Falletti, Federico & Remijsen, Bert. 2025. Ghost segments in Sengwer nouns. *Glossa: a journal of general linguistics* 10(1). pp. 1–40. DOI: https://doi.org/10.16995/glossa.16526

Open Library of Humanities

Ghost segments in Sengwer nouns

Federico Falletti, University of Edinburgh, UK, s1779844@ed.ac.uk

Bert Remijsen, University of Edinburgh, UK, aremijse@exseed.ed.ac.uk

Like other Kalenjin languages, Sengwer has a complex and understudied noun morphology which makes use of an abundance of affixes as well as word-internal tone and ATR changes. In addition, its nominal inflectional paradigms are teeming with unexpected stem-final segments which appear in certain morphological conditions but are absent in others. The phonological composition of this stem-final latent material is not predictable and varies from noun to noun—C, V, CV and even CCV are all possible.

This paper provides a first description of the form and distribution of these segments as well as a novel analysis of their behaviour which explains much of the variation and irregularity seen in nominal inflection. In particular, this paper presents evidence that, although some of these latent segments are part of the root, others are part of suffixes and others still are morphologically active suffixes in their own right. In a departure from earlier studies, we argue that these segments should not be analysed as inserted *thematic* material but rather as deleted *ghost segments* instead. Though the term *ghost* refers to a common phenomenon cross-linguistically, different labels (e.g., latent, liaison, floating) have been used to characterise it depending on the tradition and the element affected. Building on Zimmermann's (2019) definition of ghosts, we argue that these phenomena do not only occur to satisfy synchronic phonological constraints but can also be the result of historical processes no longer at work.

1 Introduction

The nominal inflection and, in particular, the number marking system of Nilotic languages is notoriously complex (Corbett 2000). Sengwer, a Kalenjin language of the South Nilotic group, is no exception to this, having a tripartite number system, tone-based nominative case marking and a plethora of number, definiteness, demonstrative and possessive suffixes (Mietzner 2016). This means that for each noun there are a large number of inflected forms. In example (1), the noun *côok* is shown in four of these: (1a) the unmarked singular, (1b) the singular definite (marked by the suffix *-ít*), (1c), the plural (marked by the suffix *-ís*) and (1d) the singular proximal demonstrative (marked by the suffix *-ɔ̀nì*).[1]

(1)	a.	côok	b.	cóok-ît	c.	cóok-íis	d.	cóok-ìnì
		dagger		dagger-SDF		dagger-PL		dagger-PXS
		‘dagger’		‘the dagger’		‘daggers’		‘this dagger’

Though suffixes concatenate predictably, stem-final vowels often undergo sandhi, yielding fusional outcomes. Therefore, although *rjàampò* (2) is marked by the same suffixes as *côok* (1) above, the surface forms of these inflections have a long vowel /ʊʊ/ in the second syllable as a result of vowel sandhi between /ʊ/ and /ɪ/.

(2)	a.	rjàampò	b.	rjàampúʊt	c.	rjàampúʊs	d.	rjàampʊ̀ʊnì
		trumpet		trumpet:SDF		trumpet:PL		trumpet:PXS
		‘trumpet’		‘the trumpet’		‘trumpets’		‘this trumpet’

Surprisingly, however, there are many nouns with a paradigm similar to that of *rjàampò* which *do not* have a stem-final vowel in their unmarked singular. For instance, the inflected forms of the noun *pèet* (3) pattern more closely with those of *rjàampò* than those of *côok* (1), even though *pèet* ends in a consonant and *rjàampò* ends in a vowel.

(3)	a.	pèet	b.	pèetúut	c.	pèetúus	d.	pèetùunì
		day		day:SDF		day:PL		day:PXS
		‘day’		‘the day’		‘days’		‘this day’

If *pèet* patterned with *côok*, inflection would yield the ungrammatical forms in (4b–d).

(4)	a.	pèet	b.	*pèet-ít	c.	*pèet-íis	d.	*pèet-ìnì
		day		day-SDF		day-PL		day-PXS
		‘day’		‘the day’		‘days’		‘this day’

[1] The tone specifications of suffixes in the text do not always match those in the examples. For instance, the singular definite suffix *-ít* is realized as *-ît* in example (1a): *cóok-ît*. This and other differences in tone specification are part of regular—though complex—tonal processes, a thorough description of which falls outside the scope of this paper. Though there is no current description of the tone system of Sengwer, more information can be found on the related Nandi language in Creider & Creider (1989).

Thus, it appears as though the noun *pèet* has a stem-final vowel /u/ which does not surface in the unmarked singular form but appears in all of its inflected forms. Such "unexpected" segments are extremely common throughout the Sengwer lexicon and, because their phonological quality is lexically specified, their appearance cannot be attributed to any synchronic phonological process such as insertion. Therefore, these segments are hypothesised to be underlying material which is absent in the surface unmarked form of the noun[2], as illustrated by the schema in (5).

(5)

SR	pèet	'day'
UR	pèetu	
	day	

To avoid two layers of representation (surface and underlying) for every example, in most of this paper we adopt the notation for this phenomenon used in Zwarts (2003), showing the deleted underlying elements within brackets in the representation of the noun. Where necessary, however, we still use the standard notation with two layers. This means that, unlike other uses of this notation, brackets in the present transcription *do not* indicate that the elements within are optionally produced by speakers. Instead, instances of a noun such as (6) are to be interpreted as shorthand for (5).

(6)

a.	pèet(u)	'day'
	day	

It is important to note that these segments only differ from the rest of material in their behaviour, as their phonological specification is equivalent to that of segments which are always present on the surface. This difference in behaviour shows that, while still present in the underlying representation, these segments are recorded as defective or floating—unlinked to the noun's syllabic structure (cf. Faust & Torres-Tamarit 2017).

As mentioned above, the phonological composition of these stem-final "unexpected" segments is not restricted to /u/; they can commonly be found as other vowels (7), consonants (8) and open syllables (9).

(7)

a.	tèr(e)	b.	tèréet	c.	térêen	d.	tèrèenì
	pot		pot:SDF		pot:PL		pot:PXS
	'pot'		'the pot'		'pots'		'this pot'

(8)

a.	kwêɛ(s)	b.	kwɛ̀ɛs-tâ	c.	kwéɛs-wɅ̂	d.	kwɛ̀ɛs-ì
	buck		buck-SDF		buck-PL		buck-PXS
	'buck'		'the buck'		'bucks'		'this buck'

[2] In this article, *unmarked form* is used to mean a noun form with no morphological marking; i.e., the starting point of morpho-phonological derivation. In Sengwer, for inherently singular or inherently plural nouns (see §2.2.1), this corresponds to the citation form. However, for *numberless* nouns there is no unmarked surface form, as both singular and plural are morphologically marked.

(9)	a.	súrûum(pʌ)	b.	súrúumpêet	c.	súrúumpʌ̀ʌn	d.	súrúumpʌ́ʌnì
		navel		navel:SDF		navel:PL		navel:PXS
		‘navel’		‘the navel’		‘navels’		‘this navel’

The paradigm of the noun *tèr(e)* in (7) is another example of an opaque inflectional pattern. While the latent vocalic segment /e/ does not surface in the unmarked form (7a), its presence is evidenced by /ee/ in all its inflected forms (7b–d). This long vowel could only be the regular outcome of a sandhi merger between /e/ and /i/. This is further corroborated by inflectional paradigms of nouns such as *séesè* (10), which have a stem-final /e/ in the surface form of the unmarked singular. Although *tèr(e)* and *séesè* have different root-final segments in their unmarked form, they pattern in the same way—compare (7b–d) and (10b–d).

(10)	a.	séesè	b.	séeséet	c.	séesêen	d.	séesèenì
		dog		dog:SDF		dog:PL		dog:PXS
		‘dog’		‘the dog’		‘dogs’		‘this dog’

The noun *kwêɛ(s)* in (8) is an example of a latent consonant. While /s/ does not surface in the unmarked form in (8a), it appears in all inflected forms (8b–d) of the noun. Since /s/ does not undergo any sandhi process, it surfaces unchanged.

Finally, the noun *súrûum(pʌ)* in (9) is an example of a whole syllable which is present in the underlying representation but only surfaces in the inflected forms of the noun. Although the onset consonant /p/ surfaces unchanged, its latent vowel /ʌ/ undergoes regular sandhi with /i/ and surfaces as /ee/ in (9b) but as /ʌʌ/ in (9c) and (9d).

Occurring in most morphologically marked contexts, these latent segments are a prominent feature of the nominal inflectional morphology of Sengwer and key to understanding its patterns. Several studies on other Kalenjin languages acknowledge this, referring to this phenomenon using a variety of labels: *thematic suffixes* (Creider & Creider 1989; Zwarts 2003; Kouneli 2021; 2022), *class suffixes* (Larsen 1991), and *thematic endings* (Toweett 1975). While these studies focus on other topics, they have contributed important observations and partial descriptions of this phenomenon in a number of languages (i.e., Nandi, Endo, Kipsikis and Sabaot). A preliminary analysis of these latent segments can be found in Kouneli’s (2021) study of noun classification in Kipsikis, where she details their possible phonological forms in the language and provides an account of their behaviour. Interestingly, Kouneli highlights the resemblance of this phenomenon to that of thematic vowels in the inflectional paradigms of Romance languages such as Spanish and hypothesizes that these segments are declension class markers (cf. Aronoff 1994). In a wider discussion of metrical structures in Nilotic, Dimmendaal (2012) also analyses these segments’ behaviour, though taking a very different angle. While

still using the term “thematic”, he treats these segments as *liaison* or *floating* elements that can be part of the root or even singular suffixes in their own right rather than declension class suffixes.

In light of the dearth of in-depth research, the present paper aims to improve our understanding of this phenomenon as a whole. Taking Sengwer as a case study for a feature shared between all Kalenjin languages, we describe and analyse the phonology, behaviour and distribution of these latent segments based on a large dataset of nouns and their inflections. The analysis includes a series of diagnostic tests to ascertain their phonological properties and interactions with the surrounding material. Based on this description, we provide a novel analysis of this phenomenon as one of *ghost segments* rather than *thematic* or *class* suffixes. Furthermore, by comparing cognates in closely related Kalenjin languages, we argue that Sengwer ghost segments are an example of historical deletion rather than synchronic insertion. These comparisons also indicate that the prerequisites for the emergence of this phenomenon must have been present as early as Proto-Kalenjin.

The dataset used for this paper was drawn from a recent lexicographic project which culminated in the creation of the first Sengwer dictionary (Falletti 2023a; b). The dataset consists of over 1232 nouns inflected for number and definiteness, with notes on etymology, alternative forms and frequency of use. The data were checked during review workshops with over 20 native-speaker consultants, three of whom had basic training in linguistics and translation.

The paper is structured as follows. Section 2 provides background on the Sengwer language, including geographical and phylogenetic information as well as introductions to its phonology (§2.1) and nominal morphology (§2.2). Within the latter are subsections on the tripartite number system (§2.2.1) and definiteness marking (§2.2.2) both of which play an important role in the analysis. Section 3 is an analysis and description of the nature, distribution and behaviour of ghost segments. In the first subsection (§3.1), we outline three complementary diagnostic tests used to determine the presence and phonological form of ghost segments. Based on these tests, we describe the phonology and distribution of these segments in the Sengwer lexicon (§3.2), including some notes on variation. The analysis of this phenomenon (§3.3) is divided into three parts: first (§3.3.1), we evaluate the prevalent *thematic suffix* analysis in the literature; in (§3.3.2), we present some critical observations on the behaviour and distribution of ghost segments which challenge the *thematic suffix* analysis; then (§3.3.3), based on these observations and other considerations on the diachrony of these segments, we propose a novel analysis centred around the concept of *ghost segment*. Section 4 is the conclusion.

2 Background on Sengwer

Sengwer (also called Cherang'any) is an endangered minority language spoken in Kenya. Along with majority languages such as Nandi, Endo-Marakwet, Tugen and Kipsikis among others, it is part of the Kalenjin language group. With linguistic surveys lacking relevant data, the linguistic community itself estimates that there are only around 20,000 Sengwer speakers spread across three Kenyan counties: Elgeyo-Marakwet, West Pokot and Trans-Nzoia. Most speakers are over the age of 40 and the language is not being acquired by the majority of children. Within the Kalenjin group, Sengwer is classified by both Distefano (1985) and Rottland (1981) as a Northern Kalenjin language, more specifically within the Markweeta branch. Based on our own preliminary cross-dialect comparisons, we call into question this classification and tentatively speculate that the Sengwer language is either part of or a sister branch to Central and Elgon-Mau Kalenjin instead. The only study on the language, Mietzner's *Grammar of Cherang'any* (2016), shows that, in line with its Kalenjin relatives, Sengwer is highly agglutinative and syntactically head-initial, with default VSO word-order and marked-nominative case alignment.

2.1 Phonology

Like other Kalenjin languages such as Nandi (Toweett 1975), Endo Marakwet (Zwarts 2003) and Pokoot (Herreros Baroja 1989), Sengwer has a relatively small consonant inventory, with only 13 phonemes. There are stops and nasals at four places of articulation—bilabial, alveolar, palatal and velar: /p t c k/ and /m n ɲ ŋ/. and In addition, there are two approximants /w j/, a sibilant /s/, a liquid /l/ and a trill /r/.

In contrast, Sengwer has a comparatively large vowel inventory, with 10 phonemes divided into two groups according to their ATR specification: -ATR /ɪ ʊ ɛ ɔ a/ and +ATR /i u e o ʌ/. All of these vowels can be short or long. Like other Kalenjin languages, Sengwer has a +ATR-dominant vowel harmony system (Casali 2008). This means that within the word domain, -ATR vowels harmonise with +ATR segments to the right and to the left. For instance, the +ATR plural adjectival suffix *-ìin* (11) triggers a change from -ATR to +ATR in the vowels of adjacent morphemes *màal* 'to paint' and the adjectival suffix *-áat*.

(11)	màal-áat	+ìin	>	mʌʌl-ʌ́ʌt-ìin	*maal-aat-iin
	paint-ADJ	+PL		paint-ADJ-PL	*paint-ADJ-PL
	'painted'			'painted (pl.)'	'painted (pl.)'

It is worth mentioning that the phonetic realisation of the low +ATR vowel /ʌ/ has merged with the round back -ATR vowel /ɔ/ in some environments, while still retaining its +ATR

phonological value in harmony processes.[3] Since this is not a factor which influences the discussion of ghost segments, in this paper, the phonologically low +ATR vowel is always represented with /ʌ/. A merger between these phonemes is not unexpected from a typological point of view; Casali (2008) notes that, in ATR-harmony languages, low +ATR vowels tend to be the most disfavoured or marked. This often leading to their disappearance and a shift from ten to nine vowel systems.

Sengwer syllable structure is flexible. The minimal syllable is a single vowel nucleus. Any of the 13 consonants can be onsets or codas. Clusters, on the other hand, have a more limited distribution: complex codas are not allowed and complex onsets are mostly found as *consonant+glide* clusters. Geminate consonants are not allowed and when they occur in a sequence across morpheme boundaries, they degeminate.

Last but not least, each syllable is specified for High (H), Low (L) or Falling (F) tone. All three tones can occur in long and short vowels alike (12–14).

(12) cáat mán
thigh castor.oil.trees
'thigh' 'castor-oil trees'

(13) kɛ̀ɛt mèn
tree.PL clay
'trees' 'clay'

(14) kâat mâ
neck fire
'neck' 'fire'

As we will see in the next two sections, aside from being specified at the lexical level, tone also plays an important role in the morphology. Notably, it expresses nominative case as well as surface-level number differences (see §3.1 and §2.2, respectively).

[3] This merger makes the ATR value of some nouns ambiguous. For instance, it is impossible to know whether the vowels in 'animal' (ia) and 'sheep pen' (iia) are -ATR or +ATR based on their surface realisation alone. However, their ATR specification can be ascertained from their inflected forms: the presence of the +ATR allomorph *-tî* for the singular definite suffix in (ib) shows that 'animal' is +ATR, while -ATR *-tô* in (iib) shows that 'sheep pen' is -ATR.

(i) a. [tjɔ̂ɔɲ] 'an animal' b. [tjɔ̀ɔn-tɔ̂] 'the animal'
animal animal-SDF
(ii) a. [ɲcɔ́ɔr] 'a sheep pen' b. [ɲcɔ̀ɔr-tâ] 'the sheep pen'
sheep.pen sheep.pen-SDF

2.2 Noun morphology

The phonological shape of nouns in Sengwer varies considerably. Monosyllabic (15), disyllabic (16) and trisyllabic (17) word shapes are all common among non-derived nouns while tetrasyllabic noun roots (18) are rare. Nouns can begin and end in any consonant or vowel in the inventory.

(15) kɛ̂ɛt ‘tree’
tree

(16) kúu.kʌ̀ ‘grandfather’
grandfather

(17) cà.wíɪ.kìk ‘spurfowl’
spurfowl

(18) à.làp.tá.nì ‘brother-in-law’
brother.in.low

The tone patterns of noun roots are equally varied, with some being more common for plural nouns and others for singular nouns.

Noun roots appear with a variety of suffixes and prefixes. The morphological template for nouns is complex, with eight slots: three prefix slots and five suffix slots. Building on Mietzner (2016: 148), the ordering and function of Sengwer affixes are summarised and illustrated in **Table 1**.

1	2	3	ROOT	4	5	6	7	8	
kâap- LOC-	cɛ̀ɛp- FEM-		méɛrɪ initiate		-íɪs(ja) -PL				káap-cɛ̀ɛ-méɛríɪs LOC-FEM-initiate:PL ‘girl-initiate’s huts’
kâap- LOC		kàa- DVB-	ɲáaj heal	-so -DVB		-ít -SDF	-ɲuu -1SG		káap-kʌ̀ʌ-ɲʌ́ʌj-sée-ɲùu LOC-DVB-heal-DVB:SDF-1SG ‘my hospital’
	kìp- MSC-		jóoŋkè baboon		-în -PL	-ík -PDF		-àap -CS	kìp-jóoŋkèen-ík-àap MSC-baboon:PL-PDF-CS ‘the baboons of’
	kìp- MSC-		kéɛj self		-jáa(nta) -SG	-ít -SDF			kìp-kéɛj-àantɛ̂ɛt MSC-self-SG:SDF ‘the selfish man’

Table 1: The morphological template of Sengwer nouns. Numbers in the first row correspond to the following morphological slots: (1) locative prefix; (2) gender prefixes; (3) nominaliser prefixes; (4) nominaliser suffixes; (5) number suffixes; (6) definiteness suffixes; (7) possessive and demonstrative suffix; and (8) construct state suffix.

2.2.1 The tripartite system of number marking

As is characteristic of Nilotic languages, Sengwer and other Kalenjin languages have a system of tripartite of number marking (Dimmendaal 2000; Di Garbo 2014; Kouneli 2021). This means that nouns are classified into three patterns of morphological marking for number: (a) *inherently singular* nouns which are unmarked in the singular and marked in the plural; (b) *inherently plural* nouns which are marked in the singular and unmarked in the plural; and (c) *numberless* nouns which are marked in both the singular and the plural. **Table 2** shows examples of nouns from all three number-marking patterns.

	Singular	**Plural**
a	môok throat 'throat'	móok-wɅ̂ throat-PL 'throats'
b	kɔ́rɔ̀ɔr-jà feathers-SG 'feather'	kɔ́rɔ̀ɔr feathers 'feathers'
c	kɛ̀pɛ́p-cà wing-SG 'wing'	kɛ́pɛ́p-âj wing-PL 'wings'

Table 2: Number marking categories in Sengwer nouns.

Morphemes marking the singular number in Kalenjin languages, such as *-jà* (b) and its allomorph *-cà* (c) have often been called *singulatives* in the literature (first used by Rottland 1981a; 1981b). Following Kouneli's (2021) interpretation of the phenomenon, however, these are best thought of as *singular* suffixes instead. This is because, while singulars are simply allomorphs of singular number, singulatives are classifier-like individuating suffixes which modify collective nouns only (Greenberg 1978; Grimm 2012; 2018). Since there is no evidence that inherently plural nouns in Kalenjin are collectives, it is best to call these singular rather than singulative suffixes (cf. Kouneli 2021).

Within the dataset (Falletti 2023a), nouns are not evenly distributed across these three patterns: (a) 53.3% (653 items) of nouns are *inherently singular*, (b) 24.5% (300 items) are *inherently plural* and (c) 22.2% (272 items) are *numberless*. These morphological classes can also be described in terms of the semantic categories of nouns they contain (Dimmendaal 2000; Kouneli 2021):

(a) The *inherently singular* class mostly contains count nouns;
(b) The *inherently plural* class mostly contains mass nouns (e.g., flour, water, seeds) and other referents which are usually found in groups (e.g., animals, plants and people).

(c) The *numberless* class contains referents found in groups (e.g. trees, plants, people, pairs of objects) but does not contain any mass nouns. A larger proportion of nouns in this class are derived compared to the other two classes.

As in other Nilotic languages (Dimmendaal 2000; Moodie 2016; 2019), the semantic characterisation of the three patterns of number marking represents a reliable generalisation rather than a consistent and regular pattern. For instance, against what one would expect from its semantics, a mass-like noun such as *mèn* 'clay' is found in the *inherently singular* class rather than in the *inherently plural class*. Moreover, there is often speaker variation regarding which number category a noun belongs to, in particular when the word is not common.[4] For instance, the word for 'cricket' is *inherently singular* for some speakers (19) and *numberless* for others (20).

(19)	a.	kìp-círìt MSC-chirp.SG	'cricket'	b.	kìp-círít-âj MSC-chirp-PL	'crickets'
(20)	a.	kìp-círìt-jʌ́ʌ MSC-chirp-SG	'cricket'	b.	kìp-círít-âj MSC-chirp-PL	'crickets'

2.2.2 Inflection for definiteness

Sengwer nouns take two suffixes for definiteness: one for the singular and one for the plural. In turn, each of these has two allomorphs. The singular definite suffix has the allomorphs *-ít* and *-tâ*, which are almost equally common. Since their vowel is underlyingly -ATR, it harmonizes with +ATR vowel(s) in the preceding root—as in (22b).

(21)	a.	sùkûʊl school	'school'	b.	sùkúʊl-ît school-SDF	'the school'
(22)	a.	tjêen song	'song'	b.	tjèen-tâ song-SDF	'the song'

The plural definite has the allomorphs *-ík* (23) and *-kâ* (24). However, the latter has a very limited distribution, with only four nouns in the whole dataset, all of which have irregular stems.

(23)	a.	kʌ̀rʌ̀tì blood	'blood'	b.	kʌ̀rʌ̀tí-ik blood-PDF	'the blood'
(24)	a.	pây millet	'millet'	b.	pàa-kâ millet-PDF	'the millet'

[4] This is not unlike variation in noun classification found in other languages; for instance, in Italian the noun 'courgette' is masculine for some speakers—*lo zucchino*—and feminine for others—*la zucchina*.

Therefore, expanding on **Table 2** and taking into account the allomorphy of the definiteness marker, **Table 3** shows that every noun can have up to four basic forms[5]: singular indefinite, singular definite, plural indefinite and plural definite.[6]

	Singular		**Plural**	
	Indefinite	**Definite**	**Indefinite**	**Definite**
a	môok throat 'throat'	móok-tɅ̂ throat-SDF 'the throat'	móok-wɅ̂ throat-PL 'throats'	móok-wêk throat:PL-PDF 'the throats'
b	kɔ́rɔ̀ɔr-jà feathers-SG 'feather'	kɔ́rɔ̀ɔr-jɛ́ɛt feathers-SG:SDF 'the feather'	kɔ́rɔ̀ɔr feathers 'feathers'	kɔ́rɔ̀ɔr-ík feathers-PDF 'the feathers'
c	kɛ̀pɛ́p-cà wing-SG 'wing'	kɛ̀pɛ́p-cɛ́ɛt wing-SG:SDF 'the wing'	kɛ́pɛ́p-âj wing-PL 'wings'	kɛ́pɛ́p-âak wing-PL:PDF 'the wings'

Table 3: Definiteness marking by number category in Sengwer nouns.

As shown in this table, definite suffixes often interact with the preceding stem, which may consist of a root or a root plus a number-marking suffix, depending on their number class. As a result the underlying forms *-ít* and ***-ík*** are rarely realised as such. Instead, these suffixes merge with preceding vowels in sandhi processes, surfacing with different vowel quality and length– see for instance the addition of the definite suffix in *kɔ́rɔ̀ɔr-jɛ́ɛt* (b) and *kɛ̀pɛ́p-cɛ́ɛt* (c). The rules governing these sandhi processes are discussed in more depth in the next section.

In Endo and Nandi, just as in Sengwer, these suffixes express specificity and definiteness (Zwarts 2001; Hollis 1909). In other Kalenjin languages, such as Kipsigis (Kouneli 2020) and Kony Ogiek, this system has partially collapsed, with the definite form becoming the only form for many nouns. Having no clear semantic role, the label "secondary" has often been used to refer to such lexicalised suffixes across this language family (Toweett 1979; Creider & Creider 1989).

3 Ghost Segments

As discussed in the introduction, some nouns in Sengwer have unexpected realisations in their marked forms which can only be explained by postulating latent stem-final phonological segments. Far from being a marginal phenomenon, these segments are found in more than 75%

[5] We conceive of these forms as basic because they are the principal parts of the nominal paradigm needed for all further morphological affixation (Stump & Finkel 2013). Suffixes with equivalent meanings may attach to either the definite or the indefinite form; for instance, the singular proximal demonstrative only attaches to the indefinite form while the plural proximal demonstrative only attaches to the definite form.

[6] Not all nouns have a plural and a singular—for instance, mass nouns may only have plural and plural definite forms.

of all nouns in the data. As noticed by Kouneli (2020), the presence and form of ghost segments do not correlate with number classes (*inherently singular*, *inherently plural*, *numberless*) nor with preceding phonological material. In the next section, we outline three types of marking as complementary diagnostic tests to investigate the phonology of ghost segments: definiteness, nominative and proximal demonstrative marking.

3.1 Diagnostic tests

So far, we have seen that the presence of ghost material (be it a segment or sequence of segments) can be determined by adding inflectional marking to the stem. However, because of sandhi processes, suffix allomorphy and speaker variation, the phonological characteristics of ghost segments are often opaque. Therefore, to ascertain the presence and nature of ghost segments in the underlying form, it is necessary to observe a noun in several morphologically marked forms: the definite, nominative and proximal demonstrative. Each of these three reveals different phonological aspects of these segments while obscuring others.

Definiteness marking is one of the most widespread and regular morphological operations in Sengwer. As shown in Section 2.2.2, definiteness applies to all nouns, both singular and plural, with phonologically symmetrical allomorphs (*-ík/kà* and *-ít/tà*). These factors make it an excellent diagnostic for the presence of both consonantal and vocalic ghost segments. In the examples below, the definite markers trigger the surfacing of three kinds of ghost material: a single coda consonant in (25), a vowel in (26), and a CV syllable in (27).

(25)	kwîi	+ tà	>	kwìis-tâ	*kwìi-tâ
	foreleg	+ SDF		foreleg-SDF	foreleg-SDF
	‘foreleg’			‘the foreleg’	‘the foreleg’
(26)	sèr	+ ít	>	sèrúut	*sèr-ít
	nose	+ SDF		nose:SDF	nose-SDF
	‘nose’			‘the nose’	‘the nose’
(27)	kérûuŋ	+ ít	>	kérúuŋkêet	*kérúuŋ-ît
	rain.cloud	+ SDF		rain.cloud:SDF	rain.cloud-SDF
	‘rain-cloud’			‘the rain-cloud’	‘the rain-cloud’

Though the surfacing of all underlying segments through the addition of a definite suffix is a reliable process, it is not always enough to determine the quality of a ghost vowel. For instance, while the definite form of the noun *kwîi* (25) surfaces with a final /s/, the definite forms of nouns *sèr* (26) and *kérûuŋ* (27) surface with long vowels /uu/ and /ee/ respectively. Rather than surfacing in their underlying form, these are the result of the coalescence between stem-final ghost vowels and the suffix-initial /ɪ/. Therefore, in order to establish the underlying vowel quality of these ghost segments, the following sandhi rules need to be taken into account:

- /u/ coalesces with /ɪt/ into /uut/
- /i/ coalesces with /ɪt/ into /iit/
- /o, e, ʌ/ coalesce with /ɪt/ into /eet/

Since three out of five vowels yield the same surface vowel quality, the nature of non-high ghost vowels remains opaque when using definite suffixes as a diagnostic. This means that, though we can establish that *sèr(u)* (26) has a final ghost vowel /u/, we cannot know which vowel is at the end of the ghost sequence in (27): is the noun *kérûuŋ(ko)*, *kérûuŋ(ke)* or *kérûuŋ(kʌ)*? In order to find out, the definite forms need to be compared with a different morphological marking operation: nominative marking.

Sengwer has marked-nominative case alignment. This morphosyntactic alignment type means that the only morphologically marked case is the nominative and the unmarked accusative is used as the citation form. This unmarked accusative is often called absolutive. Nominative marking in Sengwer does not include any suffixation. Instead, it is a process that replaces the unmarked tone pattern of nouns with a fixed nominative pattern. For unmarked singular nouns such as those in (28–30), the replacement pattern is a low tone on all syllables save for a high tone on any word-final open syllable, *be it part of a ghost sequence or not*. This means that, albeit with a different tone pattern, ghost material surfaces segmentally unchanged.

(28)	kwîi(s)		>	kwìis
	foreleg	+NOM		NOM\foreleg
	‘foreleg’			‘foreleg’
(29)	sèr(u)		>	sèrú
	nose	+NOM		NOM\nose
	‘nose’			‘nose’
(30)	kérûuŋ(kʌ)		>	kèrùuŋkʌ́
	rain.cloud	+NOM		NOM\rain.cloud
	‘rain-cloud’			‘rain-cloud’

Using this second diagnostic test, the quality of the ghost vowel of *kérûuŋ* (30) becomes immediately apparent. Unfortunately, though transparent, nominative marking does not reveal ghost segments consistently: depending on the word, speakers vary in the realisation of the nominative form, using forms with ghost segments interchangeably with forms without. For instance, while all speakers consistently use the ghost sequence in the nominative of *réɛrèɛs(-ja)* (31), there are two potential forms for the nominative of *tèr(e)* (32) and *làal(a)* (33): one with a high-toned ghost segment and one without any ghost segment at all.

(31)	réɛrèɛs(-ja)		>	rèɛrèɛs-já
	bat-SG	+NOM		NOM\bat-SG
	‘bat’			‘bat’

(32)	tèr(e)		>	tèr ~ tèré
	pot	+NOM		NOM\pot
	‘pot’			‘pot’

(33)	làal(a)		>	làal ~ làalá
	horn	+NOM		NOM\horn
	‘horn’			‘horn’

Furthermore, this variation does not apply equally to all nouns: while both nominative forms of *tèr(e)* (32) are in use by different speakers, the same speakers only use *làal* for the nominative of (33). The form with the ghost segment *làalá* is accepted as grammatical but not currently in use. The reason behind this variation is clear from a diachronic point of view: the absence of thematic segments in the unmarked form of the noun triggers a process of reinterpretation which, over time, leads to their deletion in all forms. This suggests that the loss of ghost material is happening gradually across the paradigm of individual nouns. In the case of *làal(a)*, while its ghost segment is retained in some inflections—such as the definite form *làalɛ́ɛt* (34c)—speakers are in the process of losing it in the nominative (34b).

(34)	a.	làal(a)	b.	làal ~ làalá	c.	làalɛ́ɛt
		horn		NOM\horn		horn:SDF
		‘horn’		‘horn’		‘the horn’

Being an operation that affects the tone of vowels, nominative marking would not be expected to affect purely consonantal segments such as the /s/ of *kwîi(s)* in (28). However, speakers readily produce the nominative form *kwìis*, with the consonantal ghost segment /s/ included. Other consonantal segments however, such as /p/ in *tjêe(p)* ‘girl’, never surface in the nominative. Therefore, though not reliable, the nominative can be a diagnostic for the presence of at least some consonantal ghost segments.

The third diagnostic test for vowel quality is the addition of *-ːnì*, the singular proximal demonstrative suffix. Unlike the definite suffixes, this suffix lengthens any preceding stem-final vowel but *does not* trigger sandhi. As shown in examples (35) and (36), the quality of the stem-final vowels is retained when the proximal demonstrative suffix is added.

(35)	rjàampù	+ːnì	>	rjàampùʊnì
	trumpet	+PXS		trumpet:PXS
	‘trumpet’			‘this trumpet’

(36)	táamnà	+ːnì	>	táamnàanì
	beard	+PXS		beard:PXS
	‘beard’			‘this beard’

This lack of sandhi aids in determining or confirming the vowel quality of vocalic ghost segments. For instance, the forms *sèrùunì* (37d) and *kérúuŋkʌ́ʌnì* (38d) confirm that the vowel quality of the vocalic ghost segments deduced by comparing the definite suffixation (37c–38c) and the nominative marking (37b–38b) is correct.

(37)	a.	sèr(u)	b.	sèr ~ sèrú	c.	sèrúut	d.	sèrùunì
		nose		NOM\nose		nose:SDF		nose:PXS
		‘nose’		‘nose’		‘the nose’		‘this nose’
(38)	a.	kérûuŋ(kʌ)	b.	kèrùuŋ ~ kèrùuŋkʌ́	c.	kérúuŋkêet	d.	kérúuŋkʌ́ʌnì
		rain.cloud		NOM\rain.cloud		rain.cloud:SDF		rain.cloud:PXS
		‘rain-cloud’		‘rain-cloud’		‘the rain-cloud’		‘this rain-cloud’

Similarly to the other two diagnostics, this method also has a caveat: while the quality of the high vowels /u i ʊ ɪ/ and low vowels /a ʌ/ remains unchanged for all speakers when adding this suffix, the quality of the four mid vowels /e ɛ ɔ o/ is subject to variation. Though some conservative speakers retain the quality of mid vowels, others prefer using the low-vowel allomorph of the singular proximal suffix *-ːnì*. For example, although the root-final vowels of the unmarked forms of the nouns *móosò* and *wésè* (39–40a) are mid vowels, when marking them for the singular proximal demonstrative some speakers use a mid vowel allomorph (39–40b) and others a low vowel allomorph (49–40c).

(39)	a.	móosò	b.	móosòonì	c.	móosʌ̀ʌnì	
		baboon		baboon:PXS		baboon:PXS	
		‘baboon’		‘this baboon’		‘this baboon’	
(40)	a.	wésè	b.	wésɛ̀ɛnì	c.	wésàanì	
		machete		machete:PXS		machete:PXS	
		‘machete’		‘this machete’		‘this machete’	

This variation in the inflectional pattern is also present in words ending in a ghost vowel. In examples (41–42) are two nouns marked for the singular proximal demonstrative by a non-conservative speaker.

(41)	tèr	+ːnì	>	tèrʌ̀ʌnì
	pot	+PXS		pot:PXS
	‘pot’			‘this pot’
(42)	sòt	+ːnì	>	sòtʌ̀ʌnì
	gourd	+PXS		gourd:PXS
	‘gourd’			‘this gourd’

By observing these two nouns in the forms given, it is not possible to ascertain whether their ghost vowel is (e), (o) or (ʌ). This means that to disambiguate between words ending in a *low*

ghost vowel (/a/ or /ʌ/) and a *mid* ghost vowel (/ɛ ɔ/ or /e o/), one of the other two diagnostic tests has to be used. Alternatively, since the change is in flux and both versions are still accepted by all speakers, the vowel quality can also be tested using grammaticality judgements. Examples (43–44) show the two nouns above in their unmarked form and all three inflected forms including speaker variation.

(43)	a.	tèr(e) pot ‘pot’	b.	tèr ~ tèré NOM\pot ‘pot’	c.	tèréet pot:SDF ‘the pot’	d.	tèrèenì ~ tèrʌʌnì pot:PXS ‘this pot’
(44)	a.	sòt(o) gourd ‘gourd’	b.	sòt ~ sòtó NOM\gourd ‘gourd’	c.	sòtéet gourd:SDF ‘the gourd’	d.	sòtòonì ~ sòtʌʌnì gourd:PXS ‘this gourd’

In the absence of a stem-final vowel, the suffix *-ːnì* is realised as either *-ì* or *-ìnì*. As should be expected by now, this suffix triggers the surfacing of any word-final consonants. The two examples below show the ghost segments /s/ and /p/ regularly surfacing before the singular proximal demonstrative suffix. The noun *tjêe(p)* (46)—with irregular stem *cèep*—is the only example of a ghost /p/ in the data.

(45)	kwîi(s) foreleg ‘foreleg’	+ːnì +PXS	>	kwìis-ì foreleg-PXS ‘this foreleg’
(46)	tjêe(p) girl ‘girl’	+ːnì +PXS	>	cèep-ì girl-PXS ‘this girl’

Being a singular suffix, the proximal demonstrative *+ːnì* can only be used as a diagnostic for singular nouns. Though plural nouns have a proximal demonstrative, this suffix is never added to the indefinite form of the noun and is only used after the plural definite suffix. Therefore, for plural nouns, one can only rely on nominative and definite marking, as illustrated in the *inherently plural* nouns *ŋòl(a)* (47) and *pèel(i)* (48).

(47)	a.	ŋòl(a) saliva ‘saliva’	b.	ŋòlá NOM\saliva ‘saliva’	c.	ŋòlɛ́ɛk saliva:PDF ‘the saliva’		
(48)	a.	pèel(i) elephants ‘elephants’	b.	pèelí NOM\elephants ‘elephants’	c.	pèelîik elephants:PDF ‘the elephants’		

Though all ghost segments looked at so far have been found after the noun root, ghost segments can also be regularly found after certain suffixes. In the data, these are most commonly plural

and singular suffixes. As shown in the examples (49) and (50), these ghost sequences do not behave any differently than root-final phonological material: they surface as expected in the inflected forms used as diagnostics so far, such as the nominative (49–50b) and the definite form (49–50c).

(49)	a.	pèel-jə̂ʌ(ntʌ) elephant-SG ‘elephant’	b.	pèel-jʌ́ʌntə̀ NOM\elephant ‘elephant’	c.	pèel-jʌ́ʌntêet elephant-SG:SDF ‘the elephant’
(50)	a.	cóok-íis(jʌ) dagger-PL ‘daggers’	b.	còok-íisjə̀ NOM\dagger ‘daggers’	c.	cóok-ìisjêk dagger-PL:PDF ‘the daggers’

Ghost sequences such as /ntʌ/ and /jʌ/ occur predictably after specific suffixes, evidencing that these are latent portions of these suffixes rather than separate elements. We never find an instance of an inflected stem with *-íis* without a ghost sequence /jʌ/ nor an instance of an inflected stem with *-jʌ́ʌ* without /ntʌ/. Not only that, the ghost sequence /ntʌ/ in (49) is only *ever* found after *-jʌ́ʌ* and is the only example of ghost material of this kind (i.e., CCV) in Sengwer.

In this section, we have shown how three inflections (definite marking, nominative marking and proximal demonstrative marking) can be used as diagnostic tests to reveal the underlying phonological representation of ghost segments. Conversely, this explains much of the variation in vowel quality, vowel length and unexpected consonants found in the marked forms of nouns. Other marking operations, especially plural suffixes, can also aid in determining the quality of ghost segments. However, being by far the most complex and irregular marking operation in Sengwer, the use of plural marking as a diagnostic would require a thorough description of the phenomenon, which falls outside the scope of this paper.

3.2 Phonology, distribution and variation

By applying the complementary diagnostic tests outlined above to our dataset, it was possible to determine the underlying phonological representations of most ghost segments, identify the environment in which each of them occurs and quantify their distribution. The results are summarised in **Table 4**.

The dataset used includes 1372 nouns; while some of these include only a singular form (176 nouns) and others only a plural form (78 nouns), most include both (1118 nouns). Breaking this down, it means that **Table 4** reports on around 1294 singular noun forms and 1196 plural noun forms—a total of 2490 items. Of the total number of nouns, 8% accept two or more plural or singular forms. Around 52% of the Sengwer noun forms collected were found to have ghost material.

item category	ghost material		number of tokens	number of types	environment
	-ATR	+ATR			
part of the root	a	ʌ	127	127	after a single consonant
	ɪ	i	16	16	
	ʊ	u	31	31	
	ɛ	e	9	9	
	ɔ	o	7	7	
	a/ɔ	ʌ/o	1	1	
	a/ɛ	ʌ/e	6	6	
	s		4	4	mostly after a long vowel
	p		1	1	
	j		2	2	
	c		1	1	
	-	pʌ, pe	2	2	after a homorganic nasal
	ka	kʌ, ki	6	6	
	-	tʌ, ti	1 ~ 2	1 ~ 2	
	ja	jʌ	11	11	after a vowel or a consonant
	-	wʌ	3	3	
part of a suffix	-	wʌ	15	2	part of the suffixes *-êj(wɔ)* and *-tîn(wɔ)*
	ta	tʌ	54	2	part of the suffixes *-íin(tɔ)* and *-jàn(ta)*
	nta	ntʌ	252	1	part of the singular suffix *-jáa(nta)*
	a	ʌ	55	1	part of the plural suffix *-ìin(a)*
	ɪ	i	128	1	part of the plural suffix *-în(i)*
	ja	jʌ	255	1	part of the plural suffix *-íis(jâ)*
whole suffix	ja	jʌ	164	1	as a singular suffix
	a	ʌ	15	1	
	tɔ	to	4	1	
	-	i	80	1	as a plural suffix
	-	ʌ	13	1	
	-	wʌ	2	1	
	ɪ	-	1	1	
Total			1281	244	

Table 4: Forms and distribution of ghost segments by category (part of the root, part of a suffix, whole suffix).

The ghost material in **Table 4** is classified into three categories, found on the left-most column: (a) part of the root (229 nouns), (b) part of a suffix (759 nouns) and (c) whole suffix (279 nouns). This classification is based on their distribution across the nominal inflectional paradigm, which will be discussed in more depth in Section 3.3 below. In the next column moving to the right, each ghost categories is classified by phonological shape and occurrence in -ATR and +ATR contexts. In the last two columns are the number of tokens and types found for each ghost segments as well as their environment.

Looking at the phonological shape of all the ghost material found, it is apparent that the vast majority has a short vowel and appears in a stem-final open syllable, either as single vowels (.V#) or consonant + vowel segments (.CV#). The other two kinds of latent segments are stem-final coda consonants (C#) and the ghost sequence /nta/ (C.CV#). These four kinds of ghost material are illustrated in the examples (51–54) below.

(51)	a.	lɑ̂ʌl(ʌ) bag 'bag'	b.	lɑ́ʌlɛ̂ɛt bag:SDF 'the bag'	c.	lɑ̀ʌlɑ́ NOM\bag 'bag'	d.	lɑ́ʌlɑ́ʌnì bag:PXS 'this bag'
(52)	a.	mwêeŋ(kʌ) beehive 'beehive'	b.	mwéeŋkêet beehive:SDF 'the beehive'	c.	mwèeŋkɑ́ NOM\beehive 'beehive'	d.	mwéeŋkɑ́ʌnì beehive:PXS 'this beehive'
(53)	a.	kwîi(s) foreleg 'foreleg'	b.	kwìis-tɑ̂ foreleg-SDF 'the foreleg'	c.	kwìis NOM\foreleg 'foreleg'	d.	kwìis-ì foreleg-PXS 'this foreleg'
(54)	a.	lɑ̀l-jɑ̂ʌ(ntʌ) cough-SG 'cough'	b.	lɑ̀l-jɑ́ʌntêet cough-SG:SDF 'the cough'	c.	lɑ̀l-jɑ́ʌntɑ̀ NOM\cough-SG 'cough'	d.	lɑ̀l-jɑ́ʌntɑ́ʌnì cough-SG:PXS 'this cough'

While the first two kinds (51, 52) are common across all three categories (*part of roots*, *part of suffixes* and *whole suffixes*) the last two kinds (53, 54) have a much more limited distribution. Single ghost consonants such as (53) are only found after a long vowel in a handful of irregular high-frequency nouns. The ghost sequence in (54), on the other hand, is only found at the end of the suffix *-jáa*, the most common singular suffix in the language. Looking at its distribution, it appears that this outlier is the result of the historical deletion of a .CV# syllable followed by the deletion of a C# segment. This is confirmed by comparing words like *pɑ̀ʌj-ɑ̂ʌ(ntʌ)* (55) with cognates in closely related languages such as Endo Marakwet.[7] Here we find that while Endo has retained the coda /n/, the syllable /ta/ was also lost.

[7] Endo examples are from Zwarts (2003). Tone is not always represented in the examples.

(55) Endo pʌʌ-jʌʌn(tʌ) ‘elder’
Sengwer pʌ̀ʌ-jʌ̂ʌ(ntʌ)
elder-SG

Moreover, Endo retains other stem-final single consonants which are found as ghost segments in Sengwer—illustrated in example (56) below.

(56) Endo cîic ‘person’
Sengwer cîi(c)
person

Examples (55) and (56) show that while ghost CV syllables were most likely deleted before Sengwer and Endo split from each other, the deletion of ghost consonant codas occurred separately in Sengwer. In turn, this evidences the fact that the occurrence of the unusual ghost sequence CCV in Sengwer may be the result of two consecutive deletion events: one which occurred in the common ancestor of those two languages and one which occurred later in Sengwer only.

Aside from observing variation in the realisation of ghost segments in cognate nouns between Sengwer and related languages, we also found some variation within Sengwer itself. Interestingly, most of the differences in the realisation of ghost segments were observed in the *part of the root* group. This is to be expected, as this is the group that contains the most lexical items—each item represents a separate root, while in the other two groups, each item is an instance of the same morpheme (be it a whole suffix or part of one). This variation manifests itself in two ways: ambiguity in vowel quality and ambiguity in the presence of ghost segments.

For a limited amount of nouns—shown in the *ghost material* column of the table as V/V—it was not possible to ascertain the exact vowel quality of the ghost segment. This is because of a combination of up to two reasons: the segment remained ambiguous even after applying all three diagnostic tests and speakers varied in their realisation of it or accepted more than one vowel as grammatical. In turn, the distribution of these ambiguous ghost segments falls into two categories: (a) low-frequency words which speakers were not familiar with and (b) inherently plural nouns. The occurrence of ambiguous ghost segments in low-frequency unfamiliar words is hardly surprising, given that the system relies on speakers deducing the quality and presence of ghost segments for a given lexical item based on its inflected forms. Speakers might not have learnt the exact quality of a ghost vowel if they have only heard the noun containing it a few times in their lives and only in one or two of the inflected forms. Ambiguous ghost segments found in unmarked plural nouns are a systematic case of the same phenomenon; while ghost segments in inherently singular nouns can be found by comparing the nominative, proximal demonstrative, definite and plural forms, ghost segments in inherently plural nouns can be observed in a much more limited set of inflections. First, inherently plural nouns are more likely to be mass nouns and therefore lack a singular form. Second, the proximal demonstrative forms for plurals are not

derived from the unmarked form but rather from the definite form. That is, the plural proximal demonstrative suffix *-cù* attaches to the definite plural stem suffixed with *-ík* (57b) rather than to the plural indefinite (57a). This means that *pèetùusjék-cù* (57c) is grammatical and *pèetúusjʌ́-cù* (57d) is not.

(57)	a.	pèetúus(jʌ)	b.	pèetùusjêk	c.	pèetùusjék-cù	d.	*pèetùusjʌ́-cù
		day:PL		day:PL:PDF		day:PL:PDF-PXP		day:PL-PXP
		‘days’		‘the days’		‘these days’		‘these days’

That leaves a lot of nouns, such as the mass noun *ŋòl* (58a) with only two forms to be compared, the nominative (58b) and the definite (58c) forms.

(58)	a.	ŋòl	b.	ŋòlá ~ ŋòlé	c.	ŋòlɛ́ɛk	d.	ŋòlɛ̀ɛk
		saliva		NOM\saliva		saliva:PDF		NOM\saliva:PDF
		‘saliva’		‘saliva’		‘the saliva’		‘the saliva’

Since the definite form *ŋòlɛ́ɛk* (58c) and its nominative *ŋòlɛ̀ɛk* (58d) are much more common than the indefinite nominative, speakers are likely to assume an underlying /ɛ/ rather than /a/. Moreover, because non-high vowels /a ɔ ɛ ~ ʌ o e/ all coalesce with /ɪ ~ i/ into /ɛ ~ e/, there is no other available test to disambiguate between a ghost /a/ or /ɛ/ for the noun *ŋòl*. This explains the variation found between speakers in the realisation of (58b) and the fact that both forms are readily accepted as grammatical.

In another subset of nouns of the *part of the root* group, we found that speakers disagreed in whether certain nouns had any ghost mateiral at all. For instance, the noun *ɲcɔ́ɔr* was found to have two realisations depending on the speaker: one with no ghost material (59) and one with a ghost vowel /ʊ/ (60).

(59)	a.	ɲcɔ́ɔr	b.	ɲcɔ̀ɔr-tâ	c.	ɲcɔ̀ɔr	d.	ɲcɔ́ɔr-ì
		byre		byre-SDF		NOM\byre		byre-PXS
		‘byre’		‘the byre’		‘byre’		‘this byre’

(60)	a.	ɲcɔ́ɔr(ʊ)	b.	ɲcɔ̀ɔrûʊt	c.	ɲcɔ̀ɔrʊ́	d.	ɲcɔ̀ɔrʊ́ʊnì
		byre		byre:SDF		NOM\byre		byre:PXS
		‘byre’		‘the byre’		‘byre’		‘this byre’

In a system that relies on learners memorising a large number of segments which do not surface in the unmarked form, variation of this kind is to be expected. Learners can easily re-analyse the underlying representation of a noun depending on the frequency of the word itself and its use in context. This evidence supports the view that the storage of lexical items occurs in an episodic manner during language learning (Pierrehumbert 2016). For instance, the noun *ɲcɔ́ɔr*, being a building, is mostly used in locative constructions. Contrary to English where we would expect a definite article after a preposition, locative constructions in Sengwer are formed exclusively

using the indefinite form of the noun. Example (61) shows one such sentence, where ‘byre’ is the location for the noun ‘goats’.

(61) míi ŋɔ̀rɔ́ɔr ɲcɔ́ɔr ‘There are goats in the byre.’
be goats byre

Moreover, since traditionally each homestead will have a maximum of one byre, plural and demonstrative forms are also rare; there is no need to refer to ‘this byre’ or ‘the byres’ if there is but one context-relevant byre. Finally, locations are also less frequently found as the subject of a verb and hence less frequently marked for the nominative case. These factors mean that the statistical occurrence of the noun *ɲcɔ́ɔr* is much higher in its unmarked indefinite singular form than in any other possible inflected form (plural, plural definite, proximal demonstrative or singular definite). Therefore, the presence of the ghost vowel /ʊ/ in the surface forms of *ɲcɔ́ɔr* is statistically low compared to other nouns, leading to speakers reanalysing this noun root as lacking any ghost material altogether.

3.3 Analysis

As mentioned in the introduction, the presence of “unexpected” or latent segments in the inflected forms of the noun has been reported in most of the literature dealing with the nominal morphology of Kalenjin languages—albeit under different names: *thematic suffixes* (Zwarts 2003; Kouneli 2021; 2022; Creider & Creider 1989, Dimmendaal 2012), *class suffixes* (Larsen 1991), and *thematic endings* (Toweett 1975). This nomenclature used in the literature reflects an understanding that these elements are stem-determined (i.e., *thematic*) declension class suffixes. However, out of these authors, Kouneli (2021) is the first to provide an analysis of this phenomenon in these terms for Kipsikis, a language closely related to Sengwer. In this section, we evaluate her preliminary hypothesis by applying the diagnostic tests outlined in the previous section to the Sengwer dataset. Based on the results, we outline a novel analysis of this phenomenon for Sengwer which is in line with Dimmendaal (2012) observations on Nandi morphophonology. We then propose the use of the umbrella term *ghost segments* to integrate it within our current understanding of similar phenomena in unrelated languages.

3.3.1 Evaluating the thematic suffix analysis

In her 2021 paper on number-based noun classification, Kouneli proposes that the unexpected vocalic segments which appear in the noun paradigms of Kipsikis are akin to *thematic vowels* in languages such as Latin, Spanish and Ancient Greek. While admittedly tentative and left as a topic for further research, this analysis presents interesting points to our discussion. In this view, all ghost segments are *declension class markers*: morphologically active suffixes which determine the declension paradigm of a noun (Aronoff 1994). In Indo-European languages, the presence

of these *thematic suffixes* depends on the noun stem: while some stems require them, others do not. The quality of a thematic suffix is also lexically specified, falling into a limited number of categories (i.e., declension classes).

While at first glance this is a compelling analysis for Sengwer as well, when applying our diagnostic tests, treating these segments as *thematic*—i.e., belonging to the stem—becomes problematic. Though it is true that the segments in question often appear either after roots or after suffixes—as they do in Latin for instance—they can also appear *simultaneously* after both. As shown previously, the word *pèet* (62a) has a root-final vowel /u/ that only surfaces in its inflected forms. One such form is the plural *pèetúus* (62b). Here, sandhi between the underlying stem-final /u/ and the suffix-initial /i/ of the plural marker *-íis* yields the allomorph *-úus* in the surface form.

(62)		a.		b.		c.		d.	
	SR		pèet		pèetúus		pèetúusjʌ̀		pèetùusjêk
	UR		pèetu		pèetu-íisjʌ		pèetu-íisjʌ		pèetu-íisjʌ-ík
			day		day-PL		NOM\day-PL		day-PL-PDF
			‘day’		‘days’		‘days’		‘the days’

However, the nominative plural *pèetúusjʌ̀* (62c) and plural definite *pèetùusjêk* (62d) of the same noun show that the suffix *-íis* itself has a ghost sequence /jʌ/. This means that both /u/ and /jʌ/ are simultaneously present ghost material in the inflection of the noun *pèet*. According to the model used by Kouneli (Oltra-Massuet & Arregi 2005), declension class suffixes should only occur once for every stem—whether that is after the root or after the suffix—but not after both. In other words, while this analysis predicts only one thematic suffix per stem, the data commonly includes nouns with two thematic suffixes per stem: one after the root and one after a suffix.

Moreover, since the ghost sequence /jʌ/ is *always* present after the suffix *-íis*, even in nouns such as *côok* (63) which do not have any root-final latent element in their inflected forms—i.e., are not *thematic*—we cannot attribute its appearance to a stem-dependent process. Therefore, the fact that Sengwer ghost segments appear after functional bound morphemes independently of whether the stem has a latent segment itself argues against the interpretation that they are thematic suffixes.

(63)		a.		b.		c.	
	SR		côok		cóok-íis		cóokìisjêk
	UR		côok		côok-íisjʌ		côok-íisjʌ-ík
			dagger		dagger-PL		dagger-PL-PDF
			‘dagger’		‘daggers’		‘the daggers’

Even if the ghost sequence /jʌ/ were a declension class marker introduced by the suffix *-íis*, in order to be considered the marker of a declension class, it would have to appear in more than just a single environment. However, in plural nouns, the ghost sequence /jʌ/ can only be found after *-íis*.

Another argument that challenges the analysis of these segments as *thematic suffixes* is the fact that it does not provide a descriptive distinction between root-final segments which appear in the unmarked surface form and those which do not. Nouns with a root-final /u/ which surfaces in the unmarked form such as *léŋkû* (64) take the same singular definite allomorph and the same singular proximal demonstrative allomorph as nouns with a ghost root-final /u/ such as *pèet(u)* (65). This contrasts with nouns with no root-final /u/ such as *côok* (66).

(64)	a.	léŋkû pantry 'pantry'	b.	lèŋkú NOM/pantry 'pantry'	c.	léŋkúunì pantry:PXS 'this pantry'	d.	léŋkûut pantry:SDF 'the pantry'
(65)	a.	pèet(u) day 'day'	b.	pèetú NOM/day 'day'	c.	pèetùunì day:PXS 'this day'	d.	pèetúut day:SDF 'the day'
(66)	a.	côok dagger 'dagger'	b.	còok NOM/dagger 'dagger'	c.	cóok-ìnì dagger-PXS 'this dagger'	d.	cóok-ît dagger-SDF 'the dagger'

Both /**u**/ vowels in (64) and (65) are root-final elements which do not contribute any extra meaning to the noun while influencing the choice of inflection: the only difference between them is whether or not they surface in the unmarked form. Therefore, if we were to label the /u/ of *pèet(u)* as a thematic suffix, we should use the same label for the /u/ of *lénkû*—despite their obvious differences in behaviour. Labelling root-final segments as thematic would not only make every root-final short vowel a *thematic suffix* but also every root-final /p/, /s/, /c/ and /j/—all of which are possible consonantal ghost segments. This would mean that as well as a declension class for each of the ten potential stem-final short vowels, more declension classes would have to be posited for several stem-final consonants. For instance, in this view, *éemâ(s)* (67) and *kɔ̀mɔ̀s* (68) would both have thematic suffix *-s* and belong to the same declension class.

(67)	a.	éemâ(s) longing 'longing'	b.	èemʌ̀s NOM/longing 'longing'	c.	éemʌ́s-ì longing-PXS 'this longing'	d.	éemʌ̀s-tâ longing-SDF 'the longing'
(68)	a.	kɔ̀mɔ̀s side 'side'	b.	kɔ̀mɔ̀s NOM/side 'side'	c.	kɔ̀mɔ̀s-ì side-PXS 'this side'	d.	kɔ̀mɔ̀s-tâ side-SDF 'the side'

However, as shown in more depth in Section 3.3.3, positing declension classes in Sengwer is not necessary when analysing these segments as cases of deletion rather than insertion, as the whole nominal inflectional system becomes predictable. Declension classes such as those in Latin and

Ancient Greek are nominal inflectional patterns which are not predictable by the shape of the noun and where synthetic processes make it impossible to accurately separate the root from its suffixes. While the current stage of Sengwer does have an abundance of sandhi processes which can make the boundary opaque, the root and its suffixes are still distinct.

For these reasons, we believe that there is no evidence that ghost segments are *suffixes*, as these latent vocalic and consonantal segments occur in all inflected forms of their noun without contributing any semantic content. Calling them "suffixes" in light of this would counter all definitions of the term (Crystal 1980; Hartmann & Stork 1972).

To summarise, the evidence presented so far suggests that it is not possible to analyse the phenomenon of latent segments in Sengwer as an instance of *thematic suffixes* which mark declension class based on their behaviour and distribution. First, we have shown that their presence is not always dependent on the stem, as these segments can be found (a) after suffixes, (b) after roots and (c) after both at the same time. Therefore, they cannot be called thematic. Second, the behaviour of these latent root-final segments patterns with that of counterpart segments that surface consistently in their inflected forms. Therefore, analysing them as suffixes would mean that all root-final elements should also be analysed as suffixes, which is not tenable given that neither segments carry any meaning. Third, these segments do not all form natural declension classes with segments in other nouns, since certain elements only occur after a single suffix or a single noun root.

3.3.2 Lexical and affixal ghost segments

While ghost material such as that seen in *pèet(u)* (65) and *éemʌ̂(s)* (67) discredit a *suffix* analysis, there are other cases in which the label *suffix* is warranted by the distribution of certain segments and their contribution to word meaning. A large subset of nouns in the data shows alternations in the presence of ghost segments between their singular and plural forms. The noun *múrèn* 'man', for instance, has a ghost vowel /ʌ/ in the singular (69a) but lacks a ghost segment in the plural (70a). This is evidenced by the fact that the ghost segment triggers regular sandhi in the singular definite inflection (69b) but does not in the plural definite inflection in (70b). If this ghost vowel were part of the stem, the plural definite inflection should be the ungrammatical form in (70c) instead.

(69) a. múrèn(ʌ)
man.SG
'man'

b. múrènéet
man.SG:SDF
'the man'

(70) a. múrên
man.PL
'men'

b. múrén-îk
men.PL-PDF
'the men'

c. *múrènéek
men.PL:PDF
'the men'

The absence of a stem-final vowel in the plural forms is further confirmed in the nominative inflection; compare the nominative singular *mùrènʌ́* (71b) and the nominative plural *mùrén* (72b). Once again, while we find a stem-final /ʌ/ in the singular, there is no such segment in the plural.

(71) a. múrèn(ʌ)
man.SG
'man'

b. mùrènʌ́
NOM\man.SG
'man'

(72) a. múrên
man.PL
'men'

b. mùrén
NOM\man.PL
'men'

The inverse situation is also possible; stem-final ghost vowels can appear in the plural inflection of some nouns but not in the singular. For instance, the singular noun *kêɛt* 'tree' (73a) only shows a change in tone in the nominative *kɛ̀ɛt* (73b), receiving no additional segment. In contrast, its plural form, while appearing with no stem-final vocalic segment in the unmarked form *kɛ̀ɛt* (74a), "unexpectedly" receives a high-toned /ɪ/ in the nominative plural *kɛ̀ɛtí* (75b).

(73) a. kêɛt
tree.SG
'tree'

b. kɛ̀ɛt
NOM\tree.SG
'tree'

(74) a. kɛ̀ɛt(ɪ)
tree.PL
'trees'

b. kɛ̀ɛtí
NOM\tree.PL
'trees'

This leads to the conclusion that these segments do not depend on the root but on the number specification of the noun. Therefore, while segments that surface in all marked forms of the noun must be part of the root and segments that surface in all marked forms after suffixes must be part of those suffixes, segments which only surface in the singular or plural forms of the noun must be, in fact, *number suffixes*. The difference between these three categories can be observed in their distribution among the nominative forms in the singular and plural inflection illustrated in examples (75)–(77).

(75) a. pèet(u)
day
'day'

b. pèetú
NOM/day
'day'

c. pèetúus(jʌ)
day:PL
'days'

d. pèetúusjʌ̀
NOM/day:PL
'days'

(76) a. côok
dagger
'dagger'

b. còok
NOM/dagger
'dagger'

c. cóok-íis(jʌ)
dagger-PL
'daggers'

d. còok-íisjʌ̀
NOM/dagger-PL:PDF
'daggers'

(77) a. múrèn(-ʌ)
man-SG
'man'

b. mùrèn-ʌ́
NOM/man-SG
'man'

c. múrên
man.PL
'men'

d. mùrén
NOM/man.PL
'men'

The ghost vowel /u/ is *part of the root* of the noun *pèet* (75), as it appears in all its plural and singular inflections and does not influence word meaning. The ghost sequence /jʌ/, on the other hand, is *part of a suffix* as it invariably appears after the plural suffix *-ís* in the inflection of nouns *pèetúus(jʌ)* (75) and *cóok-ís(jʌ)* (76). Finally, the ghost vowel in *múrèn(-ʌ)* (77) is a *whole suffix,* as it appears in all singular inflections but never occurs in the plural inflection and its presence marks the difference between the singular and the plural.

Following this analysis, 159 nouns which appear to be inherently singular and 23 nouns which have segmentally identical singular and plural forms can now be reanalysed as having one of the three ghost singular suffixes (*-ja, -a* and *-tɔ*) instead. Take the noun in *múrèn(-ʌ)* in (77) again, for instance; on the surface, it appears to only differ in tone from singular to plural, with no clear indication of which of the two forms is unmarked. However, after applying the diagnostic tests, the singular *múrèn(-ʌ)* is shown to have a singular ghost suffix. This makes the plural *múrên* the unmarked form and classifies this noun as *inherently plural*.

On the other hand, nouns such as *réɛrèɛs(-ja)* (78a) below appear to be unmarked in the singular in their surface form while, in fact, they have a ghost singular suffix *-ja*.

(78)	a.	réɛrèɛs(-ja)	b.	rèɛrèɛs-já	c.	réɛréɛs-âj	d.	rèɛrèɛs-áj
		bat-SG		NOM/bat-SG		bat-PL		NOM/bat-PL
		‘bat’		‘bat’		‘bats’		‘bats’

This suffix regularly appears in its singular inflected form, such as the nominative in (78b), but is completely absent from its plural form, where the plural suffix attaches directly to the root (78c). If this was not the case, and these syllables were root-final elements we would expect to find a glide /j/ surfacing in the plural forms. However, the form in (79) is ungrammatical:

(79) *réɛréɛsj-âj
bat-PL
‘bats’

Therefore, although appearing to be *inherently* singular, the noun root *réɛrèɛs* is, in fact, *numberless* in terms of its number class, as it takes suffixes both in the singular and in the plural.

Corroborating this analysis is the fact that some ghost singular suffixes such as *-jà* (and its +ATR counterpart *-jʌ̀*) are well-attested in the data as *surface* singular suffixes—as shown for *sìkìr-jʌ̀* (80a).

(80)	a.	sìkìr-jʌ̀	b.	sìkìr-jʌ́	c.	síkír-âj	d.	sìkìr-ʌ́j
		donkey-SG		NOM/donkey-SG		donkey-PL		NOM/donkey-PL
		‘donkey’		‘donkey’		‘donkeys’		‘donkeys’

Furthermore, the surface and ghost variants of *-jà* share a semantic domain distribution: both are particularly high in flora and fauna terms, while limited to a few items across other semantic domains.

By postulating ghost number suffixes, many of the suprasegmental differences in tone, length and ATR specification between surface singular and plural forms, which were previously labelled irregular, become predictable. For instance, the differences in tone and ATR between the singular and the plural of the noun *ŋéljêp* (81), can now be explained by the presence of a plural suffix ghost segment *-i*. Its presence predictably triggers three suprasegmental changes: it regularly lengthens the preceding syllable's vowel, induces a (L.)H replacive tone pattern and changes the ATR specification from -ATR to +ATR. The existence of this suffix is evidenced by its presence in the nominative plural *ŋèljèep-í* (81d) and its absence in the nominative singular *ŋèljép* (81b).

(81)	a.	ŋéljêp	b.	ŋèljép	c.	ŋèljéep(-i)	d.	ŋèljèep-í
		tongue		NOM/tongue		tongue-PL		NOM/tongue-PL
		'tongue'		'tongue'		'tongues'		'tongues'

Evidence for this plural marker (**-i**) is found in 80 noun tokens in the data (see **Table 4**). This suffix is particularly productive in the derivation of deverbal agentive nouns, as shown in (82).

(82)	al	+(i)	>	ʌ́ʌl(-i)
	buy	+PL		buy-PL
	'buy'			'buyers'

The verbal root *al*, which is an underlyingly toneless -ATR morpheme, become high-toned and +ATR with the addition of the ghost plural suffix *-i*. Since the vowel of the verb root *al* is short, it is lengthened in the agentive.

Another surface irregular change that can be explained by the presence of a ghost suffix is the alternation between palatal and velar consonants in the absence of a surface suffix.

In Sengwer, certain words—nouns, adjectives and verbs alike—have a root-final palatal which turns to velar when a suffix is added. For instance, the noun *tjʌ̂ʌɲ* (83) has a root-final /ɲ/ in its unmarked form but consistently surfaces with a root-final /ŋ/ in all its suffixed forms.

(83)	a.	tjʌ̂ʌɲ	b.	tjʌ̀ʌŋ-ì	c.	tjʌ̀ʌŋ-în	d.	tjʌ̀ʌŋ-îik
		animal		animal-PXS		animal-PL		animal-PL:PDF
		'animal'		'this animal'		'animals'		'the animals'

This same alternation can occasionally be seen even when no surface suffix is present. For example, the verbal root *rwʌʌc* has a root-final /c/ in its imperative (84a) but surfaces with a root-final /k/ in its deverbal derivation (84b).

(84)	a.	rwʌ́ʌc	b.	kíi-rwʌ̀ʌk
		IMP/try		DVB-judge
		'try (in court)'		'trial'

However, applying diagnostic tests such as the addition of the singular definite suffix (85c), we can see that this alternation is, in fact, triggered by a singular suffix -ʌ.

(85) a. kíi-rwʌ̀ʌk(-ʌ)
DVB-judge-SG
'trial'

c. kíi-rwʌ̀ʌk-éet
DVB-judge-SG:SDF
'the trial'

Therefore, by postulating the presence of ghost suffixes, many seemingly irregular segmental and suprasegmental patterns found in nouns become predictable. These alternations are evidence of their presence in the underlying representation; even when absent from the surface representation, ghost segments still influence preceding surface material.

However, while alternations in ATR harmony and length can be easily predicted by the presence of specific ghost segments changes in tone are not always as straightforward. While there is evidence for regular changes of tone and length in relation to ghost segments and ghost suffixes in particular, a study of these patterns would require a thorough discussion of the tone processes in the language which falls outside the scope of this paper.

To summarise, then, although it seems that ghost segments are never *thematic* suffixes, for some of the nouns in our data, we did find evidence of *ghost* suffixes. These appear to be *number* suffixes rather than declension class suffixes. They are a small group of morphemes (listed in **Table 4**) which appear either only in the singular or only in the plural forms of nouns. The observations presented so far in this section mean that ghost segments in Sengwer can be classified into three categories: *part of the root, part of a suffix* and *whole suffixes.*

Though differing in behaviour, all these ghost segments have common features: (a) they are absent stem-finally in the base forms of the noun but present in all or most marked forms of the noun form they appear in, either as a result of suffixation or a suppletive change in tone pattern, (b) they are either a single coda consonant C#, a short-vowelled open syllable (C) V# or a combination of both in that order C.CV#. Therefore, rather than being *suffixes* added onto marked stems to determine their inflectional class, evidence suggests that these segments are stem-final material deleted for phonological reasons. In the next section, we present some arguments for the use of *ghost segments* rather than *thematic suffixes* to refer to this phenomenon.

3.3.3 Arguments for a ghost segment analysis

The fact that some of the Kalenjin ghost segments could be part of the morphemes they appear after, rather than being separate suffixes, is proposed twice in the literature: by Bennett (1974) and by Dimmendaal (2012). Both papers deal with features of the Nilotic family at large and, therefore, do not account for this phenomenon in detail. In his description of tone in relation to the Nilotic case system, Bennett (1974) compares the ghost segment phenomenon in Kalenjin to the so-called *shadow vowels* of Teso, a related East Nilotic language. In Teso, some short root-final vowels in an open syllable are elided in the unmarked form of the noun but occur in all suffixed forms as well as before any consonant-initial word. This is in line with Zimmermann's (2019: 1) definition of the term *ghost* as:

> "segments that (1) are idiosyncratically bound to specific morphemes and (2) alternate with zero in a way that the majority of segments within this language do not."

Though Teso's shadow vowels as reported by Bennett are similar to the Sengwer phenomenon at issue in this paper, there are some important differences. The surfacing of shadow vowels in a particular word in Teso is phonologically predictable by the make-up of the following word as well as the addition of a suffix. This means that the patterning of ghost segments in Teso is sensitive to contexts that cross word boundaries as well as those which are word-internal. For Sengwer ghost segments, on the other hand, the context across word boundaries is irrelevant. More importantly, Teso's shadow vowels surface purely for phonological reasons; these ghost vowels appear to prevent consonant clusters, which are not allowed by Teso's phonology. This means that, although the presence and quality of shadow vowels are lexically specified, they surface to satisfy a phonological constraint on the phonotactics of the language. Lindsey (2019), in her discussion of ghost phenomena, calls this kind of ghost phenomenon *hero ghosts:* the shadow vowels "come to the rescue" to avoid consonant clusters. Sengwer ghosts, on the other hand, appear to be purely lexically determined; ie. they do not appear to interact with any markedness constraint in the language.

Though the term *ghost* can often be found in the literature to refer to phenomena described by Zimmermann's definition (Zoll 1993; Kiparsky 2003; Archangeli 1984; Szypra 1992), many other labels have been used depending on the element affected, its behaviour and the tradition of the field: *floating feature* (e.g., Remijsen & Ayoker 2020); *latent segments* (e.g., Tranel 1996a), *phantom consonants* (e.g., Schmidt 1994), *liaison consonants* (e.g., Adda-Decker *et al.* 1999) and *epenthetic segments* (e.g., Hyman 1972). This plethora of different terms has not always aided researchers in recognising the striking similarities between parallel linguistic features. In this section, we will show how the umbrella term of *ghost segment* allows us to draw interesting parallels between similar linguistic phenomena.

While authors such as Larsen (1991) have proposed that the Kalenjin ghost segments are also inserted for phonological reasons, there is no such evidence in our data. In his description of the nominal morphology of Sabaot, Larsen (1991: 7) states that: "the purpose of consonant insertion is in all cases to avert an unwanted vowel clash." In his view, the ghost segment /nta/ at the end of the singular suffix *-jaa* would be a thematic suffix (glossed by the author as THM) inserted "to avoid vowel fusion". Using Larsen's analysis and notation, example (86) shows that the noun *mʊr-jaa,* for instance, would receive a thematic suffix *-nta* in order to avoid a hiatus between long /aa/ and /ɪ/.

(86)	SR	mʊrjaa	+ ɪt	>	mʊrjaantɛɛt	*mʊrjaaɪt
	UR	mʊr-jaa	+ ɪt	>	mʊr-jaa-nta-ɪt	*mʊr-jaa-ɪt
		rat-SG	+ SDF		rat-SG-THM-SDF	rat-SG-SDF
		'rat'			'the rat'	'the rat'

This analysis has one obvious problem: while the hiatus between the long final vowel of the singular suffix and the initial vowel of the singular definite suffix is resolved, the addition of /nta/ gives rise to a new hiatus between short /a/ and /ɪ/. As is the case for all other instances of single adjacent vowels, this new hiatus is resolved by coalescence (i.e., sandhi) rather than the insertion of any "thematic" material. Therefore, we could restrict Larsen's claim and state that thematic suffixes are added only to resolve hiatus when long vowels are involved. Yet, hiatus between two adjacent long vowels or a long and a short vowel is extremely common in our Sengwer data, occurring in at least 100 cases both root-internally and across morpheme boundaries. Examples (87) and (88) show these two possibilities (VV.V and VV.VV) with the same vowels that would result in hiatus in the ungrammatical form **mʊr-jaa-ɪt* in (86).

(87) kwáa.ís 'to hunt solo'
hunt.solo

(88) kàa.-ìɪ.l-ɔ̂ 'oiling'
DVB-oil-DVB

However, inserting material that is not at all present is different from realising underlying "weak" material; that is, the language could be making use of underlying material when possible to avoid hiatus. This effect is called Emergence of the Unmarked (McCarthy & Prince 1994). Still, even taking this into the account, the addition of the ghost sequence /nta/ does not truly avoid hiatus in (85), as it introduces a final vowel /a/ which merges with /ɪ/ at the morpheme boundary, resulting in /ɛɛ/. This is true for the majority of our data: ghost segments in Sengwer normally contain at least one vocalic segment which would incur in hiatus with any following suffix but is instead resolved by coalescence (rather than consonant insertion). Therefore—at least for Sengwer—there is no reason to believe that ghost segments are *inserted* to avoid hiatus or any other phonological constraint. Instead, as argued in the previous section, it is more likely that the phenomenon of ghost segments is one of stem-final *deletion*. In his paper on metrical structure in the morphophonology of Nilotic, Dimmendaal (2012: 16) puts this interpretation forward in relation to the noun system of the Kalenjin language Nandi, stating that the phenomenon is one of "omitted (truncated or deleted) thematic vowels".

Although deletion appears to be the best explanation for this phenomenon, the quality and distribution of ghost segments in our data do not point towards a single phonological triggering environment in the current stage of the Sengwer language. Instead, the fact that this phenomenon is common to all branches of the Kalenjin language group (Dimmendaal 2012) suggests that this is a process which has its roots in Proto-Kalenjin at the very latest. In fact, as we have seen in the example of Teso, similar phenomena occur all over Eastern Nilotic (Dimmendaal and Breedveld 1986, Dimmendaal 1983), one of the three main branches of Nilotic. Even within South Nilotic, Rottland & Creider (1996: 1–2) state that Datooga's short vowels in final position are realised "very weakly and are generally voiceless". Dimmendaal (2012: 17) comes to the same conclusion,

stating that: “the Kalenjin system of “thematic vowel” truncation presents the end result of the kind of alternation still found synchronically in neighbouring Teso-Turkana languages”.

Moreover, it seems that different deletion triggers were active at different times in the evolutionary history of this language family. Comparative data between Kalenjin language varieties suggests that deletion from the unmarked form occurred both before and after these languages split apart. For instance, in the cognate nouns in (89), the Sabaot cognates have a ghost vowel whereas the Sengwer cognates do not.

(89)	Sabaot	cʌʌk(e)	pee-k(ʌ)
	Sengwer	cʌ́ʌkè	pèe-kʌ̀
		granary	water-PDF
		‘granary’	‘the water’

Since these languages have a common ancestor, Sabaot must have deleted stem-final vowels such as ghost /e/ *after* it split from Sengwer. Conversely, cognates such as those in (90) show that Endo (Zwarts 2003), for instance, kept certain stem-final consonants, while Sengwer deleted them.

(90)	Sengwer	cîi(c)	pèel-jʌ̂ʌ(ntʌ)
	Endo	cîic	pèel-jʌ̂ʌn(tʌ)
		person	elephants-SG
		‘person’	‘elephant’

Therefore, the deletion of certain segments stem-finally is a diachronic phenomenon which must have occurred several times, being triggered by different phonological conditions and affecting different segments. The exact triggers for such deletion events cannot be fully understood without a comparative study of the lexicon of South Nilotic which goes beyond the most recent reconstruction by Rottland (1982; 1989). However, the data suggests that syllable structure and tone were involved, therefore some preliminary hypotheses can be made.

As explained in Section 3.2, ghost material is limited to four phonological shapes, all of which occur stem-finally: C#, .V#, .CV# and C.CV#. However, the disappearance of these four shapes from the unmarked stems can be reduced to two deletion events. First, ghost sequences of the (.CV#) kind only occur after a preceding consonant, meaning that the deletion of the consonant onset was to avoid a complex consonantal coda, a constraint found in all of South Nilotic. For instance, the deletion of (.V#) from (91a) to (91b), would have produced the unallowed coda /ŋk/. To avoid this, the syllable onset was likely deleted in tandem with the stem-final vowel (91c).

(91)	a.	mwêeŋkʌ	b.	*mwêeŋk(ʌ)	c.	mwèeŋ(kʌ)	‘beehive’
		beehive		beehive		beehive	

Therefore, the deletion of .CV# sequences must have been part of the same deletion process as single vowels (.V#). Second, as explored earlier in this chapter, the C.CV# ghost sequences are the result of two deletion events, first the deletion of .CV# and then the deletion of C#. Therefore, we can say that although four phonological shapes were deleted, there were only two deletion events: an earlier one which deleted .(C)V# sequences and one which deleted C# segments.

Single coda consonants are the least common kind of ghost material in the data. They are mostly (though not exclusively) found after long vowels with a falling tone and are only present in high-frequency nouns such *tjêe(p)* 'girl', *cîi(c)* 'person' and *kwêɛ(s)* 'buck'. As mentioned in Section 3.1, some coda consonant ghost segments are found in the surface form when changes in tone are applied to mark the nominative case. This fact supports the hypothesis that the tone of the preceding vowels triggered the deletion of ghost material. However, this explanation does not work for all ghost coda consonants; only the most common of these, /s/, is found to surface in the nominative. The other ghost consonants only surface when suffixes are added (e.g., the definite suffixes or demonstrative suffixes). This suggests that either the historical deletion of /s/ had a more complex route that lead to the pattern seen today[8] or that this is an example of paradigmatic analogy, where speakers have generalised a tone rule that applies to vocalic ghost material to a more general rule that applies to the most common kind of consonant ghost material as well. Since nearly half (49%) of all the nouns in the data end in a consonant and only 8 high-frequency nouns and 1 high-frequency suffix contain a ghost consonant (see **Table 4**), we can assume that the deletion of final consonants was not a regular process.

On the other hand, all ghost material containing a vowel can be found in high-frequency as well as low-frequency nouns, after a variety of phonological contexts and after all three tones. However, although the preceding tone environment is not predictable itself, the tone sandhi interactions between the ghost segment and the following suffixes are partially predictable by

[8] By looking at the only noun with a ghost /s/ for which we can readily find cognates, we see that this is indeed the case. In example (iii), the Sengwer noun *kwêɛ(s)* is shown with its cognate forms *kwàɣá* in Pokoot (Crazzolara 1978) and *kwàrá* in Endo Marakwet.

(i)

	Sengwer	kwêɛ(s)	kwèɛs-tâ
	Pokoot	kwaɣa	kwəɣɛɛt
	Endo	kwara	kwara-ta
		buck	buck:SDF
		'buck'	'the buck'

Since /ɣ/ in Pokoot consistently corresponds to /r/ in Marakwet and /j/ in other Kalenjin varieties (Rottland 1982), this noun is a fairly aberrant, evidencing that it must have had an irregular development. Further investigation into the diachronic development of Kalenjin varieties would be required to understand this issue more in-depth. The Endo data comes from personal communication with Endo Marakwet Bible translators working with the BTL (Bible Translation & Literacy) organization.

the preceding context. If the preceding context is a low tone, the addition of a high vowel suffix to the ghost segment can yield either a level high tone or a falling tone on the sandhi syllable. Moreover, if the preceding context is a high tone or a falling tone, the addition of a high vowel suffix to the ghost segment always yields a falling tone on the sandhi syllable. In example (100), the singular definite suffix **-ít** is added to the low-toned monosyllabic noun *sòt(o)*; the ghost vowel and the suffix undergo sandhi and yield a high-toned vowel. However, when the plural definite suffix *-ík* is added to another low-toned monosyllabic noun *tʌ̀ʌk(-i)* in (101), the sandhi between the ghost vowel and the suffix yields a falling tone instead. On the other hand, both the high-toned noun *kwéen(u)* in (102) and the falling-toned noun *mbâr(a)* in (103) when suffixed with the singular definite suffix result in a falling tone on the sandhi vowel.

(100)	sòt(o)	+ ít	>	sòtéet
	gourd	+ SDF		gourd:SDF
	‘gourd’			‘the gourd’
(101)	tʌ̀ʌk(-i)	+ ík	>	tʌ̀ʌk-îik
	host-PL	+ PDF		host-PL:PDF
	‘hosts’			‘the hosts’
(102)	kwéen(u)	+ ít	>	kwéenûut
	middle	+ SDF		middle:SDF
	‘middle’			‘the middle’
(103)	mbâr(a)	+ ít	>	mbárɛ̂ɛt
	farm	+ SDF		farm:SDF
	‘farm’			‘the farm’

This patterning suggests that there is tonal contrast in vocalic ghost segments, at least for those preceded by a low tone, such as (100) and (101). To explain this kind of variation, authors have suggested that ghost vowels are underlyingly specified for tone (Kouneli 2021; Creider & Creider 1989). Although the hypothesis that all ghost segments are specified for tone is compelling, a full analysis of tone and tone sandhi interactions in Sengwer would be necessary to prove it.

Nevertheless, this patterning suggests that certain tone sequences influenced the deletion of stem-final short-vowelled open syllables. For instance, based on their sandhi behaviour, the underlying tone specification expected for (100) would be *sòt(ò)*—with a low-toned ghost vowel—while that of (101) would be *tʌ̀ʌk(-î)*—with a falling-toned ghost vowel. However, these two underlying patterns for nouns (CV̀.CV̀ and CV̀.CV̂), though not common, can be found in the dataset and the nouns in which they occur cannot be proven to be loanwords. Therefore, we must assume that the lexical tone specifications have shifted enough since this Proto-Kalenjin deletion process took place that it is no longer possible to predict the specific environment in which it took place.

Still, it is possible to make some informative observations. First, three out of four tone patterns (100, 102, 103) involve a lowering of pitch in the ghost syllable while the pattern in (101) only occurs in a handful of nouns which have the plural suffix *-î*. Since the latter almost exclusively triggers a high tone in the preceding syllable, making its environment almost always that of (102), we could consider these few cases of CV̀.CV̂ as exceptions in which the replacive high tone was blocked by the stem. Second, the deletion event did not affect nouns in the nominative case, where all targeted segments receive a high tone. Therefore, it appears that the lowering of pitch is another contributing factor in the deletion of ghost segments. This is not surprising considering that a lowering of pitch corresponds to a lowering in saliency of a particular unit.

In summary, these observations mean that, although it is not possible to outline the exact phonological environments which triggered them, there were two separate deletion events: (a) a first more widespread and regular process which targeted stem-final open syllables with a short vowel and a lower pitch than the previous vowel and (b) a second more restricted and irregular process which targeted stem-final consonants (particularly those preceded with a falling tone and a long vowel) in high-frequency morphemes.

In light of this, we argue that Sengwer ghost segments are a type of ghost segments which is lexically determined and not markedness-determined. While the deletion of Sengwer ghost segments can be traced back to phonological processes in linguistic history of Sengwer which were likely markedness determined, these are no longer active synchronically and their alternations have become fossilised. Therefore, building on Zimmermann's (2019), we can expand the definition as follows (the addition is highligted in *italics*):

> "Ghost segments are segments that (1) are idiosyncratically bound to specific morphemes and (2) alternate with zero in a way that the majority of segments within this language do not. *These alternations can either be determined lexically or be conditioned by phonological markedness constraints*."

This extension to the definition specifies the two scenarios in which ghost segments can be found, explicitly including both cases.

4 Conclusion

In this paper, we described the phenomenon of ghost segments in Sengwer nouns, including their phonology, behaviour and distribution in the lexicon. In particular, we showed that these latent segments can be part of roots and suffixes as well as whole suffixes which are morphologically and phonologically active. Following on from this description, we have demonstrated that ghost segments in Sengwer cannot be considered a case of insertion but rather one of deletion. While previous analyses have argued that these latent segments are elements added either as declension class suffixes or epenthetic forms used to avoid a constraint on vowel hiatus, we have presented

evidence that these segments are historically elided word-final elements. First, we showed that the phonology of Sengwer does not have any particular constraint against hiatus and that the latter is a common occurrence in the data. Then, by comparing cognate nouns of closely related languages, we presented evidence for the historical deletion of ghost vowels and segments. In light of this, we propose the use of the term *ghost* as a more accurate descriptor for the phenomenon at hand, one that allows us to integrate it within a wider group of similar linguistic features across the world's languages. While Sengwer ghost segments have their idiosyncrasies, parallels can be more easily drawn with other ghost phenomena than with cases of thematic suffixes. This analysis explains much of the segmental and suprasegmental irregularities found in the nominal morphology of Sengwer.

There are several avenue for further research which stem from the current paper. First, though this paper only focuses on nouns, ghost segments in Sengwer are found in at least three other lexical categories: verbs, adjectives and pronouns. However, compared to nouns, their role and variation are very limited, appearing in only a handful of items and mostly as /n/. Compare the forms in (a) and (b) of the verb *cóo(n)* in (119) and the adjective *múrjɔ̂ɔ(n)* in (110).

(104)	a.	cóo(n)	b.	ø-cóon-è
		come:IMP		3-come-IMPF
		'come!'		'he is coming'
(105)	a.	múrjɅ̂ʌ(n)	b.	múrjʌ́ʌn-èc
		dark.brown		dark.brown-PL
		'dark brown'		'dark brown'

Further research could focus on describing ghost segments in other lexical categories compared to those in nouns.

The current paper only briefly explores the role and patterning of tone in relation to ghost segments. In order to fully understand this phenomenon, however, further investigation into the patterning of tone in the language at large is required. In particular, the questions arising from the observations made here in relation to tone are: (a) Are all ghost segments specified for tone? If so, (b) how do ghost segments influence the tone patterns found in the inflected forms of the noun? And, (c) was tone one of the main factors in the deletion of ghost segments?

Finally, considering these stem-final segments as parts of roots in some cases, rather than always as thematic suffixes, could have important repercussions on the reconstructions of Proto-Kalenjin and, in a domino effect, on Proto-South-Nilotic and Proto-Nilotic. Further research could apply the present ghost segment analysis to other Kalenjin language varieties in language-specific or comparative studies in order to test the validity of current reconstructions and, if needed, amend them.

Abbreviations

1SG	1st person singular
3	3rd person
ADJ	deverbal adjectival suffix
CS	construct state suffix
DVB	deverbal nominalising affix
FEM	feminine prefix
GEN	genitive case
IMP	imperative
IMPF	imperfective aspect
LOC	locative prefix
MSC	masculine prefix
NOM	nominative case
PDF	plural definite suffix
PL	plural number
PXP	plural proximal demonstrative suffix
PXS	singular proximal demonstrative suffix
SDF	singular definite suffix
SG	singular number
THM	thematic suffix

Ethics and consent

The data collection methods were approved by the University of Edinburgh's ethics committee (PPLS Research Ethics Committee). The ID number for the project is 322-2223.

Funding information

This project was funded by the British Academy/Leverhulme Small Research Grant with the title *The sound system of Sengwer, an endangered language of Kenya (*SRG2223\231121).

Acknowledgements

First and foremost, we thank our Sengwer consultants Amos Kiprop, Emily Korir and Irine Kosgei. Without you this work would not have been possible. We also want to thank the Sengwer community at large for their hospitality, kindness and collaboration.

We are grateful to everyone who gave us their feedback and support, especially Roland Kießling, Patrick Honeybone, Pavel Iosad, Maarten Mous, Andrew Harvey, Bonny Sands, Jeroen van Ravenhorst, Alessandro Mercatelli, the *Glossa* editors and reviewers, and audiences at the University of Edinburgh's *P-workshop* and the *Rift Valley Network's Webinar*.

Competing Interests

The authors have no competing interests to declare.

References

Adda-Decker, Martine & Boula de Mareüil, Philippe & Lamel, Lori. 1999. Pronunciation variants in French: schwa and liaison. *International Congress of Phonetic Sciences*, 2239–2242. San Francisco.

Archangeli, Diana Bennett. 1984. *Underspecification in Yawelmani phonology and morphology*. PhD Thesis. MIT Press.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge: MIT Press. DOI: https://doi.org/10.2307/416331

Bennett, Patrick R. 1974. Tone and the Nilotic case system. *Bulletin of the School of Oriental and African Studies* 37(1). 19–28. In Memory of W. H. Whiteley. University of London. DOI: https://doi.org/10.1017/S0041977X00094817

Casali, Roderic F. 2008. ATR harmony in African Languages. *Languages and Linguistic Compass* 2(3). 496–549. DOI: https://doi.org/10.1111/j.1749-818X.2008.00064.x

Corbett, Greville G. 2000. *Number*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139164344

Crazzolara, Pasquale. 1978. *A study of the Pokot (Suk) language: Grammar and vocabulary*. Editrice Missionaria. Bologna.

Creider, Chet A. & Creider, Jane Tapsubei. 1989. *A grammar of Nandi*. Hamburg: Buske.

Crystal, David. 1980. *A first dictionary of linguistics and phonetics.* London: Andre Deutsch.

Di Garbo, Francesca. 2014. *Gender and its interaction with number and evaluative morphology*. PhD Thesis. Stockholm University.

Dimmendaal, Gerrit J. 1983. *The Turkana Language.* Dordrecht: Foris. DOI: https://doi.org/10.1515/9783110869149

Dimmendaal, Gerrit J. 2000. Number marking and noun categorisation in Nilo-Saharan languages. *Anthropological Linguistics* 42(2). 214–261.

Dimmendaal, Gerrit J. 2012. Metrical structures: A neglected property of Nilotic (and other African language families). *Studies in Nilotic Linguistics* 5. 1–26.

Dimmendaal, Gerrit J., & Breedveld, Anneke. 1986. Tonal influence on vocalic quality. In Koen Bogers & Harry van der Hulst & Maarten Mous (eds.) *The phonological representation of suprasegmentals*, 1–33. Dordrecht: Foris. DOI: https://doi.org/10.1515/9783110866292-002

Distefano, John Albert. 1985. *The precolonial history of the Kalenjin of Kenya: a methodological comparison of linguistic and oral traditional evidence.* PhD Thesis. University of California, Los Angeles.

Falletti, Federico. 2023a. *Sengwer Trilingual Dictionary*. Moi University Press.

Falletti, Federico. 2023b. *Sengwer Dictionary*. LivingDictionaries. Available at https://livingdictionaries.app/sengwer (accessed on the 5th of April 2024)

Faust, Noam & Torres-Tamarit, Francesc. 2017. Stress and final /n/ deletion in Catalan: Combining Strict CV and OT. *Glossa: a journal of general linguistics* 2(1). 63. DOI: https://doi.org/10.5334/gjgl.64

Greenberg, Joseph H. 1978. How does a language acquire gender markers. *Universals of Human Language* 3. 47–82.

Grimm, Scott. 2012. *Number and individuation*. PhD Thesis. Stanford University.

Grimm, Scott. 2018. Grammatical number and the scale of individuation. *Language* 94(3). 527–574. DOI: https://doi.org/10.1353/lan.2018.0035

Hartmann, Reinhard R. K. & Stork, Francis C. 1972. *Dictionary of language and linguistics*. New York: John Wiley & Sons.

Herreros Baroja, Tomás. 1989. Analytical grammar of the Pokot language. Kitapu ngala pökot nyo kikir. *Bibliotheca africana* 3. Trieste. DOI: https://doi.org/10.2307/415240

Hollis, Alfred Claud. 1909. *The Nandi: Their language and folklore.* London: Oxford University Press.

Kiparsky, Paul. 2003. Finnish noun inflection. In Nelson, Diane & Manninen, Satu (eds.), *Generative Approaches to Finnic and Saami Linguistics*, 109–161. CSLI Publications.

Kouneli, Maria. 2021. Number-based noun classification. *Natural Language & Linguistic Theory* 39. 1195–1251. DOI: https://doi.org/10.1007/s11049-020-09494-8

Kouneli, Maria. 2022. Inflectional classes in Kipsigis. *Glossa: a journal of general linguistics* 7(1). 1–33. DOI: https://doi.org/10.16995/glossa.8549

Larsen, Iver A. 1991. Sabaot noun classification. In Rottland, Franz & Omondi, Lucia Ndong'a (eds.), *Nilo-Saharan, Volume 6: Proceedings of the 3rd Nilo-Saharan Linguistics Colloquium*, 143–163. Köln: Rüdiger Köppe Verlag.

Lindsey, Kate Lynn. 2019. *Ghost elements in Ende phonology*. PhD Thesis. Stanford University.

McCarthy, John & Prince, Alan. 1994. The emergence of the unmarked: Optimality in prosodic morphology. *Proceedings of the North East Linguistics Society* 24. 333–379.

Mietzner, Angelika. 2016. *Cherang'any: a Kalenjin language of Kenya.* Rüdiger Köppe Verlag. Köln.

Moodie, Jonathan. 2016. Number marking in Lopit, an Eastern Nilotic language. In Payne, Doris L. & Pacchiarotti, Sara & Bosire, Mokaya (eds.), *Diversity in African languages*: *Selected papers from the 46th Annual Conference on African Linguistics*, 397–416. Berlin: Language Science Press.

Moodie, Jonathan. 2019. *A grammar of the Lopit language*. PhD Thesis. University of Melbourne.

Oltra-Massuet, Isabel & Arregi, Karlos. 2005. Stress-by-structure in Spanish. *Linguistic Inquiry* 36. 43–84. DOI: https://doi.org/10.1162/0024389052993637

Pierrehumbert, Janet B. 2016. Phonological representation: beyond abstract vs. episodic. *Annual review of linguistics* 2. 33–52. DOI: https://doi.org/10.1146/annurev-linguistics-030514-125050

Remijsen, Bert & Ayoker, Otto Gwado. 2020. Floating quantity in Shilluk. *Language* 96(3). e135–e156. DOI: https://doi.org/10.1353/lan.2020.0052

Rottland, Franz. 1981a. Marakwet dialects: Synchronic and diachronic aspects. In Kipkorir, Benjamin E. & Soper, Robert C. & Ssenyonga, Joseph W. (eds.), *Kerio Valley. Past, present and future. Proceedings of a seminar held in Nairobi at the Institute of African Studies, University of Nairobi, May 21–22, 1981*. 139–146. Nairobi: Institute of African Studies.

Rottland, Franz. 1981b. The Segmental Morphology of Proto-Southern Nilotic. In Bender, Lionel & Schadeberg, Thilo C. (eds.), *Nilo-Saharan Proceedings: Proceedings of the First Nilo-Saharan Linguistics Conference (1980)*, 5–18. Leiden, The Netherlands. DOI: https://doi.org/10.1515/9783110883466-003

Rottland, Franz. 1982. *Die Südnilotischen Sprachen.* Reimer. Berlin.

Rottland, Franz. 1989. Southern Nilotic Reconstructions. In Marvin Lionel Bender (ed.), *Topics in Nilo-Saharan Linguistics*, 219–231. Buske. Hamburg.

Rottland, Franz & Creider, Chet A. 1996. Datooga nominals: the morphologization of vowel harmony. *Afrikanistische Arbeitspapiere* 45. 257–268.

Schmidt, Deborah. 1994. Phantom Consonants in Basaa. *Phonology* 11(1). 149–178. DOI: https://doi.org/10.1017/S0952675700001871

Stump, Gregory T. & Finkel, Raphael A. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139248860

Szypra, Jolanta. 1992. Ghost segments in nonlinear phonology: Polish yers. *Language* 68. 277–312. DOI: https://doi.org/10.2307/416942

Toweett, Taaitta. 1975. *Kalenjin nouns and their classification.* MA thesis. University of Nairobi.

Toweett, Taaitta. 1979. *A Study of Kalenjin Linguistics.* Nairobi: Kenya Literature Bureau.

Tranel, Bernard H. 1996a. Exceptionality in optimality theory and final consonants in French. In Zagona, Karen (ed.), *Grammatical Theory and Romance Languages: Selected papers from the 25th Linguistic Symposium on Romance Languages*, 275– 291. Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/cilt.133.22tra

Zimmermann, Eva. 2019. Gradient Symbolic Representation and the Typology of Ghost Segments. *Proceedings of Annual Meetings on Phonology 2018*. Linguistic Society of America. DOI: https://doi.org/10.3765/amp.v7i0.4576

Zoll, Cheryl. 1993. Ghost segments and optimality. *Proceedings of The Twelfth West Coast Conference on Formal Linguistics,* 183–202.

Zwarts, Joost. 2003. *The phonology of Endo: a Southern Nilotic language of Kenya*. Utrecht Institute of Linguistics OTS.