Aethiopica 20 (2017)

International Journal of Ethiopian and Eritrean Studies

MAGDALENA KRZYŻANOWSKA, Universität Hamburg

Miscellaneous

A Part-of-Speech Tagset for Morphosyntactic Tagging of Amharic

Aethiopica 20 (2017), 210–235

ISSN: 1430-1938

Edited in the Asien-Afrika-Institut
Hiob Ludolf Zentrum für Äthiopistik
der Universität Hamburg
Abteilung für Afrikanistik und Äthiopistik

by Alessandro Bausi

in cooperation with

Bairu Tafla, Ulrich Braukämper, Ludwig Gerhardt,
Hilke Meyer-Bahlburg and Siegbert Uhlig

A Part-of-Speech Tagset for Morphosyntactic Tagging of Amharic*

MAGDALENA KRZYŻANOWSKA, Universität Hamburg

Introduction

Since 2014 the project TraCES: From Translation to Creation: Changes in Ethiopic Style and Lexicon from Late Antiquity to the Middle Ages has produced digital tools which allow for the collection, annotation, and display of Ethiopic texts.[1] The digital framework developed within the project for Gəᶜəz can be adapted for the analysis of written sources in other languages irrespective of their writing system and morphosyntactic structure.[2] The language which seems to be the first natural candidate for this is Amharic, a modern Ethiopian Semitic language, which has the same script as Gəᶜəz, as well as some morphosyntactic features, much vocabulary, and a shared history and culture.[3]

The first task in the computational processing of Amharic is to establish a part-of-speech tagset. Such a tagset has been defined and is currently in use for annotating Gəᶜəz texts.[4] However, it cannot be used for Amharic without making considerable changes that reflect the morphosyntactic nature of the language, which is quite different from that of Gəᶜəz. The aim

* This work was carried out within the TraCES project, funded by the European Research Council under the European Union's Seventh Framework Programme, grant agreement no. 338756.

1 For the description of the approaches and tools implemented for Gəᶜəz within the TraCES project see Vertan 2016, 33.

2 It has already been adapted to Epigraphic South Arabian, which is written in a different (but related) script to Gəᶜəz (Vertan 2016, 41).

3 I would like to express my heartfelt gratitude to Orin Gensler for providing insightful advice, criticism, and encouragement, and for the English proofreading of the last version of this article. I also thank the TraCES team members for their feedback; special thanks go to Susanne Hummel for constant support and for discussing many of the issues raised in this paper. Last but not least I am indebted to Maria Bulakh for some useful observations and to Denis Nosnitsin for reading through the article.

4 For an overview of the Gəᶜəz tagset see Hummel and Dickhut 2016.

of this article is to propose a tagset for the morphosyntactic tagging of Amharic, and to discuss it, especially those points where the author's decision may not be obvious or may seem problematic. The tagset presented below is meant to be a balanced one: not too complex for the annotators, and not so general as to slow down the user of the corpus. As in the TraCES project, it is intended that the manual annotation of texts will be performed with the help of the software programme GeTa.[5] Subsequently, the annotated texts will be searchable using ANNIS,[6] another programme that can search for any combination of tags and tokens. The ultimate target is to create an annotated corpus of Amharic texts which could in principle be used by linguists for both synchronic and diachronic investigation.

Thus far there have been two large-scale enterprises concerning the computational analysis of Amharic which are of some relevance to the present paper. The first initiative was undertaken by the Ethiopian Languages Research Center of Addis Ababa University within the project The Annotation of Amharic News Documents. The set of tags used for annotating the news documents is described by Girma Awgichew Demeke and Mesfin Getachew.[7] The process of verifying, correcting, and retagging of the corpus is treated by Gambäck.[8] The tagset was also involved, with two other tagsets, in a series of experiments testing three taggers.[9] The second initiative is HornMorpho,[10] a programme for segmenting Amharic, Oromo, and Təgrəññña nouns (treated together with adjectives) and verbs into morphemes and for generating them. According to Wintner,[11] it probably represents the state of the art for the morphological processing of Amharic.

The paper is organized as follows: in Section 1 the transliteration system for Amharic is presented; Section 2 provides some general issues concerning the morphosyntactic tagging of Amharic; Section 3 contains a table of the proposed Amharic tagset followed by a detailed presentation and discussion of the tags. These sections will be followed by the Conclusion.

[5] See Vertan 2016, 37–40.

[6] The name stands for 'ANNotation of Information Structure'. It is 'an open source, cross platform (Linux, Mac, Windows), web browser-based search and visualization architecture for complex multi-layer linguistic corpora with diverse types of annotation', see http://corpus-tools.org/annis/.

[7] Girma Awgichew Demeke and Mesfin Getachew 2006.

[8] Gambäck 2012.

[9] See Gambäck et al. 2009.

[10] Gasser 2011. The analyser is described in detail in Gasser 2012.

[11] Wintner 2014, 52.

1 The Transliteration System for Amharic

The tokenization and annotation of Amharic texts will be done on transliterated word-forms. The transliteration system proposed here basically conforms to that adopted for Gəᶜəz[12] but it follows the somewhat different representation of two vowels (first and fourth order) as found in the main grammatical and lexicographical works on Amharic and agrees with actual Amharic pronunciation.[13] In our system each consonantal grapheme receives a separate symbol even though it may be homophonous with other graphemes. As is well known, several Amharic consonants can be represented by more than one symbol: for instance, the characters ሀ, ሐ, ኀ all stand for the same sound [h] while ሠ, ሰ stand for the same sound [s]. Provided that we keep the distinction between the various signs by transliterating ሀ, ሐ, ኀ as [h], [ḥ], [ḫ] respectively, and ሠ, ሰ as [ś], [s] respectively, the corpus can also be used for research on Amharic spelling practices. To distinguish between etymological glottal አ [ʾa] and pharyngeal ዐ [ᶜa] only the sign ᶜ in front of the pharyngeal will be used; አ will have no mark. The labialized consonant will be indicated with superscript ʷ.

The parallels with the Gəᶜəz transliteration system break down in the case of the first- and the fourth-order vowels. The representation of the first-order vowel as [a] and the fourth as [ā], as has been adopted for Gəᶜəz, would only serve to obscure their pronunciation. Instead they will be represented by the symbols [ä] for the first and [a] for the fourth order, which correspond to Amharic pronunciation. Exceptional in this regard are laryngeals, for which the difference between the first-order vowel and fourth-order vowel is neutralized in favour of the fourth order, so that, in both cases, they are pronounced as [a]. However, to preserve the graphic distinction between them, our transliteration system will use the same symbols as for 'plain' consonants, that is [ä] and [a], respectively.

The table below shows the seven varieties of the characters ሀ *hä* (laryngeal) and ለ *lä* (non-laryngeal) arranged in the traditional order and transliterated according to the system adopted herein.

[12] Hummel and Dickhut 2016.
[13] Hartmann 1980, Leslau 1995, Kane 1990a, Kane 1990b.

Table 1 **Seven Varieties of the Characters ሀ and ለ**

			Orders			
First	Second	Third	Fourth	Fifth	Sixth	Seventh
ሀ *hä*	ሁ *hu*	ሂ *hi*	ሃ *ha*	ሄ *he*	ህ *h(ə)*	ሆ *ho*
ለ *lä*	ሉ *lu*	ሊ *li*	ላ *la*	ሌ *le*	ል *l(ə)*	ሎ *lo*

2 Morphosyntactic Tagging for Amharic: General Issues

Amharic morphology is very complex and highly agglutinative. Amharic word-forms may consist of as many as six morphemes (**በምትፈልገውን** *bä-mmə-ttə-fälləg-äw-ən* 'by the one [acc.] that she wants/you$_{M}$ want'). Thus, in the process of manual annotation, a given word-form (when necessary) must first be tokenized, that is, divided into a linear sequence of morphological pieces, each of which is given its own tag. In some cases, such a morphological piece may actually consist of several morphemes which for convenience are not separated. Each token is then linked to a lemma taken from the lexicon, representing its underlying form.[14]

The tagset presented below includes traditional parts of speech (POS) as well as markers of inflectional or (rarely) derivational categories.[15] Some of these are affixes, some are clitics: they will be given the cover-term 'Bound Grammatical Morphemes'. For the sake of convenience we will use the term 'parts of speech' in the traditional way. Apart from the POS tag, some classes of words take values (features) from the appropriate grammatical categories. For instance, the personal pronoun takes values from the categories of person, number, gender and politeness. Thus, each token will be character-

[14] The choice of the existing dictionaries as a basis for the lexicon will not be discussed in this paper. It seems, however, that the best would be the dictionary by Kane (Kane 1990a, 1990b), the most comprehensive bilingual Amharic–English dictionary. Because of the 'flat' structure of its entries (which can also be seen as its main drawback) it simply provides a list of words. Annotators are not obliged to conform to the lexicographer's assignment of a certain word to a given part of speech, as in Gankin's Amharic–Russian dictionary (Gankin 1969). Although in some cases it might be helpful for the team to follow the lexicographer's decision, in other cases it may generate inconsistencies in the description of word classes. Kane's dictionary does not provide entries for some bound grammatical morphemes and therefore, for the sake of tagging, the lexicon must be completed with them.

[15] The tagset draws upon the Amharic grammars written by Hartmann (Hartmann 1980) and Leslau (Leslau 1995).

ized by an underlying form, a tag by which it is assigned to an appropriate POS and, if applicable, to inflection features.

The system of annotation employed here basically ranks form over function. However, there are still many cases of form–function discrepancy and in some such cases one must rank function over form. This is most obviously true in cases of grammaticalization. Often the source morpheme may remain unchanged in form, but its morphosyntactic properties make it clear that a category change has occurred. In such cases the token will be tagged as the grammaticalized category. Thus, for instance lexemes of nominal origin functioning as postpositions will be treated as belonging to the class of postpositions rather than to nouns. This is because they cannot be pluralized and have no gender.

At present there are no technical means to account for the nature of compound or totally reduplicated words of any type at the level of morphosyntactic annotation, thus, their constituents are usually tagged separately (for detailed solutions see the appropriate sections below). However, they can be treated as compounds at a higher level of analysis. Compounds will be labelled as 'MW' which stands for a 'MultiWord', a term taken from the British National Corpus.

When consonants come together across a morpheme boundary they are very often separated by an epenthetic vowel (*ə* or *ä*). This vowel will be assigned arbitrarily to the second morpheme.

3 Tagset

Table 2 gives a synopsis of Amharic POS. It contains forty-seven tags grouped in twelve POS.

Table 2 **Tagset of Amharic POS**

POS	Tag	Full name	Example	Features
Nominals	NProp	Proper Names (personal and geographical names)	**ደሳለኝ ሕይወቴ** *Dässaläññ Ḥəywäte*	Number, Gender
		Acronyms of organizations	**ተመድ** (**የተባበሩት መንግሥታት ድርጅት**) *Tämäd* (*Yä-täbabbärut mängəśtat dərəǧǧət*)	
	NCom	Common Nouns (e.g. names of newspapers, magazines, bars, restaurants, institutions, companies, languages)	**ውሻ** *wəśśa*, **ተሞክሮ** *tämokro*, **ደግ** *dägg*, **አስቸጋሪ** *äsčäggari*, **አዲስ አድማስ** *äddis ädmas*, **ማለዳ ካፌ** *maläda kafe*	
	NAbr	Abbreviations	**ዶ/ር** *do/r*, **ወ/ሮ** *wä/ro*	
Article	Art	Definite Article	*-u*, *-wa*	Gender
Pronouns	PPer	Independent Personal Pronouns	**እኔ** *əne* **እርስዎ** *ərswo*	Person, Number, Gender & Politeness
	POobj	Object Suffix Pronouns	[C+] *-əññ*, *-äññ* [V+] *-ññ*	
	PPoss	Possessive Pronouns	[C+] *-e* [V+] *-ye*	
	PDem	Demonstrative Pronouns	**ይህ** *yəh*, **ያ** *ya*	
	PInter	Interrogative Pronouns	**ምን** *mən*, **ማን** *man*	No features
	PIndef	Indefinite Pronouns	**ምን** *mən*, **ማን** *man* **ማንኛ**- *mannəññä-*	
	PRef	Reflexive Pronouns	**ራስ**- *ras-*, **የገዛ** *yägäzza*	
	PRec	Reciprocal Pronouns	**እርስ በርስ** *ərs bärs*	
	PRel	Relative Pronouns	*yä-* *yämmə-*, *əmm-*	Perfective, Imperfective
Verbs	V	Verbs	**ይመጣል** *yəmäṭ-all*	Person, Number, Gender & Politeness Verbal forms
	VN	Verbal Nouns	**መሥራት** *mäśrat*	
	Ideo	Ideophones	**ብድግ** (**አለ**) *bədəgg* (*älä*)	
	Aux	Auxiliaries	*-allä* **ነበር** *näbbär*	Type of auxiliary Person, Number, Gender & Politeness
Quantifiers	NumCard	Cardinal Numerals	**አንድ** *änd* **ሁለት** *hulätt*	Gender Numeral symbol
	NumOrd	Ordinal Numerals	**ሁለተኛ** *hulättäññä*	Numeral symbol
	QuanInter	Interrogative Quantifiers	**ስንት** *sənt*, **ስንተኛ** *səntäññä*	No features
	QuanIndef	Indefinite Quantifiers	**አንዳንድ** *ändand*, **ብዙ** *bəzu*	

Table 2 **Tagset of Amharic POS (cont.)**

POS	Tag	Full name	Example	Features
Adpositions	Prep	Prepositions	*bä-*, *lä-*	No features
	Post	Postpositions	**ኋላ** *ḫ^{w}ala*, **ሥር** *śər*	
	PrepEmbObj	Embedded Prepositional Objects	*-bb-*, *-ll-*	
	PrepGen	Genetive Preposition	*yä-*	
	PrepDistr	Distributive Preposition	*əyyä-*	
	PostCpd	Compound Postposition	**በላይ** *bälay*	
Conjunctions	Conj	Conjunctions	*lə-*, **እና** *ənna* *bə-* + IPFV + *-mm*	No features
Adverbs	Adv	Adverbs	**አሁን** *ähun*, **ነገ** *nägä*, **በጣም** *bäṭam*, **ውስጥ** *wəsṭ*	No features
	AdvInter	Interrogative Adverbs	**የት** *yät*, **መቼ** *mäče*	
	AdvIndef	Indefinite Adverbs	**የትም** *yät-əmm*, **መቼ-ም** *mäče-mm*	
Particles	Part	Particles	**ብቻ** *bəčča*, **እንጃ** *əngǧa*, **ለካ** *läkka*	No features
	PartInter	Interrogative Particle	**ወይ** *wäy*	
Interjections	Interj	Interjections	**አሰይ** *əssäy*	No features
Bound Grammatical Morphemes	Acc	Accusative Marker	*-n*	No features
	Ass	Assertative	*-a*	
	AdvLiser	Adverbializers	*-u*, *-wə/un*, *-wan*, *-nu*	
	End	Endearment Marker	*-yye*	
	Foc	[Contrastive] Focus Marker	*-mm*	
	Inter	Interrogative (for polar questions)	*-n*	
	Indzr	Indefinitizers	*-mm*	
	Neg	Negative	*-mm*, *al-*	
	PlAs	Associative Plural Marker	*ənnä-*	
	PlEx	External Plural Marker	*-očč*	
	Pres	Presentative	*-nna*	
	Top	Topic Marker	*-mma*, *-ss*, *-ssa*	
Foreign words	For	Foreign words used to express or explain a certain meaning (for neologisms)		No features

3.1 Nouns (N)

The class of nouns (broadly construed) embraces nouns (Common Nouns and Proper Names) and adjectives. Adjectives have been subsumed under this class because they cannot be distinguished from nouns based solely on morphological criteria. However, they form a distinct class if we take their syntactic behaviour into account: they cannot occur as the subject of a clause unless they take the definite article.[16] When dealing with the syntactic representation of the corpus, a decision will have to be made as to whether there is a need to keep nouns and adjectives apart.

In principle, nouns take the features number and gender. Usually plurality is marked by a suffix *-očč* which is tokenized as such. On the other hand, plurality will be treated as an inherent feature of a noun as long as the plurality marker is non-productive and/or non-affixal; this is limited to a certain group of nouns:

1) nouns of Gəᶜəz origin ending with plural *-an* or *-at*, such as **መምህራን** *mämhər-an*, **ሕጻናት** *ḥəṣan-at*;
2) nouns of Gəᶜəz origin having broken plural forms, as **ደናግል** *dänagəl*;
3) Amharic nouns and adjectives in which plurality is marked by reduplication of consonants, for instance **ወይዛዝርት** *wäyzazərt*;
4) Amharic nouns taking markers that indicate both plurality and a social bond between people, such as *-amač*, *-am-*, as in **ወንድማማች** *wändəmm-amač*, **ጓደኛሞች** *gʷaddäññ-am-očč*;
5) Amharic reduplicated nouns with the first member followed by the vowel *-a*, conveying the meaning ‘all kinds of’, for example **ቅመማ ቅመም** *qəmäma qəmäm*.

The plurality values of the nouns in (1) and (4) will receive the label ‘Pl’ (Plural) while those in (2) and (3) will be labelled as ‘PlIn’ (Internal Plural). Reduplicated nouns as in (5) will receive the label ‘PlRed’ (Reduplicated Plural). Frequently the internal changes within the noun that indicate plurality are accompanied by an external plural marker, as in **መጻሕፍት** *mäṣaḥəft* with the *-t* ending. These will not, however, be shown separately in our analysis.

The two types of productive plural marker of Amharic, the external *-očč* and the associatives *ənnä-* and (obsolete) *əllä-* (and only these), will be treated as separable elements under the archcategory of Bound Grammatical Morphemes, and will be tokenized as such. The near ubiquitous plural marker *-očč*, which also occurs with other Amharic POS, such as indefinite pronouns and cardinal numerals, will receive a tag ‘PlEx’ (External Plural).

[16] Cotterell 1964, 36, n. 11.

Furthermore, this most common external plurality marker *-očč* can occasionally be suffixed to the plurality markers mentioned in (1) and (4) such as **ቃላቶች** *qalat-očč* and **ወንድማማቾች** *wändəmmamač-očč*. The associative *ənnä-*, labelled 'PlAs' (Associative Plural), occurs with both nouns and personal and interrogative pronouns. The associative marker in personal pronouns (**እናንተ** *ənnantä* 'you'$_{PL}$, **እነሱ** *ənnäsu* 'they') and demonstrative pronouns (**እነዚህ** *ənnazzih* 'these', **እሊህ** *əllih* 'these') will not be tokenized as a separate element but treated together with the remaining morpheme as a single lexical whole.

All nouns taking the external plural marker or associative will be tagged as 'Base form' (Bf) in respect to the category of number. If a noun has no plurality marker it will be assigned the value 'Singular' (Sg) from the category of number.

The vast majority of Amharic nouns are by default masculine. The feminine gender is lexically expressed by a limited number of nouns and is occasionally used, syntactically, for nouns to indicate diminutive or endearment. The category of gender can be specified on the noun in the following ways:

1) by nature (N), when a noun refers to a person or an animal of one of the two biological sexes, for instance **ወንድም** *wändəmm*, **እኅት** *əḫət*. These will be tagged as M.N., F.N.;
2) by pattern (P), meaning by the morphological form of the noun, such as **ኢትዮጵያዊ** *ityopyawi*, **ኢትዮጵያዊት** *ityopyawit*, **አሮጌ** *äroge,* **አሮጊት** *ärogit*. These will be tagged as M.P., F.P.;
3) syntactically (S), (a) by a preceding modifier, such as **ተባት አህያ** *täbat əhəyya*, **አንስት አህያ** *ənəst əhəyya*; by agreement with the definite article or demonstrative pronoun as in **ተማሪው** *tämari-w,* **ተማሪዋ** *tämari-wa*, **ይህ ተማሪ** *yəh tämari*, **ይቺ ተማሪ** *yəčči tämari*; (b) by agreement in the predicate, for instance **ፀሐይ ወጣ** *ṗäḥay wäṭṭa*, **ፀሐይ ወጣች** *ṗäḥay wäṭṭa-čč*; (c) by agreement in the verbal noun as in **ጓደኛዬ መምጣቱ** *g^{w}addäññaye mämṭatu*, **ጓደኛዬ መምጣቷ** *g^{w}addäññaye mämṭat^{w}a*. These nouns will be tagged as M.S., F.S.

All these kinds of gender specification will be considered as inherent features of the noun. It is possible that a noun be assigned gender both by N or P and by S. Then the tag will be (for instance) M.N., S.

Amharic has a group of compound nouns, most often of Gəᶜəz origin, whose first component (the Head Noun) takes the Gəᶜəz construct state morpheme *-ä* or, in some cases, morphological zero. There is also a small group of compounds which preserve the Amharic word order but nonetheless are built according to the construct pattern inherited from Gəᶜəz with *-ä* on the first element (the Dependent Noun), such as **አገረ ገዥ** *ägärä gäž*

'governor', **አፈ ታሪክ** *äfä tarik* 'oral history'. For both types of compound noun, the first component will be assigned the tag 'Construct State' (ConSt).

Table 3 **Nouns**

	N				
	Singular (Sg)	Plural (Pl)	Internal Plural (PlIn)	Reduplicated Plural (PlRed) (MW)	Base form (Bf)
Number	**ድመት** *dəmmät* **ግጥም** *gəṭəm* **ደግ** *dägg* **እውነተኛ** *əwnätäññä* **ሰው** *säw* **መገናኛ** *mäggänaññä* **አጻጻፍ** *äṣṣaṣaf* **ፈላጊ** *fällagi*	**ኢትዮጵያውያን** *ityopyawəyan* **ቅዱሳት** *qəddusat* **ወንድማማች** *wändəmmamač* **ጓደኞች** *gʷaddäññam-očč*	**ከዋክብት** *käwakəbt* **መጻሕፍት** *mäṣaḥəft* **ወይዛዝር** *wäyzazər* **ደጋግ** *däggag*	**ጨርቃ ጨርቅ** *č̣ärqa č̣ärq* **ፍራ ፍሬ** *fəra fəre*	**ሴቶች** *set-očč* **እነራስ** (**አሊ**) *ənnä-ras* (*Äli*)
Gender	**Masculine (M)**			**Feminine (F)**	
	By nature: **ጐረምሳ** *gʷärämsa*, **ኮረዳ** *korädda*, **ወይፈን** *wäyfän*, **ጊደር** *gidär* By pattern: **ቅዱስ** *qəddus*, **ቅድስት** *qəddəst*, **ክቡር** *kəbur*, **ክብርት** *kəbərt* **አሮጌ** *äroge*, **አሮጊት** *ärogit*, **ደግ** *dägg*, **ደጊት** *däggit* By syntax: **ወንድ ዶሮ** *wänd doro*, **ሴት ዶሮ** *set doro* (preceding modifier) **አበባው** *äbäba-w*, **አበባዋ** *äbäba-wa* (agreement with the definite article) **ይህ አበባ** *yəh äbäba*, **ይቺ አበባ** *yəčči äbäba* (agreement with the demonstrative pronoun) **ጨረቃ ወጣ** *č̣äräqa wäṭṭa*, **ጨረቃ ወጣች** *č̣äräqa wäṭṭa-čč* (pronoun on the verb) **ጓደኞዬ መምጣቱ** *gʷaddäññaye mämṭatu*, **ጓደኞዬ መምጣቷ** *gʷaddäññaye mämṭatʷa* (pronoun as possessive on the verbal noun)			
Construct State	(ConSt) (MW) **ቤተ መጻሕፍት** *betä mäṣaḥəft* **ሥነ ጥበብ** *śənä ṭəbäb* **ልበ ሰፊ** *ləbbä säffi*				

3.2 Definite Article (Art)

The Amharic definite article takes the category of gender: M *-u*, or *-(ə)w*; F *-wa*, *-itu*, or *-it^{w}a* (see Table 4). There is no indefinite article per se. The words **አንድ** *änd*, **አንዲት** *ändit* 'one' can be used as an indefinite article, but this is not mandatory; hence it suffices to abbreviate the category as simply 'Art'. The feminine definite articles *-itu*, *-it^{w}a/-ət^{w}a* might be analysed as the morpheme *-it* (feminine adjectival morpheme) followed by the masculine or feminine definite article; for the sake of simplifying the annotation this has not been done. The morphemes *-(ə)yye*, *-əyyo* that Leslau treats as '[t]he definite article with "man, woman" and kinship terms'[17] will be considered here as a derivational morpheme (belonging to the lexicon) and as such will not be given a separate tag.

The masculine definite article is homophonous with a derivational morpheme that turns nouns into adverbs (see 3.11.1).

Table 4 **Definite Article**

	Art	
	Masculine (M)	**Feminine (F)**
Gender	[C+] *-u*	[C+] *-wa*, *-itu*, *-it^{w}a/-ət^{w}a*
	[V+] *-(ə)w*	[V+] *-wa*, *-yətu*, *-yət^{w}a*
	[o/u+] *-t*	

[17] Leslau 1995, 160–161.

3.3 Pronouns

3.3.1 Independent Personal Pronouns (PPer)

Table 5 **Independent Personal Pronouns**

		PPer		
		Gender & Politeness	Number	
			Singular (Sg)	Plural (Pl)
Person	1	Communis (C)	**እኔ** *əne*	**እኛ** *əññа*
	2	Masculine (M)	**አንተ** *äntä*	**እናንተ** *ənnantä*
		Feminine (F)	**አንቺ** *änči*	
		Polite (Pol)	**እርስዎ** *ərswo*, **አንቱ** *äntu*	
	3	Masculine (M)	**እሱ** *əssu* (**እርሱ** *ərsu*)	**እነሱ** *ənnäsu* (**እነርሱ** *ənnärsu*)
		Feminine (F)	**እሷ** *əssʷa* (**እርሷ** *ərsʷa*)	
		Polite (Pol)	**እሳቸው** *əssaččäw* (**እርሳቸው** *ərsaččäw*)	

3.3.2 Object Suffix Pronouns (PObj)

Table 6 **Object Suffix Pronouns**

		PObj		
		Gender & Politeness	Number	
			Singular (Sg)	Plural (Pl)
Person	1	Communis (C)	[C +] *-əññ*, *-äññ*, [V +] *-ññ*	[C +] *-ən*, *-än* [V +] *-n*
	2	Masculine (M)	[C +] *-əh*, [V +] *-h*	[C/a +] *-aččəhu*
		Feminine (F)	[C +] *-əš*, [V +] *-š*	[e/i +] *-yaččəhu*
		Polite (Pol)	*-wo*, *-wot*, *-əwo*, *-əwot*	[u/o +] *-waččəhu*
	3	Masculine (M)	[C +] *-əw*, *-äw* [V +] *-w*, *-t*, *-ət*	
		Feminine (F)	[C/a +] *-at*	[C/a +] *-aččäw*
			[e/i +] *-yat*, [u/o +] *-wat*	[e/i +] *-yaččäw*
		Polite (Pol)	[C/a +] *-aččäw* [e/i +] *-yaččäw* [u/o +] *-waččäw*	[u/o +] *-waččäw*

3.3.3 Possessive Pronouns (PPoss)

The possessive pronouns serve primarily to express a possessor to a noun. They occur also in partitive constructions with indefinite quantifiers, where they express the superset of which a certain quantity is a part (see 3.5.4).

Table 7 **Possessive Pronouns (Suffixes)**

		PPoss		
		Gender & Politeness	Number	
			Singular (Sg)	Plural (Pl)
Person	1	Communis (C)	[C+] *-e*, [V+] *-ye*	[C/a +] *-aččən* [e/i] *-yaččən* [u/o +] *-waččən*
	2	Masculine (M)	[C+] *-əh*, [V+] *-h*	[C/a +] *-aččəhu* [e/i +] *-yaččəhu* [u/o +] *-waččəhu*
		Feminine (F)	[C+] *-əš*, [V+] *-š*	
		Polite (Pol)	*-wo*, *-wot*	
	3	Masculine (M)	[C+] *-u*, [V+] *-w*	[C+] *-aččäw* [e/i +] *-yaččäw* [u/o +] *-waččäw*
		Feminine (F)	*-wa*	
		Polite (Pol)	[C/a +] *-aččäw* [e/i +] *-yaččäw* [u/o +] *-waččäw*	

3.3.4 Demonstrative Pronouns (PDem)

Demonstrative pronouns take values from the grammatical categories of number, gender and politeness, and distance. If a demonstrative pronoun beginning with *y-* is preceded by a preposition, its form changes into an allomorph beginning with *-zz-* (e.g. *yəh* > *-zzih*). Such a demonstrative will be tokenized and be given the same tag as the canonical form of the demonstrative. As mentioned above (see 3.1), the plural forms of the demonstratives will not be further analysed into the associative plus an appropriate singular demonstrative pronoun, but will be taken as a whole lexical unit.

Demonstrative pronouns with the adjectival ending *-ñña* (always followed by a nominalizing definite article) will be considered as canonical demonstratives and analysed as follows: **ይኽኛው** *yəḵäññä-* + Art.M.

Demonstrative pronouns of the form **ይኽውና** *yəḵäwəmna* **ያቺትና** *yaččatəmna* are analysed as consisting of the demonstrative pronoun followed

by the object pronoun and the presentative -(*ə*)*nna*, for instance *yəḵ-äw-ənna* PDem-PObj-Pres.

Table 8 **Demonstrative Pronouns**

		PDem	
	Gender & Politeness	Distance	
		Proximative (Prox)	Distal (Dist)
Singular	Masculine (M)	**ይህ** *yəh*, **ይሄ** *yəhe*, **ይኼ** *yəḵe* **ይኸኛው** *yəḵäñña-* + Art.M	**ያ** *ya* **ያኛው** *yañña-* + Art.M
	Feminine (F)	**ይህች** *yəhəčč*, **ይህቺ** *yəhəčči*, **ይች** *yəčč*, **ይቺ** *yəčči*, **እች** *əčč*, **እቺ** *əčči* **ይቺኛዋ** *yəččiñña-* + Art.F	**ያቺ** *yačči*, **ያች** *yačč* **ያችኛዋ** *yaččəñña-* + Art.F
	Polite (Pol)	**እኒህ** *ənnih*, **እኚህ** *əññih* **አሊህ** *əllih*	**እኒያ** *ənniya*, **እኚያ** *əññiya* **አሊያ** *əlliya*, **እኛ** *əñña*
Plural		**እነዚህ** *ənnäzzih*, **እነኚህ** *ənnäññih*, **እነኝህ** *ənnäññəh*, **አሊህ** *əllih*, **አለዚህ** *əlläzzih* **እኒህ** *ənnih*, **እኚህ** *əññih* **እነዚህኞቹ** *ənnäzzihəññ-* + PlEx + Art.M	**እነዚያ** *ənnazziya*, **እነዝያ** *ənnazzəya* **እነዛ** *ənnazza*, **እነኛ** *ənnäñña* **እነኚያ** *ənnäññiya*, **እኒያ** *ənniya*, **እኛ** *əñña* **እነዚያኞቹ** *ənnäzziyaññ-* + PlEx + Art.M

3.3.5 Interrogative Pronouns (PInter)

Amharic has the following basic interrogative pronouns: **ማን** *man*, **ምን** *mən*, **ምንድን** *məndən*, **ምንድር** *məndər*, **ማ** *ma*, **ምንኛ** *mənəñña*. They can be pluralized by means of the plural marker *-očč* or the associative *ənnä-*. There are also interrogative pronouns with the adjectival ending *-ñña* (**ማንኛ**- *mannəñña-*), of the form **ማናቸ**- *mannačča-*, and an interrogative pronoun of the form **የት**- *yät-*. These are always followed by the definite article or the possessive pronoun, which serve as nominalizers.

Table 9 **Selected Interrogative Pronouns**

PInter	
ማንኛው/ዋ *mannəññа-* + Art	**ማንኛዎቹ** *mannəññа-* + PlEx + Art.M
ማናቸው/ዋ *mannaččä-* + Art	**ማናቸዎቹ** *mannaččä-* + PlEx + Art.M
የትኛው/ዋ *yätəññа-* + Art	**የትኛዎቹ** *yätəññа-* + PlEx + Art.M
የቱ/ዋ *yät-* +Art	**የቶቹ** *yät-* + PlEx + Art.M
ማንኛችን/ችሁ/ቸው *mannəññа-* + PPoss.Pl	

3.3.6 Indefinite Pronouns (PIndef)

Indefinite pronouns embrace lexemes of the same form as interrogative pronouns but occur in a declarative sentence (which is not an indirect question) of the type **ማን እንደሚመጣ አላውቅም** *man əndämmimäṭa älawqəmm* 'I don't know *who* will come'.

Another type of indefinite pronoun consists of an interrogative pronoun followed by an element *-əmm* glossed here as 'Indefinitizer' (Indzr).[18] For instance: **ማን-ም** *mann-əmm* 'whoever' analysed as PInter + Indzr; **ማንኛ-ው-ም** *mannəññа-w-əmm* 'any' analysed as PInter + Art + Indzr.

A third type of indefinite pronoun with the specific meaning 'so-and-so' consists of the lexemes **እንትን** *əntən*, **እንተን** *əntän*, **እንትና** *əntəna*, **እገሌ** *əgäle*, **እገሊት** *əgälit*. They will be tagged as indefinite pronouns (PIndef). Additionally, the lexemes *əgäle*, *əgälit* will take the value of gender by pattern.

3.3.7 Reflexive Pronouns (PRef)

Items with the functions of both reflexive pronouns and pronouns of insistence (also of the possessive kind 'my own') are included here. The reflexive pronouns are grammaticalized body part terms, thus, the lexemes **ራስ**- *ras-* 'head', **ቅል**- *qəl-* 'skull' and **እጅ**- *əǧǧ-* 'hand' followed by possessive pronouns are considered to be reflexive pronouns. The pronoun *əǧǧ-* is always preceded by **የገዛ** *yägäzza*, originally a relative verb, which will also be tagged as a reflexive pronoun because it has been lexicalized in this function. The same *yägäzza* may also accompany the pronoun **ራስ**- *ras-*.

[18] This *-əmm* is homophonous with the negative marker *-əmm*. I do not know if there is any connection. The status of the various clitics *-əmm* in Amharic is a much discussed and much debated topic.

3.3.8 Pronouns of Reciprocity (PRec)

The pronoun of reciprocity consists of two components **እርስ በርስ** *ərs bärs* (= *ərs bä-(ə)rs*), alternatively **እርስ በራስ** *ərs bäras,* each to be annotated as PRec. A plural possessive pronoun can be added to the second component, **እርስ በርስ**- *ərs bärs-* + PPoss.Pl (e.g. **እርስ በርሳቸው** *ərs bärs-ačč̣äw*). Each of the two components will be entered separately in the lexicon.

3.3.9 Relative Pronouns (PRel)

The relative pronoun is invariant for number and gender. Different forms exist for the perfective and imperfective; *-mmə-* is difficult to analyse and is treated here as an integral part of the imperfective relative pronoun.

Table 10 **Relative Pronouns**

	PRel
Perfective (PRel.Pfv)	*yä-*
Imperfective (PRel.Ipfv)	*yämmə-*, *əmmə-*

3.4 Verbs

The verbal system of Amharic is quite complex because of the existence of composite and compound verbs as well as auxiliaries. Additionally, a composite verb can itself also be a compound, which complicates the system of annotation even further.

The composite verbs consist of a 'fixed root',[19] being onomatopoeic, primary or derived from a verb, followed by a conjugated form of **አለ** *älä* 'say', **አደረገ** *ädärrägä* 'do' or **አሰኘ** *ässäññä* 'cause'. For the 'fixed root' the term 'ideophone' will be used irrespective of its form and origin. The second component behaves like a normal verb. Hence composite verbs will be tagged as Ideo + V. As noted, the 'V' can be a compound, which will require another layer of tagging.

By compound verbs we understand verbal forms, imperfective or gerund, to which the auxiliary verb *-allä* is attached as a bound morpheme (in the 3MSg reduced to *-all*). This *-allä* will be considered as a morpheme in its own right and it will be tokenized as such and tagged as Aux (+ per-

[19] Leslau 1995, 580.

son/number). Other auxiliaries, like **ነበር** *näbbär*, stand alone. Auxiliaries are treated together as a special subclass of verb, separate from the standard verbs because of special features like having a frozen form (**ይሆናል** *yəhonall*) or a reduced frozen form (**ነበር** *näbbär*). They have to be tokenized and separated from the main verb because an object pronoun, if there is one, will intervene between the main verb and the auxiliary; additionally, the auxiliary can occur on its own (i.e. not in a compound verb), and, in that case, it must qualify as an independent token.

There is a problem with analysing and tagging compound imperfective forms with *-allä* because these are sometimes slightly different from the transparent combination of imperfective followed by *-allä;* in particular,

1) 3MSg = **ይሰብራል**, *yəsäbr-all*, not **yəsäbr-allä*;
2) 3Pl = **ይሰብራሉ**, *yəsäbr-allu*, not **yəsäbru-allu*;
3) 2Pl = **ትሰብራላችሁ**, *təsäbr-allaccəhu*, not **təsäbru-allaccəhu*.

In other words, in the compound imperfective, the piece *yəsäbr-* does not code number at all, but only 3M: number is specified in the auxiliary (Sg *-all*, Pl *-allu*). Similarly *təsäbr-* is not 2Pl but only 2M (vs *təsäbri*, 2FSg). We will code this here by introducing a third value for the category 'Number', namely 'Unmarked' (Um). Thus the form *yəsäbr-* will be tagged 3MSg in the simple imperfective paradigm, but 3MUm in the compound imperfective paradigm. However, this only happens when the imperfective ending with *-u* is immediately followed by the *-a* of *-allä*; if there is an object suffix intervening, then the imperfective will be fully marked for number. In the second and third person plural either the verb or the auxiliary (but not both) is neutralized. If there is no object suffix pronoun, it is the main verb which is neutralized; if there is an object suffix pronoun, it is the auxiliary which is neutralized.

Compound imperfective forms will be tagged IpfvCpd-VCpd.Aux, with appropriate person/number specifications as discussed. The simple tag 'Aux' will be used for auxiliaries that are free (unbound) forms, as in **ይሰብር ነበር** *yəsäbər näbbär*.

The compound gerund form (e.g. **ሰብሮዋል** *säbro-all*) will be tagged like the compound imperfective form, but with the feature 'Verbal Forms' specified as 'Gerund of Compound Verbs' (GerCpd).

The tagging of some negative verbal forms is a further problem. Normal verb conjugation follows a standard pattern in the negative: the affirmative verb takes the prefix *al-* and the suffix *-mm*, for instance **አልሰበረም** *äl-säbbärä-mm* 'he did not break'. Under certain syntactic circumstances the *-mm* is deleted, and in the imperfective the prefix *al-* usually assimilates to the immediately following person/number marker; however, overall, the standard pattern is as described. A few verbs deviate from this pattern.

1) With the copula **ነው** *näw* 'be', the negative is not *äl-näw* but suppletive **አይደለ**- *äydällä*-.
2) With the verb **አለ** *ällä* 'exist', the negative is not *äl-ällä* but a fused portmanteau verb **የለ**- *yällä*- 'not exist'.
3) The verb **አለ** *ällä* also has a special negative relative form -**ሌለ**- -*lellä*-.

We will tag these verb stems respectively as CopNeg, ExNeg, and ExRelNeg.

Verbal nouns are grouped together with verbs. Because Amharic dictionaries normally do not register this form, in the lexicon they will be assigned to the verb from which they regularly derive.

Under the label 'Copula' we consider only the form **ነው** *näw* together with its paradigm. The tag 'Existential' refers to the verb **አለ** *ällä* and its paradigm.

The structure presented above is reflected in Tables 11 and 12.

Table 11 **Verbs**

V				
Ideophone (Ideo)		**Verbal Noun (VN)**		**Plain Verb (V)**
Number	Singular (Sg)		Plural (Pl)	Unmarked (Um)
Person	1		2	3
Gender & Politeness	Communis (C)	Masculine (M)	Feminine (F)	Polite (Pol)
Verbal Forms	Perfective (Pfv)	Imperfective (Ipfv)	Imperfective of Compound Verbs (IpfvCpd)	Gerund (Ger)
	Gerund of Compound Verbs (GerCpd)	Imperative (Impr)	Jussive (Juss)	Copula (Cop)
	Negative Copula (CopNeg)	Existential (Ex)	Negative Existential (ExNeg)	Negative Existential Relative (ExRelNeg)

Table 12 **Auxiliary**

Aux					
Type of auxiliary	Aux			VCpd.Aux	
Number	Singular (Sg)		Plural (Pl)	Unmarked (Um)	
Person	1		2	3	Frozen
Gender & Politeness	Communis (C)	Masculine (M)	Feminine (F)	Polite (Pol)	

3.5 Quantifiers

3.5.1 Cardinal Numerals (NumCard)

The class of cardinal numerals embraces lexemes that inherently possess the category of number. They occur before nouns or independently in dates, phone numbers, zip codes, statistics, arithmetical operations and the like. Fractions such as **ሩብ** *rub* 'quarter', **ግማሽ** *gəmmaš* 'half', and the analysable item **ሁለቱም** *hulätt-u-mm* meaning 'both' are also considered as cardinal numerals.

The numeral 'one' takes the category of gender (see Table 13). The numeral **አሥር** *äśśər* changes its form into **አሥራ** *äśra* when it combines with numerals 1–9, for instance **አሥራ አንድ** *äśra änd*; this is treated as an allolex of **አሥር** *äśśər* and therefore does not need to be considered in the annotation.

Any numeral can be nominalized by means of the definite article or (lower numerals only) by a possessive pronoun; as such it refers to the object that has a given numerical value.

The fractions **ሩብ** *rub* 'quarter', **ግማሽ** *gəmmaš* 'half', and round numerals (ten, hundred, thousand) can take the external plural marker. Pluralized higher numerals indicate that something occurs in a large unspecified number. If the cardinal numeral consists of more than one word, at a higher level it should be considered as MW.

Table 13 **Cardinal Numerals**

NumCard		
Gender	Masculine (M)	Feminine (F)
	አንድ *änd*	**አንዲት** *ändit*
Numeral Symbol		1, 2; 3.5; 1998; ፩, ፪, ፼

3.5.2 Ordinal Numerals (NumOrd)

Lower ordinal numerals (1st–10th) are formed by adding the adjectival ending [V] *-ñña* [C] *-äññа*. In titles, the archaic ordinal numerals ending in *-awi*, *-ay* are used and they take the category of gender as their feature.

In ordinal numerals higher than tenth, only the last digit takes the ordinal marker, and such ordinals will be annotated accordingly. For instance, **ሃያ** *haya*[NumCard] **ሁለተኛ** *hulättaññа*[NumOrd]. Fractions are to be analysed in a similar way, as in **አንድ** *änd*[NumCard] **ሦስተኛ** *śostäññа*[NumOrd].

Various special lexicalized forms such as **መጀመሪያ** *mäğämmäriya*, **የመጀመሪያ** *yämäğämmäriya* 'first', **የፊተኛ** *yäfitäññа* 'first' as well as **ዳግመኛ** *dagmäññа*, **ዳግም** *dagəm* 'second' or 'a second time' also belong to the ordinal numerals.

The ordinal numerals are often written as a numeral symbol followed by the suffix [V] *-ñña*, for instance 3ኛ 'third'.

Table 14 **Ordinal Numerals**

	NumOrd	
	Masculine (M)	Feminine (F)
Gender	**ቀዳማዊ** *qädamawi*	**ቀዳማዊት** *qädamawit*
	ቀዳማይ *qädamay*	**ቀዳማይት** *qädamayt*
Numeral Symbol		1, 2, ፪, ፫

3.5.3 Interrogative Quantifiers (QuanInter)

The interrogative cardinal quantifier, **ስንት** *sənt* 'how much', 'how many', and the interrogative ordinal quantifier **ስንተኛ** *səntäññа* will be annotated with the same tag QuanInter.

3.5.4 Indefinite Quantifiers (QuanIndef)

Amharic has the following indefinite quantifiers: **ስንት** *sənt* (with the exclamative meaning 'so many!', 'so much!'), **ስንተኛ** *səntäññа*, **ስንቱ** *sənte*, **አንዳንድ** *ändand*, **ብዙ** *bəzu*, **አያሌ** *äyyale*, **በርካታ** *bärkatta*, **ትንሽ** *tənnəš*, **ጥቂት** *ṭəqit,* **ሌላ** *lela*, and **ሁሉ** *hullu.* They can take the external plural marker *-očč* and be determined or nominalized by means of the definite article and by the plural possessive pronouns (literally 'their some' in the sense of 'some of them'). For instance, **አንዳንዶቹ** *ändand-* + PlEx (+ Art.M), **አንዳንዶች** *-aችን*/*-aችሁ*/*-aቸው* *ändand-* + PlEx + PPoss.Pl 'some of us/you/them'.

The negative indefinite quantifiers are **አንድም** *änd* + Neg 'not a single one', **አንዱም** *änd* + Art.M + Neg, **አንዳችም** *ändačč-* + Neg.

3.6 Adpositions

The class of adpositions splits into prepositions and postpositions. Almost all Amharic postpositions are derived from various POS: nouns, verbs, and combinations of prepositions with verbs and with demonstrative pronouns. However, I orient my tagging not towards the original lexical source but towards the grammaticalized function and therefore consider the majority of lexemes listed by Leslau as postpositions.[20] During the annotation these must be carefully distinguished from the corresponding nouns, verbs, combinations of some POS, and, as we will see later, from adverbs.

The class of postpositions can be defined by morphological and syntactic criteria (position within the sentence). For instance, postpositions derived from nouns can neither be pluralized nor take any determiners—as opposed to the nouns from which they are derived. Some postpositions of verbal origin, such as **ያህል** *yahəl*, **ይልቅ** *yəläq*, are frozen and do not conjugate, while others always co-occur with propositions and have a reduced form, for instance **በስተቀር** *bästäqärr*. The lexemes **ጀምሮ** *ğämmäro* and **አንሥቶ** *änśäto*, considered by Leslau as postpositions, will be annotated here as gerunds because they preserve their conjugational paradigm and agree with the subject of the sentence.[21]

The prepositions *bä- lä-* and *bästä-* can occur as a compound postposition in combination with another lexeme, for example **በኋላ** *bä-ḫʷala*, (**ፊት**) **ለፊት** (*fit*) *lä-fit*, **በስተጀርባ** *bästä-ğärba*. Such compound postpositions will be treated as a single token and assigned the tag PostCpd.

Note that *bästä-* can also occur with concrete directional terms as in **በስተምሥራቅ** *bästä-mäśraq* 'towards the east'. These are not analysed as compound postpositions because they are insufficiently grammaticalized and their meaning is too specific.

Other Amharic prepositions which perform a special function are the following:

1) Embedded Prepositional Object (PrepEmbObj): *-bb-*, *-ll-*;
2) Genitive Preposition (PrepGen): *yä-*;
3) Distributive Preposition (PrepDistr): *əyyä-*.

[20] Leslau 1995, 616–659.

[21] Another exception is the suffix **-ጌ** *-ge* listed by Leslau 1995, 652 as a postposition. Here it will be treated as a derivational morpheme belonging to the lexicon.

Table 15 **Prepositions and Postpositions**

Prepositions (Prep)	Adpositions: Postpositions (Post)	
bä- *lä-* *kä-*, *tä-* *ə-* *wädä* *əskä* *sälä* *əndä* *yalä* *bästä*	Of nominal origin: **ኋላ** *ḫʷala*, **ሌላ** *lela*, **ላይ** *lay*, **መሃል** *mähal*, **መልስ** *mäls*, **መሠረት** *mäśärät*, **ምትክ** *mətəkk*, **ምክንያት** *məknəyat*, **መካከል** *mäkakkäl* (**መሃከል** *mähakkäl*), **ማዶ** *mado*, **መጠን** *mätän*, **ሥር** *śər*, **ረገድ** *rägäd*, **ሰበብ** *säbäb*, **ሳቢያ** *sabiya*, **ቤት** *bet*, **በኩል** *bäkkul*, **ታች** *tačč*, **ትከሻ** *təkkäša*, **አማካይነት** *ämmakaynät*, **አቅራቢያ** *äqrabiya*, **አናት** *änat*, **አንጻር** *änṣar*, **አኩል** *əkkul*, **አካባቢ** *äkkababi*, **እጅ** *əǧǧ*, **እግር** *əgər*, **አግድም** *ägdəm*, **አጠገብ** *äṭägäb*, **ውስጥ** *wəsṭ*, **ውጭ** *wəč*, **ዙሪያ** *zurya*, **ዳር** *dar*, **ዳርቻ** *darəčča*, **ጀርባ** *ǧärba*, **ጊዜ** *gize*, **ገደማ** *gädäma*, **ግድም** *gədəm*, **ጎን** *gon*, **ጥግ** *ṭəgg*, **ጫፍ** *čaf*, **ፊት** *fit*, **ፈንታ** *fänta* (**ፋንታ** *fanta*) Of verbal origin: **በስተቀር** *bästäqärr* (**በስተቀረ** *bästäqärrä*), **በቀር** *bäqärr*, **በተቀር** *bätäqarr* (**በተቀረ** *bätäqarrä*), **ባሻገር** *baššaggär*, **በተረፈ** *bätärräfä*, **ያህል** *yahəl*, **ይልቅ** *yələq*, **የተነሣ** *yätänäśśa*, **ድረስ** *dəräs*	Origin unclear: **ዘንድ** *zänd*, **ጋ** *ga*, **ጋር** *gar*, **ጋራ** *gara* Combinations of Prep and PDem: **ወዲህ** *wädih*, **ወዲያ** *wädiya*, **ወደዚህ** *wädäzzih*, **ወደዚያ** *wädäzziya*

3.7 Conjunctions (Conj)

Some Amharic subordinating conjunctions have the same form as the preposition from which they originate, such as *kä-* and *əndä-*. The difference between them is that conjunctions take clauses as their objects rather than noun phrases; therefore the two classes will be distinguished and annotated accordingly. Amharic has several discontinuous conjunctions which consist of more than one component: notably a conjunction may consist of two morphemes belonging to the same word-form but separated by another morpheme, such as in *bə-* + IPFV (+ *-əmm*). Here, each of their constituent parts will be tagged as Conj. Compound conjunctions will be tagged in the same manner: thus *bə-* + IPFV (+ *-əmm*) **እንኳ** *ənkʷa* will be tagged Conj + V (+ Conj) + Conj. At a higher level, such complex conjunctions should be considered as a MW. Linking items, such as **ቀርቶ** *qarto*, **ይቅርና** *yəqərənna*, and **ተውና** *täwənna*, will be regarded as verbs because they conjugate in this function.

Table 16 **Conjunctions**

Conj
lə-, *sə-*, *səlä-*, *bä-* , *bə-*; *-əmm* (in *bə-* + IPFV + *-əmm*) (MW); *-əmm* (in *-əmm … -əmm*) (MW); **እንኳን** *ənkʷan* ... *bə-* + IPFV (+ *-əmm*) (MW); *bə-* **ቅሉ** *qəlu* (MW); *bə-* + IPFV (+ *-əmm*) **እንኳ** *ənkʷa*, **እንኳን** *ənkʷan*, **ስንኳን** *sənkʷan* (MW); *əskä-*, *əsk-*, **እና** *ənna* (*-nna*), *əndä-*, *ənd-*, **እንጂ** *ənǧi*; *əyyä-*, *kä-*, **ዘንድ** *zänd*, **ግን** *gən*, **ነገር ግን** *nägär gən* (MW), **ዳሩ ግን** *daru gən* (MW)

3.8 Adverbs (Adv)

The Amharic functional category of adverbs consists of a relatively small group of primary adverbs that refer to temporal and spatial domains and a large group of derived adverbs. The derived adverbs may be one-token items or may comprise more than one segment each belonging to different POS. Because the class of primary adverbs is quite limited they will not be further divided into semantic classes.

The following are Amharic primary adverbs: **አሁን** *ähun*, **ነገ** *nägä*, **ዛሬ** *zare*, **ድሮ** *dəro*, **ዘንድሮ** *zändəro*, **አምና** *ämna*, **ልክ** *ləkk*, **አንዴ** *ände*, **አንዳንዴ** *ändande*, **ዘወትር** *zäwätər*, **ቶሎ** *tolo*, **በጣም** *bäṭam*, and **እጅግ** *əǧǧəg*. Adverbs of nominal origin are **ውስጥ** *wəsṭ*, **ታች** *tačč*, **ላይ** *lay*, and **ጎላ** *ḫʷala*.

Adverbials of the type **እውስጥ** *ə-wəsṭ*, **ወደፊት** *wädä-fit*, **በላይ** *bä-lay*, and the like, are analysed as prepositions and nouns. Combinations of prepositions with demonstrative pronouns which function as adverbs will be treated as prepositional phrases and tokenized as such, for instance **እንደዚህ** *əndä-zzih* and **እስከዚያ** *əskä-zziya*. However, in cases where segmentation is problematic, they will be annotated as unitary adverbs and not tokenized, as for instance with **እምብዛም** *əmbəzamm*, **እንግዲህ** *əngədih*, **እንዲያው** *əndiya-w*, and **እንግዲያው** *əngədiya-w*. In the last two items the final suffix will be tagged as an adverbializer (see below). Although this approach is somewhat inconsistent, it facilitates the process of annotation.

In adverbials which can be segmented, both morphologically and semantically, each of the individual components will be tokenized. To this group belong combinations of the preposition *bä-* with nominals, like **በደንብ** *bä-dänb*, **በድንገት** *bä-dəngät*, and **በሙሉ** *bä-mulu*. Another group consists of nouns preceded by the genitival preposition *yä-*, such as **የግድ** *yägədd* and **የምር** *yämərr*.

Gerund, perfective and imperfective verb forms which are used adverbially form a potentially open-ended list, thus they will be annotated as their respective verbal forms, for example **ጅምሮ** *ǧämmərо*, **ነጋ ጠባ** *nägga ṭäbba*, and **በይበልጥ** *bä-yəbälṭ*.

3.8.1 Interrogative Adverbs (AdvInter)

Amharic has the following interrogative adverbs: **ለምን** *ləmən*, **መች** *məčč*, **መቼ** *məče*, **መች** *məč*, **ስለምን** *sələmən*, **እንደምን** *əndəmən*, **እንዴት** *əndet*, **ወዴት** *wədet*, and **የት** *yət*. Alternatively, the lexemes **ለምን** *ləmən* and **ስለምን** *sələmən* might be analysed as combinations of preposition and interrogative pronouns.

3.8.2 Indefinite Adverbs (AdvIndef)

Indefinite adverbs consist of an interrogative adverb followed by the Indefinitizer, for instance **የትም** *yət-əmm* and **መቼም** *məče-mm*.

3.9 Particles (Part)

As particles we understand independent words (not cliticized) which can modify other words belonging to different POS or scope over the whole sentence. They represent the speaker's comment about a broader linguistic expression. The class of Amharic particles embraces, among others, the following lexemes: **ለካ** *ləkka*, **ገና** *gəna*, **ብቻ** *bəčča*, **ደግሞ** *dəgmo*, **ደሞ** *dəmmo,* **እንኳን** *ənk^{w}an*, and **ምናልባት** *mənalbat*. The only Amharic interrogative particle is **ወይ** *wəy*.

3.10 Interjections (Interj)

To this class belong words and expressions that are utterances in their own right and express the speaker's emotions and reactions.[22] Examples of Amharic interjections are **ቀስ** *qəss*, **ብያ** *bəyya*, **ቼ** *če*, **እልል** *əlləl*, **ኧረ** *ʿərə*, **እስቲ** *əsti*, **እንዴ** *ənde*, **ታዲያ** *tadiya*, **እሺ** *əššə*, and **እምቢ** *əmbi*.

3.11 Bound Grammatical Morphemes

This class embraces a heterogeneous variety of bound markers (both affixes and clitics; I will not distinguish the two) that cannot easily be fitted into any other tag-category. They serve different functions: inflection, derivation and informational structure, and must be attached phonologically to some adjacent host word. They are listed in Table 2. Here only two of them will be briefly discussed.

[22] In the list of interjections Leslau 1995, 899–909 includes various lexemes that, for our purposes, will be defined as nouns, verbs, and particles.

3.11.1 Adverbializers (Advzr)

The adverbializers (Advzr) are derivational clitics which, when added to various POS, turn them into an adverb or, if added to an existent adverb, give it a certain shade of meaning. They are homophonous with the definite article and the accusative marker but their function is not grammatical but derivational. The adverbializers embrace the following items:

1) *-u*: for instance **በመጠኑ** *bä*[Prep]-*mäṭän*[NCom]-*u*[Advzr], **በየዓመቱ** *bä*[Prep]-*yyä*[PrepDistr]-*ʿamät*[NCom]-*u*[Advzr], **ደግነቱ** *dägənnät*[NCom]-*u*[Advzr];
2) *-wə/un*, *-wan*: as in **ሰሞኑን** *sämon*[NCom]-*un*[Advzr], **በጣሙን** *bäṭam*[Adv]-*un*[Advzr], **ክፍቷን** *kəft*[NCom]-*ʷan*[Advzr];
3) *-wə/-unu*: for example **ዛሬውኑ** *zare*[Adv]-*wunu*[Advzr], **አሁኑኑ** *ähun*[Adv]-*unu*[Advzr].

3.11.2 Assertative *-a*

The assertative *-a* is used for emphasizing a statement which is opposed to what was previously thought or said. The clitic is attached to verbs.

Conclusion

The proposed tagset is intended as a starting point for preparing exhaustive guidelines for annotators. It contains forty-seven tags grouped into twelve parts of speech. Additionally, some POS take values from different grammatical categories, which are tagged accordingly.

More explicit rules, perhaps in the form of an inventory of items, should be given for 'small words' such as particles and interjections. There are many black holes in the research into Amharic POS, and the 'smaller' the word is, the more difficult it is to classify. Because of this it remains to be seen whether some of the words now collected under the labels 'Particles', 'Interjections', and 'Adverbs' should be transferred to another class. It is also possible that we find homophonous lexemes that have to be placed in more than one class.

References

Cotterell, F. P. 1964. 'Amharic word classes', *Journal of Ethiopian Studies*, 2/1 (1964), 33–48.

Gambäck, B. 2012. 'Tagging and Verifying an Amharic News Corpus', in G. De Pauw, G.-M. de Schryver, M. L. Forcada, K. Sarasola, F. M. Tyers, and P. W. Wagacha, eds, *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages SaLTMiL 8–AfLaT 2012. May 22, 2012, Istanbul Lütfi Kırdar Convention & Exhibition Centre, Istanbul, Turkey* (Istanbul: European Language Resources Association, 2012), 79–84.

Gambäck, B., F. Olsson, Atelach Alemu Argaw, and L. Asker 2009. 'Methods for Amharic Part-of-Speech Tagging', in G. De Pauw, G.-M. de Schryver, and L. Levin, eds, *EACL 2009. Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages, AfLaT 2009. 31 March 2009, Megaron Athens International Conference Centre, Athens, Greece* (Athens: The Association for Computational Linguistics, 2009), 104–111.

Gankin, E. B. 1969. *Amharsko–Russkij Slovar.* **አማርኛና ፥ መስካብኛ ፥ መዝገበ ፥ ቃላት ።** (*Amharic–Russian Dictionary. Amarəññа-nna mäskobəññа mäzgäbä qalat*) (Moskva: Izdatel'stvo Sovetskaja Entsiklopedija, 1969).

Gasser, M. 2011. 'HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya', in Organizing Committee, *Proceedings of Conference on Human Language Technology for Development. 3–5 May 2011, Bibliotheca Alexandrina, Alexandria, Egypt* (online publication, 2011), 94–99, http://www.cle.org.pk/research/rep/HLTD2011.pdf, accessed 5 September 2017.

— 2012. 'HornMorpho 2,5 User's Guide' (online publication, 2012), http://homes.soic.indiana.edu/gasser/L3/horn2.5.pdf, accessed 5 September 2017.

Girma Awgichew Demeke and Mesfin Getachew 2006. 'Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges', *ELRC Working Papers*, 2/1 (2006), 1–17.

Hartmann, J. 1980. *Amharische Grammatik*, Äthiopistische Forschungen, 3 (Wiesbaden: Franz Steiner Verlag GmbH, 1980).

Hummel, S. and W. Dickhut 2016. 'A part of speech tag set for Ancient Ethiopic', in A. Bausi with assistance from E. Sokolinski, eds, *150 Years after Dillmann's* Lexicon*: Perspectives and Challenges of Gəᶜəz Studies*, Supplement to Aethiopica, 5 (Wiesbaden: Harrassowitz Verlag, 2016), 17–30.

Kane, T. L. 1990a. *Amharic–English Dictionary*, I: **ሀ–ነ** (Wiesbaden: Otto Harrassowitz, 1990).

— 1990b. *Amharic–English Dictionary*, II: **ኘ–ፐ** (Wiesbaden: Otto Harrassowitz, 1990).

Leslau, W. 1995. *Reference Grammar of Amharic* (Wiesbaden: Harrassowitz Verlag, 1995).

Vertan, C. 2016. 'Bringing Gəʾəz into the digital era: computational tools for processing Classical Ethiopic', in A. Bausi with assistance from E. Sokolinski, eds, *150 Years after Dillmann's* Lexicon*: Perspectives and Challenges of Gəᶜəz Studies*, Supplement to Aethiopica, 5 (Wiesbaden: Harrassowitz Verlag, 2016), 31–42.

Wintner, S. 2014. 'Morphological Processing of Semitic Languages', in I. Zitouni, ed., *Natural Language Processing of Semitic Languages*, Theory and Applications of Natural Language Processing (Heidelberg–New York–Dordrecht–London: Springer, 2014), 43–66.

Summary

The aim of the article is to propose a tagset for the morposyntactic tagging of Amharic and to discuss those issues which may seem problematic. The tagset contains forty-seven tags grouped into twelve parts of speech. It is hoped that it provides a starting point for more exhaustive guidelines for prospective annotators.