Three Somali Language Corpora: How can they be useful?

Morgan Nilsson

PhD, Senior Lecturer in Slavic languages (and Somali)
University of Gothenburg

4th Mogadishu Book Fair, 15 Aug 2018

Outline

- Background information about Somali in Sweden
- Somali at Gothenburg University
- What is a corpus? How can it be helpful?
- Three Somali corpora
- How to find and interpret what you are searching for

Somali in Sweden

- Sweden: 10 million inhabitants
- 1% are born in Somalia or have parent(s) who are
- 2% of the school children speak Somali
- Need for teaching materials, reference grammar, different dictionaries
  - childrens' school dictionary, terminological dictionary, Somali-Swedish dict…,
  - childrens' school grammar, exercise books…
- A corpus is a base and a tool

Somali at Gothenburg University

- (2014) Courses over one year for Somali speakers (school teachers)
- (2008) Short beginner's course for Swedish speakers
- New courses under preparation leading to a B.A. in African languages

- Reserach on Somali tonal accent and intonation
- Research on Somali grammar and lexicography
- Preparation of a Somali corpus: Bangiga Af Soomaaliga

What is a corpus?

- **Organized** collection of texts
- Record of the source (link if possible)
- Tagged (for parts of speech, syntactic functions)
- Lemmatized (all forms connected, e.g. buug, buugga, buuggaaga…)

How can a corpus be useful?

- Frequency of words — Which are the most frequent words?
- Meaning of words — What does the word ’xyz’ mean?
- Spelling of words — How is the word ’xyz’ normally spelled?
- Morphological forms of words — Which endings are most frequent for ’xyz’?
- Combinations of words — What words frequently co-occur?
- Frequency of synonyms — Which synonym is the most frequent one?
- Regional variation — Which word is used where?

etc…

Three Somali corpora

- Bangiga Af Soomaaliga, 18.9 million tokens
  - https://spraakbanken.gu.se/korp/?mode=somali#?lang=en
  - At the Swedish Language Bank, University of Gothenburg

- Kaydka Af Soomaaliga, 3.0 million tokens
  - http://www.somalicorpus.com/
  - At Redsea Cultural Foundation, Hargeysa

- Somali Web Corpus, 79.7 million tokens
  - https://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=sowac16;align=
  - NLP Center, Brno University, Czech Republic, in coop. with Oslo & Addis Abeba

Type of content

- Bangiga Af Soomaaliga, 18.9 million tokens
  - Mainly school books, online news, laws, parliamentary transcripts, Wikipedia
- Kaydka Af Soomaaliga, 3.0 million tokens
  - Mainly literature, poetry, songs, news, essays, political speeches
- Somali Web Corpus, 79.7 million tokens
  - Only automatically ”harvested” texts from the Internet

List of most frequent words

- Bangiga Af Soomaaliga (5 M)
  - http://demo.spraakdata.gu.se/martin/somali2018.txt
  - Need to produce new list based on 18 M
- Somali Web Corpus (79 M)
  - https://corpora.fi.muni.cz/habit/run.cgi/wordlist_form?corpname=sowac16
- An Crúbadán (24 M)
  - http://crubadan.org/languages/so

List of most frequent words

Bangiga Af Soomaaliga

oo	141219	25639.740754
iyo	95372	17315.753229
ku	93691	17010.550641
ka	92600	16812.468533
u	75171	13648.056934
ah	70814	12856.999425
ay	60284	10945.171199
ee	53914	9788.633137
la	53105	9641.750988
in	50452	9160.071949
soo	46234	8394.251298
uu	42717	7755.704302
waa	42151	7652.941265
)	36052	6545.606000
ayaa	30704	5574.622396
(	28971	5259.978681
aan	27089	4918.282506
waxaa	25507	4631.054372
waxa	22960	4168.620708
waxay	22953	4167.349787
lagu	22471	4079.837801
loo	21177	3844.898986
ugu	20857	3786.799743
kale	20699	3758.113242
wuxuu	19788	3592.711958

Somali Web Corpus

oo	2,130,200
ka	1,808,365
ay	1,470,184
ku	1,445,719
iyo	1,248,166
ee	1,210,830
ah	1,062,164
u	1,041,418
in	1,037,431
ayaa	985,020
uu	950,971
soo	794,868
la	720,451
lagu	397,822
ugu	365,182
waxa	357,063
kale	304,230
aan	291,053
waxaa	275,538
mid	274,672
sheegay	268,248
loo	264,442
waa	263,871
si	256,775

An Crúbadán

bo 39036
ka 29663
ku 24743
iyo 24550
ay 20215
ee 19765
u 19109
ah 18866
in 15787
uu 14155
soo 13682
la 13672
ayaa 10995
aan 7263
waa 7160
lagu 6641
ugu 5722
aad 5422
waxa 5301
loo 5065
kale 5053
waxaa 5002
mid 4808
si 4644
wax 4575
laga 4276

List of most frequent words

- Very useful if you want to produce a dictionary

How to search for the answers

Dadku waxa ay u waxyeellayn karaan deegaanka siyaabo badan. Dadku waxa ay waxyeelleeyaan quruxdiisa iyagoo ka jarjaraya dhirta oo banneynaya. Waxa ay sumeeyaan biyaha iyagoo ku daraya kimikallo sun ah iyo qashinno kale sida saalada xoolaha iyo qurubyada aadamiga. Sidaa si la mid ah, ayaa

(from Saynis, Fasalka 5aad, Muqdisho 2011)

- kimikallo (3), kimikalladani (1)
- not found in dictionaries

- Simple search
  - https://spraakbanken.gu.se/korp/?mode=somali#?lang=en
  - https://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=sowac16;align=

More complex searches

- Specify type(s) of texts
  - spelling in schoolbooks and other texts

kubbad/kubad

kubbadda/kubadda/kubbada/kubada

	wayn		weyn		
	2478	13%	16795	87%	BAS
	18459	41%	26936	59%	SomWaC
1960’s	0	5%	14	100%	BAS
School	51	5%	911	95%	BAS

	danbe		dambe		
	1501	17%	7567	83%	BAS
	5048	18%	22302	82%	SomWaC
1960’s	10	100%	0		BAS
School	10	5%	208	95%	BAS

More complex searches

- As initial part

what is the plural of abti, oday?

- **oday**, pl. **odayaal** in *Qaamuuska af-Soomaaliga* (Mansuur & Puglielli 2012).
- **oday**, pl. **odayo** in *Barashada Naxwaha af-Soomaaliga* (Puglielli & Mansur 1999).
- **abti**, pl. **abtiyo** *or* **abtiyaal** in *Qaamuuska af-Soomaaliga* (Mansuur & Puglielli 2012).

More complex searches

- As final part

  bare, pl. bar**a**yaal / bar**e**yaal

  def. barayaa**sha** / barayaal**ka**

More complex searches

- Case-sensitiveness

Days, months, seasons:

- Sabti / sabti
- Jiilaal / jiilaal

Nationalities

- af soomaali, af Soomaali, Af Soomaali
- Soomaaliyeed / soomaaliyeed

Spelling issues

SoRu 1969: weyddii; weyddiin
SoIt 1985: weyddii; weyddiin
SoEn 1991: weyddii; weyddiin; weydii
SoFr 1999: weyddii;
SoRu 2012: weydiin; weydiin
SoAr 2015: weyddii; weyddiin

SoSo 1976: Weyddiin
SoSo 2004: weyddii; weyddiin
SoSo 2008: Weyddii; Weyddiin
SoSo 2012: weyddii; weyddiin
SoSo 2013: waydii; waydiin; weydii; weydiin

Extended searches

- Several words / forms at the same time

  weyddii*/weydii*
  wayddii*/waydii*

- What words tend to co-occur?

  udgoon

- What are the proportions among synonyms?

  su’aal / weydiin
  waydiin
  weyddiin
  wayddiin

Airplane in Somali dictionaries

SoRu 1969:	dayuurad	SoSo 1976:	dayuurad (*or* diyaarad) diyaarad *see* dayuurad
SoIt 1985:	dayaarad *see* dayuurad dayuurad (*or* dayaarad, diyaarad) diyaarad *see* dayuurad	SoSo 2004:	dayuurad (*or* diyaarad) diyaarad *see* dayuurad
SoEn 1991:	dayuurad diyaarad (*or* dayuurad)	SoSo 2008:	dayuurad diyaarad
SoFr 1999:	dayuurad	SoSo 2012:	dayaarad *see* dayuurad dayuurad (*or* dayaarad, diyaarad) diyaarad *see* dayuurad
SoRu 2012:	dayuurad diyaarad *see* dayuurad		
SoAr 2015:	dayuurad (*or* diyaarad) diyaarad (*or* dayuurad) *with illustration*	SoSo 2013:	dayaarad (*or* dayuurad, diyaarad) diyaarad *see* dayaarad

Advanced searches with wildcards

- How are borrowed words written?

  ”d.*y.*rad.*”
  ”d.*s.*mb.*r”
  ”k.*mb.*t.*r.*”

- Make a search with several wildcards
- Download as Excel
- Sort
- Count

How to interpret what the computer finds

- Read the examples ”manually”
- Look at statistics
- Look at text type/source