Chapter 20

Determining word boundaries in *afaan Oromoo* (Oromic)

Tamam Youssouf
University of Toronto

Oromic, the language of the Oromo people, a Cushitic language spoken mainly in Ethiopia, has officially been using a Latin script-based orthography since 1991. The orthography is widely accepted by Oromic speakers. With the expansion of the application of this orthography to more than literacy and elementary education, the necessity for some updates on the rules has been felt at times. One such required update is the rules of determining word boundaries encompassing affixes, clitics, and words. This issue is very important in Oromic to minimize ambiguity and manage the length of the words. This paper proposes which morphemes should be affixed and which ones should be written separately, based mainly on the criteria proposed by Kutsch Lojenga (2014), and explains how this helps disambiguation.

1 Introduction

The Oromo are a Cushitic people who mainly live in the Oromia region of Ethiopia, constituting 35.8% of the total 2021 estimated population of 110.9 million (CIA, World Fact Book, Africa: Ethiopia). The Oromo also live in parts of Kenya, Somalia and Tanzania (see Janko 2007; Clamons et al. 1999). Their language is considered the third most widely spoken Afro-Asiatic language in the world after Arabic and Hausa, the second most widely used indigenous mother tongue in Africa, and a lingua franca for many groups in Ethiopia (Gragg 1982: viii). Despite its widespread use in speech, the language did not have an official writing system until 1991, when the new Ethiopian Peoples Revolutionary Democratic Front (EPRDF) government of Ethiopia at the time approved the use of the current Latin script-based orthography, referred to as *Qubee*.

Tamam Youssouf. 2024. Determining word boundaries in *afaan Oromoo* (Oromic). In Yaqian Huang, Nina Hagen Kaldhol, Jun Jie Lim, Sharon Rose & Anthony Struthers-Young (eds.), *ACAL in SoCAL: Selected papers from the 53rd Annual Conference on African Linguistics*, 517–538. Berlin: Language Science Press. DOI: 10.5281/zenodo.11210480

The orthography is based on the work of a committee formed in the neighbouring country of Somalia during the end of the 1970s. This committee agreed on basic conventions but left out the rest of the rules, apparently assuming that the English language conventions were to be followed. As time passed with no systematic follow up on the work of this committee, the need to introduce some additional conventions became apparent. One such convention is the determination of word boundaries.

This paper presents the challenges that the lack of rules on the word boundaries posed, and suggests how to overcome these challenges using criteria developed by Kutsch Lojenga (2014). As the language has not had a standard version yet,[1] this work is largely based on Eastern Oromia (Oromoland), the author's native dialect.

2 What is a word?

Before discussing word boundaries, we need to define the term "word" itself, which is a morphological concept that is difficult to precisely define.[2] However, for the purpose of this paper, Lojenga's (2014) definitions will be used. According to her, a *word* is a unit which 1) has meaning all by itself; 2) is able to move around and be put in different places in a sentence; and 3) can be split from other words by another independent word, such as the two words in *the book* being separated by another word *big* as in *the big book* (Kutsch Lojenga 2014: 137, see also Eaton & Schroeder 2012 and Sinha & Talukdar 2013). This definition distinguishes words from other morphemes which are dependent and have to remain bound.

In writing, word boundary placement relates to the attempt to locate the appropriate place to separate units of a language by a space. It is uncommon to see written languages not have units separated by spaces. However, the determination of these units is not given enough attention. Detailed rules are not set in the development of orthographies. The importance of word boundary placement may not have been foreseen.

Generally, the words that we see separated by spaces contain more than one morpheme. However, whether all morphemes that we see conjoined should be written that way or not has to be determined. The problem is that there are no universal criteria applicable to all languages detailing when to write certain morphemes or group of morphemes conjunctively or disjunctively and why. Therefore, one of the primary issues to be addressed in this paper is establishing these

[1] A committee was formed in 1992 to standardize the language but did not last long.

[2] For detailed argumentation regarding the indeterminacy of word segmentation, please see Haspelmath (2011).

criteria regarding the Oromo orthography, with the intention of minimising ambiguity.

The other motive to address criteria for identifying word boundaries is to check word length. Most Oromic[3] words, especially verbs and nominals, are complex word forms (i.e., they consist of more than one morpheme). The Oromic base is what Haspelmath & Sims (2010: 21) call a "bound stem" that cannot also function as a word form. Hence there is an inbuilt tendency towards longer word forms. Oromic is an agglutinating language and most of the words contain more than one morpheme.[4]

As observed a long time ago by Tutschek (1845), all primitive verbs

> have the property of producing, by affixing syllables, new verbs which are different modifications of the primitive signification of the radical verb. The number of members belonging to each of the verbal families so produced is, however, extremely various, and depends on the nature of the radical verb; whence it arises that in some verbs, singular forms are altogether wanting in the series: others are confined to only a few branches: and others again are capable of being extended to the sixth and even to the eighth link of the chain (Tutschek 1845: 10-11)

This number of links in the chain is further multiplied because as Tucho et al. (1996: 155) observed, a transitive verb root can have as many as four forms derived from it, namely active, passive, autobenefactive, and causative.

3 The criteria for identifying a word

Let us start by summarising Lojenga's (2014: 91-99) set of criteria, which is the primary reference in this paper for word boundary placement. She identifies three types of morphemes: independent word, a clitic and a bound morpheme. She also provides three sets of criteria for identifying words, which are 1) syntactic, 2) phonological, and 3) semantic.

[3]I prefer to use *Oromo* for the people and *Oromic* for the language following Arab — Arabic and Amhara — Amharic respectively, both of which are Afro-Asiatic.

[4]Words like the following (22 letters, 7 syllables, and 8 morphemes) are common:

(i) *gu~gurgur-sii-f-ách-uu-dhaa-f*
RED~sell-CAUS-DAT-ABEN-INF-0-GEN
'in order to have sb sell sth piece by piece for one's own benefit'

The set of *syntactic* criteria are a) mobility: the ability for a morpheme to be in different places in a sentence, b) separability: the morpheme can separate from the neighbouring lexical morpheme by the insertion of another morpheme, and c) substitutability: a grammatical morpheme can substitute a lexeme in the same syntactic slot (paradigmatic substitution).

The set of *phonological* criteria include a) pronounceability in isolation, b) phonological unity: influence from the lexical morpheme to grammatical morphemes and/or vice versa by phonological processes involving tone, vowel harmony, and the like; and c) phonological bridging: how to divide words in a phrase or sentence when in speech everything is blended.

The final set of criteria Lojenga identifies is the *semantic* one, which include a) referential independence: whether the morpheme can communicate meaning in isolation, b) conceptual unity: if words acquire a new meaning when placed together, and c) minimal ambiguities: if writing disjunctively or conjunctively helps disambiguate.

Applying these criteria to the issue of Oromic word boundaries that Eaton & Schroeder (2012: 229) call *word break placement decisions*, we find that, in the current orthography, there are morphemes which sometimes are written as an independent word, and sometimes written together with the following or preceding word. Such examples are not regulated or consistent even within the same piece of literature by the same author. There is no explanation for writing the morphemes adjoined or disjoined. However, this act of joining or disjoining at will is not free of implications for the language.

As Lojenga explains, "(w)hen it comes to word recognition in the reading process, it is easier to learn to recognize a word that has a constant visual image, *especially at the beginning of the word*" (Kutsch Lojenga 2014: 84; emphasis added). Thus, the use of prefixes can make the visual image of a word unstable, slowing down the reading process in a language like Oromic, where most of the prefixes could be written as separate words (I return to this issue below). Furthermore, scholars like Gasser (2012) have observed that Oromic is predominantly a suffixing language instead of prefixing and/or infixing one. Avoiding prefixing thus yields some advantages in the writing system as it may facilitate word recognition in reading. Accordingly, arbitrary prefixing of a morpheme can be a burden to the process of word recognition.

Besides, listing prefixed words in the dictionary and locating them can pose problems. Will the word be inserted only in its prefixed form or in both? If we insert the word with the prefix, the prefix has to go in with all the possible words to which it is prefixed. If we do not include the prefix in the dictionary entry, there is no good reason to prefix it in writing either.

The next issue is identifying those morphemes that can make sense either standing alone or joined. This identification applies mainly to clitics, encompassing most of the prefixed and some suffixed Oromic morphemes. This group is different from the inflectional or derivational morphemes that have to be suffixed.

Regarding this difference between affixes and clitics, Zwicky & Pullum (1983) write:

> The primary difference between them [affixes and clitics] is that word-clitic combinations are governed mainly by syntactic considerations. In contrast, the conditions governing the combining of stems with affixes are morphological and/or lexical, concerned with the substructure of a finite set of words. (Zwicky & Pullum 1983: 503)

However heterogeneous these clitics are, our Oromic examples overwhelmingly suggest that all or most of the morphemes that pose a challenge, as to whether to write them as bound or independent morphemes, are clitics. Their heterogeneity may contribute to the lack of spontaneous or intuitive consensus in writing them either adjoined or disjoined. Below I list the most salient morphemes around which there is disagreement on how to associate them with the adjacent words.

4 Morphemes sometimes conjoined as prefixes

Table 1 presents a list of common morphemes sometimes written as prefixes. When dialectal differences occurs, I begin with the Eastern, separating the others by the forward-slash (/). The symbol is also used to separate gender if represented by different forms, beginning with the unmarked masculine followed by the feminine.[5]

The primary justification for prefixing these morphemes is that they are phonologically a part of the following word, or when we pronounce them, we pronounce the two words without a pause. Additionally, it is argued that since the motto of the Oromic orthography is "write as you pronounce", we have to conjoin

[5]It is necessary to use IPA because of discrepancies in the choice of the letters and digraphs, such as <c, q, x, ph> and the unregulated case of digraph gemination. I will use *italics* for *Qubee* and square brackets for IPA. Geminates will be written as CC instead of C:. To save space, I skip using the *Qubee* version in the tables, and only use the IPA conventions there.

[6]Exclusively used in the East before 1991. In the other parts of Oromia *hín* or *ín* was used, distinguished from the negative only by the high tone.

Table 1: Some common morphemes sometimes used as prefixes.

Morpheme	Gloss	Category	Added on	Sample
[hin]	not	negation	_VP	hin ɗuf-ín. not come-NEG 'don't come.'
[if]/[of]/[uf]	self	reflexivity	_VP	if múr-é. self cut-3SG.PST '(he/I) cut (my/him)self.'
[ní][6]	FOC	focus	_VP	ní de:m-a. FOC go-3SG.IPFV 'he goes.'
[wal]/[wol]	RECP	reciprocity	_VP	wal fit'-án. RECP finish-3PL.PST '(they) killed each other.'

them. However, the motto seems to stress the contrast with the English spelling, where there are more graphemes in a word than the represented sounds. The syntactic and semantic criteria of creating ambiguity and other aspects were not anticipated. Besides, it is sometimes difficult to find a pause between a series of what are separate words in a phrase. This argument of pause does not even satisfy the phonological criterion to adjoin.

The other argument is the tonal distinction that a native speaker applies when pronouncing such complex words, which has its disadvantages. To use the tonal difference in reading, one must see the context, which involves looking at the following word before reading the word with an appropriate tone, which slows down the speed. Again, the orthography does not mark tone, which is a crucial component of the language. As it stands now, it is easier to use space than to introduce tone marking, which involves adding a new element to the orthography.

The tendency to adjoin the above morphemes as prefixes is more predominant (although inconsistent) in literature published inside Oromia, Ethiopia. It is not uncommon to find the same morphemes sometimes written disjunctively and sometimes conjunctively, even within the same document. In the literature produced abroad, these clitics are generally written disjunctively. The Oromo Liberation Front (OLF) and affiliated bodies were the main producers of literature in Qubee. All literature produced by this body uses the disjunctive format. One possible cause of the predominance of affixing the clitics in Oromia is the

influence of the Ethiopic abugida writing system, used by Amharic, the official language.

In Ethiopic script-written languages, many monosyllabic words, especially adpositions, are written conjunctively to the grammatical words they precede or follow. As Griefenow-Mewis (2001: 57) observed, the Ethiopian writing system does not allow writing a one-syllable word separately. From the abugida (Rogers 2005: 205) nature of the system, one syllable consists of one grapheme unless it is a closed syllable, which is rare. Since most of the users of the current orthography for Oromic were originally trained in Amharic, it is hard to deny the influence.

If we assume the motto "writing as we speak" and the Ethiopic abugida as possible factors behind conjoining the morphemes in Table 1, we may also need to present the case for writing them disjunctively. Let us see if the Kutsch Lojenga (2014) criteria support the idea of writing the morphemes in Table 1 as independent words, taking them as examples.

The first sample morpheme is the negative marker *hin* '(do) not'. The form is *hin* + V_{root} + AGR. We have the conjugation in Table 2 taking an imperfect form for *deem-* 'go'. The suffix after the hyphen marks the agreement.

Table 2: Non-past conjugation of *deem-* 'go' with *hin* 'not'

Person	Negative	Root	AGR
1SG	hin	dé:m-	-u
1PL	hin	dé:m-	-nu
2SG	hin	dé:m-	-tu
2PL	hin	dé:m-	-tan
3SG.F	hin	dé:m-	-tu
3SG.M	hin	dé:m-	-u
3PL	hin	dé:m-	-an

Let's take the syntactic criterion of substitutability. The negative *hin* can be substituted in the sentences in Table 2, which are in the imperfect form, by the jussive *haa* 'let', for example, in the same slot and the sentence will still be correct. Moreover, the mobility criteria can also apply without changing the meaning. Thus the negative imperative can be conveyed with relocated *hin* as in (1) below:[7]

[7]There is a very subtle difference in the mood of the instructor when the recipient hesitates and the sender changes his mind with hidden disappointment and withdraws the initial instruction. But, it conveys the same message, 'don't go!'

(1) Mobility of *hin* as the syntactic criterion to be written separately

a. *Hin deemin*
hin de:m-í-n
not go-IMP.2SG-NEG
'Do not go!'

b. *Deemuu hin baddin*
dé:m-ú: hin badd-í-n
go-INFIN not try-IMP.2SG-NEG
'Do not try to go!'

As for the semantic criteria of minimal ambiguities, the words in (2) have *hin* as their initial integral syllable. Thus it is difficult to guess whether the words with prefixed *hin* are prefixed complex or simple words unless one knows the meaning beforehand.

(2) Possible ambiguity created by conjoining *hin* to the next word

a.	*hinaaffaa*	[hiná:ffá:ʔ][8]	'jealousy'
b.	*hinniyyuu*	[hinníjjú:]	'negative competition'
c.	*Hindii*	[híndí:]	'India'
d.	*hirriiba*	[hirrí:ba][9]	'sleep (n)'
e.	*hilleensa*	[hille:nsa]	'rabbit'
f.	*hirroo*	[hirró:]	(said to incite bull to mate)
g.	*hirre*	[hírre]	'we distributed'
h.	*hinnaa*	[hinná:]	'red dye from plant leaves'
i.	*hinxilif*	[hint'ilíf]	'snatch (ideophone)'
j.	*hiyy-*	[hiyy-]	'poverty; become poor'
k.	*hincinnii*	[hinc'ínní:]	'tiny particles from harvest that irritate human skin'

Additionally, the semantic criterion of conceptual unity also applies. Unlike *irra* and *itti,* to be discussed later, there is no change in the meaning if we write *hin* adjoined or disjoined.

Similarly, two of the phonological criteria, namely pronounceability in isolation and phonological bridging, also apply. Thus, *hin* can be pronounced separately and, as there is a pause in speech between *hin* and the verb it negates, it can be separated by a hyphen, for example.

The negative particle *hin* typically precedes the verb it negates as a proclitic. Subsequently, there is an assimilatory change to the final /n/ which seems to

[8]Phonetically all long-vowel final words have a faint ʔ at the end, which is not written.

[9]*Hirriiba* 'sleep' (n) can be confused with a form of the verb *riib-*, namely *hirriiba* 'he does riib' (where assimilation has affected /hin + riiba/).

encourage writing *hin* conjunctively with the following verbs. This entices the proponents of writing *hin* conjunctively to invoke the presence of phonological unity. In connected speech, statements such as *hin láalu* 'he/I do not look' is pronounced [hillá:lu], and according to Oromic orthography, "you write what we hear". Additionally, if we go with this criterion, *hin* is closer to a preceding word, where in fast speech *inní hin láalu* 'he does not look' is pronounced as [innílla:lu]. On the other hand, words like *gad* 'down', *ʔol* 'up' also initiate such assimilation with the following verbs as in *gad + táaʔi* 'sit down' → [gattá:ʔi]; *ʔol + réebi* 'chase up' → [ʔorré:bi]. But no one writes these and other words of this category such as *birá* 'near', *ɟalá* 'under', *duubá* 'behind' conjunctively with the verbs that follow. Thus, attaching *hin* in any form, based on this relation of phonological unity, will require us to conjoin these other words as well, and no writer currently conjoins these words.

The next common morpheme that is sometimes prefixed is the monosyllabic reflexive marker *if* 'self'. It precedes the VP: [if + VP]. There is also a reciprocal in the form of *wal/wol*[10] 'each other' in Table 1, that requires a plural verb. Both these terms are inflected for certain cases, just like nouns, as indicated in Table 3.

[10]The spelling *wol* is mainly in the SE dialect.

Table 3: The inflection of *if* and *wal* for different cases followed by suffixes or postpositions.

	Followed	Gloss	Sentence	Sentence (IMP)
ʔif	k'ab-	hold	ʔif k'abi	hold yourself!
	-i:f	for (DAT)	ʔifi:f k'abi	hold for yourself!
	-i:n	by (INST)	ʔifi:n k'abi	use to hold yourself!
	-(i)tti	to	ʔifitti k'abi	hold to yourself, hug!
	-(i)rra:	from	ʔifirra: k'abi	hold from yourself!
	bira	near	ʔif bira k'abi	hold by your side!
wal	k'ab-	hold	wal k'aba:	hold each other, wrestle!
	-i:f	for (DAT)	wali:f k'aba:	hold for each other!
	-i:n	by (INST)	wali:n k'aba:	hold together!
	-(i)tti	to	walitti k'aba:	hold to each other, gather (TR)!
	-(i)rra:	from	walirra: k'aba:	hold from each other!
	ɟala	under	wal ɟala k'aba:	hold under each other!

The syntactic criterion of substitutability applies in this case. Nominals in the accusative case can be substituted in that slot in all conjugated forms of the verb where *if* and *wal* apply, as in *muc'aa k'aba*: 'hold the child', *isii k'abaa* 'hold her' where *muc'aa* 'child' and *isii* 'her' substitute for *wal* and so on.

The separability and the minimal ambiguity criteria require us to write *if* disjunctively. Regarding separability, postpositions can come between *if* and the following verb. The same holds true for *wal* as well. Thus in (3b), (3c), (3d), (3f) and (3g), postpositions such as *dura* 'in front', *gad* 'down', the particle *hin*, and the single phoneme dative marker *-f* intervene between *if* and *wal* and the following verb.

(3) The separation of *if* and *wal* from the verb phrase to which they are cliticised

a. *If qabi.*
?if k'áb-i
self hold-IMP.2SG
'volunteer!' (lit. hold yourself)

b. *If dura qabi*
?if dura k'ab-i
self front hold-IMP.2SG
'hold in front of you!'

c. *If hin qabinaa.*
?if hin k'ab-í-n-a:
self not hold-IMP-NEG-2PL
'Do not hold yourselves!'

d. *If gad hin qabinaa.*
?if gad hin k'ab-í-n-a:
self down not hold-IMP-NEG-2PL
'Don't hold yourselves down!'

e. *Wal qabaa.*
wal k'áb-a:
each.other hold-IMP.2PL
'wrestle (lit. hold each other)!'

f. *Wal jala qabaa*
wal ɟala k'áb-a:
each.other under hold-IMP.2PL
'don't hold under each other!'

g. *Waliif qabaa.*
wal-iː-f k'áb-aː
each.other-LV-DAT hold-IMP.2PL
'Hold for each other!'

Additionally, the semantic criterion of minimal ambiguity also supports writing *if* and *wal* disjunctively. Some words have *if* or *wal* as their integral initial syllables, including the following in (4) and (5).[11]

(4) Possible ambiguity created by writing *if* conjunctively

a.	*Ifa*	[ʔifá]	'light'
b.	*Iftaan*	[ʔiftáːn]	'after tomorrow'
c.	*Iftiina*	[ʔiftíːna]	'light'
d.	*Ifaajii*	[ʔifáːɟíː]	'energy, wealth and time spent on sth.'

(5) Possible ambiguity created by writing *wal* conjunctively

a.	*Wal'aan*	[walʔáːn-]	'medically treat (TR)'
b.	*Walaawwal*	[walaːwwál-]	'being indecisive (INTR)'
c.	*Walaloo*	[walaló:]	'poem'
d.	*Wallee*	[wallé:]/[wálle:]	'(love) song/ possibly?'
e.	*Walakkaa*	[walákkaː]	'centre'
f.	*Wallaal*	[wallá:l-]	'lose knowledge (TR)'
g.	*Waleensuu*	[waleːnsúː]	'type of tree'
h.	*Walaba*	[walábá]	'independent'
i.	*Waldaya*	[waldájá]	'association'
j.	*Walalaa*	[walálá:]	'liquified honey'
k.	*Warrana!*	[warrána]	'hey family! (VOC)'

The next morpheme in the list of those sometimes prefixed is the emphatic positive focus marker *ni*, the opposite of *hin*. See also the note on *hin* and *ni* in Table 1. Syntactically, it is *ni* + VIPFV. It can be substituted by negative *hin*, locatives like *bira* 'near', or the temporal adverb *amma* 'now' in the same slot, as shown in (6). We will take an example of the first criterion of substitutability first.

(6) Substitutability of *ni*

a. *Ni taa'an.*
ní taːʔ-an
FOC sit-3PL.IPFV
'They sit.'

[11]The hyphens used in (5) indicate that [walʔáːn-] and [walaːwwál-] are bound roots.

b. *Hin taa'an.*
hin tá:ʔ-an
NEG sit-3PL.IPFV
'They do not sit.'

c. *Bira taa'a.*
birá ta:ʔ-a
near sit-3SG.IPFV
'He sits near.'

d. *Amma teessi.*
amma te:ss-i
now sit-3SG.F.IPFV
'She sits now.'

Coming to the semantic criteria, we find ambiguity by conjoining *ni* to the verb next to it because there are words that have *ni* as their integral initial syllable, including those in (7).

(7) Words with *ni* as their integral initial syllables

a.	*Nigirtii*	[nigírtí:]	'quarrel, dispute'
b.	*Niitii*	[ni:tí:]	'wife'
c.	*Niis*	[ni:s]	'as well as'
d.	*Niin*[12]*/nan*	[ni:n]/[nan]	1SG indicator

In general, most of the criteria that apply to the negative marker *hin*, such as substitutability, mobility, minimizing ambiguity, conceptual unity, and pronounceability in isolation apply to *ni* as well.

5 Morphemes sometimes unnecessarily suffixed

Unlike the morphemes that are sometimes prefixed, which are clitics, suffixed morphemes are numerous. Moreover, most of them are generally accepted as proper inflectional or derivational suffixes. Oromic is an agglutinating language, and the affixes follow the root or the stem, except in the case of reduplication. Thus, most of the activity takes place on the right side of the root.

These suffixed morphemes have different shapes: single segments, mono- or multisyllabic suffixes. There is no question as to the single segment morpheme's

[12]Exclusively Eastern dialect. This is *ni* + *n* (emphatic *ni* and 1SG marker suffix *-n*) but written as one word.

position, and they are unanimously written conjunctively with the base they relate to.

Even though there is a consensus in writing the single segment morphemes as suffixes, there are cases where the accompanying morphophonology that requires long vowels before these suffixes create complications.[13] One such case is when the stem already has a long final vowel (actually /ʔ/-final, which is not spelt).[14] To add one of these suffixes on an already long-vowel-final base, we cannot directly add the morpheme but need an additional vowel. Adding this long vowel creates vowel hiatus, and it is necessary to resolve hiatus with an epenthesis where normally [ɗ], spelled *dh* is epenthesised.[15]

The introduction of this epenthesis also comes with a challenge. There is controversy whether *dha* is a copula or not. I have argued elsewhere (Youssouf 2019) that *dha* is not a copula, but it is a *dha* epenthesised to avoid vowel hiatus with the predicative case marker /a/ that follows it. This explanation brings us to the position of *dha* itself. Example (8)[16] below is one case. This confusion emanates from the lack of a rule to geminate digraph-represented sounds, that is, both [*ɗ*] and [*ɗɗ*] are spelled *dh*. I argue that it is easier to write *dha* disjunctively, e.g. *Boruun jabaa dha* for (8a), than making a new rule for the gemination of the digraphs (e.g. *dhdh*). Otherwise, the two statements in (8a) and (8b) cannot be distinguished. Note that the epenthetic consonant is glossed as Ø to indicate that it has no meaning.

(8) One case for the separation of *dha*

a. *Boruun jabaadha*
borúː-n ɟábaː-ɗ-a
boru-NOM strong-Ø-PCM
'Boru is strong.'

b. *Boruun jabaadha*
borúː-n ɟabaː-ɗɗ-a
tomorrow-1SG strong-ABEN-1SG.IPFV
'I become strong tomorrow.'

[13] In the Western part of Eastern Oromia and Central Oromia, the vowel is short.

[14] The existence of this /ʔ/ after long vowel-final nouns was noted, among others, by Andrzejewski (1957).

[15] It may be the case that it has to do with the final 'ʔ' rather than epenthesis. In the Western dialect of Eastern Oromia and part of Central Oromia, the final vowel length is absent. Thus *déːmuː ɗáːn* → [déːmuː ɗán]. Note also there is no sequence of two or more different vowels in Oromic. Thus Oromia → *Oromiyaa*; out → *awti*; oil → *oyli*.

[16] I coined PCM following Banti (1988: 28) who uses "case" for the nominals used as predicate, and I added "marker".

The [ɗ] in this example is an epenthetic consonant or "empty morpheme" (Lloret-Romanyach 1988), inserted between the otherwise adjoined predicative case marker /a/, as in *Boruu-n dardár-**a*** 'Boru is an adult'. Separating this single phoneme clitic from its host can be challenged. However, other factors also encourage writing *dha* disjunctively. The first is the phonological criteria related to the pause that always occurs before /ɗa/ when it is in the final position.

The final problem with [ɗa] involving affixation is the lack of consistency. For example, some writers disjoin [ɗa] but conjoin [ɗaː], sometimes itself with suffixes like the instrumental *-n*, as in [ɗaːn], and disjoining when [ɗ] is followed by the predicative or accusative case marker *-a* suffixed as in [ɗa]. It is common to see [ɗa] written separately as <dha>, as in (9a), while [ɗaːn], the longer one, is joined as <-dhaan> as in (9b). This tendency to write short [ɗa] alone disjunctively could come from the lexical copula concept especially the Amharic *nɜw* 'is' which is written disjunctively.

(9) Inconsistency in writing [ɗ]-epenthesised words disjunctively

a. *Deemuu dha fedha.*
déːm-uː ɗ-á feɗ-a
go-INF 0-PCM want-3SG.M.IPFV
'[It] is to go that he wants.

b. *Deemuudhaan fedha.*
déːm-uː-ɗáː-n feɗ-a
go-INF-PCM-1SG want-1SG.IPFV
'What I want is to go.'

It is not always the case that all multi-phoneme suffixes should better be written disjoined from the preceding word. Some are better written conjunctively. Those in this group are mainly consonant-initial. However, unlike the monophonemic ones, they need not be preceded by a long vowel. They just attach to an existing vowel of the host word form or are separated by an epenthetic short *i* in case the host word form is consonant-final. These morphemes are shown in the first six rows in Table 4. Some writers write them separately by adding or prothesising [i] at the beginning. These morphemes include conjunctive -[mmoː], locative case markers -[tti] 'to' and -[rra] 'on', and ablative case marker -[rraː] 'from'; -[jjuː] 'and', -[lleː] 'even', and the accusative focus marker -[jji], mainly in the SW Oromia dialect.

[17]Exclusively used in the SW Oromia dialect.

Table 4: List of some multi-phoneme morphemes sometimes suffixed.

Morpheme (Dial. Area)	Gloss	Category	Added on	**Example**
[(a/i)mmo:]	as for	Conj	NP_	is-á-mmó: 'as for him'
[(i)lle:]	as well	Conj/FOC	NP_	is-á-llé: 'him as well'
[(i)rra]	on	Postp	NP_	is-á-rrá 'on him'
[(i)tti]	to	Loc/Postp	NP_	is-á-ttí kénni 'give to him'
[(i)jji] (SW)[17]	it is	Focus	NP_	is-á-jjí gaafate 'it's him that he asked'
[(i)jju:]	even	Focus	NP_	is-á-jjú: ɗawe 'he hit even him'
-[uma]	just/only	Focus	NP_	/is-a-uma wa:mi/ [is-uma wa:mi] (E) [isa-ma wa:mi] (SW) 'call just him'

If we write these morphemes disjunctively with [i] prosthesis, we create ambiguity as shown in (10) by the following word-form pairs, where one is written adjoined, and the other disjoined.

(10) Morphemes for which writing conjoined or disjoined creates minimal pairs

a. *Isatti kenni.*
ʔis-á-ttí kénn-i
he-ACC-LOC give-2SG.IMP
'Give (something) to him!'

b. *Isa itti kenni.*
?is-á ?itti kénn-i
he-ACC LOC give-2SG.IMP
'Give him to (sth)!'

c. *Isarra kaayi.*
?is-á-rrá ka:j-i
he-ACC-on put-2SG.IMP
'Put (sth) on him!'

d. *Isa irra kaayi.*
?is-á ?irra ká:j-i
he-ACC on put-2SG.IMP
'Put him on (sth)!'

e. *Isarraa fuudhi.*
?is-á-rra-a: fú:ɗ-i
he-ACC-on-ABL take-2SG.IMP
'Take (sth) from him!'

f. *Isa irraa fuudhi.*
?is-á ?irra-: fú:ɗ-i
he-ACC on-ABL take-2SG.IMP
'Take him from (sth)!'

g. *Ganamollee arke.*
ganámó-llé: árk-e
N–as.well see-1SG/3SG.PST
'I/he saw Ganamo as well.'

h. *Ganamo illee arke.*
ganámó ?íllé: ?árk-é
N jaw.bone see-3SG.M.PST
'Ganamo saw the jaw bone.'

i. *Ganamoyyuu dhagaye.*
ganámó-jjú: ɗagáj-é
N-even hear-3SG.M.PST
'Even Ganamo heard.'

j. *Ganamo iyyuu dhagaye.*
ganámó ?íjj-ú: dhagáj-é
N shout-INF hear-1SG/3SG.M.PST
'I/he heard Ganamo shouting.'

Note that the suffixed morphemes have initial geminate consonants, and the short [i] put before these suffixed morphemes are not part of them but epenthesised, as mentioned above. This [i]-epenthesis can be seen in inflectional or derivational processes where consonant-initial suffixes have to be affixed to consonant-final stems. In that case, a cluster of more than two consonants is created, as in the case of *arkite* 'you/she saw' where [i] is epenthesised to avoid an [rkt] cluster so that /ark + te/ becomes [arkíte] not *[arkte]. It should be added that native speakers use tone also to disambiguate. However, since tone has not yet been marked in the orthography, as stated above, it is safer to disambiguate by using a space. Also note that the [i]-initial versions like *illee, irra, irraa, itti* and *ijjuu* are categorically different words.[18]

This group is not entirely uniform. For example, the case of -[mmo:], seems different from the group because the preceding vowel alternates between [a] in the East and [i] in the Western dialect. It is possible that underlyingly it is /a/-initial and segmental haplology applies, whereby one of two similar consecutive segments, a vowel, in this case, is deleted. Still, it seems to constitute no problem if written conjunctively like the rest in Table 4.

Again, there are other clitics consisting of multiple segments, with different phonological properties. These include -[icca], -[itti:] and -[uma]. They require the noun stems that precede them to lose their final vowel(s), if any, before attaching. Owens (1985: 96) calls -[icca]/-[itti:] particulative. They are close to the English definite article. Some people nowadays tend to use -[icca] in the English definite *the* sense, attaching it to every noun. These morphemes are normally cliticized on non-proper nouns, like collective or generic nouns.

When they attach to such nouns, they distinguish gender. For example, on [sare:] 'dog', [k'a:llu:] 'traditional religious authority', [faranɟi:] 'a white person', gender is distinguished by attaching *-icca* for the masculine and *-itti:* for the feminine. *Tokko* 'one (M)' or *takka* 'one (F)' is added to make them indefinite. Thus [faranɟicca tokko] means 'a white man, any white man'. For stems with short final vowels, [t] is epenthesised. Thus [araba] 'Arab' becomes [arabticca] not *[arabicca]; [ʃe:ka] 'sheikh' becomes [ʃe:k+ticca] which changes to [ʃe:jticca][19] not *[ʃe:kicca].

[18]For example, detached *illee* 'jaw bone' is a noun, *ijjuu* 'to make noise' is a verb. Similarly, *itti* disjoined in *muc'aa itti kanni* 'give the boy to' modifies the verb while conjoined in *muc'atti kanni* 'give to the boy' is a postposition.

[19]In the Eastern dialect, velars become palatal approximants before coronals. Example: /da:k+ta/ → [da:jta]. Palatal approximants undergo total assimilation before a coronal. Example: /da:j + ta/ → [da:tta].

The two *-icca* and *-itti:* particulative morphemes also apply to the cardinal number *tokko/takka* 'one'. Similarly, *-an* is added on the other single digits and the last digits in the tens; 10, 20, 30 and so on. If the number is vowel-final, the vowel is omitted except for *lama* 'two'. Table 5 summarises the particulative morphemes on the numbers.

Table 5: Numbers and their particulative suffixes

		-icca M -itti: F	-an	-en
1	tokk-o M takk-a F	tokkicca takkitti:		
2	lama		lama:n	(lame:n)
3	sad-ih		sadan	(sade:n)
4	afur		afran	
5	ʃan		ʃanan	
6	ɟah-a		ɟahan	
7	torb-a		torban	
8	sadde:t		sadde:ttan	
9	sagal		sajlan	
10	kuɗan		kurnan	

The next morpheme in the list being suggested to be conjoined in Table 4 above is -[uma], 'only, just'. It is added word-finally, and the final vowel of the word it is added to is dropped (see Owens 1985: 92). When the vowel(s) that precede this morpheme is/are deleted, ambiguity is created. For example, when it is suffixed on the positive focus marker *ni* and on *nu* '1PL.ACC', the [i] and [u] on the respective words are dropped, and the new term for both becomes [numa]. Thus the statement [numa ɲa:ta] can mean either 'he only/just eats' or 'he only/just eats us'. Even the tone distinction is not audible. Ambiguity generally arises when one morpheme or word is written or pronounced one way and has more than one meaning (homography or homophony). As can be seen in (11), when the vowels of focus marker *ni* and that of 1PL accusative case *nu* are deleted and *-uma* is suffixed, this results in homophones and homographs.

(11) Ambiguity created by suffixes that delete vowels before them

a. *ni-uma = numa* FOC

b. *nu-uma = numa* 1PL.ACC

However, not all dialects follow this rule. For example, in the Southeastern Oromia dialect, the suffixes lose their initial vowels instead of the stems losing their final vowels. Thus while in other dialects *tokko* 'one (M)' becomes [tokkicca], in the Southeast dialect it becomes [tokkocca]; *takka* 'one (F)' becomes [takkáttíː] not [takkíttíː] as elsewhere. Similarly, while *namá* 'human/person' becomes [námuma] elsewhere, in the Southeastern dialect, it becomes [námama]. This again creates another ambiguity as *-ma* is the passive voice suffix in Oromic. For example, [ɗabá] is 'loss/lack'. Then 'just/only loss' will be [ɗabama] which also means 'disappearance' unless we mark the tone, to distinguish [ɗábama] 'just/only loss', and [ɗabáma] 'disappearance', which cannot be marked in the current orthography. Writing *ma* alone disjunctively without its initial vowel may solve this while keeping with the Southeastern dialect.

An additional dialectal issue regarding *-uma* is that in some parts of the West, they will take a different path to disambiguate. If we suffix *-uma* to [k'ará] 'sharp' and [k'árá:] 'stalk', for example, both will become [k'áruma]. On the other hand, the examples in (12)-(13) illustrate the two words in equative sentences.

(12) *Billawni kun qara.*
billaw-ní kun k'ár-a
dagger-NOM this.M sharp-PCM
'This dagger is sharp.'

(13) *Kun qaraa dha.*
kun k'ára: ɗ-a
this.M stalk 0-PCM
'This is a stalk.'

Thus when *-uma* is suffixed to *k'áraa*, which has a final long vowel, the epenthetic [ɗ] without the following *a* is added, leading to [k'ára:ɗuma]. However, this eliminates one ambiguity and creates another one. When you pronounce [k'ára:ɗuma] there is a gap between [k'ára:] and [ɗuma] which also means 'being finished'. Besides, this does not disambiguate the case of short vowel-finals like *ni+ -uma* and *nu+ -uma* both of which end up as [numa], as indicated in (11) above. This ambiguity further supports adopting the Southeastern dialect that leaves the stem intact but deletes *-u* from *-uma* and suffix only *-ma*. Thus, /k'ára + -uma/ ⟶ [k'árama], /k'ára: + -uma/ ⟶ [k'ára:ma].[20]

[20]This version also has its own ambiguity because *-ma* is a passive voice marker on verbs unless written separately.

6 Conclusion

Oromic is an agglutinating language with affixing on the right side of words with virtually no prefixes or infixes. Yet, one of the pertinent undecided issues in the Oromic orthography is word boundary determination, meaning which morphemes to write conjunctively as affixes and which ones to write separately, with space between the affixes and the host words. This decision mainly concerns clitics, which are phonologically attached to a given host word even when they stand individually from a morphosyntactic viewpoint.

Establishing the difference between clitics that could better be written disjoined rather than adjoined to the host word has been addressed using the criteria put forward by Kutsch Lojenga (2014) as the main tool. A sample list of the clitics randomly written conjunctively with the host words has been provided, namely: *hin* 'not', *íf* 'self', *ní* FOC, and *wal* 'each other'.

The motive of this article is to suggest strategies to minimize ambiguity, which is the primary goal of standardizing orthography of a language. Possible causes of lack of convention on whether to write these clitics jointly or disjunctively are put forward. The original motto "we write as we speak", intended to emphasize the phonetic nature of the orthography, and the Ethiopic script used to write Amharic and Tigrinya as well as Oromic in the pre-1991 era, is put forward as possible causes.

Suggestions have been made, with sufficient arguments, to write the above mentioned four clitics as distinguished from the inflectional and derivational suffixes as well as single phoneme morphemes that are invariably written conjunctively. As lack of tone marking also has a share in perpetuating this ambiguity, the need to address that is also alluded to.

Abbreviations

Abbreviations in this chapter follow the Leipzig Glossing Rules, with the following additions.

E	Eastern dialect	ABEN	Autobenefactive
W	Western dialect	CONJ	Conjunction
S	Southern dialect	LV	Long vowel
SE	Southeastern dialect	PCM	Predicative case marker
SW	Southwestern dialect	RED	reduplicant

Acknowledgements

I thank members of "The Caucus to Study and Preserve the Oromo Language", *Shanacha afaanii* for short, who gave valuable suggestions. They engaged me in discussions that helped explain concepts I glossed over in the initial draft. I also thank Dr. Abdussamad, whose sharp criticisms forced me to develop more argument points. My sincere thanks also go to the anonymous reviewers, who made several valuable suggestions and to the ACAL editors, who patiently went through the different drafts and made several vital suggestions. However, any remaining mistakes or shortcomings are mine. Thank you all.

References

Andrzejewski, Bogumił W. 1957. Some preliminary observations on the Borana dialect of Galla. *Bulletin of the School of Oriental and African Studies* 19(2). 354–374. Note: This reference contains a word for the Oromo language that is now considered derogatory by the Oromo people and is no longer used.

Banti, Giorgio. 1988. Two Cushitic systems: Somali and Oromo nouns. In Harry van der Hulst & Norval Smith (eds.), *Autosegmental studies on pitch accent*, 11–50. Dordrecht: Foris.

Clamons, Robbin, Ann E. Mulkern, Gerald Sanders & Nancy Stenson. 1999. Functionalism and formalism in linguistics: Volume II: Case studies. In Michael Darnell, Edith Moravcsik, Frederick Newmeyer, Michael Noonan & Kathleen Wheatley (eds.), *Functionalism and formalism in linguistics: Volume II: Case studies*, 59–76. Amsterdam: John Benjamins.

Eaton, Helen & Leila Schroeder. 2012. Word break conflicts in Bantu languages: Skirmishes on many fronts. *Writing Systems Research* 4(2). 229–241. DOI: 10.1080/17586801.2012.744686.

Gasser, Mike. 2012. *HornMorpho 2.5 user's guide*. Research Group: Human Language Technology and the Democratization of Information. Indiana University.

Gragg, Gene B. 1982. *Oromo dictionary*. East Lansing, MI: African Studies Center, Michigan State University.

Griefenow-Mewis, Catherine J. 2001. *A grammatical sketch of written Oromo*. Cologne: Rüdiger Köppe.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Haspelmath, Martin & Andrea D. Sims. 2010. *Understanding morphology*. 2nd edn. London: Hodder Education.

Janko, Kebede H. 2007. Causative verb and palatalization in Oromo: Evidence from Oromo dialects and related Cushitic languages. *Journal of Oromo studies* 14(2). 89–104.

Kutsch Lojenga, Constance. 2014. Basic principles for establishing word boundaries. In Michael Cahill & Keren Rice (eds.), *Developing orthographies for unwritten languages*, 73–106. Dallas, TX: SIL, International.

Lloret-Romanyach, Maria-Rosa. 1988. *Gemination and vowel length in Oromo morphophonology*. Indiana University. (Doctoral dissertation).

Owens, Jonathan. 1985. *A Grammar of Harar Oromo (Northeastern Ethiopia)*. Hamburg: Helmut Buske.

Rogers, Henry. 2005. *Writing systems: A linguistic approach*. Malden, MA: Blackwell Publishing.

Sinha, Suchismita & Partha H. Talukdar. 2013. Word and syllable boundary of Sylheti phonemes/ syllables. *International Journal of Computing, Communications and Networking* 2(1). 1–5.

Tucho, Yigazu, R. David Zorc & Eleanor C. Barna. 1996. *Oromo newspaper reader, grammar sketch and lexicon*. Kensington, MD: Dunwoody Press.

Tutschek, Charles. 1845. *A grammar of the Galla language*. Munich: F. Wild. Note: This reference contains a word for the Oromo language that is now considered derogatory by the Oromo people and is no longer used.

Youssouf, Tamam. 2019. The copula in the Oromo language (Oromic). *International Journal of Innovative Research and Development* 8. 46–55.

Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Cliticization vs. Inflection: English N'T. *Language* 59(3). 502–513.