American International Journal of Social Science Research; Vol. 4, No. 2; 2019
ISSN 2576-103X E-ISSN 2576-1048
Published by Centre for Research on Islamic Banking & Finance and Business, USA

Definition of Natural Languages Using Context-free Grammer: In Context of Dari Language

Mahmood Asgharzada
Master of Science, Informatics, Technical University of Berlin
Herat University, Afghanistan
E-mail:asgharzada85@gmail.com

Abstract

Formal definition of a natural language lets computers understand it; the languages have challenging complexity to come under representation of any formal language. Modeling assumptions, using formal languages, we can approximate the languages; and better approximation we do, more complexities of formal languages are resolved. Objective is to study current researches and possibilities to model Dari as a context-free language. Researchers have worked for more than 6 decades on definition of syntax of natural languages using context-free grammer. It is crucial that computer scientists to be able to thoroughly understand about which category of formal languages the natural language exist in in. In overall, our skill to decide the grammar type of natural languages performs a significant role in our capability to parse it. This review article concludes that discussed issue of the context-freeness of a certain language is regularly reliant on upon the level of complexity of the language. The common of languages investigated in this regard are languages of European family and Indo-Aryan languages liker Dari, Arabic and Pashto have not been adequately fortunate to be devoted thorough researches in this regard.

Keywords: Context-Free, Grammer, Formal Language, Modeling.

1. Introduction

A context-free grammar is a system to define languages. It is a formal grammar to produce all possible strings in a specified formal language; and in formal language theory, it is a specific form of formal grammar: a group of production instructions which describe all probable strings in a given formal language (Klabunde, R. 2002).

Context-free grammer is a potent formalism to define syntactic structure of formal language; it lets individuals to stipulate rules which state how a component can be segmented into smaller and smaller components, up to level of individual words. Context-free grammer can be used to provide a formal explanation of syntax. Any language potential to be defined by some context-free grammer is context-free language.

"Natural language processing is mainly working with the understanding of nature (syntax) and the meaning (semantics and pragmatics) of natural languages" (Klabunde, R. 2002).

"To make a computer understand a natural language, it is first necessary to define a language formally, but natural languages are much too complex for representation by any formal language" (Sundararajan, 2007).We can approximate natural languages by formal languages simplifying assumptions; better the approximation we need, less relaxed the assumptions we can make and more complex the formal language model becomes (Klabunde, R. 2002).

Among different types of formal grammars, context-free grammar gives the best replacement between complexity and degree of approximation.

Along with multiple challenges which parsing natural languages have, Dari requires further research to develop models to be analyzed. The language is a Perso and Arabic language and its morphological system has essential linguistic forms and terminology of Arabic and Urdu language (Mukhtar, N., Khan, M. A., & Zuhra, F. T. 2011).

2. Results and Discussion

Context-free grammars have been widely used by computer scientists and linguist to describe the syntax of programming languages and natural languages. Context-free grammars are used in the Backus-Naur Form when they are used to describe the syntax of programming languages (Earley, 1968).

Defining the syntax of natural languages using Context-free grammars has been the object of interest of computer scientist and researchers for now more than 6 decades. From the very early days of computer programming, computer scientists had predicted the necessity of translating natural languages using computing power and parsing techniques. Gazdar et. al argue that it is not only desirable but absolutely crucial for human-beings to bring the command of natural languages under their own control as the computers increase in their complexity and functionality (Gazdar & Pullum, 1985).

For this purpose, computer scientists need to be able to parse a certain natural language to be able to enable computers to interact with that specific language. This is the exact point when the theory of parsing of natural languages come into our equation. For computer scientists to be able to parse a natural language, it is crucial that they pedantically understand about which class of formal languages that specific natural language resides in. It is why Gazdar et.al argue that "There can be no parsing without a grammar." (Gazdar & Pullum, 1985). For computer scientists to be able to come up with programs that can compile and interpret natural languages, it is necessary to have parsing algorithms and grammars. (Earley, 1968). Therefore, Jay Earley believes that the development of automatic parsing algorithms is highly significant. (Earley, 1968). All in all, our ability to determine the grammar type of natural languages plays an important role in our ability to parse it as well.

The Chomsky hierarchy of languages (Type 0 or recursively enumerable languages, Type 1 or context-sensitive languages, Type 2 or context-free languages, Type 3 or finite state languages) is accepted as the standard linguistic division of languages and "the question of where the natural languages are located on the 'Chomsky hierarchy' is an intrinsically interesting one" as pointed out by Pullum et. al (Pullum & Gazdar, 1982). A large number of scientists and linguists have been searching for universal limitations on the class of natural languages and in this search; they have investigated some formal linguistic properties including context-freeness (Sheiber, 1988).

Though the initial consensus among computer scientists has been the context-freeness of all natural languages, many later on have objected to the validity of this notion. Some scientists find context-freeness to be a rather restrictive and weekly class to define natural languages as Shieber (Sheiber, 1988) puts it in this way: "Soon after Chomsky's categorization of languages into his well-known hierarchy, the common conception of the context-free class of languages as a tool for describing natural languages was that it was too restrictive a class – interpreted strongly (as a way of characterizing structure sets) and even weekly (as a way of characterizing string sets)."

Some languages like English, Swiss and German have been tested against the properties of context-free grammars to see if it is possible to conclude that natural languages are a set of context-free grammars or not. Noam Chomsky, James Higginbotham and Gerald Gazdar have been the pioneering scientists in this regard. Stuart M. Shieber in his paper "Evidence Against the Context-Freeness of Natural Languages" (Sheiber, 1988) attempts to prove that categorizing natural languages as a class of context-free grammars is completely problematic and he presents some data from Swiss-German language to counter the idea and provide a proof of the "weak non-context-freeness of Swiss German" (Sheiber, 1988). Therefore, Sheiber believes that it not a hard task to prove that the set of natural languages is not included by the bigger set of context-free grammars and suggests that he proves it to be so by making "as few (and as uncontroversial) linguistic assumptions" and making no particular assumption about the structure or semantics of Swiss German. Throughout this paper, subordinate clauses have been his main object of interest to show non-context freeness of Swiss German language. In the same paper, he tackles some counterarguments against his claim in order to make sure to convince the readers that his theory is sound enough to resist the counter claims.

"English Is Not a Context-Free Language" is the title of another paper by James Higginbotham in the series of rebuttals of the idea of inclusion of the natural languages under the set of context-free grammars, though this time scrutinizing English language specifically. He began by accepting that the question of whether English is a context-free language "has for some time been regarded as an open one." (Higginbotham, 1984).

Later on in the same paper, he emphatically argues against this idea and goes on to prove his thesis using the properties of context-free grammars. Traditionally, computer scientists against the idea of inclusion of natural languages under the set of context-free grammars have used two main properties to prove their notion: pumping lemma and closure properties of context-free languages under intersection. Higginbotham in this paper uses the idea that a regular set L (this is a set that can be generated using a finite automaton) under intersection with English language produces a non-context free grammar. Therefore, given that "context-free languages are closed under intersection with regular sets, that L ∩ English is not a context-free language proves that English is not a context-free language either." (Higginbotham, 1984). His proof consists of empirical and mathematical parts and putting a lot of stress on "such that" construction and its use in English language to prove his point.

Another attempt in this regard to prove that English is not a context-free language is done by Postal in which he uses sluicing to refute that English is a context-free language (Postal, 1984). Sluicing is a type of ellipsis where expected or predicted worlds are removed from the end of the clause.

Later in the same year, a famous paper named "On Two Recent Attempts to Show that English Is Not a CFL" is published by Geoffrey K. Pullum (Pullum, 1984). He argues that proofs in this regard do not prove to have the required rigor by stating that "many of the purported demonstrations of the non-CF-ness of various languages that have appeared over past twenty-five years have been replete with both mathematical errors and empirical shortcomings." (Pullum, 1984).

In this paper, he concentrates on the attempts to show that English is not a CFL and claims they have substantial logical flaws. He goes on to say that "They fail to make their case because closer attention reveals that their empirical claims about English are incorrect." (Pullum, 1984). Throughout the paper, he discusses why sluicing clauses and such that clauses fail to prove that English is not a context-free language. He also argues that it is important for linguists to determine "at least whether English – the most studied language in the era of generative grammar and the primary focus for natural language processing efforts – is context-free or not." (Pullum, 1984). The reason for importance and necessity of such a decision is, as he puts it, "the availability of a vast fund of information about efficient parsing of CFL that could in principle be put to work on parsing English". (Pullum, 1984).

It is also important to say that Pullum et. al always have defended the idea that the question whether natural languages come under class context-free grammars or not should be an open question. What Pullum and Gazdar often do (Pullum & Gazdar, 1982) is to demonstrate that the negative answers to the question above often lack some level of logical integrity. They propose that the question should be taken up again and again since they believe "all the arguments for the negative answer that have been provided in the literature are either empirically or formally incorrect." (Pullum & Gazdar, 1982).

3. Conclusion

Therefore, we can conclude that the debated issue of the context-freeness of a particular language is often dependent upon the level of complexity of the language. It looks like the majority of languages investigated in this regard are languages of European family and Indo-Aryan languages liker Dari, Arabic and Pashto have not been lucky enough to be dedicated thorough researches in this regard.

Dari is the official language of Iran, Afghanistan and Tajikistan with more than a hundred million speakers around the globe (Shamsfard, 2011). The lack of researches in this regard have rendered Dari a "less-resourced" language and more researches about the grammar classification and text-process-ability of Dari language will convert this popular language to "a machine processible one with efficient resources and processing tools." (Shamsfard, 2011). For us to be able to come up with tools to parse the language, it is necessary to have a good grasp of distinctions of the language and challenges it might pose. Some of the challenges that Mehrnoush Shamsfard counts that potentially proposes challenges for researches in the grammar classification and parsing of Dari language might be summarized as following:

"Dari is a pro-drop language with canonical SOV word order. There are a lot of frequent exceptions in word order, which are caused by processes such as scrambling, verb preposing, postponing, dislocation, clefting, pseudoclefting and topicalization movement (Mahootian, 1977). The frequent exceptions have turned Dari to a free word order language." (Shamsfard, 2011)

Very frequently, we might have some of the words or even phrases omitted because of the semantic or syntactic properties of the language. It is also very popular to omit the subject as well due to the fact that the agreement (person and number) embedded in the world plays an important role in determining the subject. (Shamsfard, 2011)

"Dari is a derivational and generative language in which many new words may be built by concatenating words and affixes. So the possibility of facing a new world that is not available in the system's lexicon is high." (Shamsfard, 2011)

References

Berstel, J. (2013). *Transductions and context-free languages*. Springer-Verlag.

Chomsky, N., & Schützenberger, M. P. (1963). The algebraic theory of context-free languages. In *Studies in Logic and the Foundations of Mathematics* (Vol. 35, pp. 118-161). Elsevier.

Ginsburg, S. (1966). *The Mathematical Theory of Context Free Languages.[Mit Fig.]*. McGraw-Hill Book Company.

Ginsburg, S., & Greibach, S. (1965, October). Deterministic context free languages. In *6th Annual Symposium on Switching Circuit Theory and Logical Design (SWCT 1965)*(pp. 203-220). IEEE.

Klabunde, R. (2002). Daniel jurafsky/james h. martin, speech and language processing. *Zeitschrift für Sprachwissenschaft*, *21*(1), 134-135.

Knuth, D. E. (1968). Semantics of context-free languages. *Mathematical systems theory*, *2*(2), 127-145.

Mukhtar, N., Khan, M. A., & Zuhra, F. T. (2011). Probabilistic context free grammar for Urdu. *Linguistic and Literature Review*, *1*(1), 86-94.

Pullum, G. K., & Gazdar, G. (1982). Natural languages and context-free languages. *Linguistics and Philosophy*, *4*(4), 471-504.

Pullum, G. K., & Gazdar, G. (1987). Natural Languages and Context-Free Languages. In *The Formal complexity of natural language* (pp. 138-182). Springer, Dordrecht.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. In *Philosophy, Language, and Artificial Intelligence* (pp. 79-89). Springer, Dordrecht.

Shamsfard, M. (2011). Challenges and open problems in Persian text processing. *Proceedings of LTC, 11*.

Sundararajan, S. (2007). PROBABILISTIC CONTEXT-FREE GRAMMARS IN NATURAL LANGUAGE PROCESSING.

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n3. *Information and control*, *10*(2), 189-208.

Wang, Y. Y., Mahajan, M., & Huang, X. (2000, June). A unified context-free grammar and n-gram model for spoken language processing. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)* (Vol. 3, pp. 1639-1642). IEEE.

Copyrights