions. Another rule deletes formulas that duplicate each other in full or partially. One more resource for constructing bigger units from elementary Sits is the standard structure of textual linguistic SIT (the part of it may be seen in the example above). Complex SITs can be constructed outside the limits of a sentence as well as inside a sentence.

I would like to emphasize the idea that the coherent text analysis is an analysis through generation of complex units, this generation beginning inside analysis. The processes of analysis and generation within text overlap much more tightly than one might think. Several elementary Sits may be combined into a full SIT according to the canonical linguistic SIT-structure being applied to every simple sentence of a text. The most effective is gathering of isolated, incomplete and parceled sentences. Ex. from A. Blok: «*Ночь. Улица. Фонарь. Аптека*» will be analyzed at the SIT level as follows: *Ночь* 'night' is the TIME of some SIT that is absent. *Улица* 'street' is LOC of some SIT that is absent. *Фонарь* 'lantern' is an object of type 'device' that may give light (LF from Dict). *Аптека* 'pharmacy' is some place (LOC) of some SIT that is absent, it is bigger than 'lantern', but 'street' is bigger than 'pharmacy' (from pragmatic knowledge introduced manually). So we may deduce that all those entities must be parts

of the same SIT that is absent (no any other alternative). But we can bind them to this unknown SIT (?SIT) using their own semantic characteristics that predict the SemRels. The whole description will be as such: Time (*Ночь*,?SIT); Loc (*Улица*,?SIT);. Obj (*Фонарь*,?SIT='action'); Loc (*Аптека*,?SIT); Loc (*Улица*,*Аптека*); Ref (?,SIT). The last formula permits to look for the referent of the unknown SIT further in this text (it is a created valency of this complex utterance under analysis). Redundant units in the global SemSpace may be eliminated, using rich dictionary resources of modern lexical semantics schools (ECD, see Mel'čuk, 1984-1999 and Apresjan *et al*., 2000, not to mention the number of special Thesauri). But this way needs further study.

If we collect all SITs being built in the course of analysis of a given text, this will be the SIT-Representation that adds one more structure to the set of compressed products. Each SIT must be accompanied by its own frame, which can be empty to a certain extent (any absolutely empty SIT will be deleted from SemR). The main SIT (if it is meaningful) with its subordinate Sits may turn to be a textual fact (TF) representing the principal content of the text under analysis (if no TF has been built). Other variants of summarization may be possible by agreement (only SITs, SIT + EVENT, SIT + some important lexemes etc.). As for EVENT-unit it may be built as linguistic unit on the base of dictionary prediction (see field EVENT) or by the procedure of generalization of several SITs.

7 Relative Analysis and TF-Representation

Primary SemSpace is a linguistic structure indifferent to a notional sphere. All transformations applicable to SemSpace aim at constructing meaningful units – SITs and TFs. If built from linguistic material, they remain purely linguistic entities. Meanwhile the ultimate task of **specialized/professional texts analysis** is receiving a KnowBase consisting of **conceptual entities**. How to examine lexical units for conceptual status? The simplest way to prove that a linguistic unit (LU) may be transformed into conceptual entity, or concept (CNC), is to match the LU, be it simple or complex, against some existing Domain sources introduced as **counter-text**. The results of comparison being successful (according to certain conditions), the LU may be declared a conceptual entity (LU ⇒ CNC). In this way we could obtain (theoretically) one more kind of SemR consisting mostly of concepts, notions, etc. It may be called the naïve Knowledge Base (KnowBase), in our case the Base of Textual Facts (BTF). It may be a common BTF or individual KnowBases for different specialists. Each of them forms a new dimension of text content.

The specialized text SemAn is a type of Pragmatic analysis (PragmAn). Such SemAn has been implemented (in ISCRAN), matching syntactic noun phrases against the Thesaurus of Russian political life. The results obtained were far from ideal: syntactic and terminological units boundaries were often in contradiction.

The following experiment consisted in matching semantic groups from SynSemR against lists of domain objects such as Names of politicians, Geographical notions, Institutions, and Political functions. These domain objects have to receive their own syntactic or semantic information before the matching procedure, so that the process was "top-down" one. Those groups (noun phrases) that succeeded in matching received the status of denotation and a greater information weight and may enter the final DBs.

The information about denotational status of textual units is very important when procedures of further global SemAn proceed within the limits of Semantic Space. The completeness of DBs taken as counter-texts will lead to the success of PragmAn and further BTF-construction. If there are no matches (the PragmAn, or domain Analysis, fails, has zero results), the initiative is passed to proper LinguAn.

Formally, a "Textual Fact" is a multi-term predicate where the terms are maximally meaningful notions (i.e., they have maximum informational weight for the given text and/or for the given domain and pragmatic orientation). It's worth mentioning that a TF as a multi-term predicate is not the same as that of syntactic structure: the terms of the former has to be gathered across the whole text – usually not according to syntactically "strong" actants of an appropriate word-predicate. Moreover, many literary texts (being analyzed without "counter-text") will have as their TF the name of a simple object, not of a predicate. Ex.: in I. Bunin's novel "The Gentleman from San-Francisco," the title turns to be a TF, in our metalanguage – "Start_point(*S.-F.*, *Gentleman*)", and as its terms would be listed the words = "actions" of that Gentleman. I see here the similarity with some DBs where the name of some place, or animal, bird etc. is announced as the main predicate of a table, and all properties (can fly, etc.) of such an entity form fillers in DB fields.

The TF-structure in linguistics-based systems is to be built by summarizing all SitRs constructed in conformity with rules of global semantic analysis and with properties of the coherent text structure. As for the practical tasks, TF and its terms are to be computed taking into account pragmatic orientation – what Events you are looking for. Below is one of 6 TFs built manually on the Corpus of about 100 short texts from the newspaper "Obshchaja gazeta" in 1991. It was the first attempt to formulate some rules of TF-analysis (introduced in a man-machine scenario), the terms were taken from the verbal material.

TF = COUP D'ETAT (1,2,3,4,5)
Variant: *seizure of power*
1. Agent = *the USSR State Emergency Committee (SEC)*
   Variant = t*he Soviet leadership*
   Identification = *G.Yanayev, V.Pavlov, O.Baklanov, B.Pugo, V.Starodubtsev, A.Tizyakov, V.Kryuchkov, D.Yazov*
2. Counter-agent = [*former power, President Gorbachev*]
3. Cause = *destabilization of political and economic situation in the USSR*
4. Goal = *to overcome economic and political crisis in the USSR*
5. Time = *from August 19, 1991*

It is obvious from this description that it was a "relative fact," true only for the texts in question, for the given data, etc. To enter the historical DB of real FACTs, our **Textual Fact** as a linguistic unit must be compared with many TFs from other reliable textual sources. I regard this task as a serious challenge to linguists.

The final representation of a text is to be supplied by the text-description with its own title (or number), source (author) of information, composition, etc. Actually, it will be the frame of the document. As for the content, each document may be represented by Situations (SIT), Events (EVENT) and/or Textual Facts (TF), each accompanied by its own frame, temporal data, names of places of events etc. So, the BTF for a given corpus of texts has to be a condensed structure, a result of comparison and generation of new units.

8 Conclusion

Based on the belief that a complete computable understanding of a whole text is impossible to achieve, I conceive the **text compressing ability** as an obligatory property of NLP systems. We have tested our SemAn and SemDict mostly on real documents (Decrees of Russian President, criminal reports, newspaper messages and titles, etc.). SemSpace was the first whole text structure being implemented, and this experimental structure merits to be regarded attentively. An uncertainty of different kinds – redundancy, incompleteness, ambiguity, and contradiction to the SemGrammar rules – appears in this initial SemR in an explicit form. **The task of eliminating text "defects" and that of text compression meet each other** in the design of ILM, thus creating the moving force for the next stage – the proper SemAn. The phenomenon of the **multiplicity of possible views** on the same thing, be it text or political life, made me abandon the intention of building one unique well-formed structure of a text. Maybe such semantic structure as our "handicapped" SemSpace will ensure a new solution for many NLP problems.

Many formal definitions of separate word meanings seem not corresponding to the real usage in the huge realm of texts, general and specific, scientific and metaphoric, rather vague than exact, etc. (cf. Altman & Polguere, 2003 and Wanner, 2003, who discussed the difficulties of programming the ECD definitions). Being too formal, they involve great difficulties in text automatic treatment by NLP systems. We have proposed the **more "soft" dictionary descriptions** of words, expressions, symbols etc., using SemRel language. As a subset of NL it is a semi-formal one, but it has many advantages of NL "naturality." Our SemRel language was tested on difficult examples, rich in linguistic violations and distortions, which no purely linguistic theory would be able to manage. Meanwhile the "invalid" SemSpace built in the course of SemAn leaves the possibility to use the whole semantic context. The choice of that or another variant of lexeme meaning may be based on common sense rules such as: *One person can't be in two places at the same time*, or *Agent and its action are not to be separated* (*at the same time and space*), and the like. This way would relieve an applied system of many bottleneck problems in the proper linguistic works.

The more cardinal idea consists in automatic analysis of classical dictionary descriptions in terms of our SemRels. By building SemSpace for Dictionary entries we may obtain natural and powerful resources of lexico-semantic knowledge. Using this information as additional counter-text would advance the problem of textual SemAn. But this task as a whole needs some new technique (of analysis and matching).

The **counter-text** idea seems to make the task of **including Special Knowledge** in NLP not so unreal. We admit that Knowledge may be involved in the NLP system by small portions **in a usual textual form** that can be formulated by the User (his request, list of desired terms, etc.) – to examine the idea. To involve the formal descriptions of different Domains in NLP-system may constitute the next, not immediate task (see Nirenburg, 2004). But I suppose it will be more realistic to face the domain problem as a kind of machine translation problem using the same metalanguage for initial text analysis as well as for domain text analysis followed by comparison (and adaptation) of those "foreign-like" languages.

Acknowledgements

Two linguistic phenomena that have led me out of limits of SynR to the whole text problems are syntactic ellipse and semantics of prepositions in the field of Machine Translation. Both topics were launched by Igor Mel'čuk, the supervisor of my Ph.D. dissertation, 50 years ago. I am very grateful as well to Lidija Iordanskaja for her important remarks on this text, and I appreciate her interest in this work.

References

Altman, Joel & Alain Polguère. 2003. *La BDef: base de definitions derivée du Dictionnaire explicative et combinatoire* // MTT 2003, Paris, 16-18 juin 2003, 43-54

Apresjan, Ju, *et al*. 2000. *Новый объяснительный словарь синонимов русского языка.* Jazyki slavjanskix kul´tur, Moskva

Grishman, Ralph. 1999. *Information extraction: Techniques and Challenges*. Internet

Leontyeva, Nina.1987. Stages of Information Analysis of Natural Language Texts. *Int. Forum Inf. and Docum*. 12: 4, 8-14

Leontyeva, Nina. 1995. ROSS: Semantic Dictionary for Text Understanding and Summarization. *META*, 40: 1, 43-54

Mani, Inderjeet. 2001. Automatic Summarization. In R. Mitkov, ed., *Natural Language Processing*, vol. 3. Benjamins, Amsterdam/Philadelphia, 286 p.

McKeown K. *Text Generation*. Cambridge, 1988.

Mel'čuk, Igor *et al*. 1984 - 1999. *Dictionnaire explicative et combinatoire du francais contemporain. Recherches lexico-semantiques. Vol. I - IV*. Les Presses de l'Université de Montréal, Montréal

MUC - Message Understanding Conferences, 1-7.

Nirenburg, Sergey & Victor Raskin. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA

Semënova, Sofia. 2000. Word Taxonomy Fields of one Russian General-Purpose Semantic Dictionary: Descriptor Selection, Analysis of Representation Possibilities. In *DIALOG-2000*, vol. 2: 308-316

Sokirko, Alexej. 2001. *Semantic Dictionaries in Automatic Processing of Texts*. Ph.D. dissertation. Moskva

Wanner, Leo. 2003. Definitions of Lexical Meaning: Some Reflections on Purpose and Structure. In *MTT 2003*, Paris, 55-65