Cliticization and Endoclitics Generation of Pashto Language

Azizud Din
aziz621@gmail.com
Department of Computer and Information Sciences, Al Jouf University, KSA
Faculty of Computer Science and Information Technology, University Malaysia Sarawak, Malaysia

***Abstract*---Pashto is one of the national languages of "Afghanistan", and the home language of Pushtuns living in the "Khyber Pakhtoonkhwa Province" of "Pakistan" and many Pushtuns living in Baluchistan. Pashto language allows pronominal clitics to be inserted into morphological words. The clitics with this property are called endoclitics. This paper describes an account of Pashto Endoclitics generation which is an early stage of generation, Cliticization rules and the unique challenge posed by these clitics to the traditional syntactic theory. Pashto endoclitics are interesting, because they cannot be completely accounted for by syntax or prosody alone, but transcend different levels of grammar framework. In a natural generation task, the problem of clitic generation has to deal with syntax, prosody, and discourse constraints.**

***Index Terms*---Clitics, Cliticization, Endoclitic, Prosody , Syntax.**

I. INTRODUCTION

Pashto is spoken by about 13 million people in the south, east and a few northern provinces of Afghanistan and over 28 million in the province of Khyber Pakhtoonkhwa, Federally Administered Tribal Areas, and Baluchistan. Smaller, modern "transplant" communities are also found in Sindh (Karachi, Hyderabad). In the linguistics literature clitics are described as morphemes that are neither independent words nor morphological affixes. Syntactically and phonologically clitics follow the host word to which they are attached. Clitics are grouped into four types: proclitics, enclitics, meesoclitics, and endoclitics. Proclitics are prefixed to host word; enclitics are suffixed to host word; and mesoclitics appear between the stem of the host word and other affixes. Endoclitics are inserted inside the host root stem by splitting the root stem into semantically deficient parts. Pashto allows all of these types of clitics to occur in sentences. Pashto has been written in a variant of the Persian script (which in turn is a variant of Arabic script) since the late sixteenth century [1].

Pashto clitics normally occurs in the second position (2P) of a clause or sentence [2], however they may occur in various other positions in sentences as well, but never occurs at the beginning of a sentence as the following examples show.

دي	ورور	**دي**	وقاص
de	*wroor*	*dee*	*waqas*
aux	brother	CLT(yours)	Waqas

Waqas is your brother.

لو ستو	دي	کتاب
lwasto	*dee*	*kitaab*
were reading	CLT	book

(You) were reading a book.

The following table gives a complete list of clitics used in Pashto language [3]. However, endoclitics generation in Pashto language occurs only with pronominal clitics *mee*, *dee* and *yee*, *am*.

Table I
Pashto clitics

Pashto Clitics	Gloss	Type
مي	**mee**	**Pronominal**
دي	**dee**	**Pronominal**
ي	**yee**	***Pronominal***
ام	**am**	**Pronominal**
مو	**mo**	**Pronominal**
به	ba	Modal
دي	de	Modal
خو	kho	Adverbial
نو	no	Adverbial
را	Ra	Oblique Pronominal
د ر	der	Oblique Pronominal
ور	*wer*	Oblique Pronominal

Pashto clitics display properties commonly attributed to *post-lexical* clitics as they are *prosodically* dependent on an adjacent prosodic element and co-occur with hosts from a limited set of syntactic categories. Tagey [4] derives the generalization that 2P Clitics appear after the *"First stress bearing"* phrasal constituent in the Pashto clause. The phrasal host must be stress-bearing and must contain at least one primary *accent*. 2P Clitics normally are not hosted by unaccented constituents. In general, it has been demonstrated in work done so far by other authors, that clitic placement in a phrase or a sentence is driven by syntactic, morphological and prosodic rules. The following example shows clitic occurring after a phrasal constituent. The unstressed material infront of the verb makes the clitic appear at the very right *edge of the phrase*.

[اغه د شلو کالو دنګه او خائسته پیغله]
$_{NP}$ [*Peeghla khaaysta aw danga kaloo shaloo da aagha*]
[Girl pretty and tall years twenty postp that]$_{NP}$

The 4th Workshop on South and Southeast Asian NLP (WSSANLP), International Joint Conference on Natural Language Processing, pages 77–82, Nagoya, Japan, 14-18 October 2013.

ولیده بیا نن دي
Wa'lida Bya nen dee
Saw again today CLT-you

You saw that twenty years old tall and pretty girl again today.

The rest of the paper is organized as follows. In section II, we describe the related works about Pashto endoclitics generation with examples. Section III reviews Syntactic and Phonological Features of Clitics. In section IV, we presented clitics placement rules. Conclusions are presented in section V.

II. PASHTO ENDOCLITICS

Pashto allows clitics to be inserted into morphological words. The clitics with this property are called endoclitics. By definition endoclitics are inserted inside a word (verb in Pashto is split by endoclitic) by splitting the word into separate nonadjacent and semantically vacuous pieces. Endoclitics may not be regarded as morphological inflections as their semantics are unrelated to the host word in most of the cases. Morphologically endoclitics violate principle of *Lexical Integrity* (which states that syntactic operations may not interfere with morphology of words) [5]. The following example from [4] shows the occurrence of an endoclitic in a Pashto sentence with imperfective verb form.

اخستل ما
Akhist-el maa
buy_{3sg} 1sg
I was buying them.(Tagey 1977:89)

Pashto is strictly a verb final language (word order in Pashto is SOV). The verb [akhist-el] appears non-finally and clitic [*mee*] occurs after it, because the clitic needs a host element if the strong pronoun ***maa*** is deleted. Sentences can thereby consist of simply a verb and a clitic.

مي اخستل
mee akhist-el
1SG $buy3_{sg}$
I was buying them.

Tagey observes that *a-initial* verbs can be split apart by clitics. Specifically, in the presence of a clitic the initial [a] of these verbs can split off from the rest of the verb root rendering the above sentence as show below. It is important to note that the part of verb appearing before the verb cannot be classified as either affix or an independent word.

خستل مي ا
Khist-el mee a
buy_{3sg} CLT(I) ??
I was buying them.

Similarly, in the perfective form of the verb, the verb [akhistal] is prefixed with [wa] perfective marker resulting the following sentential form.

اخستل وا ما
Akhist-el wa maa
buy_{3sg} PERF 1SG
I bought them.

In the above sentence, deleting the strong pronoun [maa] introduces the clitic [mee]. This is shown by the sentence below.

خستل مي ا وا
Khist-el mee a wa
buy_{3sg} CLT ?? PERF
I bought them.

For explanatory purpose another example in which a clitic introduces as endoclitic is demonstrated by the following sentences.

خستل نه ا وا مي هغه
Khist-el na a wa mee agha
buy_{3sg} not ?? PERF CLT(1sg) 3SG
I did not buy it.

ه واه و ي هغه
ah waha wa yee agha
AUX_{3SG} beat PERF CLT(3sg) 3SG
He beats him.

Clitics always maintain second position. For example, if the strong pronoun [agha] is deleted from the second sentence above, the endoclitic would still be in second position after the perfective marker [wa], resulting in a sentence in which perfective marker [wa] (suffix) is no longer attached to the verb.

ه واه ي و
a *Waha yee wa*
Aux_{3SG} best CLT(3sg) PERF
He beats him. (the pronoun agha deleted)

If the perfective marker [wa] is removed, the endoclitic is again placed in the second position, and moves to the last position in the sentence.

ي ه واه
yee a waha
CLT Aux_{3SG} beat
He was beating him.

There is another example which illustrates the insertion of clitic between perfective marker and verb.

ولوله ي ته
walwala *yee* *ta*
read it(CLT) you
You read it.

When the strong pronoun [ta] is deleted, a new sentence is generated with endoclitic as shown below.

لوله ي و
lwaala yee *wa*
read it(CLT) PERF
You read it.

Pashto verb has been identified to play important role in clitic placement. Kopris describes following five different classes of verb that have different behaviors in the presence of endoclitics [5].

1. Imperfective and Perfective verb
2. *a*-initial verb
3. Simple verb
4. Derivative verb
5. Doubly irregular verb

In Bogel's analysis, endoclitics are subject to prosodic as well as syntactic constraints [6]. Prosodically, a clitic is placed after the first item bearing lexical stress in a sentence. Pashto is classified as an argument-dropping language, which is made possible by the syntactic agreement system on verbs and nouns. The endoclitics appear after aspect-caused stressed constituents. With regard to stress, Pashto verbs fall roughly into three classes, depending on their word-internal structure [6]. Bogel defines three classes of verbs with respect to clitics and endoclitics.

Class 1 Verbs**: **Monomorphemic imperfective verbs bear stress on the last syllable; the clitic is placed after the verb. *The perfective monomorphemic verbs* take on a perfective prefix [wa] that bears the main stress and the clitic occurs after the prefix. . The following shows an example.

مي تښنوله
me *texnawala*
CLT tickle
I was tickling (her). (Tagey 1977: 86)

In the perfective aspect the [wa] marker attaches to the verb as a prefix and clitic occurs after it. In this case [wa] prefix is stressed.

تښنوله مي و
texnawala *me* *wa*
tickle CLT PERF
I tickled (her). (Tagey 1977:92)

Class 2 Verbs: (**compound prefix + root):** These verbs form the perfective by means of a stress on the first syllable of the verb. A class-2 verb is bi-morphemic and is formed by a derivational prefix and a root. Syntactically these verbs are viewed as one unit.

Class 3 Verbs: (**compound lexical item + auxiliary verb**): They are similar to class-2 verbs, but are complex predicates (light verb + adjective/adverb/noun). These verbs are also split by clitics as shown by the next two example sentences.

مي پوري وسته
mee *pore wasta*
1SG carry across(3sg,FEM,PAST)
I carried her across.

وسته مي پوري
Wasta *mee* *pore*
PERF 1SG Carry across
I carried her across.

It has been suggested by Tagey [4], that there is a separate group of *a-initial* verbs, which has nine verbs that start with vowel [a]. These verbs show a very distinct behavior with regard to optional stress in the imperfective aspect. These verbs are: [akhistal] 'to buy', [aleyal] 'to singe', [acawal] 'to throw', [agustal] 'to put on', [alwtal] 'to fly', [astawal] 'to send', [arawal] 'to turn over', [azmeyal] 'to test', and [awral] 'to hear'.

Some researchers have concluded that [a] was originally a prefix clitic [7], though [a] is no longer a recognizable prefix in Pashto. The class-2 and class-3 verbs can be thought of allowing clitic to be inserted post-lexically (at phonological level) into verb, without violating the principle of Lexical Integrity.

In the perfective tense, *a-initial* verbs take the perfective prefix [we] like all other class-1 verbs. Perfective *a-initial* verbs display vowel coalescence, a process that is assumed to take place in the lexicon. The *a-initial* verbs in class-1 undergo vowel coalescence when they are preceded by a particle ending in a vowel i.e. [we] [na] and [ma].The Pashto rule of vowel coalescence (VC) and its interaction with clitic placement was studied by Tegey [4]. The following example illustrates the vowel coalescence.

واخله ي ته
waxla *yee* *ta* **ta yee w**aa**xla*
buy$_{PERF}$ it you
You buy it.
مه اخله ي ته
maxla *yee* *ta* **ta yee m**aa**xla*

not-buy it you
Don't buy it.

نه اخلي ي ته
naxla *yee* *ta*
no-buy it you **ta yee naaxla*
Don't buy it.

The interaction of clitic and vowel coalescence is shown by the sentence below, as the clitic is inserted between vowel coalesced parts [wa] and [*staw-el*].

ستول وا مي نن
staw-el *wa* *mee* *none*
sent PERF CLT today
I sent them today.

ستول مي وا
staw -el *mee* *wa*
sent CLT PERF
I sent them.

Tegey supposed that a syntactic rule for clitic placement applied after phonological rule (vowel coalescence). According to Kassie the phonological motivation of VC is the elimination of haitus (phonological gapping) [8]. She suggests the following process for VC.
$[ə]_{particle} + [a, ɑ]_{verb} \rightarrow [ɑ]$

Kassie concludes that VC is a type of lexically restricted phonological process and only *a- for a-initial verbs* undergo VC [8]. Therefore *a-* is considered as a morphological prefix, thereby claiming that no verb stem begins with a vowel. The *a-initial* verbs are described as midway between class-1 and class-2, as they take the perfective particle, but contain a stressable prefix. Clitics never move in the syntax, but may only move in the phonology to find a host to their left by the process of *prosodic inversion*. Bogel concludes that clitics are inserted into the *morphological word* post lexically, and are subjected to prosodic constraints and stress [6]. Moreover she assumes that prosody inserts clitics post lexically after an accent-bearing element, thereby asserting that attachment to a host is *a strong prosodic constraint*.

III. Syntactic and Phonological Features of Clitics

The first detailed study of Pashto clitics was carried out by Tagey [4] in his Phd dissertation. Tagey proposed that the clitic placement was syntactic, without elaborating on the exact syntactic mechanisms that determine clitic placement. Kassiere affirmed that the Pashto clitics can be dealt with only syntax and morphology. In Tagey's analysis "clitics are placed after the first major surface constituent that bears at least one main stress". Apparently the suggestion posits that phonology interacts with syntax inorder to place clitics in correct position in sentences. In a later publication Muhammad and Babrakzai proposed that clitic placement can be treated as syntactic agreement [2]. According to Dost clitics placement within sentences and clauses is governed by constraints on syntax, prosody, lexical and sublexical levels, thereby blurring the distinction and interaction between these different levels [9].

In the analysis of Roberts, clitics are divided into two groups: one appearing in the *second position* of the clause, and another that appearing *nearer to the verb* [10]. In Robert's analysis Pashto 2P clitics identify oblique-case NPs (in ergative, accusative and genitive cases) and license null oblique-case arguments. Clitics do not intervene among conjuncts, and among the parts of any clause-initial constituent.

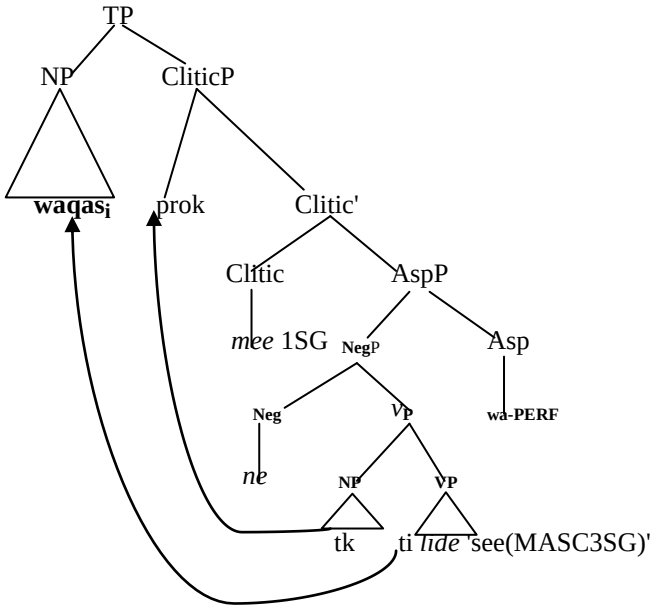
خو مي و اخستل [کتاب او کاپی]
Kho *wakhist-el* *mee* *ConjP[copy* *aw* *kitaab]*
Adv.CLT bought CLT-I notebook and book
I bought a notebook and a book but ….'

But the native speakers cannot speak it as below:

کتاب مي او کاپی واخستل
wakhist-el *copy* *aw* *mee* *kitab*
Bought notebook and CLT-1sg book
Or
کتاب او مي کاپی واخستل
wakhist-el *copy* *mee* *aw* *kitaab*
bought notebook CLT-1sg and book

The ordering of pronominal clitics within a cluster (a series of adjacent clitics) is determined by person feature *syntactically* instead of a morphological template. Clitics bear person and number features which are not unique. Possessive clitics are dislocated from overt nominal with which they are semantically associated. There is a strong relationship between strong pronouns and pronominal clitics as stated by Roberts [10]. Strong pronouns occur at the same positions as the full NPs, but discourse neutral (topic) pronouns tend to appear in the form of second position Clitics. Pashto clitics have been studied from pure phonological aspect as well [10]. Roberts attempted to incorporate Pashto clitics into Chomsky's Minimalist Program. He states that 2P pronominal clitics are agreement morphemes based on the observation (also made in [2]) that pronominal 2P Clitics are in complementary distribution with verbal agreement morphology. This leads to the prediction that only ergative and accusative arguments may be criticized, whereas nominative or absolutive arguments cannot be criticized. Each clitic heads an agreement projection, whose specifier licenses a null pronominal argument. As an example the constituent tree for the

sentence [waqas me wanalide] *'I didn't not see Waqas (a person).'* is shown below from [9].



The NP leaves a trace and overtly moves to the subject position in the sentence thereby positioning itself with respect to clitic.

Dost [9] objects to the proposal brought forward by Roberts and identifies three problems with the syntax only analysis.

1. 2P clitics should be *double overt pronominal arguments* based on the agreement hypothesis ("clitics as agreement hypothesis"), but that do not.
2. The second problem is with endoclitics (that clitic in-fixation). Tagey in his analysis, identifies three classes of main verbs in Pashto, that sometimes require 2P Clitics to appear after the first *accented syllable* of the verb, rather than after the verb itself.When stress occurs on the first syllable of the verb (in perfective aspect), 2P Clitics obligatorily occur after the stressed syllable. According to Dost, Robert's analysis does not consider clitics to be another natural word class, but rather considers them agreement morphemes and wrongly predicts the behavior of endoclitics.
3. The third problem identified by Dost is that 2P Clitics are prosodically required to have hosts. This property is lost by the syntactic treatment and EPP (Extended Projection Principle: the requirement that clauses have subjects), which is a syntactic requirement and not a prosodic one.

According to Dost the hypothesis, that 2p clitics are agreement morphemes is wrong. The suggestion made by earlier studies that clitics should co-occur with full nominal arguments, is not supported by the actual linguistic data [6]. As an alternate Dost describes a *Domain Based Approach* to clitic placement in Head-Driven Phrase Structure Grammar (HPSG).

IV. Clitic Placement

Pashto 2P clitics are unaccented, monosyllabic words that need single stressed phonological structure to their immediate left. According to existing clitic theories, Pashto clitics occur after first syntactic/prosodic structure in a sentence, which may be a word, phrase or prosodic unit. The first structure has to bear primary stress. Therefore, when placing clitics, syntax as well as prosodic constraints on phrases, words and units smaller than words have to be considered. Below Fig.1shows the sentence with strong pronoun on the top down parsing tree on the left side and with clitic on the right.

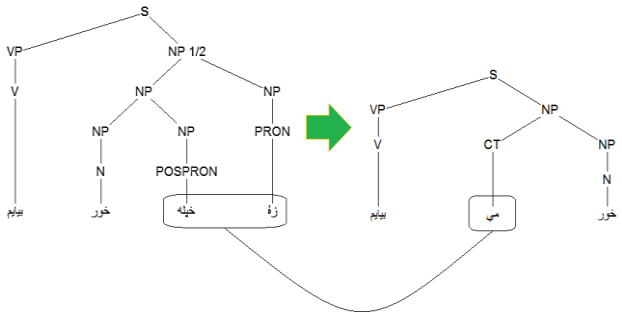


Figure 1. Cliticization

Clitics occur with wide range of syntactic hosts, and show low degree of selection with respect to hosts.

Table II specifies the rules for replacing strong pronouns with 2P (second position) clitics based on syntactic constraints only. This process is called cliticization. The abbreviations used in Table II are defined in Table III.

Table II
Rules for 2P Clitics

Rule	***Strong Pronoun***	***Syntactic Constraint***	***Replacement Clitic***
1	زه خپله	verb(sng, p1, present)	مي
2	زه خپل	dob(mas), verb(sng, p1, present)	مي
3	ما	verb(past)	مي
4	زما	verb(sng, p3, past)	مي
5	تاسو خپل	verb(past), mod(qst)	ام
6	ته خپل	dob(mas), verb(sng, p2, present)	دي
7	ته خپله	dob(fem), verb(sng, p2, present)	دي
8	تا	verb(past)	دي
9	ستا	verb(past)	دي
10	دَ هغه	verb(sng, p1, present)	ي
11	دَ هغوئ	verb(sng, p1, present)	ي
12	دَ هغي	verb(sng, p1, present)	ي

The rules specified in Table II use only syntactic knowledge of clitic placement to introduce a clitic at second position (2P)

into a sentence by replacing an already existing strong pronoun within the sentence. These rules do not include Pashto endoclitics. Pure syntactic rules such as given in Table II cannot be formulated for the generation of endoclitics in a sentence, because of afore mentioned problems. To meet up with the challenge of endoclitic generation it is necessary to allow interaction between most notable linguistic levels; syntax and prosody during the generation of Pashto language text from any knowledge represented in computational form.

TABLE III
Abbreviations used in Table II

Abbreviations	***Definition***
Dob	Direct object in the current clause
Fem	Feminine
Mas	Masculine
Sng	Singular
Pl	Plural
p1	First person
p2	Second person
p3	Third person
Present	Present tense
Past	Past tense
qst	Question

V. CONCLUSION

In this paper, I have presented few examples of Pashto endoclitics generation and some rules for cliticization process. The generation mechanism can be situated into any existing of linguistic theories such as Lexical Functional Grammar (LFG), Head-Driven Phrase Structure Grammar (HPSG) and Combinatory Categorial grammars (CCG). The Combinatory Categorial Grammar formalism will be 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 theoretical work has been shown in this paper for cliticization and endoclitics generation of Pashto sentences. Later on OpenCCG toolkit which is an open source natural language processing library will be used for Pashto clitics generation.

REFERENCES

[1] Wikipedia: "The History of Pashto Language", (30 November, 2007). Retreived From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Pashto_language.

[2] Babrakzai, F. (1999). "Topics in Pashto Syntax" , Ph.D Thesis, Linguistics Department, University of Hawaii.

[3] Din, Khan (2007), " Syntax Based De- Cliticization of Pashto text for Better Machine Translation" in the proceedings of Conference on Language and Technology (CLT07) at Bara Gali campus, University of Peshawar (August 7- 11, 2007) Page no.1.

[4] Tagey, Habibullah. (1977) "The Grammar of Clitics: Evidence from Pashto and Other Languages" PhD Dissertation University of Illinois.

[5] Kopris, A Craig & Davis Anthony R. (2005) "Endoclitics in Pashto: Implications for Lexical Integrity" Presented at the Fifth Mediterranean Morphology Meeting, Sept. 15-18, 2005, Fréjus, France.

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

[7] Anderson, S. (2005) "Aspects of the Theory of Clitics" Oxford University Press.

[8] Kaisee, Ellen. (1981) "Separating Phonology from Syntax: A Reanalysis of Pashto' Journel of Linguistics, Col. 17, No. 2 (Sep., 1981), pp. 197-208 Cambridge University Press. Dost, A. (2005) "A Domain-Based Approach to 2P Clitics in Pashto" in the Proceedings of the Texas Linguistics Society IX Conference

[9] Dost, A. (2005) "A Domain-Based Approach to 2P Clitics in Pashto" in the Proceedings of the Texas Linguistics Society IX Conference

[10] Roberts, T. (2000) "Clitics and Agreement" PhD Dissertation Massachusetts Institute of Technology