ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-5206, Impact Factor: 12,23

American Academic publishers, volume 05, issue 02,2025

Journal: https://www.academicpublishers.org/journals/index.php/ijai

AMERICAN ACADEMIC PUBLISHER OPEN ACCESS JOURNAL

SYNTACTIC ANALYSIS IN COMPUTATIONAL LINGUISTICS: EXPLORING COLLOCATIONS AND VALENCY

Nurmuxammedova Nurjaxon Istam kizi

Master's student at Samarkand State Institute of Foreign Languages

nurmuxammedovanurjaxon@gmail.com

Asadov Rustam Muminovich

Samarkand State Institute of Foreign Languages PhD, dotsent

rustamasadov1972@gmail.com

Annotation: This article examines the significance of syntactic analysis in computational linguistics, focusing specifically on the concepts of collocations and valence. It discusses how these linguistic phenomena contribute to the understanding of sentence structure and meaning, and highlights their applications in natural language processing (NLP) and artificial intelligence (AI). The article also outlines recent advancements in the field and suggests directions for future research.

Keywords: Syntactic analysis, computer linguistics, collocations, valiancy, natural language processing, language models.

Introduction

Syntactic analysis is a fundamental component of computer linguistics, which seeks to understand and model human language through computational means. The structure of sentences—how words combine and interact—is essential for various applications, including machine translation, information retrieval, and text mining. Among the critical aspects of syntactic analysis are collocations and valence.

Collocations refer to the habitual co-occurrence of words that form phrases commonly used by native speakers, while valence pertains to the capacity of verbs to combine with a specific number and type of arguments. This article explores the roles of collocations and valence in enhancing syntactic analysis, elucidating their importance in natural language processing tasks and their implications for future research.

In December 2003, the World Summit dedicated to the problem of building an information society was held in Geneva (Switzerland). The following slogan was introduced: "Building an information society is the global challenge of the new millennium." Two important documents were adopted at the summit: the Declaration on the principles of building an information society and the work plan for building an information society. According to these documents, the development of society will be closely related to computer technologies. It should be noted here that computer linguistics is considered the most important direction that defines the perspective of information technologies and it will be of decisive importance in the development of the information society.

Main Part

One of the most promising areas of computer linguistics is artificial intelligence. On the pages of Wikipedia, which is considered a virtual dictionary, artificial intelligence is defined as follows: "Artificial intelligence is a feature of a computer or a robot that is mainly aimed at solving issues related to human mental activity, in particular, thinking, understanding the content of speech, and summarizing data. This term is also used for a branch of computer technology related to the development of systems with these features."

Until the 70s of the 20th century, research on artificial intelligence was carried out within the framework of cybernetics and informatics. Since the 80s and 90s of the 20th century, artificial intelligence has become the object of study of many sciences. In particular, such sciences as neurolinguistics, psychology, informatics, neurophysiology, epistemology (the doctrine of knowledge in philosophy), cognitology, cognitive linguistics, and computer linguistics deal with the problem of artificial intelligence.

Computational linguistics is a logical continuation of mathematical linguistics, which is the most important part of applied linguistics. Computational linguistics began to form as a field in 1954 at Georgetown University in the United States during the world's first experiment on machine translation, and by 1960 it was formed as an independent science. Computer linguistics is a copy of the English word "computational linguistics". Until the 80s of the 20th century, this science was called by different names: computational linguistics, mathematical linguistics, quantitative linguistics, and engineering linguistics. The main goal of this science is the development of computer programs for solving linguistic problems, optimization of human-machine (computer) communication, natural language processing. NLP involves computer analysis and synthesis of natural languages in computational linguistics. In this case, analysis refers to the computer understanding of natural language using morphological, syntactic and semantic analysis, and synthesis means the grammatical formation and generation (production) of text in a computer. Software developed for NLP are: AlchemyAPI, Expert System S.p.A.. General Architecture for Text Engineering (GATE), Modular Audio Recognition Framework, Monty Lingua, Natural Language Toolkit (NLTK). Computational linguistics has many functions.

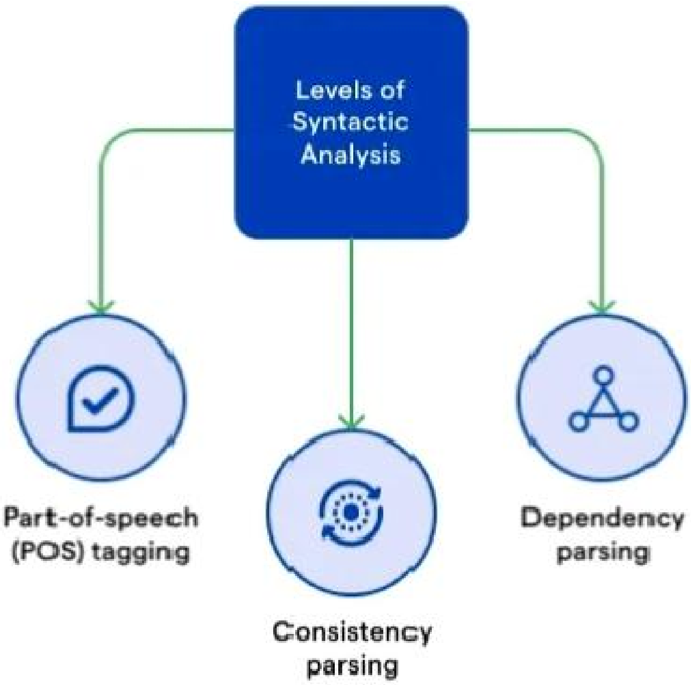
Collocations are combinations of words that frequently occur together in a language. Word combinations are syntactic units formed from the combination of words based on certain logical and grammatical rules. It is clear from this that the internal meaning, external grammatical compatibility and their appropriateness are required for the words to enter into a relationship. They can be classified into several types, including:

- Adjective + Noun: e.g., "strong coffee," "heavy rain."
- Verb + Noun: e.g., "make a decision," "give a presentation."
- Noun + Noun: e.g., "data analysis," "traffic jam."

Syntactic parsing is the automatic study of natural language syntactic structure, particularly tagging spans of constituents (in constituency grammar) and syntactic interactions (in dependency grammar). The challenge of structural ambiguity in natural language serves as its impetus. A sentence may be assigned more than one grammatical parse, necessitating knowledge beyond computational grammar rules to determine which parse is meant. One of the key jobs in computational linguistics and natural language processing is syntactic parsing, which has been studied since the middle of the 20th century when computers were invented.

Syntactic analysis in computational linguistics converts text into computer-process able structured representations using algorithms. Computer science processes like code compilation and natural language processing tasks like sentiment analysis depend on parsing algorithms. Because it can adequately express the majority of computer language syntax, the context-free grammar is frequently employed. For instance, the majority of contemporary programming languages, including Python, Java, and C++, implement parsers using context-free grammars. A collection of recursive rules or productions that specify which symbol strings make up syntactically proper strings in the language are used to define these grammars.

Different formalisms are proposed by various grammar theories to describe the syntactic structure of sentences. These formalisms can be categorized as constituency grammars or dependency grammars for computational purposes. Different methods are needed for parsers for each class, and different strategies have been used to tackle the two issues. Together with the development of new parsing algorithms and techniques, the process of creating human-annotated treebanks using different formalisms (such as Universal Dependencies) has continued.



Picture 1. Levels of syntactic analysis

Part-of-speech tagging (which resolves some semantic ambiguity) is a related problem, and often a prerequisite for or a sub problem of syntactic parsing. Syntactic parses can be used for information extraction (e.g. event parsing, semantic role labelling, entity labelling) and may be further used to extract formal semantic representations.

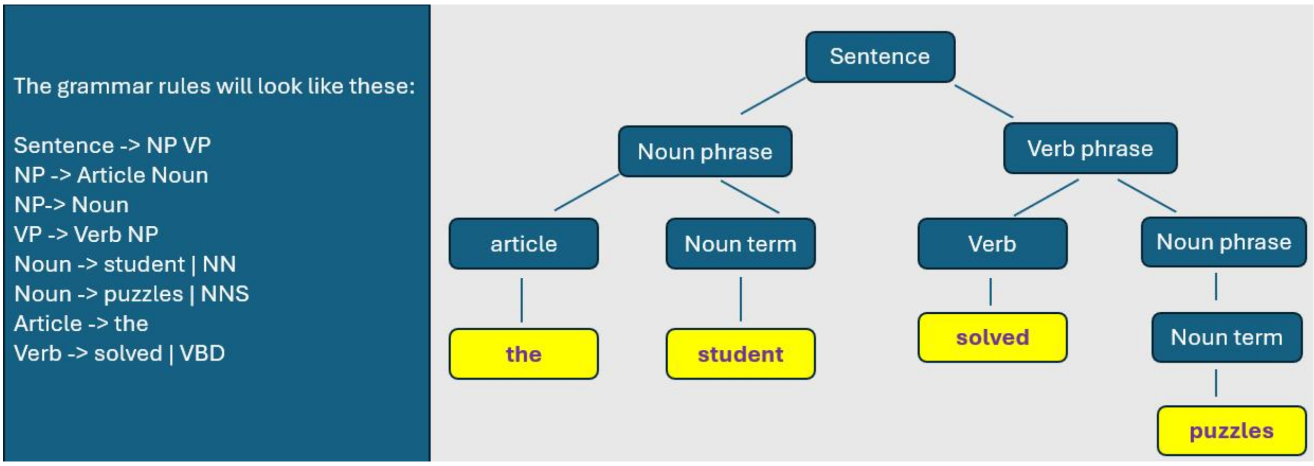
Constituency parsing analyzes sentences by breaking them down into hierarchical constituents (phrases), revealing the nested structure of a sentence. For example, a sentence might be divided into a noun phrase and a verb phrase. Constituency parsing involves parsing in accordance with constituency grammar formalisms, such as Minimalism or the formalism of the Penn Treebank. This, at the very least, means telling which spans are constituents (e.g. [The man] is here.) and what kind of constituent it is (e.g. [The man] is a noun phrase) on the basis of a context-free grammar (CFG) which encodes rules for constituent formation and merging.

Dependency parsing focuses on the direct grammatical relationships (dependencies) between individual words, representing these relationships as a directed graph. Parsing using a dependency grammar formalism, such Universal Dependencies (a project that creates multilingual dependency treebanks), is known as dependency parsing. By giving each token a head (or several heads in some formalisms, such as Enhanced Dependencies, for example, in the case of coordination) and a matching dependence relation for each edge, a tree or graph covering the entire phrase is ultimately created. There are basically three recent paradigms for describing dependency parsing: transition-based, grammar-based, and graph-based.

These techniques form the basis for extracting a deeper level of linguistic information, such as identifying collocations and determining valiancy.

In natural language processing (NLP), collocations play a crucial role in:

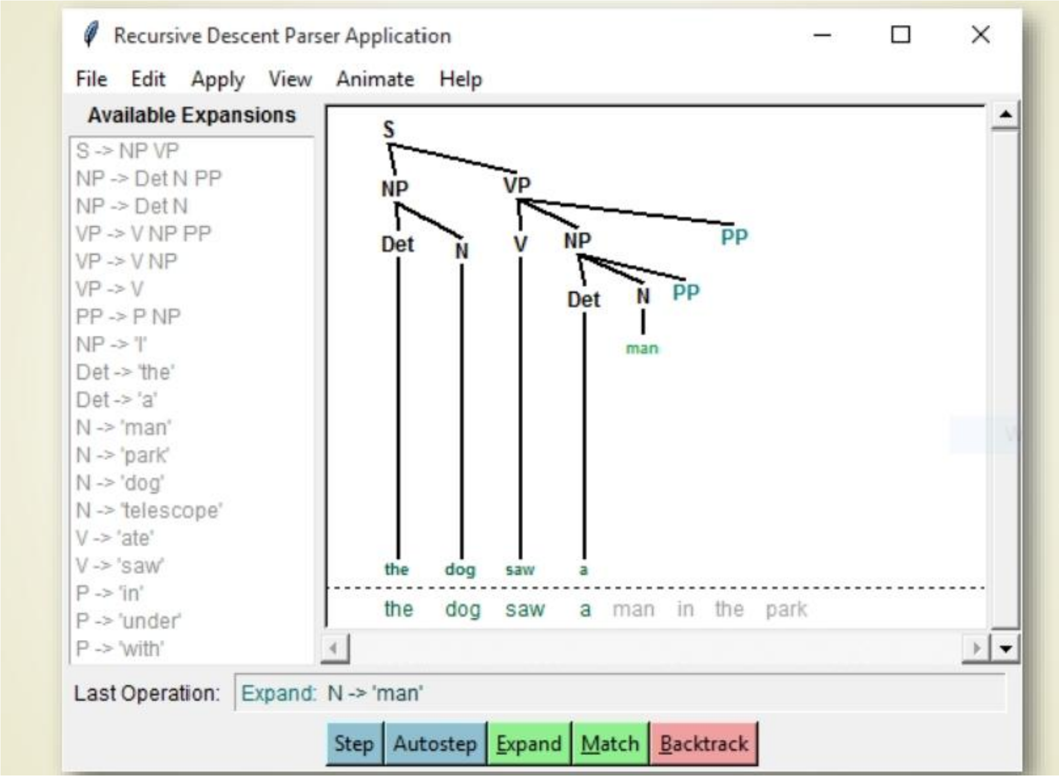
- Natural language generation. Algorithms that generate text must select words that conform to native usage patterns. By incorporating collocation data, systems can produce more fluent and contextually appropriate sentences.
- Semantic understanding. Collocations often convey meanings that are not directly inferred from individual words. For example, the phrase "kick the bucket" signifies death, illustrating how collocations can alter semantic interpretation.
- Language modeling. PModern language models utilize collocation data to enhance predictions about word sequences. This is vital for applications such as speech recognition and text completion, where accuracy in word choice significantly impacts user experience.



The grammar rules will look like these:

Sentence -> NP VP
NP -> Article Noun
NP-> Noun
VP -> Verb NP
Noun -> student | NN
Noun -> puzzles | NNS
Article -> the
Verb -> solved | VBD

Picture 2. The grammar structure of sentence



S -> NP VP
NP -> Det N PP
NP -> Det N
VP -> V NP PP
VP -> V NP
VP -> V
PP -> P NP
NP -> 'I'
Det -> 'the'
Det -> 'a'
N -> 'man'
N -> 'park'
N -> 'dog'
N -> 'telescope'
V -> 'ate'
V -> 'saw'
P -> 'in'
P -> 'under'
P -> 'with'

Last Operation: Expand: N -> 'man'

Picture 3. The syntactic analysis in computer linguistics

Valence refers to the capacity of verbs to combine with a specific number and type of arguments. Understanding valence is essential for parsing sentences accurately.

Valence can be categorized into three main types:

- ✓ Intransitive Verbs: Require only a subject (e.g., "He sleeps").
- ✓ Transitive Verbs: Require both a subject and an object (e.g., "She reads a book").
- ✓ Ditransitive Verbs: Require a subject, a direct object, and an indirect object (e.g., "He gave her a gift").

Valence analysis aids in:

• Argument Structure Identification. Different verbs have unique argument structures that dictate how they interact with other sentence elements. Recognizing these patterns allows for more accurate parsing.

• Semantic Role Labeling. By understanding valence, systems can identify the roles played by different constituents within a sentence (e.g., agent, patient), which is crucial for tasks like information extraction and question answering.

• Grammar Induction. Knowledge of valence contributes to developing grammar models capable of generalizing across languages and dialects, enhancing multilingual NLP systems.

Recent advancements in machine learning and deep learning have significantly improved the ability to analyse collocations and valence. Techniques such as word embedding's (e.g., Word2Vec, GloVe) capture semantic relationships between words, facilitating better identification of collocations. Additionally, neural network architectures (e.g., Transformers) have enhanced the modelling of syntactic structures by enabling context-aware representations.

Future research should focus on:

• Cross-Linguistic Studies. Investigating how collocations and valence differ across languages can inform machine translation systems and improve their accuracy.

• Integration with AI Systems. As AI becomes more integrated into everyday applications (e.g., chatbots), understanding syntactic nuances through collocation and valency analysis will be critical for enhancing user interactions.

• Exploration of Contextual Variability. Examining how context influences collocation usage and verb valency can lead to more sophisticated language models that adapt to varying linguistic environments.

Conclusion

In conclusion, syntactic analysis is integral to the field of computational linguistics, with collocations and valence serving as pivotal components in understanding sentence structure and meaning. Their relevance spans various applications within natural language processing, from improving text generation to enhancing semantic understanding. As research continues to advance, further exploration of these concepts will lead to more robust language models and improved AI systems capable of interacting with human language more naturally and effectively. The ongoing study of collocations and valance will undoubtedly contribute to the evolution of computational linguistics and its practical applications in technology and communication.

The scientific significance of the research results lies in clarifying the features of the manifestation of valence theory at the linguistic levels and demonstrating that the role of syntactic connections in determining the valence of linguistic units at the syntactic level is important. Syntactic connections in sentences can be considered as the semantics of syntactic units involved in the structure of a sentence, by highlighting the junctional models and differential syntactic features of the components and ways of expression using componential models in terms of functional syntax. As part of the analysis of simple sentences in English and Uzbek, this approach to the analysis of simple sentences helps to solve some controversial problems encountered in comparative typology and theoretical grammars.

References:

1. Асадов Р.М. Синтаксическая валентность на примере синтаксемного анализа трехвалентных элементов в позиции неядерного оппозитивного предицирующего компонента (nap2) Вестник Челябинского государственного …, 2016. С. 25-35.
2. World summit on the information society. Plan of Action. Document WSIS-03/ GENEVA/DOC/5-E. 12 December 2003

3. Рассел С., Норвиг П. Искусственный интеллект: современный подход /Artificial Intelligence: a Modern Approach / Пер. с англ. и ред. К. А. Птицына. 2-е изд. М.: Вильямс, 2006. - С.8.
4. Новое в зарубежной лингвистике. Вып. 24. Компьютерная лингвистика. М.: Прогресс, 1989. - С.10.
5. A.K. Zulpukarova INFORMATION TECHNOLOGIES IN LINGUISTICS // Международный журнал гуманитарных и естественных наук. 2024. №4-1 (91).
6. Bozorov S.M. INFORMATION COMMUNICATION TECHNOLOGY FIELD COMPUTER LINGUISTICS AND DATA MINING SYSTEM // Экономика и социум. 2022. №3-2 (94).
7. Abdullaeva Nargiza Erkinovna International classification of proverbs in computational linguistics // International scientific review. 2019. №LV.
8. Nurmuxammedova, N. (2024). The Importance of Syntactic Analysis in Computer Linguistics (in Research Examples of Word Combinations and Valency). Journal of Language Pedagogy and Innovative Applied Linguistics, 2(5), 74-77. https://doi.org/10.1997/c8eqhc44
9. Karimova M., Asadov R., Aminova N., Proficient English users' vocabulary Dictionary for advanced and proficiency English learners 2020. http://www.morebooks.shop/
10. Rustam Asadov, Shohista Mardiyeva, Syntactic Valency of Predicative Components Conference Proceedings: Fostering Your Research …, 2024 C. 159-162.