Endoclitics in Pashto: Can They Really Do That?

Craig Kopris
AppTek, Inc.
6867 Elm Street, Suite 300
McLean, VA 22101, USA
kopris@apptek.com

Abstract

A cross-linguistically very rare type of clitic, the endoclitic, occurs in Pashto. Like infixes, endoclitics can be inserted inside of a word, but by splitting words apart into separate non-adjacent pieces which themselves might not have any meaning. Unlike infixes, however, endoclitics are not inflections; their meaning is unrelated to that of their host word. This paper discusses some of the problems endoclitics cause for processing Pashto, both written and spoken.

1 Introduction: What are Endoclitics?[1]

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 at the same time independent words, like particles. An English example would be the possessive *'s*. Instead of attaching to a noun referring to the possessor, it actually attaches at the end of the whole possessor noun phrase. For example, *the Queen of England's hat* places *'s* at the end of *England*, not at the end of the possessor noun *Queen*.

The two most common types of clitic across languages are enclitics, which attach at the end of their host (parallel to suffixes or postpositions), and proclitics, which attach at the beginning (parallel to prefixes or prepositions). Pashto has several proclitics, including و (PERF), نه (NEG), را (1), در (2), and ور (3), but it is the next type that is of interest here.

The third type is the endoclitic, which attaches inside a word (similar to infixes). These do not simply get inserted within 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 other words. In linguistic theory they are generally considered to be an impossibility, violating 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). Instead of a theoretical discussion of how endoclitics can exist at all, the focus here will be on practical problems of encountering them in Pashto, especially in the written language.

2 Data Sources

Data sources include online Pashto news from sources such as the BBC (www.bbc.co.uk/pashto), the VOA (www.voanews.com/Pashto), Deutsche Welle (www.dw-world.de/dw/), and Pashtun sites such as Benawa (www.benawa.com/) and Tolafghan (www.tolafghan.com/), various publications, and materials produced in-house for corpus building and linguistic analysis, totaling around 1.8 million words. Online sources are from a mixture of dialects, while in-house materials are predominantly from the Western (Kandahari, Southern)

[1] Abbreviations: 1 first person, 2 second person, 3 third person, sg singular, pl plural, FUT future, NEG negative, PERF perfective, - morpheme boundary, = clitic boundary

dialect, with substantial amounts of Eastern (Jalalabad, Northern) and to a lesser extent Southern (Khost, Central) dialects as well. Dialect differences can affect the membership and behavior of endoclitics in ways beyond the scope of this paper. For instance, for some Kandahari speakers at least the negative proclitic has some endoclitic properties.

3 Pashto Endoclitics

Pashto endoclitics are of three types: pronominal, modal and adverbial (the latter are not fully endoclitic for some Kandahari speakers).

type	*Pashto*	*meaning*
pronominal	مې	1sg
	دې	2sg
	يې	3
	مو	1pl, 2pl
modal	به	future, 'will'
	دې	'must, should, let'
adverbial	خو	'indeed, but'
	نو	'then, so'

Table 1. Pashto endoclitics

When multiple endoclitics occur, they follow a strict internal ordering (Tegey, 1977):

خو به مو مې دې يې نو

Although there are two different endoclitics with the shape دې, only one may appear at a time. The type of the endoclitic has no bearing on its ordering. The two adverbials are at opposite ends of the list, and the two modals are interspersed among the pronominals.

4 Second Position

Pashto endoclitics prefer to be in second position in a sentence, with the caveat that "second position" may be defined in various ways. There are four different classes of verb that behave differently in the presence of endoclitics, especially in distinguishing imperfective and perfective forms: simple, derivative, A-initial, and doubly irregular.

Simple verbs such as وهل 'beat' distinguish perfective from imperfective forms by the addition of the perfective proclitic, و.

If a non-endoclitic pronoun like هغه '3sg' comes first, the endoclitic (here مې) will follow immediately:

هغه مې وواهه				
ه	واه-	و=	مې	هغه
3sg	-beat	=PERF	1sg	3sg

Table 2. *I beat him*

Note that the verb وواهه is contiguous, and parsing is straightforward. If the simple pronoun هغه is removed, the endoclitic must still be in second position. To accomplish this, it is inserted between the (stress-bearing) perfective proclitic و and the rest of the verb.

و مې واهه			
ه	واه-	مې	و
3sg	-beat	1sg	PERF

Table 3. *I beat him*

Note that now the perfective marker و is no longer attached to the verb, although the rest of the verb is still contiguous, and easily parsable. If even the perfective marker و is removed, resulting in imperfective aspect, the endoclitic will still be in second position. This time, the basic syntax rule that verbs are final will be violated, and the endoclitic will be last.

واهه مې		
مې	ه	واه-
1sg	3sg	-beat

Table 4. *I was beating him*

Although the unusual word order needs to be addressed, the verb is still contiguous and readily parsable.

Derivative verbs (Tegey and Robson, 1996) incorporate a noun or adjective into an auxiliary in the imperfective, but split them apart in the perfective, creating a type of splitting verb.

ما وراناوه			
ه	او-	وران-	ما
3sg	-do	-worse	1sg

Table 5. *I was making it worse*

In table 5, the imperfective ofورانول 'make worse' incorporates the adjective وران 'worse' into a shortened form of the auxiliary کول 'do', resulting in وراناوه.

ما وران کړ		
کړ	وران	ما
do.PERF.3sg	worse	1sg

Table 6. *I made it worse*

In the perfective however, as in table 6, the adjective وران is separated and there is a full auxiliary کړ. Unlike simple verbs, there is also no perfective و.

If the 1sg endoclitic مې is used in place of the corresponding simple pronoun ما, the state of incorporation due to the aspect is preserved.

وراناوه مې			
مې	ه	او-	وران-
1sg	3sg	-do	-worse

Table 7. *I was making it worse*, endoclitic

وران مې کړ		
کړ	مې	وران
do.PERF.3sg	1sg	worse

Table 8. *I made it worse*, endoclitic

In the imperfective (table 7), the endoclitic takes second position after the verb (which incorporates the adjective), violating basic word order, while in the perfective (table 8) the endoclitic appears after the non-incorporated adjective. In terms of parsing, derivative verbs pose no particular problems, as long as incorporation in the imperfective can be handled.

A-initial verbs (Tegey, 1977) are also a type of splitting verb, but not in a semantically or morphologically natural manner. In the presence of an endoclitic, the initial ا of these verbs can split off from the rest of the root. As with simple verbs, A-initial verbs also take و in the perfective.

ما اخیستل	
اخیستل	ما
buy.3sg	1sg

Table 9. *I was buying them*

ما واخیستل		
اخیستل	و=	ما
buy.3sg	PERF	1sg

Table 10. *I bought them*

Note in tables 9 and 10 that the imperfective and perfective forms are parallel to those of simple verbs. However, when an endoclitic is added, unexpected changes occur.

اخیستل مې	
مې	اخیستل
1sg	buy.3sg

Table 11. *I was buying them*, endoclitic final

ا مې خیستل		
خیستل	مې	ا
buy?.3sg	1sg	?

Table 12. *I was buying them*, endoclitic medial

وا مې خیستل			
خیستل	مې	ا	و=
buy?.3sg	1sg	?	PERF

Table 13. *I bought them*, endoclitic

The underlining in tables 11 through 13 indicates the stressed syllable. In the imperfective, if the final syllable of the verb is stressed, the endoclitic assumes second position after the verb (table 11). However, if the first syllable is stressed, the endoclitic again appears after it, but by forcing that syllable to separate from the rest of the verb (table 12). The initial ا, which is not a meaningful prefix, stands on its own. This causes problems for parsing, in that two meaningless strings from different positions in the sentence must be identified as parts of one whole. In the perfective (table 13), the marker و pulls the ا so that both form a new single initial string, وا. This pull even occurs when an endoclitic can appear second without causing a split.

هغه مې وا نه خیستل					
خیستل	نه	ا	و=	مې	هغه
buy?.3sg	NEG	?	PERF	1sg	3sg

Table 14. Further A-initial split

Although in table 14 there is a simple pronoun in first position, allowing the endoclitic مې to be second without affecting the verb, the ا of the verb is still pulled away from the rest of the verb to attach to the perfective proclitic, leaving the negative proclitic to intervene. There is an additional change in pronunciation, in that the vowel of the perfective [ə] and the vowel of the verb [a] merge into a new vowel [α]. Parsing written text is not affected by the pronunciation change, but speech recognition is.

Doubly irregular verbs, as called by Tegey and Robson (1996), are like derivative verbs in that they do not take و in the perfective, and like A-initial verbs in that the first part of a root can be split off (even though not ا). Unlike the other categories, these verbs use a stress shift to indicate perfective aspect.

Compare the verb بوزې 'you take' in table 15 (infinitive بيول) with the sentence بو به مې نه زې 'you won't take me' in table 16.

بوزې	
ې	بوز-
2sg	-take

Table 15. *you take*

بو به مې نه زې					
ې	ز-	نه	مې	به	بو
2sg	-take?	NEG	1sg	FUT	take?

Table 16. *you won't take me*

Note that the root بوز- 'take' is split into two separate partials, not at a morpheme boundary but at a syllable boundary. Not only are they split apart, but three other words occur between them, the endoclitics به and مې, and the negative proclitic نه. It is especially important to indicate that the بو partial has no meaning of its own, nor does the remaining ز of the root. This makes Pashto endoclitics distinct from potentially similar phenomena from better known languages, such as English verb particles (*look at*) or German separable prefixes (*anschauen*). Of course, it also renders parsing of the verb difficult.

Although often the tokens intervening between the partials are only a small set of particles, in extreme cases an entire clause can be wrapped between two partials, as in څه به يې ستا پلار هم نه ملوي 'even your father won't pin him', where more than an entire noun phrase intervenes.

څه به يې ستا پلار هم نه ملوي			
ي	ملو	به يې ستا پلار هم نه	څه
3sg	-pin?		pin?

Table 17. *Even your father won't pin him*

Between the two partials of the verb, څه and ملوي, come a pair of endoclitics, future به and 3rd person يې (which cause the split into partials), the noun phrase ستا پلار 'your father', emphatic هم, and the negative proclitic نه.

5 Tokenization and Segmentation

From one perspective, tokenization (finding sentence and word boundaries) is not affected by the presence of endoclitics. They normally are set off by white space in writing, and so are easily identified as individual strings, with the caveat that due to the nature of Pashto script in using both connecting and non-connecting letters, endoclitics ending in a non-connecting و may be written without a space character, relying on the reader to see the non-connection as space. From another perspective, however, they are more difficult in that they create problems for segmentation (morphology, finding roots) by the creation of non-word strings (partials). Using lexical look-up to determine if a string is a word will fail because the word partials created by endoclitic insertion will not normally be in the lexicon, and those found in the lexicon will be homographs. The بو of table 16 is a homograph of a female name, and the څه of table 17 is a homograph of a word meaning 'some'.

Simply applying morphology is not effective because a word is split into separate words, rather than affixes being added. Segmenting زې from table 16 might find a substring corresponding to the second person singular suffix, ې, but the remaining ز cannot be used for finding the verb بيول in the lexicon (despite the morphology operations already required to recognize irregular بوز- as بيول).

Treating the partials as a simple compound, like English *blackbird*, is also not effective, since the partials have no meaning to be compounded, in addition to the same morphology problems as before.

Another problem sometimes appears due to the nature of the Pashto writing system. Since it is a variant of Arabic script, many vowels are unwritten, especially word internally. At the ends of words, where suffixes for person, number, gender, case, tense and aspect are found, attempts are made to indicate otherwise unwritten vowels. When an endoclitic splits a word, it is possible that a vowel which is unwritten in the whole word becomes written at the end of the first partial. Compare these two variants of the doubly irregular (stressshifting) verb څملول 'knock down', one with an endoclitic and one without:

without endoclitic	ما څِملول
with endoclitic	څه مې ملول

Table 18. *I knocked them down*

Note that in the first example there is no vowel indicated between the consonants څ and م. However, in the second, where the endoclitic مې has split the verb into two partials (after the stressed syllable), the first partial now ends in the vowel letter ه. Whether treated as simple morphology or as compounding, the extra letter in the partial must be taken into account. Fortunately, that letter is usually (perhaps always) ه. Of course, since this change only applies to writing, speech recognition would not need to address the "new" vowel.

Segmenting partials as unfound strings can be successful, as long as there are methods in the following parsing stage to recover the words that have been split.

6 How to Parse Them?

Assuming a satisfactory stage of tokenization and segmentation, one possible approach to parsing the verb partials resulting from Pashto endoclitics is to treat them as discontinuous strings. Reuniting the partials while undoing potential spelling changes is straightforward, as long as the partials can be identified as such. Section 5 suggested that morphology alone will fail, and that is because it cannot deal with multiple word tokens at one time. However, if partials can be identified as such, rather than say as unknown proper names, then there is the opportunity to put them back together.

The problem then is how to identify what to put together? How to know that unfound strings are partials rather than other unknowns? The key is the occurrence of one or more endoclitics. If no endoclitics are found, than unfound strings cannot be partials, and must be treated in the normal manner (e.g. as proper names). If endoclitics are found, then unfound strings have the potential to be partials, especially if one of the unfounds is at the end of the sentence and the other unfound is before the endoclitics. The likelihood is increased if the unfound preceding the endoclitic(s) is short, particularly only a single syllable. Short recognized strings preceding the endoclitics might also be in fact partials, with only homographs recognized, if the string in the verb position is unfound.

If unfounds fulfill these requirements, then they can be tested as partials. If there are two unfounds, they need to be merged in order and then tested with standard morphological processing, including testing both a string with all characters of the partials and a string with the last ه of the first partial dropped. If only the final word is unfound, then it needs to be tested the same way, but with an otherwise recognized string positioned before the endoclitics as the potential first partial.

Returning to table 16, the string زې will be unfound, while the string بو will be recognized as a proper name, 'Bow'. Between them, the parser will recognize two endoclitics, به and مې. The existence of the endoclitics and a final unfound string can then trigger the merging of that unfound زې with the short string بو preceding the endoclitics (even though already recognized as a name). Applying morphology to the resulting بوزې allows the unfound to be segmented as an inflected verb. If transitivity information is included in the lexical entry, the resulting sentence will be syntactically sound as the removal of the proper name will reduce the number of arguments to two, matching the transitivity of the verb.

Fortunately for text analysis, and unfortunately for speech recognition, creating partials through the use of endoclitics is more common in spoken than written Pashto. Formal written text has a low frequency of partials, while speech has a higher frequency. On the other hand, speech recognition does not need to address spelling changes of certain partials, except in so far as transcribing them directly.

The fact that partials are less frequent in writing means that speakers can find ways around using them. This raises another possibility, that of converting sentences with partials into equivalents without. One way is to avoid using endoclitics, and the other is rephrasing such that the verb is not split into partials.

with endoclitic	ا مې خيستل
without endoclitic	ما اخيستل

Table 19. *I was buying them*

In table 19, the 1sg endoclitic مې is replaced by the simple 1sg pronoun ما. Because there is no endoclitic, there are no partials, and the verb becomes contiguous. Where the first example has contiguous. Where the first example has two partials, ا and خيستل, the second just has the complete verb اخيستل.

with partials	ا مي خيستل
without partials	هغه مي اخيستل

Table 20. *I was buying them*

Table 20 conversely shows a rearrangement due to the addition of another pronoun, هغه, initially. This allows the endoclitic to appear in second position without needing to split the verb.

Although these methods are the ones presumably used by speakers in avoiding generating partials, attempting to use them in parsing existing sentences runs into the same basic problem as before: how to identify partials and merge them back together. Rearrangement or alternate choice of pronoun in an existing sentence does not touch the partials in written text, only the minds of the speakers.

7 Conclusion

Endoclitics are cross-linguistically an exceedingly rare phenomenon, but they exist in Pashto and when encountered must still be parsed.

Although no single specific solution has been provided in this paper, various workable approaches have been presented involving recognizing unfound strings (especially single syllables) in the presence of endoclitics as potential partials, allowing them to be remerged for lexical lookup. As endoclitics exist on the boundary of morphology and syntax, the parsing of endoclitics must also involve both morphology and syntax.

References

Alice C. Harris. 2002. *Endoclitics and the Origins of Udi Morphosyntax*. Oxford University Press, Oxford.

Ethelbert Emmanuel Kari. 2003. *Clitics in Degema: A Meeting Point of Phonology, Morphology and Syntax*. Research Institute for Languages and Cultures of Asia and Africa, Tokyo.

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

Habibullah Tegey. 1977. *The Grammar of Clitics: Evidence from Pashto and Other Languages.* International Center for Pashto Studies, Kabul.

Habibullah Tegey and Barbara Robson. 1996. *A Reference Grammar of Pashto*. Center for Applied Linguistics, Washington, DC.