Constituent structure representation of Pashto Endoclitics

Azizud Din
Faculty of Computer Science and Information Technology
UNIMAS, Malaysia
University of Peshawar, Pakistan
aziz621@gmail.com

Bali Ranaivo-Malancon
Faculty of Computer Science and Information Technology
UNIMAS, Malaysia
mbranaivo@fit.unimas.my

M. G. Abbas Malik
Faculty of Computing and IT – North Jeddah Branch
King Abdulaziz University
Jeddah, Saudi Arabia
mgmalik@kau.edu.sa

Abstract

Pashto is one of the widely spoken languages in Pakistan, Afghanistan and by the diaspora around the world, especially in Middle East. This paper presents an initial study of Pashto endoclitics, the constituent structure representation of Pashto endoclitics and the context free grammar rules. Subsequently, this study will help in the development of clitics generation system for Pashto language. The special focus is on the development of CFG rules for the generation of endoclitics which would finally be incorporated into a Text Generation System for Pashto language.

1 Introduction

This paper examines the phenomenon of Pashto second position (endo)clitics as described by Tegey (1977). Pashto is an Eastern Iranian language spoken in parts of Afghanistan, Pakistan and by the diaspora around the world, especially in Middle East. It is the mother tongue of more than 40 million people. It is the official language of Afghanistan. In Pakistan, it is mainly spoken in the province of Khyber Pakhtunkhwa, but it also significant number of mother tongue speakers in Sindh (Karachi, Hyderabad) (Din *et. al.*, 2012).

Clitics have been defined in many ways, both phonologically and syntactically, often as semi-independent forms which attach to phrases rather than words. The technical details of different definitions are not relevant for this paper; here clitics can be described simply as a part of speech somewhere between affixes and particles, attached to hosts like affixes, yet behave at the same time as independent words, like particles.

The two most common types of clitic, found in other languages besides Pashto, are enclitic and proclitic. Enclitic is attached at the end of the host (parallel to suffixes or postpositions) whereas, proclitic is attached at the beginning (parallel to prefixes or prepositions). Pashto has several proclitics, including و (waw,perf), نه (na, not), ړا (raa, 1p),) در (der, 2p), and ور (wer, 3p) (Kopris, 2009). They are also classified as oblique pronominal clitics.

The third type of clitics is the endoclitic, which attaches itself to the middle of a word like an infix. Endoclitics are not simply inserted in a word at a grammatical boundary (in which case they would simply be affixes) but rather they can split morphemes into separate chunks (called *partials* here). Part of a morpheme may end up in one *partial* while the rest of the morpheme may end up in another, potentially separated by multiple words in between. In linguistic theory, they are generally considered to be impossible, a violation of lexical integrity (Kopris and Davis, 2005). This theoretical impossibility may explain why the only languages claimed to have endoclitics are Pashto, Udi (Harris, 2002) and Degema (Kari, 2003). The existence of Endoclitics is still a major controversial issue in today's linguistic theory. Putting aside the status of Endoclitics in linguistics, this paper focuses on the develop-

Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 74–79,
Dublin, Ireland, August 23-29 2014.

ment of Context Free Grammar (CFG) rules for endoclitic generation in natural language generation systems.

2 Context-free Grammar and C-Structure

Different models in computational linguistics have been used to process natural languages at lexical, syntactic and semantic levels. CFG is one such model that is commonly employed to represent language syntax (grammar) rules for natural languages (Jurafsky and Martin,2002). In this paper, a CFG for parsing and generating Pashto endoclitics is presented. In the contemporary linguistic approach, a sentence is an acceptable if native speakers accept the sentence as sounding good. Thus, if a majority of native people accept a sentence to be valid, then the sentence is considered good and valid. In the view of formal language theorists, the sentence is grammatical with respect to a grammar under consideration, if the grammar permits it, by generating the parse tree of the sentence. A good grammar should not only accept valid sentences but also reject invalid sentences. A grammar is also required to have fewer rules and the derived parse trees should be compact. Languages like Pashto, Urdu and Sindhi have complex syntactic structures. Formal language models like CFGs are not able to handle all the syntactic complexities for these languages. Despite these limitations, CFG can be used as an effective tool for dealing with the most sentence types in these languages (Rehman and Shah, 2003). Like Lexical Functional Grammar (LFG) (Joan, 2001), my approach for the creation of context free grammar is based on the idea of phrase structure and constituent structures. In the LFG, the phrase structures are in the form of a tree diagram, called c-structure (Joan, 2001).

3 PASHTO VERBS FORMS AND CFG PRODUCTION RULES

Pashto is an argument-dropping language (Miriam, 2007); therefore sentences may consist of only a verb and a clitic. The endoclitics mainly appear in short sentences in the context of a stress alternation that accompanies a difference in the aspect as shown in the example (1) parse tree in Figure 1 with constituent structure (c-structure) representation. Two syntactically aligned terminal nodes may share one lexical item, a representation similar to Figure 1.

وهو	مي	ټک	(1)
Waho	mee	ta'k	
Shook	I	shook-1(Perf)	

I shook it.

Figure 1. Clitics as postlexical elements

In Pashto language, the perfective aspect of the verb is accompanied by a verb-internal stress shift, placing the main stress on the first foot of the verb, while the verb in the imperfective aspect carries the main stress on the last foot of the verb. Verbs are split by clitics and each sub part of the verb is called root. With regard to the stress shift, Pashto verbs fall roughly into three classes based on their *word-internal structure*. Since these structures are essential to the correct placement of the clitic, it is necessary to analyse them more closely in order to find the appropriate (prosodic or syntactic) unit on which the clitic depends. Based on the morphological structure of different verbs, verb classes have to be defined. These verb classes and their internal characteristics and behaviour concerning the placement of clitic are described next.

3.1 Monomorphemic Verbs

Monomorphemic verbs are classified as **Class-I** verbs. Imperfective monomorphemic verbs bear stress on the last foot; the clitic is placed after the verb as shown in (2a). On the other hand, the perfective Class-I verbs take on a perfective prefix (و,wa) - which receives the main stress. In this case, the clitic occurs between the prefix and the stem, as shown in (2b).

Imperfective (2a)

مې	تېښنوله
me	*texnawala*
I	tickle

I was tickling (her).

Perfective (2b)

تېښنوله	مې	و
Texnawala	*me*	*wa*
Tickle	CLT	PERF

I tackled her

The infinitive of تېښنوله "təxnaw´ əla" is تېښنول "təxnaw´ əl". To make it masculine, morpheme و "wa" is added to the infinitive and ه "ha" is added to make it feminine in Pashto. The CFG production Rules for monomorphemic verbs are specified below:

```
S    → VP*
VP   → Perf NP*
NP   → CL V*
V    → V_sg_mas
V    → V_sg_fem
V    → V_pl
```

Each production consists of a rewrite rule. Each symbol on the left hand side of arrow (→) called non-terminals can be replaced with symbols on the right hand side of the arrow. The Kleene star (*) denotes zero or more repetitions. The Symbol S stands for sentence, NP for noun-phrase, CL for endo-clitic, VP for verb phrase and Perf for perfective marker. The verb V in Pashto is usually a derived form of the stem, called "مصدر" [maSder] in Pashto, using morphological rules. It contains information about tense, gender and number. The words or lexical items like w'ə, mee and təxnawəla are terminals. Each non-terminal must be replaced with a terminal to generate a sentence. Using bottom-up parsing technique, the phrase structure tree of sentence (2b) is shown in Figure 2.

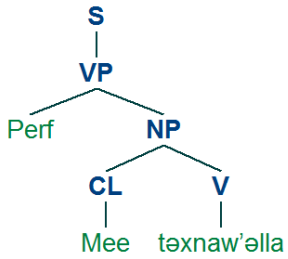

Figure 2. Phrase structure tree of sentence (2b)

The prosodic surface representation of (2b) is shown in Figure 33.

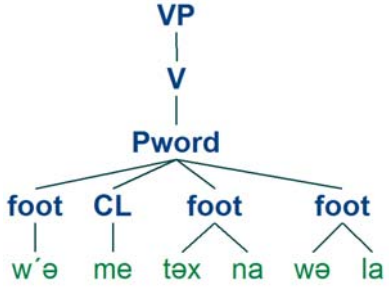

Figure 3. Prosodic surface representation of (2b)

3.2 Bimorphemic Verbs

Class-II and III are bimorphemic verbs. They form the perfective verb by means of a stress shift from the last to the first foot of the verb without adding a perfective prefix. Class-II verbs consist of a deri-

vational prefix and a root. In the bimorphemic imperfective verb, the stress is on the second foot of the verb and the clitic is placed after it as exemplified in (3a). The perfective verb is formed via a stress shift from the last to the first foot of the verb. The clitic is then placed after the first foot, i.e. after the derivational prefix as shown in (3b).

Imperfective (3a)

مي تيل وهو

mee telwah´ə

CLT-I Pushing-him

I was pushing.

Perfective (3b)

وهو مي تيل

Waho mee t'el(Perf)

Push-past I Push

I pushed it.

Class-III verbs are complex predicates consisting of an adjective, adverb or noun and a light verb. They form the largest group of verbs in Pashto. Noun-verb combinations such as obə lagawəl 'to water' (see Figure for parse tree), suč wahəl 'to guess, think', paysé lagawəl 'to deposit money', and qadám wahəl 'to walk' are frequently used in Pashto.

These noun-verb combinations are structured as compound verbs or verbal phrases and often correspond to simple verbs in English. Their behaviour with respect to clitics is the same as with the Class-II verbs, because there is also a verb internal stress shift that goes along with a change in the aspect. The clitic is positioned after the first foot in the perfective, as shown in (4), and after the whole verb in the imperfective.

Perfective (4)

کو مي پوښ

Ko mee pox'

Did I cook

I cooked (it)

Tree models are helpful in the derivation of phrase structure rules. Such a tree model is also called a phrase marker. Here forth, we will give the additional CFG rules using phrase markers.

کو مي اوبه

Ko mee obə

Did I water

I watered (it).

With Class-III verbs, one can easily identify the individual elements of a word because they are complex predicates. Thus, an analysis in favour of treating all three elements as post-lexically independent items seems likely. On the other hand with Class-II verbs, the separation of the elements is not clear, but one could argue that the derivational prefix might itself be a 'lexical word', e.g. a clitic (Stephen. 2005). Assuming that clitics are post-lexical elements that occupy separate syntactic nodes, the Class-II verb in (3b) would thus lead to a c-structure representation similar to **Error! Reference source not found.**.

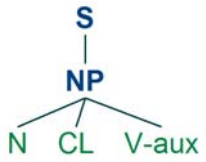

Figure 4. Phrase Stucture Tree for Noun phrase

S → NP*

NP → N CL V-aux

Figure 5. Clitics as Post-Lexical elements

Similarly, Adjective-verb combinations are structured as compound verbs or verbal phrases and often correspond to simple verbs in English. The phrase structure tree for this is given in figure 6.

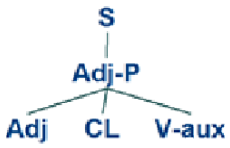

Figure 6. Phrase Stucture Tree

S → Adj-P*
Adj-P → Adj CL V-aux

However, there is a group of verbs within Class-II, which do not contain any identifiable derivational prefix, i.e., the element after which the clitic is placed as in the (5a) 'bay', does not constitute a morpheme with a separate meaning. It is therefore rather difficult to argue in favour of a clitic status of 'bay' as in Figure 1, if the morpheme is not clearly identifiable and furthermore holds a unique position within the language, i.e. it cannot be found in any other word.

Imperfective (5a)
مي بيلود
Mee bəylod´e
I Lose
I was losing (it).

Perfective (5b)
لود مي بي
Lode mee b´əy
Lose2- I -lose(1)
I lost (it).

3.3 The Special Class of A-initial Verbs

Apart from the three classes introduced above, there is a small group of verbs that can have alternating stress in the imperfective, but form the perfective with the perfective prefix of Class-I verb (wa-), thus adopting properties of all three classes. Within this group, there are verbs that begin with consonants, which do not show any special behaviour in the imperfective: even if the stress is on the front vowel, the clitic is placed after the verb.

However, there is a small number of verbs in this group with an initial vowel a- which show a very distinct behaviour with respect to the alternating stress shift in the imperfective. If the stress falls on the second foot, the clitic is placed after the verb as in (6a). If it falls on the initial vowel a-, the clitic is placed directly after the vowel as in (6b), thus acting similar to class of bimorphemic verbs.

Imperfective — stress on the second foot (6a)
مي ا غستل
Mee a˘gust´əl
I wear
I was wearing it.

Imperfective — stress on the first foot (6b)
غستل مي ا
˘gust´əl mee a
Wear-2 I wear-1
I was wearing it.

Partials are much less in use in writing and in speaking. It indicates that speakers can find a way around using them. This raises the possibility of converting sentences with partials into equivalent sentences, which do not contain partials. One way is to avoid using endoclitics altogether, and the other is to rephrase such that the verb is not split into partials. We present a solution to the above problem in 6, which is commonly used by speakers of Pashto. Table *1* shows the rearrangement based on the addition of strong pronoun هغه, at the start of the sentence. This allows the endoclitic to appear in second position without necessitating verb-splitting. Parsing of such sentence with strong pronoun at the start can be done easily.

Original	With partials	ا مي غستل I was wearing them.
Modified	Without partials	هغه مي اغستل I was wearing them.

Table 1. Solution for a'Intial words

Production for such modified sentence is given as follows.

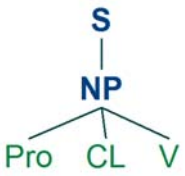

Figure 7. Phrase structure tree for modified sentence in table.1

S → NP*
NP → Pro CL V

4 Conclusion

This paper presented endoclitics in Pashto language, solution to A-initial Verbs, Constituent structure representation and CFG rules for Pashto endoclitics. However, this paper is only a first proposal for the development of CFG grammar for the grammatical phenomena of the Pashto language. Further research is necessary, especially implementation of these rules for the generation of Pashto endoclitics.

Reference

Anderson, Stephen. 2005. Aspects of the Theory of Clitics. Oxford University Press.

Bresnan, Joan (2001). Lexical Functional Syntax. Blackwell. ISBN 0-631-20973-5

Butt, Miriam. 2007. The Role of Pronominal Suffixes in Punjabi. In Annie Zaenen, Jane Simpson, Tracy HollowayKing, Jane Grimshaw, JoanMaling and ChrisManning (eds.), Architecture, Rules, and Preferences, pages 341–368, CSLI Publications.

Craig Kopris, AppTek, Inc. Endoclitics in Pashto: Can They Really Do That?

Craig A. Kopris and Anthony R. Davis. 2005. Endocliticsin Pashto: Implications for Lexical Integrity.Presented at the Fifth Mediterranean Morphology Meeting, Sept. 15-18, 2005, Fréjus, France.

Daniel Jurafsky & James H. Martin. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognit ion. Pearson Education Series 2002.

Din, A. Malancon, B and Yeo, A. 2012. Pashto Endoclitic Generation. In proceedings of IEEE conference, Pages 248-252, ICCIS Publication (Volume:1).

Rahman,M and Shah,A. Grammar Checking Model for Local Languages. In proceedings to the SCONEST (Student Conference on Engineering Sciences and Technology) 2003. SCON-S15, Hamdard & Bahria University Karachi Pakistan, October 2003.

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