B.V. BABYCH

SYSTEMS OF SYNTAXEME GROUPS AND THEIR PROCEDURAL SEMANTICS

Abstract. In this article we propose a framework of representing deep syntactic structures of a sentence in the form of a syntaxeme graph. In functional syntax a syntaxeme is an elementary constructive and meaningful unit of the deep structure, which receives additional structurally motivated meaning within a deep syntactic relation. This meaning interacts with lexical semantics of word forms in the sentence. We show how these structural meanings can receive procedural interpretation, and participate in updating knowledge about a situation described by a sentence. We conclude that functional syntactic analysis is a productive framework for NLP tasks of Slavic languages, and proper formalisation techniques can be effectively developed within this framework. The research was conducted within the project "*Theoretic linguistic foundations of automatic text processing*" at the Ukrainian Language Information Fund of the National Ukrainian Academy of Sciences.

Introduction. It is generally accepted that formal methods of analysis, used in natural language processing technology, need to be extended in order to preserve the explanatory power and flexibility of traditional linguistic approaches. This is also true for different frameworks of describing syntactic structures: formal generative models are less expressive, than functional descriptions, but they are closer to the needs of NLP applications, such as systems of machine translation, information retrieval, automatic grammar control, etc. A common way of improving the performance of such systems is to augment generative models of syntax with new features, which originate in less formal paradigms. The problem with this approach is that new features are borrowed and implemented very often in unsystematic way. In this article we explore an alternative solution to this problem: giving formal interpretation to concepts of functional syntax. The central concept in functional syntactic theory is *syntaxeme* – an elementary building unit of a deep syntactic structure, which is interpreted semantically on the basis of *deep syntactic relations* (DSR) between word forms. We suggest that such formalisation will allow coming closer to modelling cognitive processes of understanding text.

In applied NLP systems, a syntactic structure of sentences is usually represented by dependency trees or by systems of constituents. A.Gladky[1] proposed the formalism that combines these two types of representations – *the systems of syntactic groups*. This formalism is close to traditional non-formal methods of syntactic analysis in respect to flexibility and naturalness of representation, since it allows establishing dependency relations not only between word forms, but also between phrases that form the same constituent, the later are allowed to be discontinuous. Constituents can also have internal structure of dependencies between word forms and smaller constituents. The system of syntactic groups can be easily mapped both to the system of constituents and to the dependency tree[2].

In this paper we propose extension to the format of the systems of syntactic groups, which allows representing functional level of the deep syntactic structure. We will refer to the extended format as *the*

---

[1] *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. – М., 1985. – 144 с. In addition, there are formats used in individual models: systems of ranked constituents of M.Halliday (*Halliday M.A.K.* An introduction to functional grammar. – London., 1994. – 434 p.), used in SHRDLU system of T.Winograd (*Виноград Т.* Программа, понимающая естественный язык. – М., 1976. – 294 с.), structures of conceptual dependencies, used in MARGIE та BORIS systems of R.Schank's group (*Шенк Р.* Обработка концептуальной информации. – М., 1980. – 361 с.; *Ленерт В., Дайер М.Г., Джонсон П.Н., Янг К.Дж., Харли С.* BORIS – экспериментальная система глубинного понимания повествовательных текстов // Новое в зарубежной лингвистике. – Вып. 24. – М., 1989. – С. 106-160).
[2] *Гладкий А.В.,* ibid., p. 9, 49.

systems of syntaxeme groups. In this format constituents and dependencies are interpreted as DSRs rather than surface syntactic links.

Formally, the system of syntaxeme groups is a graph; its nodes are formed either by word forms (possibly, with explicated lemas and sets of morphological features), or by elementary or complex syntactic relation structures. Word forms and the structures of syntactic relations are jointly referred to as syntaxemes. The edges of the graph are directional and represent the type and the direction of the deep syntactic relation between separate syntaxemes. A logical meraoperator OR can also form a node, joining two alternative syntaxemes describing the same sequence of words, or alternative mark-up of a DSRs between the syntaxemes. This feature allows representing syntactic ambiguity locally, within a single graph, rather then by sets of alternative graphs. Since the system of syntaxeme groups represent DSR structure of the sentence, the directions of edges on their graphs correspond to the directions of deep syntactic expectations.

A structural interpretation of the system of syntaxeme groups – is a graph of a deep syntactic structure without syntactic ambiguities. The structural interpretation represents one of possible way of understanding (in contrast to the complete system of syntaxeme groups, which represents several possible ways of understanding the sentence).

The proposed format of syntactic representations has procedural semantic interpretation, specifying the order and the type of processes that change the system of extralinguistic knowledge of a recipient or have other intended cognitive effects. This interpretation of the systems of syntaxeme groups can be used in NLP technology. It is based on the ideas of procedural semantics of P.Johnson-Laird[3]. The theory of procedural semantics interprets linguistic meaning operationally. In the process of text interpretation, the meanings create or modify a mental representation model of a situation, which is described by a discourse. The syntactic structure of a sentence is used for specifying a sequence and arguments of the procedural interpretation of lexical meanings within the extralinguistic knowledge representations.

Extensions to the deep syntactic representation format. The systems of syntaxeme groups are based on the format of the systems of syntactic groups of A.Gladky and on the concept of a *syntaxeme,* developed in functional linguistics. The syntaxeme is an elementary syntactic unit that receives additional meaning (related to its lexical meaning, but conditioned by the syntactic structure) within a DSR. The format of syntactic representation that is proposed in this article differs from the systems of syntactic groups with respect to the following:

1. The systems of syntaxeme groups represent DSRs rather than surface syntactic links, so the meaning of syntactic structures correlates with *lexical semantics* of the conjoined elements. On the contrast, the systems of syntactic groups of A.Gladky are marked by names of the syntactic links that correlate with *grammatical features* of the word forms, many of which are uninterpretable in the semantic structure. For example: the Ukrainian phrases *приймати ухвалу* (*to take a decision*), *прийняття ухвали* (taking the decision), *прийнята ухвала* (the decision taken) all form an *object* syntaxeme, but have different markup of syntactic links: *object, compositional, attributive.*

2. The graphs representing the systems of syntaxeme groups are allowed to be cyclic, since some DSRs are formed between elements and groups, which are already joined by some other DSR. Non-cyclic graphs, or representation of syntactic structures in the form of trees, do not allow representing all necessary relations. For example, a co-referential relation between an anaphora and an antecedent, as well as coordinating constructions, where a conjunction joins syntaxemes of different types, form cycles on deep syntactic graphs: *Хто і кому це сказав?* ('Who and to whom told this?'); *Вони нікому і нічого не обіцяли* (lit.: 'They promised nothing and to nobody'), *Не дерево, а з листками, не сорочка, а зшита, не людина, а говорить* (Not a tree, but with leaves/sheets, not a shirt, but sewed, not a man, but talks).

3. The systems of syntaxeme groups represent syntactic ambiguity, which is realised as structural ambiguity or ambiguity or of the deep relation mark-up. The ambiguity appears when the same sequence of word forms is related to several interpretations of the deep syntactic structure. In these cases the alternative syntactic organisation sub-graphs[4] or alternative names of DSRs[5] are connected with a disjunctive metaoperator "OR". The disjunctive substructures and the disjunctive mark-up on deep

---

[3] Cited by the translation of P.Johnson-Laird. Procedural semantics: *Джонсон-Лерд Ф.* Процедурная семантика и психология значения // Новое в зарубежной лингвистике. – Вып. 23. – М., 1988. – С. 234-257.

[4] "Dependency" and "constituent" syntactic ambiguity (*Гладкий А.В.*, ibid., p. 111).

[5] "Mark-up" ambiguity (*Гладкий А.В.*, ibid., p. 111). For our format the term "syntaxeme ambiguity" is more appropriate.

syntactic graphs allow representing ambiguity in the local areas of the graph, so there is no need to build separate representation for every possible interpretation of the sentence, for example:

*(1) "Йому не завадить сказати дещо". (*It is desirable for him to say something |OR| It is desirable that somebody tells him something*)*

*"Йому потрібно позичити гроші". (*It is necessary for him to borrow money |OR| It is necessary that somebody lends money to him*)*

*"Йому потрібно приготувати завдання" (*It is necessary for him to prepare an assignment |OR| It is necessary that somebody prepares an assignment for him*)*–

– are represented as:

*{йому ↔ {потрібно/не завадить → приготувати/позичити/сказати…}}*subject. OR recepient.

– In these cases dative form of *він (йому)* – 'to him' can be interpreted either as a sytaxemmes of subject or of a recipient; the verbs in (1) allow both syntaxemes.

*(2) "Він заробляє гроші на землі". (*lit.: He earns money on land: working on land |OR| re-selling land*)*

– is represented as: *{заробляє ↔ {на → землі}}*locative OR instrument.

*(3) "Будемо їсти хліб до обіду". (*We will eat bread for lunch |OR| We will eat bread before lunch*)*

– is represented as: *{їсти ↔ {до → обіду}}*duration OR time of action

*(4) "Запрошення письменника". (*An invitation of a writer – A writer invites somebody |OR| Somebody invites a writer |OR| An invitation letter which belongs to a writer*).*

– is represented as: *{запрошення ← письменника}*subject. OR object. OR possessive .

*(5) "Квіти на вікні розквітнуть" (*Flowers on the window will bloom |OR| Flowers will bloom on the window – different attachment of the prepositional phrase*).*

– is represented on the diagram 1:

Diagram 1
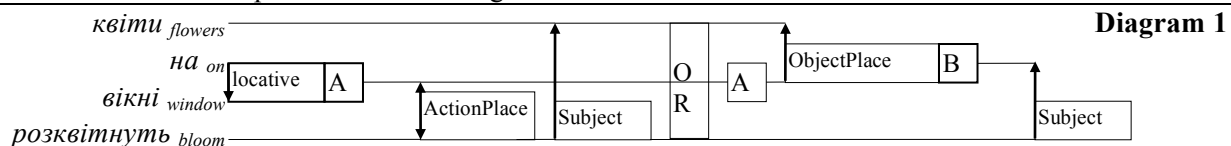
The bracket notation: *{на →вікні}*locative *= A; <(розквітнуть ↔ A*actionPlace*; розквітнуть → квіти*subject.*)* *OR (A →квіти*objectPlace*= B; розквітнуть →B*subject.*)>*

(6) *"Є для серця така покута*   There is a following penance for the heart
– забувати скоріше зло,   <to forget the evil quicker |OR| to forget the quicker evil>
аніж те, що мусило бути   than those things which could happen
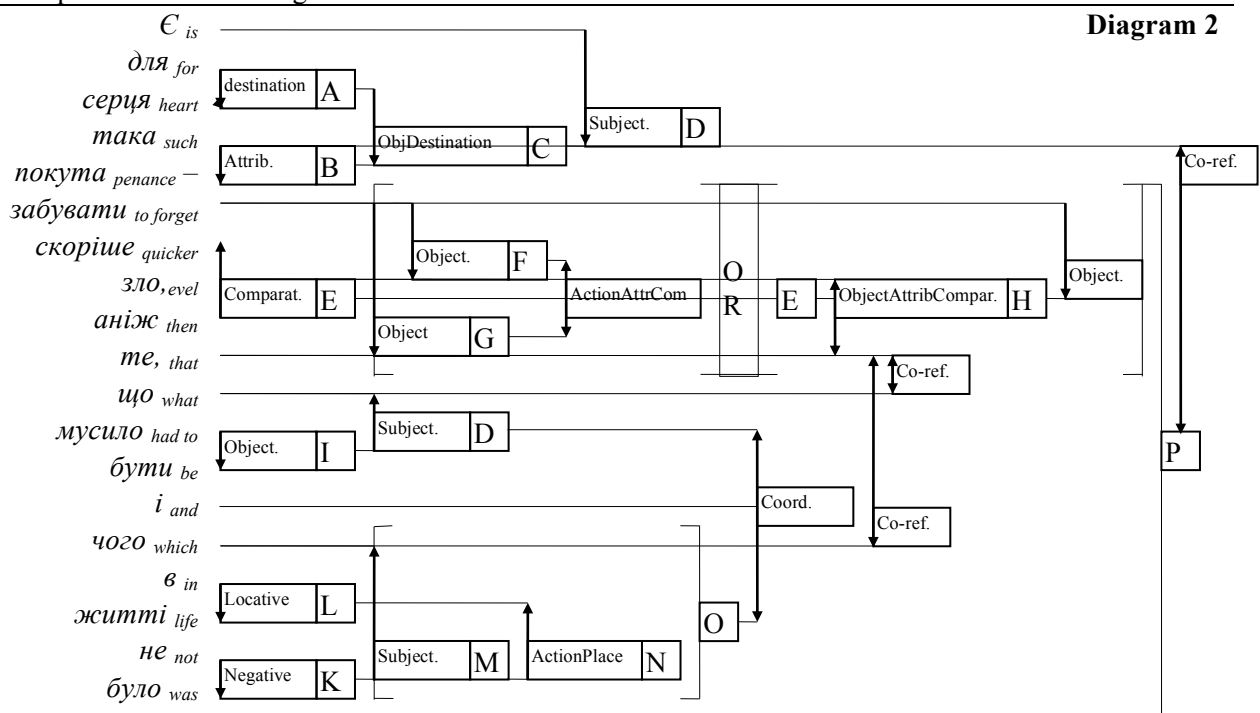*і чого в житті не було"*   but which did not happen in one's life

      (Л.Костенко)   (Lina Kostenko)

– is represented on the diagram 2:

Diagram 2

The bracket notation: {для →серця}destination = A; {така →покута}Attrib. = B; {A →B}ObjDestination = C; {є →C}Subject. = D; {скоріше ←аніж}Comparat.(Adj. OR Adv.) = E; <({забувати →зло}Object. = F; {забувати →те}Object. = G; {F ←E →G}ActionAttrCom) АБО ({зло ←E →те}ObjectAttribCompar.= H; {забувати →H}Object..)>; {мусило →бути}Object = I; {що ←I}Subject = J; {не →було}Negat. = K; {в →житті}Locat. = L; {чого ←K}Subject = M; {L ←K}ActionPlace = N; {J ←i →(N,M)}Coord.; {те ↔що}Co-ref.; {те ↔чого}Co-ref.; {така ↔(забути... було)}Co-ref.

– The two structural interpretations of the ambiguous area on the graph: (1)*"The evil is forgotten quicker than…"* (2) *"The evil is quicker than…"*.

The following reasons can be given why there is a need to represent DSRs on syntactic graphs:

1. DSRs are based on extralinguistic and lexical knowledge[6]. Lexical semantics can be interpreted and accessed in a direct manner from the deep syntactic structure, because the syntactic links between lexemes (the edges on syntactic graphs) are marked by semantically related features. A syntaxeme is an elementary unit of the deep syntactic structure, which holds an elementary meaning, and on the same time it is a constructive element of more complex syntactic constructions[7]. It functions on a level, where the interaction of the lexicon and grammar takes place. Syntactic graphs in this case represent the way of structural organisation of syntaxemes in sentence.

2. The direction of formal syntactic links on dependency trees is based on intuitive criteria; for example, there are different views if a pronoun is dependent on a noun or vice versa[8]. At the same time the criteria for determining the direction of DSRs arises from objective characteristics – the direction of prediction of some features in word forms by the features in other word forms. DSRs are directed from a dependent to a main syntaxeme[9], which corresponds to the direction of prediction of grammatical and semantic features. The DSRs are formed on the basis of such predictions; for example, an attributive DSR is formed as a result of finding a name of a concept in a contact position to the name of the attribute of this concept (since the concept can exist without a particular attribute, but the attribute cannot exist without a corresponding concept). On the isomorphic surface level an adjective predicts the existence, the syntactic position and morphological features of the noun phrase, which implements DSR. If the surface and deep levels are not isomorphic, an attribute predicts semantic features of a concept, but the direction of surface prediction can change[10]. The syntaxeme can form external DSRs with other syntaxemes. In the attributive DSR the way of functioning is inherited from the predicted element – the concept (e.g., a noun or a noun phrase on an isomorphic surface level). In the process of forming DSRs there is a division of functions between the participating elements: an attribute determines the type and the internal structure of the syntaxeme, finds and relates the required concept; the concept on the other hand determines the choice of the external DSR of the syntaxeme.

A grammar of deep syntactic predictions can be based on these properties of lexical and grammatical features. Such grammar would describe combinatorial properties of concrete semas (or features) of the lexical and grammatical meaning. Deep syntactic predictions project the general syntactic structure of the sentence from valencies of word forms with certain lexical and grammatical features.

On the other hand, some aspects of the syntactic structure are derived not from context-independent combinatorial properties of word forms, but from global properties of the sentence. For example in (7):

(7)"Наукову конференцію привітав представник президента"
'Scientific conference(accusative) greeted a representative(nominative OR accusative) [of a] President(genitive OR accusative).'

– *A presidential representative greeted the scientific conference*:

– combinatory properties of the last four word forms allow constructing an object syntaxeme out of the morphologically ambiguous word form *привітав президента* ('… greeted the president') and a subject syntaxeme *привітав представник* ('a representative greeted…'). But in this case the first two word forms could not be included into the deep syntactic structure of the sentence. The global requirement of the structure coherency does not allow such interpretation, so the second interpretation of

---

[6] *Вихованець І.Р.* Граматика української мови. Синтаксис. – К., 1993. – С. 92

[7] *Золотова Г.А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. – М., 1988. – С. 4.

[8] *Гладкий А.В.* Ibid., p. 8

[9] *Вихованець І.Р.* Ibid., p. 246

[10] Predictions of grammatical and lexical features allow disambiguating the system of the syntactic groups on morphosyntactic level. E.g., the degree of ambiguity of adjective word forms with respect to case and number features is smaller than of noun word forms, so only the features predicted by adjective need to be checked in noun position. The features predicted by adjective and found within a noun word form remain in the derivation, but alternative features of the noun and an adjective are rejected.

this locally ambiguous segment is selected: *представник президента* (a presidential representative) is a possessive syntaxeme[11].

A DSR can be bi-directional in case, if its syntaxemes predict certain features of each other. E.g., in the first structural interpretation of the sentence (5) an optional valency of the verb *розквітнуть* ('will bloom') predicts the prepositional phrase *на вікні* ('on the window'), but it is also predicted by the prepositional phrase itself. (The valency on the basis of this prediction is obligatory, since the form of this sentence excludes the isolated use of the locative syntaxeme and the only possibility is its functioning as a component within a phrase or a sentence[12]). The prepositional phrase predicts a verb (in the first interpretation of the sentence) or a noun (in the second interpretation). The type of the syntaxeme depends on the type of the element, which fills the valency; when both possibilities are found the sentence becomes globally ambiguous. So, directions of edges (which represent DSRs) in systems of syntaxeme groups are based on objective criteria – the direction of prediction of lexical and grammatical features.

Procedural semantics of the systems of syntaxeme groups. The semantics of syntaxeme structures correlates with lexical meanings of word forms in the nodes of representations. This property allows interpreting the syntactic structure procedurally, that is as a specification of a sequence of operations which build or modify a mental model of a situation represented in a text. The procedural semantics was proved to efficiently describe lexical meanings. We propose a way of using it for description of the syntaxeme meanings and of relating these meanings to the procedural lexical semantics of the word forms[13].

The procedural model of semantics assumes that there are two stages in the process of text understanding: "… on the first stage, the surface understanding builds a propositional representation, which is close to the surface form of a sentence. […] The second stage, which is optional, uses prepositional representations as a basis for constructing a mental model, the structure of which is close to the state of affairs described by a discourse. […]. The utterance is more a key to the construction of a model than a draft for building it"[14].

Using the terminology of the procedural semantics theory, the systems of syntaxeme groups describe prepositional representations of sentences. These representations can be used for constructing or correcting mental models of corresponding situations. This approach allows using the same syntactic representation in different mental models, and hence, will specify different sequences of procedures in the system of recipient's situational knowledge. Ambiguities in the systems of syntaxeme groups can be resolved (via receiving different structural interpretations) with respect to a particular mental model or a communicative situation in which it is interpreted. E.g., in (8):

(8) "Команда з Києва виграла в Іспанії" – formally: 'A team from Kyiv won against Spain' or 'A team won against Spain from Kyiv '

– the mental model of a soccer game allows only the first interpretation of the syntaxeme 'from Kyiv' – within the attributive DSR with the meaning 'Kyiv team'; the alternative interpretation 'won, playing form Kyiv', is not allowed, since no distant games are possible in soccer (in contrast with e.g., chess competition over the internet). On the other hand, the ambiguity if not resolved for the syntaxeme '*в Іспанії*', which can mean: 'against Spain' or 'in Spain'; in some situations the mental model will not contradict any of these interpretations. One of these facts in this case is based on the explicit mentioning in the utterance, and the other – on the previous context.

---

[11] If incorrect syntactic structure is build before the end of sentence processing, it is left behind and a new path is explored: *Представник президента привітав // наукову конференцію.* ('A representative of the president greeted' // scientific conference – the first three word forms can be interpreted as 'A representative greeted the president'. The example is similar to the English example *The horse raced passed the barn // fell.*)

[12] *Золотова Г.А.* Ibid., – p. 4.

[13] The problem of architecture of syntactic representations and the problem, how these representations are used, are closely related to each other. There are different solutions for the later problem. For example, dependency trees have been used for finding terms in text (*Севбо И.П.* Структура связного текста и автоматизация реферирования. – М., 1969. – 135 с.), the normalized dependency trees are used in machine translation systems (*Апресян Ю.Д.* Этап-2 с птичьего полета // Апресян Ю.Д. Избранные труды. – Т. 2. Интегральное описание языка и системная лексикография. – М., 1995. – С. 554-583; Лингвистическое обеспечение системы Этап-2. – М., 1989.– 294 с.). Theoretical models of the linguistic competence, e.g., "Sence ⇔ Text" and "the Universal Grammar" of N.Chomsky use the syntactic graphs for deriving 'deep' or 'logical form' representations (*Апресян Ю.Д.* Типы информации для поверхностно-семантического компонента модели "Смысл ⇔ Текст" // Апресян Ю.Д. Избранные труды. – Т. 2. Интегральное описание языка и системная лексикография. – М., 1995. – С. 8-101).

[14] *Джонсон-Лерд Ф.* Ibid., – p. 235.

In order to illustrate the concept of using the systems of syntaxeme groups to build up a mental model of a situation, let's exam the representation of the sentence (9):

(9) "Лідери України і Німеччини вирішили підтримати аерокосмічну промисловість цих країн у її спробах конкурувати на ринку, де поки що домінують Сполучені Штати".

– Ukrainian and German leaders resolved to strengthen the aerospace industries of their countries in their efforts to compete on the market now dominated by the United States.

After a parsing algorithm builds the system of syntaxeme groups (diagram 3) for the linear string (9), a sequence of procedures has to be defined for constructing a mental model of the described situation. This sequence of procedures is again linear, but the order of processing word forms in it can differ from the linear order of word forms in the sentence. The procedures specify operations, which are performed in the knowledge base as well as the arguments for these operations.

Diagram 3

From the point of view of the recipient we have to process syntactic representations in the bottom-up direction. For cyclic representations this means the requirement that the inner structure of a certain object has to be created before this object is used for forming a more complex object or an event. In this way a system of syntaxeme groups determines the sequence of operations in a situational knowledge base. This sequence can be described by a *protocol of operations*, which has linear structure, but specifies the addresses of objects and event to be used in other operations.

The systems of syntaxeme groups are disambiguated because certain structural interpretations are incompatible with the structure of the subject field, or with the content of prior operations or the general context of the situation.

The sequence of operations for the sentence (9) is determined by checking if objects, needed for a certain operations, are already formed on prior stages. E.g., the procedural processing of the possessive DSR "*лідери України і Німеччини*" – 'leaders of Ukraine and Germany' cannot start before the procedure, which forms the co-ordinate group "*Україна і Німеччина*" – 'Ukraine and Germany'. The procedural processing of the subjective DSR "*лідери ... вирішили*" – 'leaders… decided' can start only after two previous operations are completed. Non-linear syntactic representation specifies the linear sequence of procedural operations, symmetrically to the process of creating a non-linear representation from the linear sequence of words on the prior stage of parsing, but the initial and final linear sequences – of word forms and of procedures – can be very different.

The following protocol of operations in the knowledge base is specified for the first part of the sentence (9):

1. Forming a concept: Country.Ukraine /1
2. Forming a concept: Country.Germany /2
3. Forming a list [1., 2.] /3
4. Modifying a list: [leader.1., leader.2.] /4

    5. Forming an Event: Decision /5

    6. Filling the terminal of the event: Event.5.Actor=(4) /6

(On this stage the other terminal in the Event.5 – Object cannot be filled, since the proper concept is not yet formed, according to the system of syntaxeme groups).

    7. Forming an Event: Support /7

    8. Forming a concept: Industry.ATTR:Aerospace /8

    9. Forming a pointer: Countries.these /9

    10. Finding a co-referent for the pointer: 9:: Countries.[1., 2.] /10

    11. Forming a concept: [8.10] /11

    12. Filling the terminal of the event: Event.7.Object=(11) /12

    13. Filling the terminal of the event: Event.5.Object=(12) /13

    14. Filling the terminal of the event: Event.7.Subject=(Event.5.Subject)=(4) /14

(The last operation uses a global rule that an event, which is an object of another event, inherits the same subject, unless an alternative subject is specified explicitly, as in: "*я хочу, щоб він сказав*" – 'I want him to tell').

A syntactic structure of a sentence determines the necessary sequence of operations, which creates a mental model of a situation represented by the sentence. The same mental model can be created with different morphosyntactic means of synonymous sentences. In this case, even if the way of concept formation is different, the terminals of events are filled in a similar way, so the mental model does not change. For example, the protocol of operations for a phrase "*український і німецький лідери*" (the Ukrainian and German leaders) is the same as the protocol for the phrase "*лідери України і Німеччини*" ('the leaders of Ukraine and Germany'), since given the present lexical content, an attributive and a possessive syntaxemes determine the same cognitive operations in the protocol. The phrase "*лідер України та лідер Німеччини*" (the Ukrainian leader and the German leader) omits the operation /3 in the protocol, since the content of the following operation is determined explicitly in the sentence. The phrase "*своїх країн*" (their countries) instead of "*цих країн*" (these countries) determines a different way of forming the concept /11: 'the countries of the Ukrainian leader and the German leader' = 'Ukraine and Germany'. The way of filling terminals of events does not change.

It can be shown, that English translation of the sentence (9) specifies similar protocol of cognitive operations and the same mental model, despite the fact that it has different morphosyntactic structure.

The procedures of formation of mental models are specified by the systems of sytaxemme groups; these procedures can be considered to be a meta-language for representing a universal extralinguistic meaning of texts.

It can be concluded that systems of syntaxeme groups represent the structural organisation of syntaxemes in text and are the means of formal modelling of deep syntactic structure of sentences. They allow representing lexical and syntaxeme semantics as well as syntactic ambiguity in a homogeneous way. The systems of syntaxeme groups can be used for creating of modifying a mental model of a situation described by the text, specifying a sequence and types of changes in extralinguistic knowledge structures.