КОМП'ЮТЕРНА ЛІНГВІСТИКА

УДК 811.161.2'42:004
https//doi.org/10.17721/um/49(2019).117-129

Natalia Darchuk, Doctor of Philology, Prof.
Taras Shevchenko National University of Kyiv, Ukraine

COMPILING OF THE ELECTRONIC DICTIONARY OF MODELS OF THE UKRAINIAN LANGUAGE MULTICOMPONENT COMPLEX SENTENCES

The purpose of this study is to construct an automatic syntactic analysis (ASA) and, as a result, to compile a dictionary of models of multicomponent complex sentences for studying the fectures of the linear structure of Ukrainian text. The process includes two-stages: the first stage is an automatic syntactic analysis of the hierarchical type which results in building of a dependency tree (DT), in the second stage, the sentence structure information is automatically extracted from the obtained graph.

ASA is a package of operations performed with a string of morphological information (the result of AMA work) representing the incoming text for determination of syntactic relations between text units. The outgoing text for the ACA is a string of information reduced after the AMA to wordforms.

We have studied features of the linear structure of 2000 Ukrainian language sentences in journalistic genre (selection of 52000 words use). Based on the obtained results, we have constructed the real models of the syntactic structure of sentences, in which the relations between simple clauses were presented. All grammatical situations of the linear context were possible manifestations of models in the text. Based on that data, the algorithm for the automatic generation of a complex sentence model was created.

These models are linear syntax grammar. All types of syntactic connection between the main and subordinate clauses are recorded algorithmically. Thus, it is possible to build the interpretations of the linear structure of the Ukrainian language sentence almost not using lexical-semantic information. The theoretical value of the paper is in extension of our knowledge about the structure of the syntactic level of the language and the variety of mechanisms functioning at that level. The applied value, is first of all, in creation of the dictionary of compatibility of compound (coordinated) and complex (subordinated) sentences, and in the possibility of constructing requests to the Ukrainian language Corpus in order to mine from the text definite models sentences, creating own dictionaries of authors and styles.

Keywords: *dependency tree, automatic syntactic analysis, models of multicomponent complex sentences, phrase, frequency dictionary.*

The construction of systems of automatic text analysis (ATA) has more than fifty-year history, but any stable system of morphological, syntactic, morphemic and semantic analysis has not been found due to the ambiguity and vagueness of language units. However, researches are in process, already created systems are under development and, for example, AGAT-grammar has been developed in the laboratory of computer linguistics at the Institute of Philology of Taras Shevchenko National University of Kyiv [Дарчук / Darchuk]. First, the theoretical basis of the ASA itself was created, and then its experimental implementation was performed. In the theoretical aspect, within the AGAT computer grammar, such a type of ASA was developed, which made it possible to extract various information about functioning of grammatical syntactical units and their categories, in particular, formally-syntactic: predicativity, subordination, coordination, as well as category of subject, predicate, auxiliary part of sentence, etc. Linguistic software has been developed, which, in the first stage, provides automatic phrase mining from any text Corpus, which makes it possible to build automatically alphanumeric dictionaries of compatibility of all parts of the language, and to make syntactic analysis of the whole sentence in form of dependency tree in the second stage. As a result of AGAT-grammar rules functioning the linear morphological structure of the sentence is transformed into a two-dimensional tree-like syntactic structure.

Semantic information has not been practically included to the model for the following reasons: 1) linear order and tree-like principle which are typical for syntactic structure are not important for semantic information; 2) the essence of the syntactic rules is such that global semantic problems must be divided into smaller ones, which can be parsed at the lower, syntactic level; 3) for semantic information, it is important to comprehend sentence in a context to study the string of the semantic presentation of a sentence as a general representation of the text.

In general, ASA is a package of operations performed with a string of morphological information (the result of AMA work) representing the incoming text for determination of syntactic relations between text units. The outgoing text for the ACA is a string of information reduced after the AMA to wordforms. In our effort to develop a comprehensive and clear technique of linguistic research, we were focused on the

external linguistic form, applying distributive analysis, when each phenomenon is studied in the environment of other phenomena and in interaction with other phenomena. Linguistic units were identified and classified not on the basis of meaning, as it was in the traditional grammar, but on the basis of their distribution in speech.

The AGAT-syntax task is to identify all varieties of compatibility – **predicative, subordinate, and coordinate** – of each word in the text. The grammatical characteristics of the phrase directly depend on which part of the language its keyword belongs to. The lexical and grammatical nature of the word determines its compatibility to the other words. Accordingly, phrases can be divided into **substantive, adjective, pronouns, numeral, verbal and adverbial.** According to our concept of ASA, selection of phrases is based on the grammatical valency, namely, sub-grammar for verb (**31 206 rules**), sub-grammar for noun (**40 023**), sub-grammar for adjective (**6 205**), and idioms vocabulary (about **2720** units). Computer sub-grammars of valencies of the said parts of the language are built by us on a single principle: a lexema is indicated, preposition that participates in government and a case of a substantive word form in the shape of a two-letter code. In theory, according to their composition words combinations (phrases) are divided into **simple, complex and combined** [Загнітко / Zahnitko]. We study only simple binary phrases that can be transformed into complex forms, since definition of their composition requires analysis of the semantic structure. For example, there is no formal reason to separate common phrases of the complicated type, which contain the reference nouns with dependent adjective agreement and not in agreement substantive in a genitive case: *напружена праця вченого* (intense work of a scientist), *урвистий берег річки* (steep bank of river), etc. They combine such forms of subordination as agreement and parataxis, agreement and government. According to our method of analysis the following will be separated: *напружена праця* + *праця вченого* (intense work + work of a scientist); *урвистий берег* + *берег річки* (steep bank + bank of river). Since in the process of analysis the number of each wordform is preserved in database (BD), it is possible to show at the screen “glued” from elementary, simple phrases a complex world combination. Instead, for a complex phrase,

in which a reference noun is combined with two dependent adjectives, there are two degrees of division: the first division step is the separation of the substantive phrase with a dependent adjective; second division step is: subordination of the first adjective not to the noun, but to the elementary substantive phrase, for example, *пахуче / скошене сіно* (sweet-smelling / dry mowed hay). The formal sign for the selection of such a complex phrase is the absence of a comma or cumulative conjunction. Then we get the phrase: *пахуче скошене сіно* (an arrow from *сіно*). At the same time, coordination between *пахуче* і *скошене* is not determined.

If in the word combination the headword is informationally insufficient (lexicalized) word and the dependent word compensates for this insufficiency (so-called complementary or completive relations), they are considered as a phrase acting as a member of the sentence, for example, *дехто з присутніх* (some of the present)*, четверо з них* (four of them)*, почав працювати* (began to work) and similar.

For each word, the following relations are defined: **subordinate, predicative, and coordinate**, since they correspond to the reproduction of the general system of relations between components of the described situation in the sentence. We have also denied the traditional types of subordinate relations – agreement, government and parataxis, based on a wide understanding of the concept of syntactic relation: it is any combination of case form of a noun with the head word. At agreement, the dependent word takes the grammar forms of the headword, and at parataxis, having no changes of the word form it joins the headword under the context. Recently, some syntax researches have interpreted the cases of combinations with the headword of the dependent form of the noun with attributive or contingent meanings as a parataxis [Русская / Shvedova : 21]. However, such a concept, in which the type of relation is determined through the analysis of meaning of each part of the combination, is not suitable for automatic systems. For example: *працювати лікарем* (to work as a physician) is a government relation, *прогулюватися парком* (to walk around the park) is a parataxis; *допомога матері* (help to mother) is a government, and *пам'ятник поетові* (monument to poet) is a parataxis.

Extracting the phrase algorithmically within the ASA, we must be based exclusively on the morphological forms of words, for example, on the instrumental (ablative) case of the subordinate word in the first pair or on the dative case in the second one.

In the AGAT computer grammar, subordinate connections are divided into **kernel** and **non-kernel**. **Kernel** connection is a connection, in which the analyzed word is governing, the head. For example, in a sentence *Від економічної кризи сильно постраждали майже всі європейські держави* (Almost all European states have been severely affected by the economic crisis) we define such kernel connections: ***кризи*** dominates over *економічної*; ***постраждали*** dominates over *від*; ***від*** dominates over *кризи*; ***постраждали*** *dominates over сильно;* ***всі*** dominates over *майже*; ***держави*** dominates over *всі*; ***держави*** dominates over *європейські*.

Non-kernel connection is a connection in which the analyzed word is subordinated, governed. In the above example, non-kernel connections can be observed between the words *економічної* (is subordinated to *кризи*), *сильно* (is subordinated to *постраждали*), *європейські* (is subordinated to *держави*), etc.

Predicative is the connection between the main components of the sentence "subject – predicate", which is based on mutual subordination of the main parts of the sentence, that is, on their interdependence.

Coordinative is a connection in which words are neither dominant nor dominated in relation to one another. It is considered that two words are in coordinative connection, if each of them is subordinated to the same third word, or if they are joint by connective (conjunction) or separated by comma. The formal reason for finding of coordinative string is a check of codes of the analyzed pair of words according to the table of coordination.

At the automatic compiling of collocation dictionary in any Ukrainian text Corpus we receive **four types of models: kernel; non-kernel** (adjunctive which reflect subordinate relations**); coordinative; predicative.**

Since dependency grammar is based on the theory of word-combination and syntactic relations, the main concept of which is government, and hence a valency, we can characterize the second stage, the construction of dependencies tree with the following theoretical principles.

1. The tree is constructed from units of the same taxonomic level, from the wordforms. Connective words, like those ones having a full meaning are considered to be minimal elements of the syntactic structure.

2. Wordforms in the sentence are connected by the relation of subordination, which is a generalization of the traditional syntactic relations of government, agreement, parataxis (coordination and homogeneity fits in with the general system of subordination).

3. The sentence has one absolutely independent wordform – a predicate. All other wordforms are subordinated to a certain "master" and only to one (wordform can not have two simultaneous governances). However, one wordform can subordinate more than two wordforms. In the system of subordination two types of relations in the group of wordforms can be defined:

– branching, when several wordforms are subordinate to one wordform;

– a string of subordination, when each subsequent wordform is a dependent of the previous one.

The structure of dependencies ends with the wordforms, which do not subordinate any other wordforms.

4. For each wordform you can specify a set of subordinates that reflects its behavior in speech at construction of a certain sentences. Such models describe the valency of wordforms, in other words, potential syntactic connection of a certain wordform with its dependent wordforms.

5. The hierarchical structure of dependencies of wordforms in a sentence can be graphically illustrated using brackets, arrows or as a dependency tree [Севбо / Sevbo : 8–9].

It should be noted that we are focused on the last two statements (4–5): the calculation of valency models; building a dependency tree. Actually, the first stage of the AGAT-syntax was a construction of the direct components tree (which are a word combination (phrase), and the second one was based on the rules of the transition from the grammar of the direct components to the dependence tree, particularly, because on the first stage the kernel member (master) was determined.

Fig. 1 represents a table of subordination which is built by the program **automatically** and which corresponds to a dependency tree – it can be inverted into a graphical view (Fig. 2). Above the working table of Fig. 1 the sentence with a morphological annotation of the parts of speech and categorical characteristics is presented. The table has three columns: in the first there is the main member of the binary compound (the master), in the second there is the member of the compound (servant) subordinated to it, and in the third the syntactic information about the type of compound is given. The entire registry is shown in the additional table that pops up when you click on the triangle at the end of the line (its fragment is visible in Fig. 1 in the upper right corner). It should be noted that each type of syntactic connection has its internal code traced in Fig. 2 (the last column). This user-friendly interface allows you to edit a tree manually.

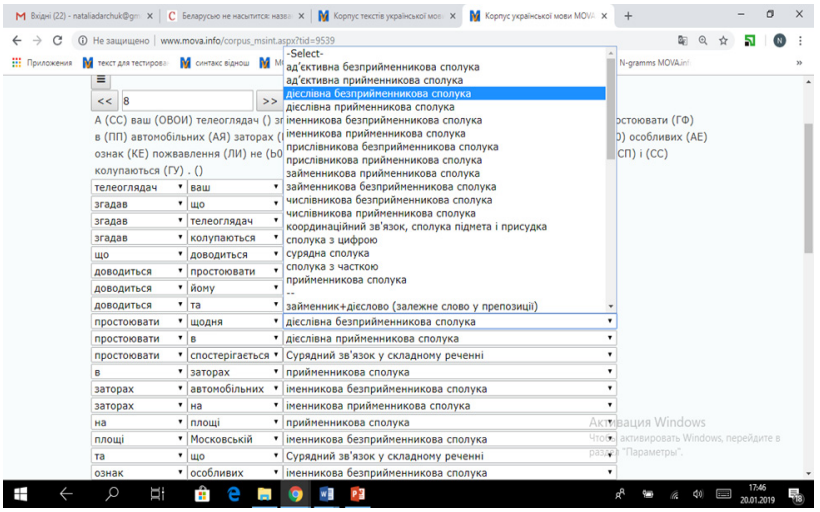


Fig. 1. Program interface of DT working table construction

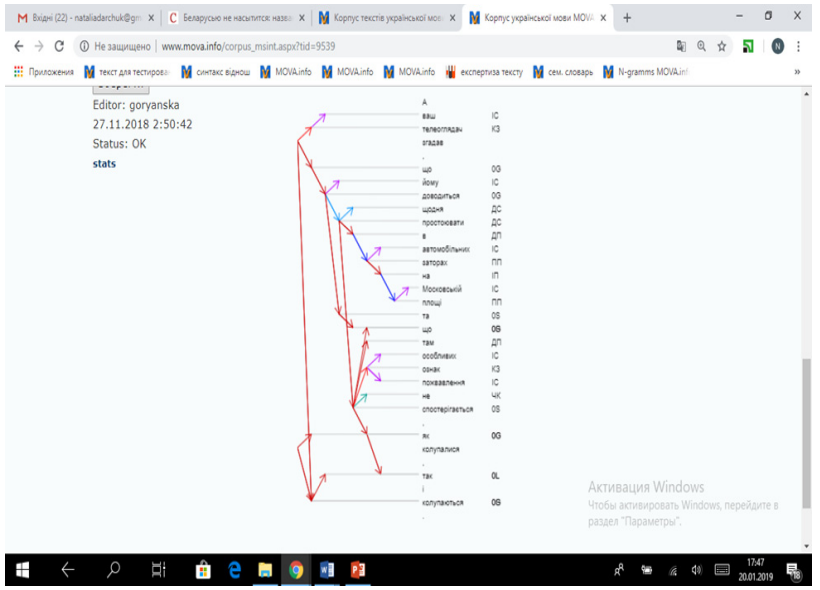


Fig. 2. Program interface with built dependency tree

As you can be seen in Fig. 2, the tree is continuous, that is, all the connections between the predicative parts of the complex sentence are visualized.

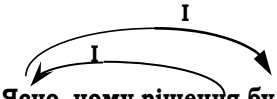
Five models of representation of connections of predicative parts in a complex subordinated sentence have been proposed.

Model 1. The rules of Model 1 work in case when the predicative parts are connected by a subordinate connective (СП (SC) code). Then the SC is subordinate to the root (kernel) of the main clause, and the connective subordinates root (kernel) of a subordinate clause, that is the predicate of the main clause becomes the absolute top of the complex sentence. Communication through the subordinate connective is marked with letter **G.** For example,

Припускається, що подібні розподіли зберігають зв'язки

(It is assumed that such division preserve connections)

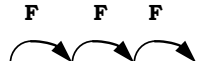
Model 2. When the subordinate clause is attached to the main clause by connective word, then the absolute top is the root (kernel) of the main clause, which subordinates the root (kernel) of the subordinate clause and the latter, in its turn, the connective word. Communication through the connective word is marked with letter **I**.



Ясно, чому рішення було єдиним.

(It is clear why the decision was the only one)

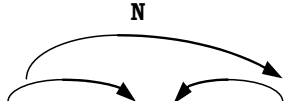
Model 3. In accordance with the rules of this Model 3, the main clause, which comprises a correlative word, is connected with a subordinate clause, which has a subordinate connective in its structure. Then the root (kernel) of the main clause subordinates the correlative word in the main clause, the correlative word subordinates the connective in the main clause, and the connective subordinates the root (kernel) in the subordinate clause. For the purpose of this model, the subordinate connection is marked with letter **F,** for example,



Модель описана так, що стає зрозумілою.

(Model is explained in such a way that it becomes clear)

Model 4. In accordance with the rules of Model 4, the main and the subordinate clauses are connected with the connective word with such an anaphoric connection, which correlates this conjunctive word with its antecedent and the connective word. For the representation of the anaphoric connection in the dependency tree, the following relations are proposed: subordinate connection (N) attaches the root (kernel) of the main clause to the correlative word in it and the root of the subordinate clause, and the latter, in its turn, a connective word, for example,

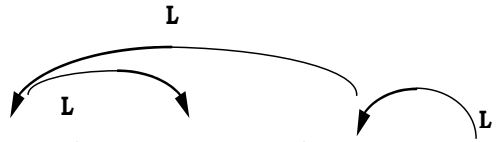


Програма аналізується тоді, коли з'являється помилковий оператор

N N

(Program is analyzed in case when false operator occurs)

Model 5. Use of separate subordinate connectives is accompanied in the corresponding main clause by the wordforms, the correlates. In this case, the first part of the connective in the main clause is subordinated to the root (kernel) of the main clause, the second part, which is in the subordinate clause, is subordiante to that part of the connective which is contained in the main clause, and in its turn, subordinates the root of subordinate clause (the type of subordinate connection **L**), for example,



Якщо добре описаний принцип роботи, то полегшується розуміння деяких вузлів.

(If the operation principle is described properly it facilitates the understanding of certain units functioning)

The proposed models are the basis of constructing of dependency tree (DT) of a complex sentence. Thus, it provides the possibility to consider one more gnoseological problem, namely, the compatibility of the types of relations between simple sentences in a complex sentence, and in ontological terms - to compile electronic frequency dictionary of the compatibility of simple sentences within the complex sentence. The formal indicators used by the program are connectives, punctuation marks, correlative words.

The models of government are composed according to the aforementioned in the presented figures taking into account the order of simple sentences. If for coordinative sentences only linear order is possible, a subordinate clause allows linear and insertion.

We propose to compile such an electronic dictionary of complex sentences models that would reflect the number of simple sentences within the complex, the order of simple sentences, the compatibility of the types of relations between simple sentences [Кулагина / Kulagina; Цимmерлинг / Tsimmerling]. An experiment was conducted for this purpose.

1. We have studied features of the linear structure of 2000 Ukrainian language sentences in journalistic genre (selection of 52000 words use).

2. Based on the obtained results, we have constructed the real models of the syntactic structure of sentences, in which the relations between simple clauses were presented.

3. All grammatical situations of the linear context were possible manifestations of models in the text.

4. Based on that data, the algorithm for the automatic generation of a complex sentence model was created.

The proportion of simple and complex sentences in the studied material is as follows: 39% are simple sentences, 61% are complex.

The statistics on the quantitative structure of complex sentences: two-clause – 760 (64.6% of 1177; number of models – 9); three-clause – 268 (22.77%; models – 29); four-clause – 100 (8.49%; models – 38); five-clause – 29 (2.46%; models – 13); six-clause – 14 (1.18%; models – 12); seven-clause – 2 (0.162%; models 2); eight-clause (2%; 2); ten-clause – 1 (0.08); eleven-clause – 1 (0/08).

Models of complex sentences:

a) for two-clause sentences:
P1+P2 (130)
P1> P2 (428)
P1> {P2}*P1 (83)
P1& P2 (39)
P1 < P2 (37)

b) for three-clause sentences:
P1+P2+P3 (21)
P1 + P2 > P3 (19)
P1 > P2 + P3 (37)
{ P1 + P2}+ P3 (4)
P1 > {P2}* P1> P3 (19)
P1 +P2 > {P3}*P2 (2)
P1 > P2 > P3 (69)
{ P1 > P2} + P3 (3)
P1 > {P2 > P3}* P1 (7) etc.

c) for four-clause sentences:
P1 + P2 + > P3 +P4 (1)
P1 + P2 + P3 > P4 (3)
P1 +P2 > P3 + P4 (1)
P1 + P2 > P3 > {P4}* P3 (1) etc.

The models are read as follows: P1, P2 Pn stand for the number of a simple sentences in the complex. Plus between the symbols demonstrates coordination, the signs “more” or “less” illustrate subordination, for

example, P2> P3 means that the second simple sentence subordinates the third. And the model P1> {P2}* P1 illustrates the discontinuity of the main clause, the second clause is "wedged" into the first, and then it is continued (for example, *Книга* (P1), *яку я придбала вчора* (P2), *виявилася дуже цікавою* (P1)) (Book (P1), which I bought yesterday (P2), was very interesting (P1)). All in all we have found 107 models. These models are linear syntax grammar. All types of syntactic connection between the main and subordinate clauses (see Models 1–5) are recorded algorithmically. Thus, it is possible to build the interpretations of the linear structure of the Ukrainian language sentence almost not using lexical-semantic information. The theoretical value of the paper is in extension of our knowledge about the structure of the syntactic level of the language and the variety of mechanisms functioning at that level.

The applied value, is first of all, in creation of the dictionary of compatibility of compound (coordinated) and complex (subordinated) sentences, and in the possibility of constructing requests to the Ukrainian language Corpus in order to mine from the text definite models sentences, creating own dictionaries of authors and styles.

REFERENCES

1. Darchuk N.P. (2013) Kompyuterne anotuvannia ukrainskoho tekstu: rezultaty i perspektyvy [*Computer Annotation of Ukrainian Text: Results and Prospects*]. Kyiv: Osvita Ukrainy. 543 p. (in Ukrainian).

2. Zahnitko A.P. (2004). *Osnovy ukrainskoho teoretychnoho syntaksysu [Fundamentals of Ukrainian Theoretical Syntax*]. Part 1. Gorlovka: GDPIIM, 227 p. (in Ukrainian).

3. Kulagina O.S. (2001) Ob odnom podhode k ustanovleniyu otnosheniy mezhdu prostymi predlozheniyami v sostave slozhnogo pri avtomaticheskom analize tekstov [*One approach to define relations between simple sentences as part of a complex at automatic analysis of texts*]. *Mathematical questions of cybernetics*, no 10, pp. 15-34 (in Russian).

4. Shvedova N.Yu (ed.) (1980) Russkaya Gramatika [*Russian Grammar: in 2 Vols*]. Moskva: Nauka, vol. 2, 709 p. (in Russian).

5. Sevbo I.P. (1981) Graficheskoe predstavlenie sintaksicheskih struktur i stilisticheskaya diagnostika [*Graphical representation of syntactical structures and stylistic diagnostics*]. Kiev: Naukova dumka, 192 p. (in Russian).

6. Tsimmerling A.V. (1999) Poriadok slov i sintaksicheskie pozitsii [*Words order and syntactic positions]. Proceedings of the international seminar "Dialog'98" on computer linguistics and its application.* URL : https://antonzimmerling.files.wordpress.com/2013/06/turus.pdf (in Russian).

ЛІТЕРАТУРА

1. *Дарчук Н. П.* Комп'ютерне анотування українського тексту: результати і перспективи / Н. П. Дарчук. – К. : Освіта України, 2013. – 543 с.

2. *Загнітко А. П.* Основи українського теоретичного синтаксису : у 3 ч. / А. П. Загнітко. – Горлівка : ГДПІІМ, 2004. – Ч. 1.– 246 с.

3. *Кулагина О. С.* Об одном подходе к установлению отношений между простыми предложениями в составе сложного при автоматическом анализе текстов / О. С. Кулагина // Математические вопросы кибернетики. – 2001. – № 10. – С. 15–34.

4. Русская граматика : в 2 т. / редкол.: Н. Ю. Шведова (гл. ред.). – М. : Наука, 1980. – Т. 2 . – 709 с.

5. *Севбо И. П.* Графическое представление синтаксических структур и стилистическая діагностика / И. П. Севбо. – Киев : Наук. думка, 1981. – 192 с.

6. *Циммерлинг А. В.* Порядок слов и синтаксические позиции / А. В. Циммерлинг // Труды международного семинара "Диалог'98" по компьютерной лингвистике и её приложеним / А. С. Нариньяни (ред.). – Казань, 1999. URL : https://antonzimmerling.files.wordpress.com /2013/06/turus. pdf (дата звернення: 10.10.2018).

Наталія Дарчук, д-р філол. наук, проф.
КНУ імені Тараса Шевченка, Київ

Щодо укладання електронного словника моделей багатокомпонентних складних речень української мови

Метою дослідження є побудова автоматичного синтаксичного аналізу (АСА) і як наслідок – укладання словника моделей багатокомпонентних складних речень для вивчення властивостей лінійної структури українськомовного тексту. Процес укладання двоетапний: на першому етапі працює автоматичний синтаксичний аналіз ієрархічного типу, який завершується побудовою дерева залежностей (ДЗ), а на другому – з одержаного графа автоматичного здобувається інформація про модель речення.

АСА – це сукупність операцій, які виконуються над послідовностями інформації морфологічного характеру (результатом роботи АМА), що представляють вхідний текст, для встановлення синтаксичних зв'язків між текстовими одиницями. Вихідним текстом для АСА є редукована після АМА послідовність інформації до словоформ.

Досліджувалися властивості лінійної структури 2000 українськомовних речень публіцистичного стилю (вибірка у 52000 слововживань). На підставі цих результатів будувалися реальні моделі синтаксичної структури речень, у яких відображені відношення між простими реченнями. Усі граматичні ситуації лінійного контексту були можливими маніфестаціями моделей у тексті. На підставі цих даних будувався алгоритм автоматичного творення моделі складного речення. Отримані моделі є граматикою лінійного синтаксису. Усі види синтаксичного зв'язку між головним і підрядними реченнями фіксуються алгоритмічно. Таким чином можна будувати інтерпретації лінійної структури українського речення, майже не використовуючи лексико-семантичну інформацію. Теоретичне значення роботи полягає в поглибленні наших уявлень про будову синтаксичного рівня мови і різноманітності механізмів, які діють на синтаксичному рівні. Прикладне значення вбачаємо у створенні словника сполучуваності складносурядних і складнопідрядних речень, у можливості побудови запитів до Корпусу української мови з метою здобувати з тексту речення певних моделей, створюючи свої власні словники авторів, стилів.

Ключові слова: *дерево залежностей, автоматичний синтаксичний аналіз, модель багатокомпонентного складного речення, словосполучення, частотний словник.*

Стаття надійшла до редколегії 21.11.18