A Review on Urdu Language Parsing

Arslan Ali Raza[1], Asad Habib[1], Jawad Ashraf[1], Muhammad Javed[2]
[1]Kohat University of Science and Technology, Kohat, KPK, Pakistan
[2]Gomal University, Dera Ismail Khan, KPK, Pakistan

***Abstract*—Natural Language Processing is the multidisciplinary area of Artificial Intelligence, Machine Learning and Computational Linguistic for processing human language automatically. It involves understanding and processing of human language. The way through which we share our contents or feelings have always great importance in understanding and processing of language. Parsing is the most suited approach in identifying and scanning what the available sentences expressed? Parsing is the process in which syntactic structure of sentence is identified using grammatical tags. The syntactically correct sentence structure is achieved by assigning grammatical labels to its constituents using lexicon and syntactic rules. Phrase and Dependency are two main structure formalisms for parsing natural language sentences. The growing use of web 2.0 has produced novel research challenges as people from different geographical areas are using this channel and sharing contents in their native languages. Urdu is one of such free word order native language which is widely shared over social media sites but identification and summarization of Urdu sentences is challenging task. In this review paper we present an overview to recent work in parsing of fixed order (i.e. English) and free word order languages (i.e Urdu) in order to reveal the most suited method for Urdu Language Parsing. This survey explored that dependency parsing is more appropriate for Urdu and other free word order languages and parsers of English language are not useful in parsing Urdu sentence due to its morphological, syntactical and grammatical differences.**

Keywords*—Natural Language Processing; Machine Learning; Urdu Language Processing and Dependency Parsing

I. INTRODUCTION

In linguistic, human behavior can be assessed by considering three key aspects; speaking, writing and communication. The rise of machines like computer gave birth to the concept of communicating human with non-human devices. This particular concept proved as preliminary base for Natural Language Processing. Natural Language Processing is multidisciplinary area of Artificial Intelligence, Linguistic and computer science. The basic aim is to develop such system which can understand and generate natural human language. Normally, Machine Learning (ML) algorithms are used in recognizing and creating human language to achieve meaningful information. The growing use of web enabled technologies has changed the general trends in research and academia by endorsing huge availability of informative contents. Initially the extraction and interpretation of available contents was difficult but the progressive growth of Natural Language Processing presented easy and systematic solutions. In last few decades, NLP proved as an active research area by providing effective applications such as; Language Translation, Information Retrieval, Data Mining, Text Summarization, Sentiment Analysis, Speech Recognition. Cambria, E et al [1] reviewed the recent trends in natural language processing and stated that NLP evolved from the age of batch processing and punch cards to the era of social networking websites. In the era of batch processing single sentence might take 5 to 7 minutes while on the other hand novel technologies have changed the trends as millions of websites can be processed in seconds. Numbers of NLP tasks as Information Extraction, Text categorization, Named Entity Recognition, Parts of Speech Tagging, Word Sense Disambiguation and Parsing are addressed through Machine Learning algorithms [2]. In fact, NLP is backbone for Data Mining, Human Computer Interaction, Emotion Detection and Data Warehouse. Nowadays NLP is facing several challenges due to the advent of Web 2.0 and other social networking websites. Multi Linguistic text adaptation like Chinese, Japanese, Bengali, Arabic and Urdu is one of big challenge which gave birth to hundreds of other issues. Urdu is an Indo-Aryan language which now became the part of web contents. The availability of these valuable contents attracts the language engineers to utilize this data for the sake of analysis but not enough experiments have been performed for Urdu language processing due to lack of resources. The way through which we share our contents or feelings has always great importance in processing text for the sake of analysis. Parsing is the most appropriate method used in the interpretation of natural language sentences. Basically it is the process in which syntactic structure of sentence is identified using grammatical tags. The syntactically correct sentence structure is achieved by assigning grammatical labels to its constituents using lexicon and syntactic rules. Generally parsing generates a logical tree of sentence to eliminate the interpretation ambiguity as shown in Fig.1. Numerous techniques are available for English language parsing but there is lack of parsers for free word order languages i.e. Urdu & Hindi. In this article we have presented a review on parsing for English (Fixed word order) and Urdu (Free Word Order) Language. Much research has been done in NLP for English language. Limited work exists for free word order languages such as; Urdu and Hindi so the main focus of this review is to explore the most suited method for Parsing Urdu Language Sentences. The rest of the article is organized as; Section 2 provides brief literature on Past, Present and Future of Parsing, Section 3 presents URDU: A novel challenge for Natural Language Processing. Section 4 explores the basic idea of Dependency Parsing and Section 5 provides the conclusive remarks.

II. PAST, PRESENT AND FUTURE: PARSING

Parsing is mean of automatically identifying and building sentence syntactic structure. In general, grammar and lexicons

are used in the construction of parse tree, where the grammar is set of rules that govern the overall structure of any given language and lexicon is list of words along with associated tags. While on the other hand Parser is a tool which is responsible for generating parse tree. In fact it is a procedural component which remains same throughout the generation of parse tree irrespective of the language but grammar does not remain same for all languages. Parser and grammar primarily depend on formalism being used. Mainly there exist two formalisms; ***Top Down Parsing (TDP)*:** In TDP a tree is generated from root/parent to leaf/terminal node. TDP formalism can also be viewed as expansion process as tree is expanded at each step. As it is preorder parsing so it has some merits and demerits. In the absence of start node such formalism never wastes time in generating parse tree which is advantageous while the limit of top down parsing is backtracking. ***Bottom Up Parsing (BUP)*:** In BUP a tree is generated from down (leaf/terminal) to up (root/parent) node. BUP formalism can also be viewed as reduction process as tree is reduced at each step. As it is post order parsing so it has its own merits and demerits. In the absence of leaves/tokens such formalism never generates parse tree which is helpful while the limit occurs in case of absence of root node. These formalisms are also considered as directionalities. Similarly Parsers have two search strategies either it searches one branch at a time which is depth first strategy or it follows breadth first strategy in which all possible branches are searched in parallel.

A. Parsing Techniques & Applications

It is obvious that tree can be generated in one of two ways; Top-Down or Bottom-Up. These two formalisms are considered as directional strategies. Parsing techniques generally based on the criteria of directionality and non-directionality. Non- directional methods generate the parse tree with different input order based on the criteria it fits but this specific strategy can't generate parse tree if the entire input is not available in the memory. In simple, parsing techniques are divided into two classes [3] directional and non-directional. In addition, parsing techniques in directional strategy are divided into Top-Down and Bottom-Up. Initially non-directional method was first proposed in 1968 by Stephen, H Unger [4], although this method has not gained noticeable attention but it opened the doors for parsing in Natural Language Processing. Cocke, J [5], Younger, D.H [6] and Kasami, T [7] (CYK) method was developed for generating syntactic structure of sentence. It is non-directional bottom up method in which parser goes from shorter to longer string by deriving substring through non-terminals. Later on Grune, D et al [3] proposed a more optimized form of CYK parser. The birth of corpora, lexicons and treebank has changed the way of generating syntactic structure of sentence. Treebanks and corpora proved as good sources for modern parsing algorithms. Pulman, S.G [8] presented a survey on modern techniques of parsing and stated that an efficient parsing algorithm be sound, complete and robust. It is fact that one cannot generate semantically correct structure without proper knowledge of language constructs. An ideal algorithm is one which generates efficient tree with minimum computational effort. Statistical and modern lexicalized statistical methods were introduced to meet these requirements. It is probabilistic parsing in which probability value is used to remove the structural ambiguity in language. Statistical model expresses the probability as P (T|S). Here T, S and P represents parse tree, sentence and probability respectively. Similarly statistical parser utilizes PCFGs (Probabilistic Context free grammars) and corpus of hand parsed text. Penn Tree Bank is the most well-known English corpus developed in 1993 for Natural Language processing [9]. Collins, M [10] proposed a state of art system for parsing the text of Wall Street journal and presented the improved understanding of various statistical parsers by testing their performance and concluded that head-driven statistical parser is dominant among others. Vadas, D et al [11] proposed a Noun Phrase (NP) bracketing model with numbers of lexical features. It is the first large scale NP parsing experiment. They explore the difficulties in parsing Noun Phrases with Bikel, D.M [12] implementation of Collins, M [10] model. They attained 93.8% and 89.14% F-score over simple and complex task respectively. Charniak, E [13] reviewed the task of sentence syntactic structure and stated that part of speech tagging is the preliminary step in parsing. This study explores that statistical information regarding to sequence of words helps parser in the generation of syntactic parse tree. Additionally comparative results disclose the fact that statistical parsers have good performance in comparison with tree bank style parsers. Tree bank parsers have few limitations as; Labor intensive, lack of head to head information, problem of conjoined words and speed of processing. The progressive development of parsers has shifted the trends of parsing from syntactic to semantic. Semantic structure provides more inner and fine grained information about the language because it follows both domain knowledge and linguistic in order to generate semantic structure instead of syntactic. Jia, R et al [14] proposed a novel framework which utilizes prior knowledge and recombines the data to achieve more semantic information. The results show that data recombination refined the Recurrent Neural Network (RNN) model by producing promising outcomes on standard GeoQuery dataset. In last few decades Artificial Intelligence is also playing active participation in Parsing & Natural Language Processing. Bowman, S.R et al [15] proposed artificial neural network based stack augmented parser-interpreter neural network. This system achieved the fine-grained semantic information through the combination of parsing with interpretation. The system outperformed the existing methods on Stanford NLI entailment task. Liang, P et al [16] presented a framework which demonstrates the learnability of statistical parser. The issues associated to generate automatic learnability of parser are also addressed. Their study aims to provide a paradigm for the automatic development of semantic parser from data. In modern linguistic, sentence of natural language has multiple writing styles which impose big challenge to semantic parsers. Han, B.C et al [17] proposed a rewriting based semantic parser to capture the semantic information in sentences without any care of language style because if a sentence to be parsed has different structure from that of target logical form then this system rewrite the sentence into its desired form. Semantic data of WEBQUESTIONS is parsed to assess the system

performance and their system achieved satisfactory results with an average F1-Measure of 83.9%.

Parse Trees act as backbone in number of natural language processing tasks so parsing has numerous applications as *Sentence Recognition:* parsing provides different algorithms and software in the identification of sentence according to its grammatical sequence. *Information retrieval (IR)*: Parsing is the essential step in IR whereas information retrieval is the field of extracting desired information from stored data. *Sentiment Analysis:* is the problem of natural language processing in which public views, opinions and sentiments are mined in the analysis of desired entities and recognition of syntactic structure is the most necessary step in identification of desired opinions. *Summarizer:* Summary generation and compressed sentence creation needs proper implementation of syntactic structure so an effective summarization is highly dependent on the parsing strategy. *Plagiarism Detection:* Exact location of target document cannot be detected without appropriate sentence structure. *Word Sense Disambiguation*, *Machine Translation*, *Transformation & Topic Modelling* are also based on the syntactic nature of words/lexemes. This section covers the key contributions from past to present about sentence identification to automatic parser generation. In short, Wang, Y et al [18] stated that we can build an effective parser in just few hours even for new domain with zero learning examples. Although extensive research has been done for resource rich languages due to the availability of linguistic resources like corpora, wordnet, gazetteer lexicons, dictionaries and ML classifiers but at the same time unavailability of such resources is valid reason behind the lack of research for Urdu.

III. URDU: A NOVEL CHALLENGE FOR NATURAL LANGUAGE PROCESSING

The proliferation of social networks and microbloging websites permits billions of online users to publish text over these sites in their own languages. In past, many languages have been used and retrieved for multilingual and cross lingual information retrieval and data mining tasks but due to the lack of resources few languages have not examined properly. Urdu is one of widely used language over social networking websites. It is the prominent language of east with an average 300 million speakers all over the world. It is the official language of Pakistan. In recent past, Urdu Language Processing has become the hot topic of research as various NLP tasks are experimented for Urdu language as; Tokenization [19], Part of speech tagging [20], Morphology orientation [21], rule based stemmer [22, 23], Urdu grammar checker [24], word segmentation, Sentence Boundary Detection [25] Urdu text classification [26], Urdu WordNet [27], Named Entity Recognition [28], and Urdu corpus construction [29]. Instead of going to the details of other problems here our aim is to explore the recent experiments performed for Urdu language parsing. Recently few experiments are performed for parsing Urdu sentences but serious effort is required to extract more syntactic and semantic information for Urdu. Kabir, H et al [24] implemented a two pass parsing strategy for Urdu language text by applying phrase structure grammar and movement rules in order to reduce the redundancy of phrase structure grammar rules. Nivre-arc-eager algorithm and Maltparser system is used for parsing Urdu sentences with Urdu Dependency Treebank (UDT) [30]. Rizvi, S.J et al [21] proposed language oriented parsing algorithm for Urdu sentence identification. The algorithm generates syntactic structure of Urdu sentences through morphological closed word classes as verb morphemes tags, postpositions and conjunctions. This Parser is based on chunking which is achieved by applying grammar rules for performing shallow parsing to generate unambiguous syntactically correct Urdu Sentence. Mukhtar, N et al [31] offered a method for developing Urdu parse tree. Their proposed technique is based on multipath shift reduce strategy. The optimized parser is selected on the basis of probability value as numbers of stack were utilized for evaluating probability values of parse tree. Abbas, Q [32] proposed Earley algorithm based Urdu parser. This parser uses the morphological rich context free Urdu grammar. The parsing tasks are performed through syntactic and functional information accessed from Urdu Kon Treebank (UKTB). UKTB contains 1,400 tagged sentences. This study explores that parsers having high morphological information produce better results in comparison with other strategies for Urdu parsing. Previous studies validate that there is need of appropriate parsing techniques for Urdu language as it has many practical and potential applications; Urdu Language Understanding, Urdu Summarization, Urdu Plagiarism, Urdu Opinion Mining and Urdu text modelling etc.

IV. DEPENDENCY PARSING

In natural language processing, the syntactic structure of sentence can be described in two ways; Phrase Structure (PS) in which whole sentence is tokenized into constituents or phrases and a tree is generated as output shown below in fig.1 while the second way is Dependency Structure (DS) in which individual tokens are connected through links by ensuing dependency relations as shown in fig.2. Past experiments [33, 34, 35, 36] explored that phrase structure is effective for fixed order languages while dependency structure is better for free word order languages. As Urdu language is free word order so we discussed the dependency parsing, terminologies and its associated concepts to find appropriate direction for Urdu Language Parsing (ULP). In dependency parsing, each individual token has one of two labels; Head or Dependent. Head and dependent are connected through a link/relation. Dependent is also considered as modifier or child and similarly head is referred as parent or regent. Head is actually governor in dependency parsing. Mainly there exist two methods for dependency parsing, Grammar driven dependency parsing and data driven dependency parsing. In first method [37], grammar parsing algorithms are used for evaluating and analyzing the input string. Grammar driven parsing is sub classified into two types; context free grammar and constraints based grammar. Most of the grammar driven parsers are based on constraints based parsing. Nivre, J [38] presented the detailed study on data driven and grammar driven dependency parsing. Main focus of this study is to consider full parsing representation instead of Head driven and partial representation of parsing. In comparison with constituency parsing three key advantages of dependency parsing are underlined; Dependency parsing is more close to

semantic relationship, more straightforward and word at a time operation instead of waiting for phrases.

Input Sentence: Economic news had little effect on financial markets.

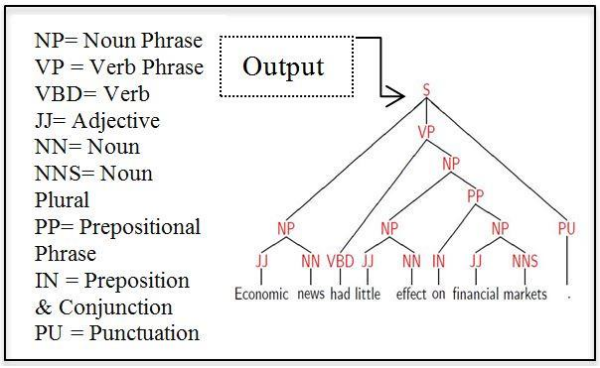

Fig. 1. Phrase Structure Parsing

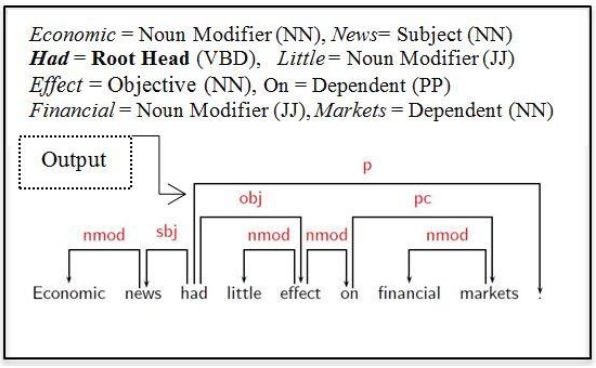

Fig. 2. Phrase Structure Parsing

Covington, M.A [39] proposed a novel algorithm for parsing English language sentences into the dependency trees. Instead of scanning and handling whole sentence collectively this algorithm deals with single word at a time on the basis of heads and dependents. This study followed six basic assumption in parsing as; *Unity:* The single structure is output, *Uniqueness*: Each word follows exactly one head, Projectivity*:* No crossing branches, *Word at a time operation*: One word is operated at time, *Left to right single pass*: No backtracking and *Eagerness*: connects the words as soon as possible. This study validates that a parser with these six assumptions produces more effective outcomes especially when direct object is missing in target sentence. For Example; *Sang Loudly*. Ali, W et al [30] proposed a data driven dependency parsing method for parsing Urdu text. Maltparser system was used to train Urdu parser using Urdu dependency treebank. UDT is annotated corpus of 2853 sentences at three different levels; POS, Chunks and Dependency Relation. Maltparser is used for performing all parsing tasks. Their results demonstrate that data driven dependency parser achieved promising results with an average labeled accuracy of 74.48%.

V. CONCLUSION

Linguistic Processing is one of prominent field of Computer Science which involves understanding and generation of human natural language to contribute in multiple real life applications; Information Retrieval, Sentence Recognition, Plagiarism Detection, Topic Modelling, Text Summarization, Sentiment Analysis and many more. Modern linguistic style and heterogeneous language adoptability produced number of challenges in NLP. Urdu is one of such language which got great attention in last few decades. We see that social sites are full of Urdu contents as millions of online users belonging from Asian countries especially Pakistan, Afghanistan and India are sharing informative contents about various interests. What they have shared is always a curious question for those who are interested in knowing these contents for the sake of analysis but there exists no such proper systems available which automatically identify and investigate the Urdu Language Text. Few essential tasks are always required in processing human languages such as Words Identification (Tokenization) Grammatical Labels (POS), Normalization (Punctuation, Stop words, Stemming & Coreference Resolution), Syntactic Structure (Sentence syntactic tree) and Semantic Information. In the light of these prerequisites, Urdu language is not handled accordingly due to resources scarcity so this language needs serious attention in comparison with others languages. Urdu language parsing is key problem which still have not been handled up to the satisfaction. Recognizing syntactic structure of sentence is one of key phase in every language processing task. Therefore, keeping in view the importance of Urdu language processing we have reviewed various experiments in parsing Urdu language text. To achieve suitable method for Urdu language parsing we reviewed parsing for both English and Urdu language and concluded that dependency parsing is more appropriate for Urdu and other free word order languages but parsers developed for languages like English are not workable for Urdu due to its morphological, syntactical and grammatical differences. We must encourage researcher community to develop algorithms using dependency parsing formalism for Urdu language sentences.

REFERENCES

[1] Cambria, E. and White, B., 2014. Jumping NLP curves: a review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), pp.48-57.

[2] Olsson, F., 2009. A literature survey of active machine learning in the context of natural language processing.

[3] Grune, D. and Jacobs, C., Parsing Techniques–A Practical Guide. 1990. VU University. Amsterdam.

[4] Unger, S.H., 1968. A global parser for context-free phrase structure grammars. Communications of the ACM, 11(4), pp.240-247.

[5] Cocke, J., 1970. Programming languages and their compilers

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

[7] Kasami, t., 1965. An efficient recognition and syntaxanalysis algorithm for context-free languages (no. Scientific-2). Hawaii univ honolulu dept of electrical engineering.

[8] Pulman, S.G., 1991. Basic Parsing Techniques: an introductory survey.

[9] Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), pp.313-330.

[10] Collins, M., 2003. Head-driven statistical models for natural language parsing. Computational linguistics, 29(4), pp.589-637.

[11] Vadas, D. and Curran, J.R., 2008, June. Parsing Noun Phrase Structure with CCG. In ACL (pp. 335-343).

[12] Bikel, D.M., 2004. On the parameter space of generative lexicalized statistical parsing models (Doctoral dissertation, University of Pennsylvania).

[13] Charniak, E., 1997. Statistical techniques for natural language parsing. AI magazine, 18(4), p.33.

[14] Jia, R. and Liang, P., 2016. Data recombination for neural semantic parsing.

[15] Bowman, S.R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C.D. and Potts, C., 2016. A fast unified model for parsing and sentence understanding.

[16] Liang, P., Jordan, M.I. and Klein, D., 2011, June. Learning dependency-based compositional semantics. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 590-599). Association for Computational Linguistics.

[17] Han, B.C.L.S.X. and An, B., Sentence rewriting for semantic parsing.

[18] Wang, Y., Berant, J. and Liang, P., 2015, July. Building a Semantic Parser Overnight. In ACL (1) (pp. 1332-1342).

[19] Durrani, N. and Hussain, S., 2010, June. Urdu word segmentation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 528-536). Association for Computational Linguistics.

[20] Anwar, W., Wang, X., Li, L. and Wang, X.L., 2007, August. A statistical based part of speech tagger for Urdu language. In Machine Learning and Cybernetics, 2007 International Conference on (Vol. 6, pp. 3418-3424). IEEE.

[21] Rizvi, S.J., Husssain, M. and Qaiser, N., 2004, December. Language oriented parsing through morphologically closed word classes in Urdu. In Engineering, Sciences and Technology, Student Conference On (pp. 19-24). IEEE.

[22] Ali, M., Khlid, S. and Saleemi, M.H., 2014. A novel stemming approach for Urdu language. J Appl Environ Biol Sci, 4(7S), pp.436-443

[23] Gupta, V., Joshi, N. and Mathur, I., 2013, September. Rule based stemmer in Urdu. In Computer and Communication Technology (ICCCT), 2013 4th International Conference on (pp. 129-132). IEEE.

[24] Kabir, H., Nayyer, S., Zaman, J. and Hussain, S., 2002, December. Two pass parsing implementation for an Urdu grammar checker. In Proceedings of IEEE international multi topic conference (pp. 1-8).

[25] Rehman, Z. and Anwar, W., 2012. A hybrid approach for urdu sentence boundary disambiguation. Int. Arab J. Inf. Technol., 9(3), pp.250-255.

[26] Ali, A.R. and Ijaz, M., 2009, December. Urdu text classification. In Proceedings of the 7th international conference on frontiers of information technology (p. 21). ACM.

[27] Adeeba, F. and Hussain, S., 2011. Experiences in building the Urdu WordNet. Asian Language Resources collocated with IJCNLP 2011, p.31.

[28] Singh U , 2012, Named entity recognition system for Urdu. In: Proceedings of international conference on Urdu, (pp. 2507–2518).

[29] Becker, D. and Riaz, K., 2002, August. A study in urdu corpus construction. In Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12 (pp. 1-5). Association for Computational Linguistics.

[30] Ali, W. and Hussain, S., 2010. Urdu dependency parser: a data-driven approach. In Proceedings of Conference on Language and Technology (CLT10), SNLP, Lahore, Pakistan.

[31] Mukhtar, N., Khan, M.A. and Zuhra, F.T., 2012. Algorithm for developing Urdu probabilistic parser. *International journal of Electrical and Computer Sciences*, *12*(3), pp.57-66.

[32] Abbas, Q., 2015. Morphologically rich Urdu grammar parsing using Earley algorithm. Natural Language Engineering, pp.1-36.

[33] Hudson, R.A., 1984. Word grammar. Oxford: Blackwell.

[34] Shieber, S.M., 1985. Evidence against the context-freeness of natural language. In The Formal complexity of natural language (pp. 320-334). Springer Netherlands.

[35] Bharati, A., Chaitanya, V., Sangal, R. and Ramakrishnamacharyulu, K.V., 1995. Natural language processing: a Paninian perspective (pp. 65-106). New Delhi: Prentice-Hall of India.

[36] Mel'čuk, I., 1988. Dependency Syntax: Theory and Practice, State University of New York Press. Arabic Generation in the Framework of the Universal Networking Language, 209.

[37] Kübler, S., McDonald, R. and Nivre, J., 2009. Dependency parsing. Synthesis Lectures on Human Language Technologies, 1(1), pp.1-127.

[38] Nivre, J., 2005. Dependency grammar and dependency parsing. MSI report, 5133(1959), pp.1-32.

[39] Covington, M.A., 2001. A fundamental algorithm for dependency parsing. In Proceedings of the 39th annual ACM southeast conference (pp. 95-10)

.