So-Called “Prepositions” in Somali are Not Prepositions

A Linguistic Approach for Somali POS Tagging

Chihiro Taguchi, Taro Watanabe

`{taguchi.chihiro.td0, taro}@is.naist.jp`

Nara Institute of Science and Technology (NAIST)

1 Introduction

- 4 lexemes in Somali are to be discussed
  - *u, ku, ka, la* (“UKKL” henceforth)
- Inconsistency in their linguistic status
  - prepositions (Puglielli 1981, Saeed 1993, Morgan 2020)
  - prepositional indicators (El-Solami-Mewis 1987)
  - adpositional verbal particles (Mansur 1988)
  - verbal adpositions (Saeed 1999)

This study provides:

- Linguistic finding: they are best analyzed as **preverbal applicative particles**
- Proposal to NLP: more consistency in the universal POS tagging based on linguistics

2 Overview: Somali

- Bio:
  - Somali < Cushitic < Afroasiatic
    - cf. Arabic, Hebrew, Amharic (Ethiopia)
  - approx. 15 million speakers
- Syntax:
  - SOV with scrambling
  - Rigid word order of VP
- Morphology:
  - Nominal: number, gender, case, definiteness
  - Verbal: person, number, gender, tense, aspect, mood, polarity, **focus**
- Information structure:
  - Focus

BOUNDARY REPRESENTATION IS NOT NECESSARILY AUTHORITATIVE
Port Sudan
SAUDI ARABIA
RED SEA
SUDAN
OMAN
ERITREA
KHARTOUM
Kassalā
KAGNEW STATION
Mits'iwa
Āsmera
YEMEN
ṢAN'Ā'
(ṢAN'Ā')
TIGRAI
AFAR
YEMEN (ADEN)
Blue Nile
White Nile
SHANKELLA
Āseb
FRENCH TERRITORY OF THE AFARS AND ISSAS
ADEN
SOCOTRA (Yemen-Aden)
Bab el Mandeb
DJIBOUTI
Ar Ruṣayriṣ
Lake Tana
AMHARA
GULF OF ADEN
Ayshā
Berbera
GALLA
ADDIS ABABA
Dirē Dawa
Hargeisa
ETHIOPIA
GURAGE
OGĀDĒN
SIDAMO
GALLA
administrative line
SOMALIA
Lake Rudolf
Lugh Ganane
INDIAN OCEAN
UGANDA
Giuba
Lake Kyoga
KAMPALA
Wajir
MOGADISCIO
KENYA
Chisimayu
Lake Victoria
NAIROBI
THE HORN OF AFRICA
Approximate area inhabited by Somali people
AFAR Other ethnic group
Arusha
Mombasa
TANZANIA
0 100 200
MILES

Distribution of Somali speakers
Source: Wikimedia Commons

Nara Institute of Science and Technology NAIST ®

2 Overview: Focus in Somali

cf. Japanese *-ga*: “*Dare-ga kita?*” — “*Taroo-ga/*wa kita*”

- Focus by auxiliaries:
  - *ayaa*, *baa*: focus on the preceding noun
    -

Maxamed	(*baa*|*ayaa*)	*bariis*	*cunay*.
M.	FOC	rice	ate

“**Mohammed** ate rice.”

  - *waxa*: focus on the noun after the inflected verb
    -

Bariis	*waxa*	*cunay*	***Maxamed***.
rice	FOC	ate	M.

“**Mohammed** ate rice” (or “It is **Mohammed** who ate rice”)

- No focus:
  - *waa*
    -

Maxamed	*bariis*	*wuu*	*cunay*.
M.	rice	AUX.3SG.M	ate

2 Overview: *u, ku, ka, la*

- *u*, *ku*, *ka*, *la*:
“for” (benefactive), “in” (locative), “from” (elative), “with” (comitative) respectively

- UKKL and the verb must not be intervened:

Maxamed	*baa*	***ka***	*yimi*	*Soomaaliya*.
M.	FOC	from	came	Somalia

Maxamed	*baa*	*Soomaaliya*	***ka***	*yimi.*
M.	FOC	Somalia	from	came

Soomaaliya	*Maxamed*	*baa*	***ka***	*yimi*.
Somalia	M.	FOC	from	came

“Mohammed came from Somalia.”

roughly the same meaning

**Maxamed*	*baa*	***ka***	*Soomaaliya*	*yimi*.
M.	FOC	from	Somalia	came

ungrammatical (intended: “Mohammed came from Somalia.”)

2 Overview: Ambiguities caused by UKKLs

- The position of UKKLs’ argument is free as long as it is grammatical
    - UKKLs select their argument based on context information

- Ambiguity may emerge:
    -

Cali	*baa*	*Maxamed*	*i-u*	*dilay*.
A.	FOC	M.	me-for	hit.

“Ali hit Mohammed for me.”

“Ali hit me for Mohammed.”

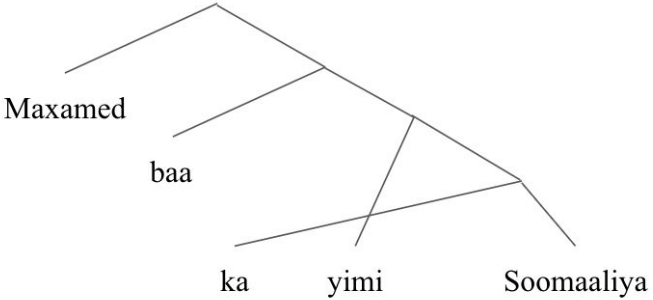
3.1 UKKLs are not adpositions

- *Prepositions* ⊂ *Adpositions*
- Consider:
  -

Maxamed	*baa*	***ka***	*yimi*	*Soomaaliya*.
M.	FOC	from	came	Somalia

“Mohammed came from Somalia.”

- ***ka*** “from” and *Soomaaliya* “Somalia” are intervened by the verb *yimi* “he came”
  - Produces a crossing tree



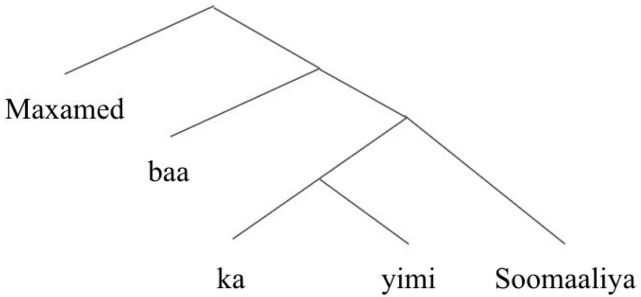
3.2 UKKLs are Applicatives

- **Applicative**: one of verbal voices
  - Voices in Japanese: passive (-*sareru*), middle (Hokkaido dialect -*rasaru*) , etc.
- Applicative promotes an oblique argument to an object (core argument)
  - Widely reported in Bantu languages (Africa), etc.

- Somali UKKLs play a role of augmenting an object with a particular semantic role
  - *u*: benefactive (“for”)
  - *ku*: locative (“in”)
  - *ka*: source, elative (“from”)
  - *la*: comitative (“with”)
- Example:
  - *yimi* “came”: **monovalent** verb, requiring only its subject
  - *ka yimi* “came from”: **divalent** verb, requiring its subject and source object

3.3 UKKLs are Particles

- Linguistic status of UKKLs:
  - Verbal prefix, clitic, particle?
- Theoretical linguistics may prefer them as verbal prefixes
  - Phonogically they are pronounced together
- In NLP, it is better to treat them as **particles**
  - one token needs one POS tag
  - no POS tag for verbal prefixes, because they are a part of verb
  - yields an uncrossing tree



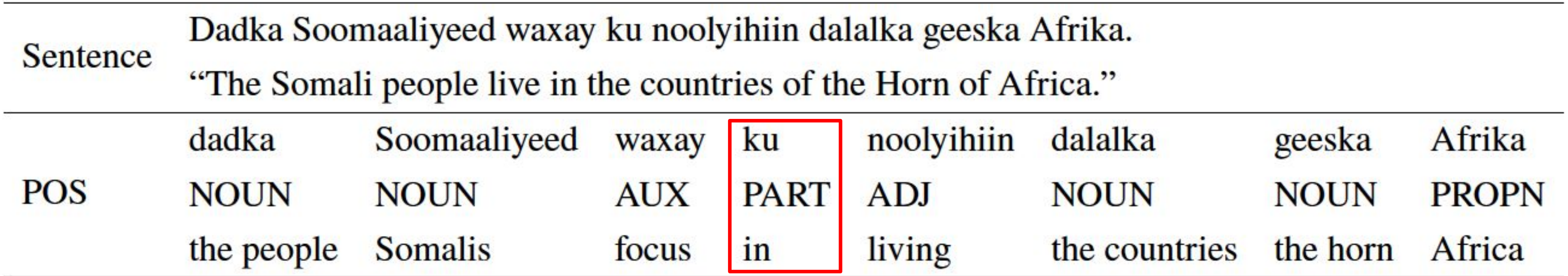
4 Universal POS Tags and Dependency Tree

- Universal Dependency (UD):
  - universal framework for annotation of grammatical information in different languages
  - covers more than 120 languages (as of 2020)
  - ongoing project
  - Somali is not included

- We propose a way of annotating UKKLs in UD based on our linguistic findings

4 Universal POS Tags and Dependency Tree

- Example of a manual POS Tagging below:
  - **PART** is tagged to the UKKL *ku* “in”

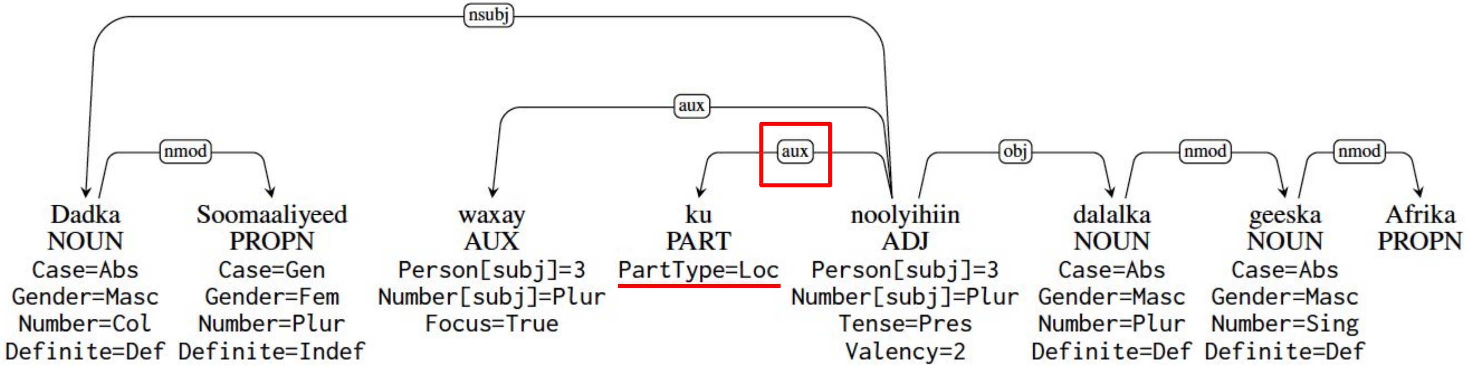


Sentence	Dadka Soomaaliyeed waxay ku noolyihiin dalalka geeska Afrika. “The Somali people live in the countries of the Horn of Africa.”							
POS	dadka	Soomaaliyeed	waxay	ku	noolyihiin	dalalka	geeska	Afrika
	NOUN	NOUN	AUX	PART	ADJ	NOUN	NOUN	PROPN
	the people	Somalis	focus	in	living	the countries	the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

Table 1 An example of POS tagging (1)

4 Universal POS Tags and Dependency Tree

- Example of a manually drawn dependency tree below:
  - The relation of the predicate (*noolyihiin*) and the UKKL (*ku*) is **`aux`**
  - The feature of the UKKL (*ku*) is specified as **`PartType=Loc`**

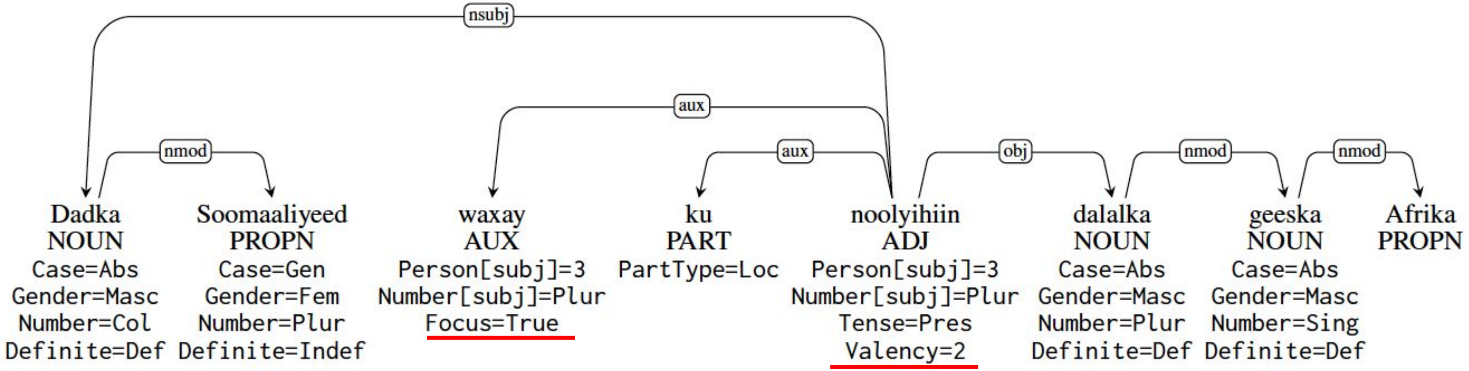


Dadka	Soomaaliyeed	waxay	ku	noolyihiin	dalalka	geeska	Afrika
NOUN	PROPN	AUX	PART	ADJ	NOUN	NOUN	PROPN
Case=Abs	Case=Gen	Person[subj]=3	PartType=Loc	Person[subj]=3	Case=Abs	Case=Abs	
Gender=Masc	Gender=Fem	Number[subj]=Plur		Number[subj]=Plur	Gender=Masc	Gender=Masc	
Number=Col	Number=Plur	Focus=True		Tense=Pres	Number=Plur	Number=Sing	
Definite=Def	Definite=Indef			Valency=2	Definite=Def	Definite=Def	

Figure 3 An example of dependency tree (1)

4 Universal POS Tags and Dependency Tree

- Two new features:
  - **`Valency`**: specifies the number of the predicate’s core arguments (cf. Senuma&Aizawa 2017)
  - **`Focus`**: specifies that the token introduces focus
    - Its syntactic relation is specified by the arrow



Dadka
NOUN
Case=Abs
Gender=Masc
Number=Col
Definite=Def

Soomaaliyeed
PROPN
Case=Gen
Gender=Fem
Number=Plur
Definite=Indef

waxay
AUX
Person[subj]=3
Number[subj]=Plur
Focus=True

ku
PART
PartType=Loc

noolyihiin
ADJ
Person[subj]=3
Number[subj]=Plur
Tense=Pres
Valency=2

dalalka
NOUN
Case=Abs
Gender=Masc
Number=Plur
Definite=Def

geeska
NOUN
Case=Abs
Gender=Masc
Number=Sing
Definite=Def

Afrika
PROPN

Figure 3 An example of dependency tree (1)

5 Conclusion

This study has shown:

- Linguistic findings:
  - *u*, *ku*, *ka*, *la* are not prepositions (nor adpositions!)
  - *u*, *ku*, *ka*, *la* are either prefix, clitic, or particle, which functions as applicative
  - the best candidate is particle for the sake of NLP

- Proposals:
  - for relatively minor languages, UD heavily relies on descriptive linguists
  - an example of the Somali UD based on the linguistic findings

References

- Catherine El-Solami-Mewis. *Lehrbuch des Somali*. Verlag Enzyklopädie, 1987.
- Ethnologue, 2019.
- Abdalla O. Mansur. *Le lingue Cuscitiche e il Somalo*. Ministero AA. EE., 1988.
- Morgan Nilsson. *Beginner's Somali grammar*. 2020. Available at: http://morgannilsson.se.
- Martin Orwin. *Colloquial Somali*. Routledge, 1995.
- David A. Peterson. *Applicative constructions*. Oxford University Press, 2007.
- Annarita Puglielli. *Sintassi della lingua somala*. Ministero AA. EE., 1981.
- John I. Saeed. *Somali reference grammar* (2nd ed.). Dunwoody Press, 1993.
- John I. Saeed. *Somali*. John Benjamins B.V., 1999.
- Daisuke Shinagawa. *A grammatical sketch of Chaga-Rombo (Bantu E623).* Research Institute for Languages and Cultures of Asia and Africa, 2014.
- Universal POS Tags, 2020. https://universaldependencies.org/pos.
- Arnold M. Zwicky. Clitics and particles. Language, 63(2):283–305, 1985.

To Do

- ページ番号の挿入
- 発表者名（自分の名前）をボールドで示すなどしてわかりやすく
- Overview: どの言語に近いのかを示すとわかりやすいかも
- 強調の仕方（ボールド、色分け）に一貫性を持たせる
- Focus をわかりやすく説明（日本語では …？）
    - 誰が鹿を見ましたか。
        - *太郎は見ました。/ 太郎が見ました。(focus)
- 三つの文（scrambling）が permutation のみで同じ意味を示していることを明確にする
    - 波カッコを用いるなど
- Crossing に関して質問が出るかも
- NLPではParticleで分析されるべきであることを強調する
    - 言語学のツリーとの比較、など
- UD, POS では kuの位置をboldにするなどしてわかりやすくする
- 図のどこに言及しているかをわかりやすくする