So-Called "Prepositions" in Somali are Not Prepositions: A Linguistic Approach for Somali POS tagging

Chihiro Taguchi Taro Watanabe
Nara Institute of Science and Technology
{taguchi.chihiro.td0, taro}@is.naist.jp

1 Introduction

In Somali, the four lexemes *u*, *ku*, *ka*, *la* (UKKLs henceforth) are called in various terminologies such as "prepositions"[1, 2], "prepositional indicators"[3], "adpositional verbal particles"[4], and "verbal adpositions"[5]. However, as this polyonymy indicates, the morphosyntactic status of the lexemes is still controversial. This study first demonstrates that the lexemes are neither prepositions nor adpositions, but are either clitics, verbal prefixes, or particles that function as an applicative[6] to augment an extra argument. Particularly, in view of Universal POS tags[7], we argue that they should be categorized in particles (PART). Then, we propose a solution to implement the Somali Universal Dependencies (UD) by practically applying our POS tagging rules.

2 Overview

2.1 Overview of the Language

Somali (< Cushitic < Afroasiatic) is spoken in the Horn of Africa by approximately 15 million people[8]. The basic word order is Subject-Object-Verb (SOV), but it may change with respect to the information structure. In contrast, the order of verbal components is rigid. The nominal morphology distinguishes number (singular and plural), gender (masculine and feminine), case (subject and absolute), and definiteness (definite and indefinite). The verbal morphology includes inflections by person, number, and gender of the subject, tense (present and past), aspect (unmarked and continuous), mood (indicative, subjunctive, and imperative), polarity (positive and negative), and focus. As for the information structure, focus plays a significant role in Somali. Auxiliaries *ayaa* and *baa* puts focus on the preceding noun (cf. (1, 2)), and *waxa* put focus on the noun after the verb phrase (cf. (3)). When there is no focus on a particular noun, the auxiliary *waa* is selected (cf. (4)). Note that the long vowel *-aa* in the auxiliaries is changed to *-uu* when succeeded by the third person masculine subject marker.

(1)*Maxamed* (*baa*|*ayaa*) *bariis cunay*.
Mohammed FOC rice ate
"**Mohammed** ate rice."

(2)*Maxamed* *bariis* (*buu*|*ayuu*) *cunay*.
Mohammed rice FOC:3SG.M ate.
"Mohammed ate **rice**."

(3)*Bariis waxa cunay Maxamed*.
rice FOC ate Mohammed
"**Mohammed** ate rice."

(4)*Maxamed* *bariis wuu* *cunay*.
Mohammed rice AUX:3SG.M ate
"Mohammed ate rice."

2.2 *u*, *ku*, *ka*, *la*

Sentences below (5)-(7) are examples with *ka* "from". They basically share the same meaning, only differing in the word order. As apparent in the ungrammaticality of (8), the lexeme *ka* and the verb *yimi* must not be intervened.

(5)*Maxamed* *baa* ***ka*** *yimi* *Soomaaliya*.
Mohammed FOC from came Somalia

(6)*Maxamed* *baa Soomaaliya* ***ka*** *yimi*.
Mohammed FOC Somalia from came

(7)*Soomaaliya Maxamed* *baa* ka *yimi*.
Somalia Mohammed FOC from came
(1)-(3): "Mohammed came from Somalia."

(8)**Maxamed* *baa* ***ka*** *Soomaaliya yimi*.
Mohammed FOC from Somalia came
intended: "Mohammed came from Somalia."

The other UKKLs, *u* "for", *ku* "in", *la* "with", also syntactically behave similarly to *ka*. Below are examples containing them.

(9)*Cali shaah* ***u*** *samee*!
Ali tea for make
"Make tea for Ali!"

(10)*Caano koob-ka* ***ku*** *shub*!
milk cup-DET in pour
"Pour milk in the cup!"

(11)*Maxamed waa-n* ***la*** *joogay*.
Mohammed AUX-1SG with stayed
"I stayed with Mohammed."

Interestingly, the four UKKLs can be combined with each other (12) as well as with object pronouns. In addition, because of the relatively free word order, ambiguities may occur as exemplified in (13). This kind of ambiguities is in most cases resolved by contextual information.

(12)*Maxamed guri-ga baa-n* ***ku-la*** *kulmay*.
Mohammed house-DET FOC-1SG in-with met
"I met with Mohammed in the house."

(13)*Cali baa Maxamed i-u dilay*.
Ali FOC Mohammed me-for hit
"Ali hit (Mohammed for me | me for Mohammed)."

3 Linguistic Analysis of UKKLs

This section discusses the grammatical status of UKKLs in more detail, chiefly to show that UKKLs are not adpositions and that they function as applicatives. In addition, in view of application to the Universal POS Tagging, we argue that it is suitable for the UD to analyze UKKLs as particles (PART).

3.1 Why UKKLs are not adpositions

Adpositions roughly comprise two subcategories: prepositions and postpositions. Adpositions form an adpositional phrase when combined with a nominal phrase (NP) adjacent to it, and prepositions specify that the NP succeeds them. In other words, in an adpositional phrase, an adposition is the head and the adjacent NP is the dependent.

In this sense, sentences (5)-(8) clearly show that they are not adpositions, because *ka* does not necessarily come next to the dependent. If we assume that they are adpositions, we will end up allowing for dependency trees in which branches cross arbitrarily. Figure 1 is an outline showing crossing branches in (10). Therefore, it is inappropriate to deem them as adpositions.

3.2 Why they are applicatives

Applicative is a grammatical voice by which an oblique argument is promoted to a core argument of the predicate. (14) is an example of the applicative construction reported in Rombo (< Chaga < Bantu)[9]. As English never allows "child" to be an object in this case, the argument *mwaná* "child" is a semantically marginal argument to which is assigned a benefactive semantic role. However, *mwaná* "child" is treated as an object because of the applicativization in the verb.

(14)*Ksali é-le-m'-koṛ-i-a mwaná klálo*.
Kisali SM.3SG-PST-OM.3SG-cook-APPL-F child food
"Kisali cooked food for her child."

In Somali, UKKLs are applicatives because of the following two facts. First, UKKL is a verbal component that cannot be separated from the main predicate as shown in (13). Second, UKKL augments a new object that is semantically marginal; in particular, *u*, *ku*, *ka*, and *la* promote benefactive ("for"), locative ("in"), source ("from"), and comitative ("with") arguments to objects respectively. Without using any UKKLs, *yimi* "came" is a monovalent predicate that only takes a subject (agent). Adding *ka* to *yimi* changes it to a divalent predicate requiring an additional object pertain-

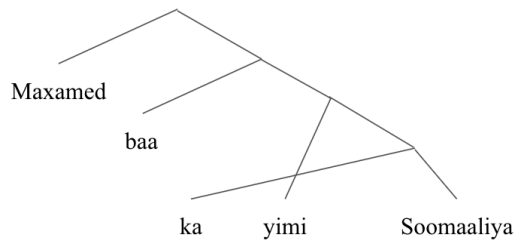

Figure 1 Crossing tree

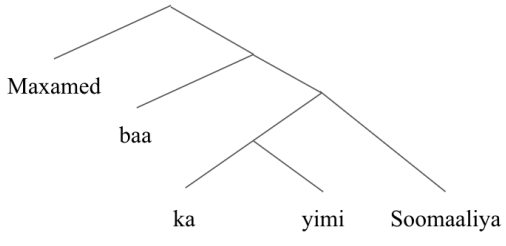

Figure 2 Uncrossing tree

Sentence	Dadka Soomaaliyeed waxay ku noolyihiin dalalka geeska Afrika. “The Somali people live in the countries of the Horn of Africa.”							
	dadka	Soomaaliyeed	waxay	ku	noolyihiin	dalalka	geeska	Afrika
POS	NOUN	NOUN	AUX	PART	ADJ	NOUN	NOUN	PROPN
	the people	Somalis	focus	in	living	the countries	the horn	Africa

Table 1 An example of POS tagging (1)

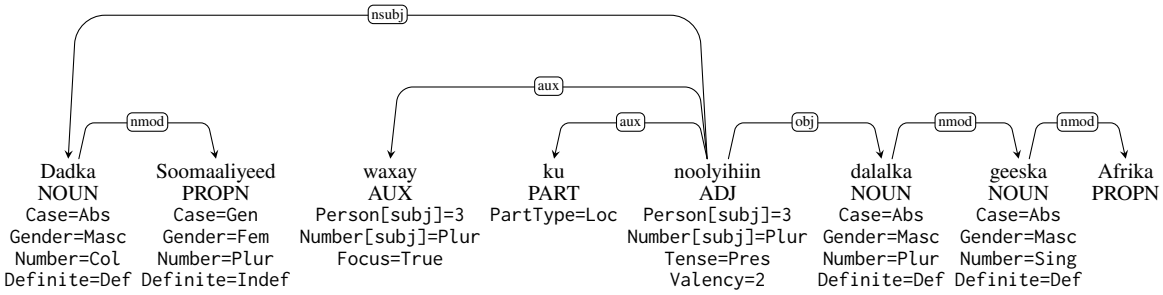

Figure 3 An example of dependency tree (1)

ing to source. Since Somali is a pro-drop language, even when the augmented object is not pronounced as in (16), the sentence is interpreted as entailing implicit information of source.

(15)*Maxamed baa yimi.*
Mohammed FOC came
“Mohammed came.”

(16)*Maxamed baa* ***ka*** *yimi.*
Mohammed FOC from came
“Mohammed came from there.”

In formal semantics, the structural difference between *yimi* and *ka yimi* is represented in (17) and (18). Combining multiple UKKLs adds more arguments as in (19).

(17)$[\![yimi]\!] = \lambda x.\text{come}(x)$

(18)$[\![ka\ yimi]\!] = \lambda y \lambda x.\text{come-from}(x, y)$

(19)$[\![kala\ yimi]\!] = \lambda z \lambda y \lambda x.\text{come-from-with}(x, y, z)$

This representation solves the problem of dependency mentioned in the previous subsection, and allows for word order scrambling seen in (1). Taking these semantic representations into account, the ideal branching of (1) would look like Figure 2.

3.3 Why they are particles

Having demonstrated that UKKLs are at least not adpositions and that they derive applicative construction, it is still unclear as to which part-of-speech they are covered in. Given the strong combination of UKKLs and predicates, there are three possibilities: verbal prefix, clitic, and particle. For the sake of compatibility with UD, we argue that UKKLs are particles (PART).

Formally distinguishing prefixes, clitics, and particles is a disputable matter. As Zwicky noted that so-called particles are in fact either clitics or affixes or something that are difficult to be labelled[10], particles can even be an unnecessary label. In theoretical linguistics, it is possible, or even suitable, to assume that UKKLs are verbal prefixes in speakers' linguistic knowledge. When a morpheme is an affix, it is attached to its stem and cannot appear on its own. Even though they look like an independent lexeme in the orthography, they are phonologically pronounced together with the succeeding predicate.

However, this interpretation would require text-based NLP to prepare some additional pre-processings. In light of the definition of particles (PART) given by UD[1)], it is more reasonable for UD to label them as particles.

4 Universal POS Tags and Dependency Tree

UD is an ongoing project to establish a universal framework for annotation of grammatical information in different languages. As of 2020 it covers over 120 languages, and more UD languages are being prepared. At the time of this writing, Somali is not included in the list of UD languages. Given this situation, this section briefly dis-

1) “Particles are function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech”[7].

Sentence	Jawaahir gurigeeda baannu idinkula kulmi doonnaa. “We will meet you at Jawaahir’s house.”					
POS	Jawaahir	gurigeeda	baannu	idinkula	kulmi	doonnaa
	PROPN	NOUN	AUX	PART	VERB	VERB
	Jawaahir	her house	focus	you-in-with	meet	will

Table 2 An example of POS tagging (2)

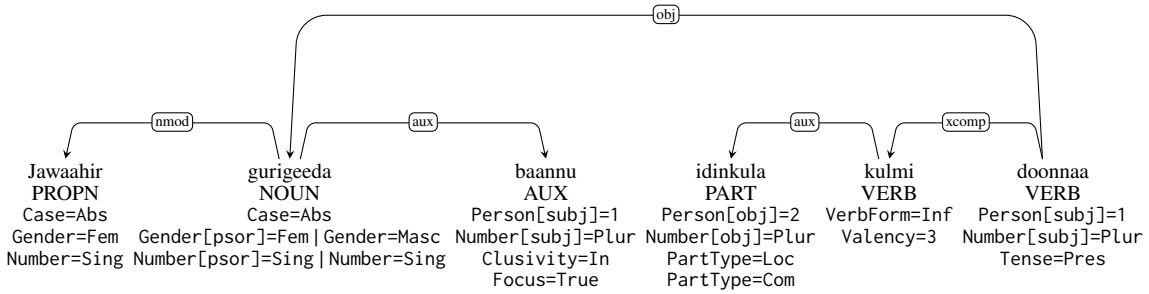

Figure 4 An example of dependency tree (2)

cusses what POS tags, features, and syntactic relations are needed for the establishment of the Somali UD based on the observation we have made in the previous sections.

As we have already seen that UKKLs should be categorized in PART, adposition (ADP) is no longer necessary for the Somali POS tags. A practical example of POS tagging is shown in Table 1[2)]. Based on this POS tagging, Figure 3 is an instance of a manually drawn dependency tree. The syntactic relation binding the predicate and the particle *ku* is assumed to be aux (auxiliary). The feature of the particle is specified as `PartType=Loc`, as it adds an object with a locative semantic role.

Specifying the morphological features in this way resolves the morphological complexity of Somali; for example, in (13), the morphemes *i* “me” and *u* “for” are combined into one token even though they do not necessarily share a direct syntactic dependency. An example of this kind of particle combination is shown in Table 2, where *idinkula* is a combination of *idin* “you (plural)”, *ku* “in”, and *la* “with”, and its corresponding dependency tree in Figure 4. Since applicativization and object pronouns are specified in the features, the dependency trees correctly predict the desired interpretation.

These dependency trees contain two novel features: `Valency` and `Focus`. `Valency`, requiring the number of core arguments as its value, is a feature proposed by Senuma & Aizawa for a morphological analysis of Ainu, which also has applicative construction[12]. The fact that the valency is augmented to 3 (subject, locative object, comitative object) is represented by this feature. `Focus` is a language-specific feature for Somali. It takes a boolean value, and when `Focus=True`, the token puts a focus on a certain phrase which is uniquely determined syntactically.

5 Conclusion

This study clarified that *u*, *ku*, *ka*, and *la* in Somali are neither prepositions nor adpositions, but are functional morphemes for applicativization, each playing a role of augmenting a new core argument with different semantic roles. While verbal prefixes might be a suitable categorization for UKKLs from the viewpoint of theoretical linguistics, PART is reasonable for the UD POS tagging for the sake of consistency. In addition, we made a tentative proposal for practical POS tagging and dependency parsing of Somali by defining POS tag sets, features, and syntactic relations. This study established a ground for the Somali UD, since it is indispensable to examine the grammar of low-resource languages before starting to apply NLP to them.

6 Future Work

How to resolve the ambiguity of grammatical functions mentioned in (13) still remains unclear. In human’s natural language understanding, this kind of ambiguity is usually resolved by contextual information. Embedding knowledge graph information or hypernym–hyponym relations might improve the accuracy for machine to resolve the ambiguity.

2) The sample sentences in Tables 1 and 2 were collected from Orwin (1995)[11].

References

[1] Annarita Puglielli. *Sintassi della lingua somala*. Ministero AA. EE., 1981.

[2] John I. Saeed. *Somali reference grammar (2nd ed.)*. Dunwoody Press, second edition, 1993.

[3] Catherine El-Solami-Mewis. *Lehrbuch des Somali*. Verlag Enzyklopädie, 1987.

[4] Abdalla O. Mansur. *Le lingue Cuscitiche e il Somalo*. Ministero AA. EE., 1988.

[5] John I. Saeed. *Somali*. John Benjamins B.V., 1999.

[6] David A. Peterson. *Applicative constructions*. Oxford University Press, 2007.

[7] *Universal POS tags*, 2020. https://universaldependencies.org/u/pos/.

[8] Ethnologue, 2019.

[9] Daisuke Shinagawa. *A grammatical sketch of Chaga-Rombo (Bantu E623)*. Research Institute for Languages and Cultures of Asia and Africa, 2014.

[10] Arnold M. Zwicky. Clitics and particles. *Language*, 63(2):283–305, 1985.

[11] Martin Orwin. *Colloquial Somali*. Routledge, 1995.

[12] Hajime Senuma and Akiko Aizawa. Toward Universal Dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139, Gothenburg, Sweden, May 2017. Association for Computational Linguistics.

A Appendix: Abbreviations

- APPL: applicative
- AUX: auxiliary
- DET: determiner
- F: final suffix
- FOC: focus
- M: masculine
- OM: object marker
- PST: past
- SG: singular
- SM: subject marker