The International Journal of Soft Computing and Software Engineering [JSCSE], Vol. 3, No. 3, Special Issue:
The Proceeding of International Conference on Soft Computing and Software Engineering 2013 [SCSE'13],
San Francisco State University, CA, U.S.A., March 2013
Doi: 10.7321/jscse.v3.n3.40 e-ISSN: 2251-7545

Generating Pashto in Combinatory Categorial Grammar

Aziz-Ud-Din[1,2] , Bali Ranaivo-Malancon[2], Alvin W. Yeo[2]

aziz621@gmail.com, mbranaivo@fit.unimas.my, alvin@fit.unimas.my

[1]Department of Computer and Information Sciences, Al Jouf University, KSA

[2]Faculty of Computer Science and Information Technology, University Malaysia Sarawak, Malaysia

***Abstract*- This paper is related to the field of Natural Language Generation. NLG is a subfield of NLP, which is itself a subfield of AI. This paper describes the development of Pashto language generation system, which is in early stages of development. The system is based on the combinatory categorial grammar which is derived from categorial grammar. The system is being implemented using open source OpenCCG toolkit. The special focus of generation process is on the generation of clitics and endoclitics which would be incorporated into the final system when it is complete.**

Index Terms: **Clitics, Endoclitic, Natural Language Generation**, **Prosody, Syntax**

I. Introduction

Pashto, as a native language is spoken by about 28 million people in parts of Pakistan such as North-West Frontier Province, Federally Administered Tribal Areas, Karachi and Balochistan. It is also spoken by over 13 million people in the south, east, west and a few northern provinces of Afghanistan. Pashto is the first official language of Afghanistan. Smaller, modern "transplant" communities are also found in Sindh (Karachi, Hyderabad). Other smaller communities, peopled by Pashtun invaders in the past centuries, exist in Northern India (Pathankot, Rampur) and north eastern Iran. Pashto has been written in a variant of Persian script (which in turn is a variant of Arabic script) since the late sixteenth century (Wikipedia, 2007) [1].

Pashto is a Subject-Object-Verb (SOV) language, while in English language the word is Subject-Verb-Object (SVO). In Pashto language adjectives come before nouns. Nouns and adjectives are inflected for gender (Masculine/Feminine), number (Singular/Plural) and case (Direct/Oblique). Direct case is used for subjects and direct objects in the present tense. Oblique case is used after most pre- and post-positions as well as in the past tense as the subject of transitive verbs. There is no definite article, but instead there is extensive use of the demonstratives this/that. The verb system is very intricate with Simple Present, Subjunctive, Simple Past, Past Progressive, Present Perfect, and Past Perfect. In any of the past tenses (Simple Past, Past Progressive, Present Perfect and Past Perfect), Pashto is an ergative language, i.e. transitive verbs in any of the past tenses agree with the object of the sentences. Pashto has a pattern of split-ergativity similar to that of Hindi/Urdu, except that Pashto has generally been thought to define the split on tense, rather than on aspect. Pashto is split ergative only in the past tense.

II. Combinatory Categorial Grammars

The concept of categorial grammars was established by Kazimierz Ajdukiewicz (1935) [2], a Polish logician and philosopher. He developed a system for describing combinations of expressions which are admissible and well-formed expressions in a natural language. His work is in succession on semantic categories by his colleague Le´ sniewski and work of Husserl, Russell's theory of types Frege's mathematical analysis of language.

Ajdukiewicz's categorial grammar is a parsing mechanism that is motivated by syntactic connections between words in a sentence. The grammar uses primitive categories and a category functor, '|'. Complex categories are constructed from two atomic categories N (proper names) and S (sentences) and a recursive procedure which states that: If A and B are categories, then so is (A|B). A sentence is valid in this grammar if and only if some ordering of its word types reduces to a goal type by successive cancellations of functors and atomic categories.

Combinatory Categorial Grammar (CCG) is a derivative of Categorial Grammar (Steedman, 2006) [3], which is a highly lexicalized grammar are commonly used in NLU and also in NLG (Stede, 1996). Therefore all the syntactic information is contained in lexicon. It combines intonation structure with surface structure, and can handle long-range dependencies in relative clauses and coordination. CCG is founded on type-driven rules from categorial logics such as the Lambek calculus, Type-Logical Grammar, Linear logic and Modal Logic. CCG have been implemented in as open source project OpenCCG, and is available under open source license for research and academic purposes.

III. Pashto Text Generation

Natural language generation in Pashto language is non-existent at this time and there are no reported works in progress or completed on natural language generation in Pashto language. A Pashto language generation system is under development using Combinatory Categorial grammar and implemented in OpenCCG toolkit, by me as my doctoral research. The system focuses on Pashto language generation in general and Pashto clitic generation in particular. The high level organization of the system is illustrated in FIGURE 1.

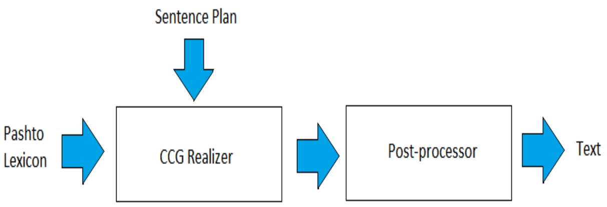


FIGURE 1

There are two inputs to the generation system. The first input is the language specific Categorial lexicon which contains lexicalized entries of most frequent word from Pashto language. The second input to the system is semantic representation of the text to be generated by the system. The semantic representation is language independent and most suited to logical processing and inferences by the computational processes. The Realizer's task is to generate sentences from semantic representation and lexicon. The output of the Realizer is fed to postprocessor whose task is to implement those rules that cannot be accommodated into realizer for theoretical reason. The output of postprocessor is natural language text suitable for human dialog.

The main reason for using CCG is that, it is lexicalized grammar which allows for efficient parsing and natural language generation, particularly surface realization. Features such as unbounded dependencies including object relative clauses and right node arising are easily accommodated in CCG. CCG supports transparent semantics which can be integrated into HLDS (hybrid logic dependency semantics).

IV. A Test Run of Pilot Pashto Grammar in OpenCCG

A small pilot lexicon containing lexical CCG grammar for Pashto has been developed. FIGURE 2 and FIGURE-3 show an example sentence parse from the system containing syntactic and semantic tags. In the later stages of the project grammar for parsing would be used for to generate sentences.

Ahmad khat lyikyee.

Ahmad letter writes

Ahamd writes a letter.

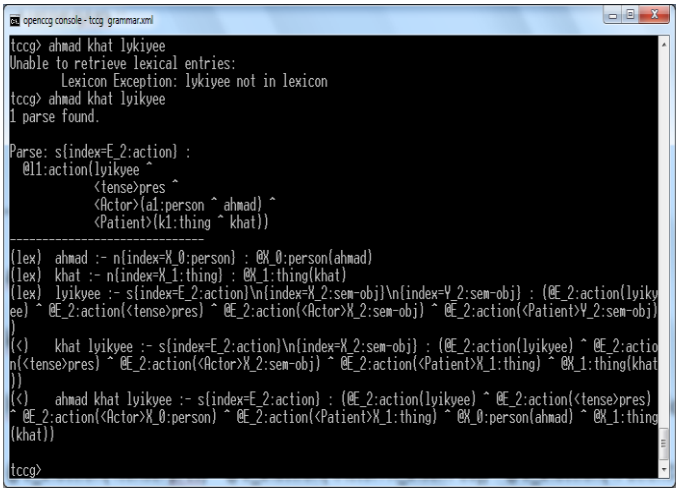
```
Parse: s{index=E_2:action} :
@l1:action(lyikyee^
                              <tense>pres ^
                    <Actor>(a1:person ^ ahmad) ^
                    <Patient>(k1:thing ^ khat))
-------------------------------------------------------------
-------------------
(lex)  ahmad :- n{index=X_0:person} : @X_0:person(ahmad)

(lex)  khat :- n{index=X_1:thing} : @X_1:thing(khat)
(lex)      lyikyee    :-s{index=E_2:action}\n{index=X_2:sem-
obj}\n{index=Y_2:sem-obj} :
                    (@E_2:action(lyikyee) ^
                    @E_2:action(<tense>pres) ^
                    @E_2:action(<Actor>X_2:sem-obj) ^
                    @E_2:action(<Patient>Y_2:sem-obj))

(<)      khatlyikyee  :-s{index=E_2:action}\n{index=X_2:sem-
obj} :
                    (@E_2:action(lyikyee) ^
                    @E_2:action(<tense>pres) ^
                    @E_2:action(<Actor>X_2:sem-obj) ^
                    @E_2:action(<Patient>X_1:thing) ^
                    @X_1:thing(khat))

(<)    ahmadkhatlyikyee :-s{index=E_2:action} :
                    (@E_2:action(lyikyee) ^
                    @E_2:action(<tense>pres)^
                    @E_2:action(<Actor>X_0:person) ^
                    @E_2:action(<Patient>X_1:thing) ^
                    @X_0:person(ahmad) ^
                    @X_1:thing(khat))
```

FIGURE 2. Example parse of Pashto sentence



```
tccg> ahmad khat lykiyee
Unable to retrieve lexical entries:
        Lexicon Exception: lykiyee not in lexicon
tccg> ahmad khat lyikyee
1 parse found.

Parse: s{index=E_2:action} :
  @l1:action(lyikyee ^
             <tense>pres ^
             <Actor>(a1:person ^ ahmad) ^
             <Patient>(k1:thing ^ khat))
------------------------------
(lex)  ahmad :- n{index=X_0:person} : @X_0:person(ahmad)
(lex)  khat :- n{index=X_1:thing} : @X_1:thing(khat)
(lex)  lyikyee :- s{index=E_2:action}\n{index=X_2:sem-obj}\n{index=Y_2:sem-obj} : (@E_2:action(lyiky
ee) ^ @E_2:action(<tense>pres) ^ @E_2:action(<Actor>X_2:sem-obj) ^ @E_2:action(<Patient>Y_2:sem-obj)
)
(<)    khat lyikyee :- s{index=E_2:action}\n{index=X_2:sem-obj} : (@E_2:action(lyikyee) ^ @E_2:actio
n(<tense>pres) ^ @E_2:action(<Actor>X_2:sem-obj) ^ @E_2:action(<Patient>X_1:thing) ^ @X_1:thing(khat
))
(<)    ahmad khat lyikyee :- s{index=E_2:action} : (@E_2:action(lyikyee) ^ @E_2:action(<tense>pres)
^ @E_2:action(<Actor>X_0:person) ^ @E_2:action(<Patient>X_1:thing) ^ @X_0:person(ahmad) ^ @X_1:thing
(khat))


tccg>
```

FIGURE 3. Screen-shot of OpenCCG parse output

V. Clitic Generation in CCG

The concept of clitics is an integral part of Pashto language. Native speakers of Pashto language use clitics extensively in daily discourses. Clitic generation is the process of incorporating clitics into computer generated natural language text. In Pashto, clitics are morphemes that are neither independent words nor affixes. Clitics follow the host word to which they are attached. In general, their placement in a phrase or a sentence is conditioned by syntactic rules of the language. Clitics occur in various positions in sentences, except in the start. Normally, a clitic occurs in the second position of the clause, i.e second position from the right of the clause (shown in the following examples mostly taken from the work of Babrakzai (Babrakzai, F. 1999) [4].

Example 1

وهي	نه	زيا ت	مې	اسلم
[wɒh i:]	[nɒ]	[zeIɒt]	[meI]	[ɒslɒm]
Beat	Not	Very	me (clitic)	Aslam

Aslam does not beat me a lot.

Example 2

وهي	نه	مې	زيا ت
[wɒh i:]	[nɒ]	[meI]	[zeIɒt]
Beat	Not	me (clitic)	Very

(He/She) does not beat me a lot)

Example 3

وهي	مې	نه
[wɒh i:]	[meI]	[nɒ]
Beat	me (clitic)	Not

(He/She) does not beat me

``In particular generation of Pashto endoclitics create several problems for the generation process, because placement of these clitics in generated text is governed by multiple linguistic levels, including phonology, prosody, and syntax (Bogel, 2010) [5]. Incorporation of endoclitics into any grammatical formalism complicates and blurs distinction among classical linguistic levels such as syntax, and prosody. The proposed Pashto language generation system would incorporate generation of endoclitics into the generation system at the successful completion of the project to asses how cleanly it can be implemented in the final version of system.

Clitic generation for the Spanish language has been studied using discontinuous grammars by Charles Grant Brown (Brown 1987) [6] in his PhD dissertation. Clitic generation has not been studied so far for the Pashto language. Even no system for Pashto language generation has been developed so far by researchers. Brown (1987) applied combines Discontinuous grammars and Government Binding theory to express free word order in rewriting rules. He used a grammar for the generation of tensed Spanish sentences with object clitics. Castel (Castel 2008) [7] used a micro-grammar *of River Plate Spanish clitics* to addresses the word order constraints underlying the combinatory potential of clitics with other clitics, and clitics with their governing verbs. Clitics are defined as functor signs that seek for arguments (verbs or other clitics) in forward direction. Spanish is different from Pashto because it does not have endoclitics. Pashto clitics and endoclitics have not been studied in combinatory categorial grammar framework in our knowledge. Later part of the project will focus onseeing how Pashto clitics with all their constraints can be generated within the CCG formalism without violating formal characteristics of the generative system.

VI. Conclusion

Numerous natural language generation systems have been developed and applied to practical problems. Some example systems include Generation of Textual Weather forecasts from graphical weather maps (FOG), Statistical data summarizer and Medical information explanation generator for patient etc. The generation mechanism can be situated into any of existing linguistic theories such as LFG, HPSG, GPSG and Categorial grammars. The Categorial Grammar formalism is chosen because of clear syntax-semantics interface particularly suited to natural language generation systems. As no theoretical and computational work has been done so far in NLG for Pashto language, therefore we have shown in this paper how to generate Pashto sentence using toolkit OpenCCG.

References

[1] Wikipedia: “The History of Pashto Language”, (30 November, 2007). Retrieved from Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Pashto_language.

[2] Ajdukiewicz, Kazimierz. (1935). *Die syntaktische Konnexität*. In Storrs McCall, editor, Polish Logic 1920–1939. Oxford University Press, Oxford, pages 207–231. Translated from StudiaPhilosophica, 1, 1–27.

[3] Steedman, M. Baldridge J. (2006) *Combinatory Categorial Grammar*, Encyclopedia of Linguistics Elsevier (pp 610).

[4] Babrakzai, F. (1999). *Topics in Pashto Syntax*.Ph.D. Dissertation, University of Hawai'i at Manoa.

[5] Bogel, T. (2010) *"Pashto Endoclitics in Parallel Architecture"* in the Proceedings of LFG10 Stanford: CSLI Publications

[6] Grant Brown, C. (1987) *Generating Spanish Clitics with constrained Discontinuous Grammars* Ph.D Dissertation Simon Fraser University

[7] Castel Victor M. (2008). *River Plate Spanish Clitic Packages: An OpenCCG Account of Order Constraints*. Revista argentina de lingüística. Mendoza: FFyL, UNCuyo, 2008.