LINGUISTICA ANTVERPIENSIA, 2021 Issue-1
www.hivt.be
ISSN: 0304-2294

The Morphological Analysis and Synthesis of Word Forms in the Linguistic Analyzer

Bakhtiyor Mengliev [1],
Shohida Shahabitdinova [2],
Shahlo Khamroeva [3],
Shakhnoza Gulyamova [4],
Adiba Botirova [5]

[1] Doctor of philology, professor of Tashkent State University of Uzbek language and literature
Editor at Portal International Scientific Journals Tadqiqot.uz
[2] Doctor of philology, professor of Andijan State University,
[3] PhD, Senior researcher of Tashkent State University of Uzbek language and literature
[4] PhD, Senior researcher of Tashkent State University of Uzbek language and literature
[5] PhD, Navai State Pedagogical Institute

Issue Details
Issue Title: Issue 1
Received: 08 October, 2020
Accepted: 19 November, 2020
Published: 25 December, 2020
Pages: 703 - 712

Abstract

This article is dedicated to the issue of morphological analysis and synthesis of word forms in a linguistic analyzer, which is a significant feature of corpus linguistics. The article discourses in detail the morphological analysis, the creation of artificial language, grammar and analyzer, the general scheme of the analysis program that "recognizes" the natural language, the types of analysis (identifying) software, the purpose of the analysis, some approaches in the information search engine. The possibility of using a number of tools in the design of a morphological analyzer, their function and features are described in detail.

Keywords

linguistic analyzer, lexical, morphological, syntactic, semantic analysis, query language, morphological analysis, morphoanalysis software, artificial language, language semantics and syntax, database (D), information resource (IR), automated information system (AIS), information search engine (ISE), expert system (ES), automatic design systems (ADS), automated scientific and technical information system (ASTIS), information search thesaurus (IST) and the Internet.

INTRODUCTION. Types of analysis such as lexical, morphological, syntactic, semantic must be implemented to perform a perfect linguistic analysis of text / information. It is well known that sequential analysis is a linguistic analysis of a text (information) in a logical sequence, in which each task seems more difficult than the previous one. The function of a linguistic analyzer arises from the nature

of the language being analyzed (the text being presented). Therefore, the information search engines to be created should include a software and hardware package that implements a complete list of tasks for the linguistic analysis of text information. The best information search engines available perform the function of morphological analysis of text information (text indexing, user query), as well as execute some syntactic analysis of the sentence [11].

The performance algorithm of information search engines is based on the keywords of pre-indexed texts and the user query is performed on the basis of the keyword. Query language-based searches can find relatively accurate results, in which case the query language has its own characteristics for each information search engine. Therefore, the user often does not refer to it; which reduces the quality of the search results, because the search results will find a large amount of material, the user will have to sort them. This material remains inadequate for the analysis to be performed.

THE MAIN PART. Morphological analysis (Part of Speech tagging). The task of morphological analysis is to automatically determine which category each word in the text belongs to; is to determine which lexical-grammatical class the words belong to. Due to the development of formal morphology in the Russian language, there are ample opportunities to do so. In English, too, the algorithm is relatively simple: although there is lexical polysemy, morphoanalysis software (tagger) can accurately identify a set of words in a text by 90 percent. To perform morphological analysis of Russian texts, a computer version of A. Zaliznyak's grammatical dictionary was used, and the morphoanalysis of English was based on Mueller's grammatical dictionary. The experience of creating Russian and English morphoanalysis suggests that formal grammar must have been created to perform morphological analysis.

Two types of algorithms are used to increase the quality of morphological analysis in the process of determining the set of polysemous words: statistical-probabilistic and word/code running instructions. Most statistical-probability algorithms are based on 2 sources of information:

1) a dictionary of word forms with explanations of lexical-grammatical affiliation, available in each word form of a particular language. It is also called information about all possible circumstances of lexical-grammatical forms. The bigram, trigram, and quadrigram models differ depending on how this information is reflected;

2) The rule-based algorithm is automatically detected from the language corpus or based on material prepared by experts.

Both approaches show the same result. As a result of their use unaccompanied or in various combinations, lexical-grammatical analysis is 96-98 percent qualitative. Even the lexical-grammatical analysis carried out manually, up to 0,5-2 percent of cases of error were observed. It can be said that automatic lexical-grammatical analysis is equal to the level and quality of morphological analysis performed independently by the individual [11].

Syntactic analysis of textual information allows to distinguish semantic elements of a sentence - belonging to a group, predicative basis. This enhances the intelligence of working with generalized semantic elements in the process of dispensation the information specific to it in the text. Text information reprocessing systems, of course, require the use of expert systems, artificial intelligence systems to separate semantic information. The lack of a perfect semantic analysis system today, the problems in creating such a system are due to the fact that the scientific direction of the construction of artificial intelligence systems is not fully established.

The lexical analysis refers to the analysis of textual information in the form of head of speech, a sentence, word forms that represent the expressions of the national language, and it identifies the forms of speech, lexical expressions (assimilation, slang words, etc.) that reflect the specificity of each language. In practice, its implementation does not cause difficulties.

Creating artificial language: language syntax and semantics. Each language is made up of thousands of special characters arranged in a specific order. There is also the rule of links outside the alphabet, because not all the chain of characters in a particular alphabet belongs to a language. Symbols can be combined in an elementary construction of language - a word or a lexeme. Based on them, a relatively more complex construction - sentences are formed. Both cases are considered a chain of characters and follow the rules of construction. Thus, it is necessary to specify these rules, or more precisely, to define the language. In general, language can be defined in three different ways [10]:

1) to count all links / chains available in the language;

2) to show methods to continue the chain / valence; to determine the laws of formation of connections in the language (with the construction of grammar);

3) to find methods to identify chains.

The peculiarity of this logical device (identifier – распознаватель (Russian) is that it reads the question as a chain of characters at the input, and answers it at the output: it determines whether a particular chain belongs to this language. For example, when we read a text, we also act as a morphological analyzer (identifier): we confirm that the text is in Uzbek language.

When thinking about a language, of course, its syntax and semantics are separated. The analyzer / translator also works with the lexical construction (lexeme) of the language. Below we explain the tools that make up such a morphological analyzer.

Language syntax is a set of device rules that are likely to exist in a language. The syntax defines the "form of language", indicating the sum of a chain of characters belonging to a language. As a rule, strict syntactic rules for formal language are developed. Most often, programming languages require the additional introduction and replenishment to the rules of formal language. The syntax of natural language (speech communication) works on the principle of "exceptions confirm the rule" [11].

It is well known that in traditional linguistics, a lexicon is a set of verbal structures of a language. A word or lexical unit of a language (lexeme) is a structure consisting of elements of the alphabet; they do not store any other device. In other words, a lexical unit consists only of elementary characters, no other lexical unit belongs to it. For example, the lexicon of the Uzbek language consists of words in the Uzbek language, and punctuation and spaces are considered to be separating them. Just as lexical units of algebra are mathematical operation symbols denoting a unit of number, function, and measure, in a programming language, a keyword, identifier, constant, and mathematical operation symbol are considered as lexical units. This lexical unit also includes dividers (commas, parentheses, dots, semicolons, etc.).

Grammar and analyzer. Grammar is a description of the sentence construction methods of a particular language; a mathematical system that defines language. While describing grammar, we show the rules for the emergence of a chain of characters belonging to that language. This is why grammar can be called a language chain generator. It shows the separation and identifying of the language − the second type of the emergence of a chain of characters. The language grammar can be described in a variety of ways. For example, Russian grammar can be represented by a more complex set of rules taught in school. For some languages (e.g., the syntactic construction of a programming language) it

is possible to create on the basis of a formal mathematical description based on a system of rules.

The general scheme of the analysis program "identifying" the natural language. An analysis program is a special automated program that determines which language a particular character chain belongs to. The main task of the analysis program is to identify whether a particular character chain belongs / does not belong to the selected language.

Types of analysis software (identifier). Such a program can be described depending on the components that make up it, such as the counting device, the control device, as well as external memory. Depending on the type of counting device, one-way and two-way analysis software is available. This program reads only from one side (left to right) to identify the character chain and does not go back. The two-way analysis program "reads" the character chain from right to left and from left to right [11].

According to the control device, the programs divided into deterministic and non-deterministic programs. The following types of programs are distinguished according to the availability of external memory:

1) analyzer with external memory;
2) analyzer with limited external memory;
3) analyzer with unlimited external memory.

Without external memory, there will be no memory at all in the analyzer. During its operation, only the memory of the control unit works. In a limited external memory analyzer, the size of the memory depends on the length of the character chain. The external memory of such a program works on the basis of a list, model example. In the third type of program, the external memory is "unlimited" and can "read" a chain of characters of any length. Based on this classification, a program is created that combines the three characters of the classification. For example, *a limited external memory analysis program with a two-way identifier.*

The more complex the program of analysis, the more the algorithm that will carry out its work, the more perfect the structure will be. It is more difficult to develop a two-way analyzer than a one-way analyzer.

The purpose of the analysis. For each programming language it is necessary not only to compile the program text in this language, but also to determine whether the existing text belongs to this language. Among other things, the compiler performs the same function. If the compiler performs the function of language recognition, the person who created the particular programming language acts as a character chain generator for

that language. Grammar and the analyzer (identifier) are two independent ways of "knowing" a language.

In general, the task of an analysis program is to create an analysis program for that language based on the existing grammar of the language. Both of these practices must be alternate: it is important that both "identify" and distinguish the same language. The compiler should not only determine whether the character chain belongs to a particular language, but also clarify its semantic load. If the word in the query does not match one of these compiler character chains, then it is negative to the user, that is, will respond in the form that, *such a word does not exist in this language*. So what causes the program to come to such a conclusion? Of course, the origin of this problem should be sought in the linguistic supply. This means that there is no such character chain in the database. The more complex the language, the more difficult it is to develop a compiler. For some languages, it is not at all possible to develop a compiler that analyzes texts in a language based on mathematical resources.

About some approaches in the information search engine. V.V. Ponomarev's "Linguistic support and sociolinguistic specifics of the problem of auto-indexation actualization of information systems" [13] dedicated to the analysis of the problems of activation of information systems, its sociolinguistic specifics, the analysis of the process of development of linguistic support, the research described the principles of descriptive-analytical, distributive-contextological, comparative analysis, component analysis, content analysis, logical analysis, the use of algorithmic modeling methods. The following information systems are distinguished in the study: the database (D), information resource (IR), the automated information system (AIS), the information search engine (ISE), the expert system (ES), the automated design systems (ADS), the automated scientific and technical information system (ASTIS), the information searching thesaurus (IST) and the internet.

The article by T.B.Boltayev and S.I.Ibragimov "On the project of the program system of morphological analysis of the Uzbek language" focuses on the automation of the morphological analysis of the Uzbek language, in particular, nouns and adjectives. Since word formation in Turkish languages, especially noun formation, is very productive, this series of morphological analysis requires knowledge of several types of analysis methods [8]. The expert suggests the use of mathematical apparatus in the process of designing a formal language analysis system as one of the solutions to this problem. Such devices are widely used in high-level programming languages, such as programming languages or

language processing programs (compiler, analyzer, specifier, etc.) [7]. By the 60s of the last century, the use of language processing software became a tradition. In the work on the ETAP system [5], all analytical tools of formal languages were used for the analysis of inflected languages (e.g., Russian). D.S.Yuravsky and J.G.Martin also analyzed the issues of formal English language analysis program [2].

This section discusses the implementation of the traditional tool of formal language analysis in the morphological analysis of the Uzbek language. It is the result of research on modeling the process of automatic (computer) analysis of word groups in analytical languages. The work set itself the task of applying the methods of the theory of formal languages in the analysis of texts in the Uzbek language. There is also experience in using this method: in English [3] and Russian [12] language this method has given effective results. When the text is given in the introductory part of the article, the output of which describes the process of creating a mechanism, which provides the output of the same text with a chain of lexemes, a complete morphological characteristic of the lexemes. Here, morphological analysis in the processing of formal languages is compared with lexical analysis. The multi-scheme of word-forming in the Uzbek language allows the use of the lexical analysis apparatus of legalization with the help of the addition of a word-building apparatus, in which the apparatus of continuous expressions can also be used in the formalization of the word-building scheme.

A relatively complex scheme of word formation: compound word, rooting, lexemization of a phrase goes beyond the analysis of automatic and conventional expressions; they are analyzed on the basis of free-text grammar. When we study computer morphological analysis (morphological parser - MP), we consider not only the necessary mathematical models, but also the information structures used in this process.

MP can be described as follows: a word form is given during the MP input process; in the output process, the stem of the word with all the morphological features of the word is obtained. This information will be the basis for the next stage of analysis (syntactic, semantic, pragmatic analysis, word formation, translation). The morphological description of a word indicates some of its meanings, indicating to which morphological category it belongs, which word group the word belongs to.

For example, a Noun-N is attached to a singular-SG, a plural-PL number category, a Case, and so on. For the Verb-V will be an indicator of the categories of Tense-TS, Person-PR.

The following implements are used in the design of the morphological analyzer:

1. Lexicon: information about the basis, additional, information structure of the stem and suffix needed to form a morphological analyzer. The lexicon is the repository of MP words. The stem, suffix, and morphotacts are included in the lexicon, but it is not possible to include all units in the lexicon. This method is very handy for tracking new word formation.

2. Morphotactics - information model processes that indicate which morpheme category a word form belongs to. They regulate the relationship between morphemes specific to a particular word form. For example, in the noun category, the plural form is always placed after the stem, while other forms are added after it in strict order.

3. Orthographic rules: A set of rules that describe changes in a morpheme belonging to two categories. For example, when a suffix CS (dative) is added to a noun ending in “к” or “қ”, there is a change in these stems: элакка=элак+ CS (dative).

4. Stem Semantics Information Structure-SSUS. Such units are usually referred to as semantic attributes. Based on these attributes, semantic-morphologic rules are formed that check the semantic correctness of the word being analyzed. For example, (Proper (N) & Singular (N)) | (Abstract (N) & Uncount (N)) Þ The set of SG (N) tags denotes a singular, plural noun (Sun, Earth, Moon, Tashkent, Navoi). In addition, SSUS is also used in syntactic and semantic analysis to distinguish morphological characters.

5. Finite State Automata (FSA) [6] is used in the modeling of morphotactics. Another special version of the FSA - transducers [1] is used to model orthographic rules.

IN CONCLUSION, all research on grammar and lexicography in world linguistics relies on the text in the illustrative language corpus. The development of modern intelligent software systems for text processing in natural languages requires a large experimental linguistic base. The language body is designed for repeated use by the user, so the system of tags used in it and the linguistic supply should be brought to uniformity. The corpus layout is based on a standard, a set of layouts called a “coding standard”.

REFERENCES:

1. Akhmedova D., Mengliev B. Semantic Tag Categories in Corpus Linguistics: Experience and Examination // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-3S, October 2019. – P. 208-212.

2. Aho A., Ullman J. The theory of syntactic analysis, translation and compilation. In two volumes. – M.: Mir, 1978, trans. from English.
3. Aho Alfred V., Lam Monica S, Seti Ravi, Ullman Jeffrey D. Compilers: principles, technologies and tools, 2nd ed.: trans. from English – M.: OOO "ID Williams", 2008. – 1184 p.
4. Apresyan Yu.D., Boguslavsky I.M., Iomdin L.L. and others. Linguistic support of the ETAP-2 system. – Moscow: Nauka 1989. – p. 296.
5. Boltaev T.B., Ibragimov S.I. On the project of a software system for morphological analysis of the Uzbek language // (http://ziyonet.uz/uploads/books/473012/5afbccb8e9f55.pdf (31.10.2020)
6. Boltaev T.B., Kuzminov T.V., Pottosin I.V. On the structural design of programs and tools for its support // Programming environment: methods and tools. – Novosibirsk, 1992. – P.22-37.
7. Khamrayeva Sh.M. Specific and prevalent peculiarities of the authorship corpus // IMPACT: International Journal of Research in Humanities, Arts and Literature. – ISSN (P): 2347-4564; ISSN (E): 2321-8887. – Vol. 6, Issue 6, Jun 2018. – P. 431-438.
8. Khamroeva Sh. Morphotactic rules in the morphological analyzer of the uzbek language // Middle European scientific bulletin. VOLUME 6, NOVEMBER 2020. – ISSN 2694-9970 – P. 45-49.
9. Khamroeva, Sh.M., Gulyamova, Sh. K. (2020). Electronic dictionaries − the product of applied linguistics. ISJ Theoretical & Applied Science, 07 (87), 463-466.
10. Mohri M.A. Finite-state transducers in language and speech processing. Computational Linguistics, 23B) – p. 269-312.
11. Nozhov I.M. Morphological and syntactic text processing (models and programs): dissert. Candid. philol. sciences. – Moscow, 2003. // https://docplayer.ru/26110069-I-m-nozhov-morfologicheskaya-i-sintaksicheskaya-obrabotka-teksta-modeli-i-programmy-1.html (application date: 02.01.2021)
12. Place and purpose of linguistic support in information systems. The concept of an information system // https://gigabaza.ru/doc/116551-pall.html (application date: 02.01.2021)
13. Shahabitdinova Sh., Sayfullaeva R., Sadullaeva N., Ernazarova M., Inogamova N. General philosophical issues of language research // Journal of Critical Reviews. Vol 7, Issue 9, 2020. ISSN- 2394-5125.
14. Ponomarev V.V. Linguistic support and sociolinguistic specifics of the problem of auto-indexation actualization of information systems: dissert. abstr. cand. philol. sciences. – Moscow, 2005.
15. Daniel S.Jurafsky, James H.Martin. Speech and Language Processing. Contributing writers: Andrew Kehler, Keith Vander Linden, Nigel Ward 2000 y. Prentice Hall, Englewood Cliffs, New Jersey 07632. – pages: 950.
16. The Lexical Semantics of a Machine Translation Interlingua. Rick Morneau // http://www.eskimo.com/~ram/lexical_semantics.html 2006. (26.10.2020)

17. The Structured Constructing as a Discipline of Safe Programming and Instruments Supporting It /Aniskov M.I., Boltaev T.B., Kochetov D.V. at al//Instrumental Congress on Computer Systems and Applied Mathematics CSAM'93. St-Petersburg. July 19-23.
18. Xamrayeva Sh.M. Morphological markup ang linguistic model // American Journal of Research. – USA, Michigan, 2018. – № 9-10. – P.187-198. (SJIF: 5,065. № 23)