D6.2 & D6.3:
FINAL USABILITY REPORT AND SYSTEM EVALUATION

Grant Agreement number	7F14047
Project acronym	HaBiT
Project title	Harvesting big text data for under-resourced languages
Deliverable title	**D6.2 & 6.3: Final Usability Report and System Evaluation**
Responsible partner	**Janne Bondi Johannessen, UiO and Björn Gambäck, NTNU**
Dissemination level	Public
Due delivery date	30 April 2017 (+ 30 days)
Actual delivery date	1 June 2017

Principal Investigator	**Karel Pala**
Project Promoter	**Masaryk University**
Tel	+420 549 49 5616
E-mail	**pala@fi.muni.cz**
Project website address	www.habit-project.eu

Table of Contents

1. Introduction

There are several challenges with respect to the corpus resources developed in the HaBiT project – Harvesting big text data for under-resourced languages) financed by EEA/Norway Grants), in cooperation with the project Linguistic Capacity Building — Tools for the inclusive development of Ethiopia, financed by the Norwegian NORAD, Norwegian Agency for Development Cooperation.

This evaluation assesses usefulness and the usability of the resources and tools developed for Norwegian (both Bokmål and Nynorsk) and all the Ethiopian languages in the project: Amharic, Oromo, Tigrinya and Somali. The corpora are very different in size given the amount of material that is actually available on the web. The types of text also vary considerably. For some of the languages a problem in the corpus creation process was found already in the harvesting process, given that there are varieties that are very close to each other linguistically. For Norwegian this is true of the two varieties, Bokmål and Nynorsk, with the additional problem that Danish is also a variant that is easily confused with Bokmål. Further, Amharic can be confused with Tigrinya and with Ge'ez (the older clerical language still used by the Ethiopian church).

The evaluations reported here were performed on the versions of the resources and tools towards the end of the project. An initial evaluation of the usability of the tools and the reliability of the produced resources was also conducted during a workshop in Brno in Ferbuary 2017, with participants from the teams at Addis Ababa University, Ethiopia, as well as the Norwegian groups from University of Oslo and NTNU.

This deliverable evaluates the resources and tools for each language in turn, and concludes with an overall summary in a table.

2. Norwegian

Evaluated by: Janne Bondi Johannessen (UiO)

For Norwegian, the most challenging task must be the *language recognition part*. In order to make internet corpora, the texts on the internet must be recognised as being from the intended language. Norwegian Bokmål and Nynorsk are very similar to each other, and Bokmål is also very similar to Danish. In addition, old Norwegian texts were written in Danish. A corpus that contains large chunks of texts from the other languages must be considered unsuccessful. This is therefore something that must be checked.

200 most frequent words from the wordlist

Norwegian Bokmål

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=notenten15_4_bokmal&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

We have evaluated the Habit Bokmål frequency list by comparing it with the NoWaC frequency list (available at the Text Laboratory at UiO, see Guevara 2010), see Table 1.

As Table 1 shows, the most frequent words are more or less the same in both lists, with pronouns and grammatical words topping the list: *og* 'and', *i* 'in', *er* 'is', *å* 'to' (infinitive marker), *på* 'on', *det* 'it', *som* 'which', *en* 'a', *til* 'to', *av* 'of' etc. These results are expected and shows that the method used has been successful. The fact that each word has approximately the double frequency of that of the existing list shows that the corpus is much bigger (double) than the existing one, which is a very welcome result. The Habit Bokmål corpus has 1,178,357,993 words.

Table 1: Comparison of the Bokmål frequency lists from HaBiT and NoWaC

Habit project Bokmål list		NoWaC frequency list
word	**Freq**	
.	66,302,248	16145498 og
,	48,783,706	15256386 i
og	38,459,956	12206184 er
i	29,244,572	11698509 det
er	26,587,096	9571416 på
å	20,257,955	8630777 som
på	20,092,511	8293212 en
det	19,741,244	8168237 til
som	19,704,300	8131491 å
en	19,532,414	7993433 av
til	17,771,732	7491272 for
av	17,219,280	6670288 med
for	15,515,275	6639161 :
med	15,026,187	6531429 -
har	12,734,327	6301028 har
at	11,518,852	6030608 at
ikke	10,277,938	5235776 ikke
jeg	9,448,321	5073332 jeg
de	8,835,895	4612638 de
den	7,746,306	4146748)
kan	7,156,786	4020263 den
du	7,094,265	3980207 om
et	6,959,830	3678914 (
om	6,950,325	3239821 "
var	6,057,924	3199327 et
så	5,799,457	3102444 men
Det	5,634,376	2935861 vi
		2879552 kan
		2874527 du
		2747899 fra
		2632350 så
		2471901 var
		2375553 ?
		2203208 ...
		2104178 dette
		2083048 han
		2051824 seg
		1858213 skal
		1848715 vil
		1818427 !
		1710749 eller
		1690743 også
		1547255 mer
		1508222 _
		1410569 man
		1376300 ut
		1349778 ble
		1331468 da
		1326745 være

Norwegian Nynorsk

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=notenten15_4_nynorsk&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

Habit Nynorsk frequency list

word	Freq
.	3,277,542
,	2,329,138
og	1,967,701
i	1,709,665
er	1,136,473
på	1,007,837
det	943,786
til	916,306
som	905,752
å	747,507
av	736,146
med	730,444
for	683,801
ein	665,297
har	587,469
at	512,242
dei	460,646
var	424,225
ikkje	412,301
eg	392,888
om	362,124
den	321,121
frå	320,828
eit	317,410
Det	289,144
)	278,553
:	271,313
ei	246,774
han	245,719

kan	243,963
(	233,281
så	221,023
seg	220,810
men	214,5

The first part of the Nynorsk list shows the same expected words as the Bokmål one, again indicating that the method is right. There is a lot less written in this variety of Norwegian than in the other, and therefore there is also a lot less in the web. Thus, while there are 38 million occurrences of the most frequent word, the conjunction *og* (‘and’ in Bokmål), there are only around 2 million occurrences of this word in the Nynorsk corpus. Some of the most frequent words have different form in Bokmål and Nynorsk, and we find them in the same position in the respective frequency lists, which is a good sign. For example, the article *ein* (‘a’) is number 12 in the Nynorsk list (with more than 700 000 occurrences) and the equivalent Bokmål *en* (‘a’) number 8 in the Bokmål list (with nearly 20 million occurrences). The Nynorsk version *ikkje* (‘not’) is number 17 in its list (with more than 400 000 occurrences), while the Bokmål negation *ikke* (‘not’) is number 15 in its list (with more than 10 million occurrences). The similarities between the position of the frequent words on the lists therefore indicate that the lists and therefore these two corpora are good reflections of the two languages.

Language recognition

While the language recognition of Bokmål and Nynorsk look right when considering the top positions of the frequency lists, we need to check the corpora for the "wrong" words. Thus, we do not want to find a high number of Nynorsk words in the Bokmål corpus or vice versa. Neither do we want to find Danish words in the Bokmål corpus.

Looking for the Bokmål *ikke* (‘not’) in the Nynorsk corpus reveals that there are more than 4000 occurrences of this word. But given that there are more than 400 000 of the correct equivalent *ikkje* (‘not’), we can accept that the Nynorsk corpus has 1 % Bokmål versions of this word. We also find nearly 20 000 occurrences of the Bokmål *en* (‘a’), but this is still a small number compared with the Nynorsk *ein* (‘a’) of more than 700 000. Given that *en* (‘a’) in Bokmål to some extent replaces the feminine article *ei* 'a' as well, we must also look for this feminine article in the Nynorsk corpus, and find 263 000 occurrences. Thus the Bokmål article in the Nynorsk corpus makes up 20 000 compared with 963 000 equivalent Nynorsk articles, which is a low number, i.e. ca 2 %.

One might wonder why there are Nynorsk words in the Bokmål corpus at all. The reason is linked to the fact that the Nynorsk language norm is in minority in Norway, so that many of those that write Nynorsk also use Bokmål words in their Nynorsk texts as a result of influence by the majority. The texts show this clearly: Many are from informal blogs with a mix between the two norms.

Looking for Nynorsk words in the Bokmål corpus gives the following results. The nynorsk negation *ikkje* (‘not’) occurs nearly 38 000 times, but given that there are more than

10 million occurrences of the Bokmål *ikke* (‘not’), only 0,4 % are Nynorsk. There are 31 000 occurrences of the Nynorsk *ein* (‘a’) in the Bokmål corpus, which makes up 0.1 % of the 31 million occurrences of the correct Bokmål occurrences of *en* (‘a’). Again, we find that blogs are the cause of many of the occurrences from the other norm. The result is therefore very good.

We must also check the results compared with Danish. Searching for the Danish *næsten* (‘almost’) in the Bokmål corpus gives under 300 hits, as against the Bokmål *nesten* (‘almost’), which occurs nearly 500 000 times. There are, then, 0.15 % of this Danish word in the corpus. There are less than 100 occurrences of the Danish *fuld* (‘full’) in the corpus, as against 420 000 occurrences of the Bokmål *full* (‘full’), giving 0.0025 %. The amount of Danish is negligible. The few occurrences that do occur are found in texts like blogs and film titles.

It can be concluded that the language recognition module works well and has produced good quality web corpora.

Random concordance pages

To check the concordance production, concordances for at least five words or phrases were generated for each language.

Norwegian Bokmål

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=notenten15_4_bokmal

***arbeide* 'work', as a verb:**
More than 10 000 hits, that look right, though there are some instances of wrongly tagged words, so that there are some nouns in between the verbs.

***arbeid* 'work', as a noun:**
More than 470 000 hits, that look right and rightly tagged.

***snill* 'kind', as an adjective:**
More than 65 000 hits. These include many inflectional forms: *snill, snille, snilt, snilleste, snillere*.

***glad for* 'happy about', i.e., a phrase:**
More than 54 000 hits that are good.

***tsjekkisk* 'Czech', as a lemma:**
Nearly 4000 hits that look right. Both inflectional forms: *tsjekkisk, tsjekkiske*.

Norwegian Nynorsk

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=notenten15_4_nynorsk

The Nynorsk corpus is not tagged.
***arbeide* 'work':**
Over 5000 hits. Only this word form, since the corpus is not tagged.

***arbeid* 'work':**
20 000 hits that look right. Only this word form.

***snill* 'kind':**

1300 hits, only this form.

***glad for* 'happy about':**
2700 hits. Only these word forms.

***tsjekkisk* 'Czech'**
200 hits. Only this word form.

Word sketches

Word sketches are one page summaries of collocations and grammatical behaviour of words. To test these, word sketches for at least five frequent words (varying in the part of speech) for each language were generated (names have to be lowercased, e.g., “norge”, “oslo”)

Norwegian Bokmål

http://corpora.fi.muni.cz/habit/run.cgi/wsketch_form?corpname=notenten15_4_bokmal

tegne 'draw', word sketch:
The **modifiers** (postpost) are very good: *dyster* 'gloomy', *deretter* 'then', *ferdig* 'finished', *først* 'first'
The **objects** are good: *forsikring* 'insurance', *membership* 'medlemskap', *portrett* 'portrait',: *tegne* -> psykolog,
The subjects are mostly, but not always good. Sometimes the head of a relative clause is misinterpreted as a subject:

en ***tegneserie*** **tegnet** for hånd = en cartoon drawn by ...

The **coordination** (tegne and/or) are very good: write, paint, explain, ...
The **pronominal** subjects are good.

synge 'sing', verb, word sketch:
The (postposed) **modifiers** are good: choir, solo, live, together...
The **objects** are good: song, duet, birthday song, melody, opera...
The **subjects** are good: bird, choir, artist, ...
Coordination and/or is good: dance, play, speak...
Pronominal subjects are good

elev 'pupil', noun, word sketch:
Modifiers are good: single, minority-linguistic, good...
Verbs with *elev* as object are good: teach, help, let, motivate
Verbs with *elev* as subject are good: learn, participate, work, experience, make...
Coordination and/or is good: teacher, parent, student...
Elev **is**: eager, motivated, engaged, good...

***lat* 'lazy', adjective, word sketch:**
Modifiers are good: too, intellectually, quite, very...
Coordination and/or is good: fat, useless, spoilt, stupid
Subjects of 'be lat' are good: youth, student, Norwegian...
Verbs modified by 'lat' are slightly strange: understand, call, film, feel...
Nouns modified by 'lat' are good: summer's day, holiday, morning...

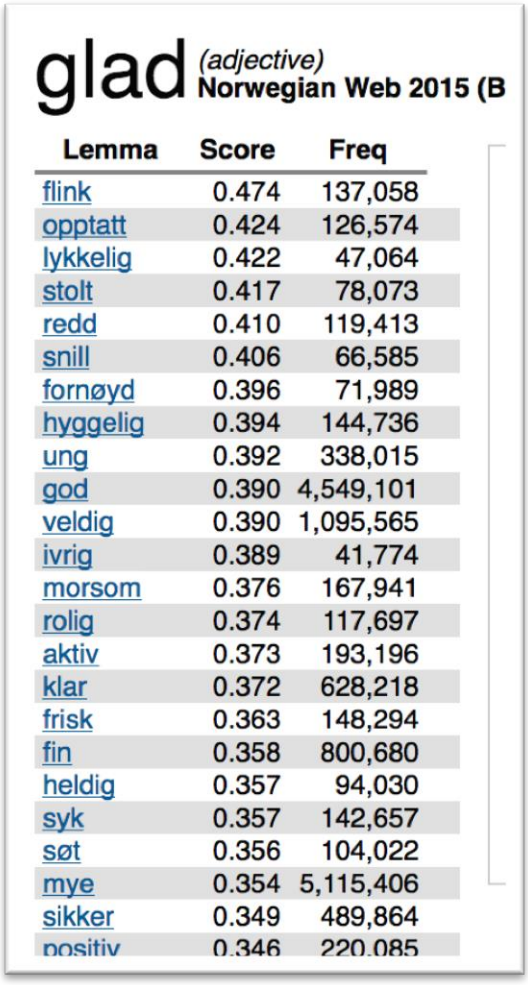
***blå* 'blue', adjective, word sketch:**
Modifiers are good: dark, glowing, light...

Coordination and/or is good: white, green, yellow, pink...
Subjects of 'be blå are good: sky, colour, sea...
Nouns modified by 'blå' are good: sky, eye, sea...
Verbs modified by 'blå' are slightly strange: lighten, colour, unite...

Thesaurus

Norwegian Bokmål

http://corpora.fi.muni.cz/habit/run.cgi/thes_form?corpname=notenten15_4_bokmal



glad *(adjective)* Norwegian Web 2015 (B

Lemma	Score	Freq
flink	0.474	137,058
opptatt	0.424	126,574
lykkelig	0.422	47,064
stolt	0.417	78,073
redd	0.410	119,413
snill	0.406	66,585
fornøyd	0.396	71,989
hyggelig	0.394	144,736
ung	0.392	338,015
god	0.390	4,549,101
veldig	0.390	1,095,565
ivrig	0.389	41,774
morsom	0.376	167,941
rolig	0.374	117,697
aktiv	0.373	193,196
klar	0.372	628,218
frisk	0.363	148,294
fin	0.358	800,680
heldig	0.357	94,030
syk	0.357	142,657
søt	0.356	104,022
mye	0.354	5,115,406
sikker	0.349	489,864
positiv	0.346	220,085

The thesaurus is good for finding alternative words, but the words are not necessarily synonyms. For example the adjective *glad* 'happy, glad' have as the most frequent suggestions: clever, busy, happy, pround, afraid, kind, contented. Another example is *skrive* 'write', with the "synonyms": read, tell, mean, see, say, speak.

Overall, the Norwegian resources have high quality.

3. Oromo

Evaluated by: Fede Negesse (AAU)

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=orwac16&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

Most frequent words

After checking, the most frequent words in the web corpus of Oromo, were identified as being words that are mostly function words which exist in the language. The results make sense, but there are bound morphemes (hin-, ni-, nu-, haa-, -e, -a, -u) that were separated from other morphemes, likely during tokenization.

Examples:

hin- hin-deemu (I/he won't go)
ni- ni-deemu (They will go)
nu- nu-beeku(They know us)
haa- haa-deemnu(Let's go)
-e/-a/-u haa-taa'-u (Let him sit)
taa'-e (He/ sat)
taa'-a (Sit or take a chair)

-ti (copula) Kun mana kee-ti (This is your house)
in- (allomorph of /ni/) An in-deema (I will go) for some dialects of Oromo.

The wrong tokenization has to do likely with the glottal stop /?/, orthographically represented by an apostrophe /'/. The problem is significantly reduced now. It was rampant before the evaluation workshop in Brno.

Concordances

An example of the concordance is given below.

Query **hin, deemu** **29** (5.70 per million)

Page 1 of 2 Go Next I Last

bilisummaa.com	bara keessa jirru kana waliinis gonkumaa	hin deemu	. Odoo seerri humnaan amantii ofii
omicwomo.info	ittiin yaalamus argachaas hin jiru argatuufis	hin deemu	ka nama sodaachisu mootummaan warra du'a
finfinnetribune.com	hanqinni kun kan Hogganaatii? kana irra	hin deemu	. Yaadoota kanan kaasuuf, Ani Hoggana
ayyaantuu.net	kee lameenuu, mata duree kenniteef waliin	hin deemu	. Tasuma barreeffama kee ijaa fi ijoo hin
oromedia.net	otoo guyya sadan tokko waliin hidabaarsin	hin deemu	.Baqalaa Nadhii anaf obboleecha jechuu dha
qeerroo.org	keenyaa hamma Wayyaaneen jiru hiikkachuuf	hin deemu	,kanaaf manni Hidhaa keessii fi alli isaa
qeerroo.org	qabeenyaa habashaa kun naannoo kana gadhiisee	hin deemu	taanaan fuuldura lubbuu isaaf sodaachisaa
gubirmans.com	kaanne qaba. Bilisummaa sabicha irra fagaatee	hin deemu	. Kan hojiin farraa ta'e illee afaaniin
blogspot.cz	akka danda'amu mul'isaa jirti. Kana irra	hin deemu	. Shira diinaa kana fashalsuuf waan hojjachuu
gadaa.net	qawweedhaan dhufe, humnaan malee namarraa	hin deemu	. Kana isaanuu ifaa ifatti dubbachaa turan
voaafaanoromoo.com	dhihaatanii, akkasitti immoo haasaan nagaa	hin deemu	. . . " jedhan. sagataan keessan
jw.org	ifni dungichaa dhaamu ifni sun eessayyuu	hin deemu	. Salphaadhumatti dungoon sun ifa kennuusaa
ulfo.org	ergarama mana hidhaatti irra geyerraa haasawuuf	hin deemu	. Mana hidhaa keeysatti addaachuu, daaruu
oromoliberationfront.org	kophee malee deema; garuu mataa gadi qabatee	hin deemu	, jedhe. Yeroo sanatti qabsoo bilisummaa
voaafaanoromoo.com	dhabsiisan. Walitti bu'insi sunis wal waraansatti	hin deemu	jedhan. Yunivarsitii Jimmaa fi Mormii
gadaa.com	hari'uun Ulee lafa hin godhu. Harka duwwaa	hin deemu	. Saroonni yoo itti dhufte qarrifaa ishii
oromiatimes.org	waraana naquun maal nu akeeka ??? kana irra	hin deemu	. garuu waan hamileen yaadinu fi saganteeffannee
gadaa.net	akkuma yaadannu warreen amantaa keenya waliin	hin deemu	jedhaa turan ni jirani. Isaan kunneen maaliif
readbag.com	nam tokkeenis wal abaaruun furmaata ta'uuf	hin deemu	. Hegereen Oromoo bareeduuf, ar'a hujii
gadaa.com	garboomsaa kamiyyuu waliin araara uumuuf	hin deemu	; araarri kan dhufu, mirgii fi dantaan ummata

Page 1 of 2 Go Next I Last

Word sketches

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=orwac16

The search for words or phrases generally produces good results.

The hits from a search for word sketches is illustrated below.

deemu **Oromo WaC [2016] freq = 696** (136.69 per million)

modifiers of "deemu"	356	1.00
khana	3	8.08
taasisa	3	7.57
sakatta'aa	2	7.48
alaalatti	2	7.48
mara	5	7.45
ammatti	2	7.31
rasaasaan	2	7.14
hundatti	3	7.08
gochuuf	5	6.98
ennaa	3	6.95
eessa	2	6.85
godhanii	2	6.68
godha	3	6.52
qeerroon	4	6.46
qabanii	2	6.44
daandii	2	6.24
godhe	2	6.06
sababa	2	5.86
tahuu	3	5.81
turan	8	5.05
waliin	4	4.84
kanatti	3	4.80
danda'	2	4.65
humna	2	4.47
nama	7	4.25

verbs with "deemu" as object	35	0.90
jechuu	2	4.19
qabnu	2	4.19
qaba	5	4.03
yommuu	4	3.97
kanaan	4	3.86
gochuu	2	3.79
jira	5	3.00
jechuun	2	2.79
qabu	2	2.13

verbs with "deemu" as subject	19	1.00
qaba	4	4.09
jira	8	3.95
wal	2	3.17

"deemu" is ...	3	3.90
kanatti	3	4.96

"deemu" is a ...	42	1.40
alaalatti	2	10.02
dhugumatti	2	9.19
ammatti	2	8.71
hundatti	3	7.11
godha	3	6.57

nouns modified by "deemu"	429	1.30
ta'uuf	8	8.44
guddachaa	6	8.40
dabalaa	6	8.30
tahuuf	4	8.04
dhihaachaa	4	8.00
jabaachaa	4	7.85
baadhatee	3	7.82
dhalachuuf	3	7.78
dhufuuf	4	7.77
garamitti	3	7.77
ennaa	6	7.75
karaarra	3	7.70
fudhatanii	4	7.56
yommuun	3	7.53
galuuf	3	7.51
fuudhee	3	7.40
dabree	3	7.34
baatee	3	7.29
kheessa	2	7.24
qottoon	2	7.24
babaldhachaa	2	7.23
diigamuuf	2	7.23
oolfamuuf	2	7.23
babal'ataa	2	7.21
fudhatamuuf	2	7.20

4. Amharic

Evaluated by Derib Abo (AAU)

Word frequency list

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=amwac16&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

There are several punctuation marks that are generated as words. Three different letters that should have been written together with a word are presented separately and a symbol for addition is also generated as the most frequent word.

The words that are generated as most frequent are likely to be most frequent specially based on the type of data collected. On the top are the stop words and the verb ነው, which is likely to be the case for any corpus on Amharic. Since online sources in Amharic have political and religious content, the content words are from these two domains and related subjects.

Concordances

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=amwac16

The concordance is very good. I have tried to display concordance for the following words: መንግስት ‘government’, ቤት’house, home’, አንዱ ‘the one’, ሳይንስ ‘science’, የተለያዩ 'different’.

All sentences on the first page of the concordance are correct.

Word sketches

http://corpora.fi.muni.cz/habit/run.cgi/wsketch_form?corpname=amwac16

Generally, the modifiers (modifiers and modified by) and the and/or sketches are good. The *subject* sketches and the *verbs with* sketches seem to be a bit problematic because Amharic sentences are complex sentences with several verbs in one sentence. There are spelling errors in the data by the writers and that has contributed its share, specially for the word ሰራ'[he] worked'. Compound nouns could be treated as modifiers sometimes when they do not refer to the specific proper or common noun.

Word sketch of a noun:

ሰው 'man, human being'

Amharic WaC [2013 + 2015 + 2016] freq = 41,861 (2,063.41 per million)

modifiers of "ሰው"	21,181	1.20	verbs with "ሰው" as subject	39,971	2.10	"ሰው" and/or ...	663	0.80	nouns modified by "ሰው"	19,963	1.20
አንድ 'one'	4,345	11.15	ይችላል 'can be'	441	7.70	ድርጅት 'organization'	178	9.89	ሰራሽ A compound noun with ሰው to mean man made	191	8.27
ማንም 'anyone'	490	9.41	ሆኖ 'being'	469	7.68	አንድ 'one'	208	8.68	መሆን 'being'	263	8.19
አንድን 'one, object'	251	8.49	እንደሆነ 'that being'	372	7.12	እንስሳ 'animal'	88	8.55	“ punctuation	353	7.44
የአንድ 'that of one'	235	8.31	የለም 'not'	227	7.10	ተቋም 'institution'	88	8.40	‘ punctuation	124	7.12
በአንድ ‘by one'	253	8.06	ብሎ 'he having said'	319	7.01	ነገር 'thing'	118	8.35	ነኝ 'I am'	98	7.12
ዓይነት ‘type, kind'	232	7.91	አልነበረም 'it was not'	188	7.01	አካል 'body'	98	8.34	* **symbol**	100	7.09
ለሌላ 'for other'	126	7.56	አይደለም 'it is not'	278	6.96	ቡድን 'group'	98	8.24	መሆኑን 'that he is'	133	7.03
ከአንድ 'from one'	140	7.53	አለ 'he said', ‘he is present'	236	6.92	”	75	5.67	በሞት 'by death'	83	7.01
ብዙ ‚many'	201	7.46	አይችልም 'it cannot'	171	6.86				የለም 'there is not'	73	6.74
ታላቅ 'great'	134	7.45	ቢሆን	200	6.83				” punctuation	221	6.68
የአንድን 'that of one'	113	7.41	ሳይሆን	265	6.79				በመሆኑ 'by being'	71	6.59
ለአንድ 'for one'	117	7.35	ከሆነ	185	6.65				አንድ 'one'	117	6.57
ፍጹም 'perfect'	106	7.28	ይሆናል	202	6.62				ሕይወት 'life'	73	6.56
እኚህ	102	7.27	ሊሆን	170	6.60				ማን 'who'	65	6.52
አንድም	101	7.22	ነበር	143	6.44				ልብ 'heart'	63	6.44
የመጀመሪያው	104	7.16	ሆነ	143	6.37				ከእግዚአብሔር 'from God'	56	6.44
ከፉ	103	7.16									
የሚባል	102	7.15									
ሌላ	100	7.13									
አምላክ	119	6.96									

Word sketch of a verb:

ሰራ 'worked'
The ones highlighted in yellow are the result of the spelling error that the writers made. They wrote ሰሪ 'worked, verb' in stead of ስራ 'work, noun'

Amharic WaC [2013 + 2015 + 2016] freq = 194 (9.56 per million)

modifiers of "ሰራ"	**108**	**1.20**		**nouns modified by "ሰራ"**	**65**	**0.80**
በእርሻ 'in farming'	2	8.16				
ህበረት 'in unions'	2	5.70	ፈጣሪ 'creator, God'	2	6.00	
ታሪክ 'history'	6	5.32	አስኪያጅ 'excutive'	2	5.03	
የልማት 'developmental'	4	4.66	.. punctuation	2	4.39	
ስራ 'work, Noun'	3	4.14	? punctuation	12	3.81	
አይነት 'kind, type'	2	3.47				
ጥሩ 'good'	2	3.38				
ነገር 'thing'	5	3.30				
ሥራ , 'work, Noun'	3	2.87				
ከፍተኛ 'higher'	2	2.14				

subjects of "ሰራ"	**66**	**0.70**		**verbs with "ሰራ" as subject**	**161**	**1.60**
በሌሎች 'by						
This is instrumental, not subject others'	2	5.42				
ወገኖች 'sides'						
This is instrumental, not subject	3	4.75				
ስራ ;work, Noun'	4	3.56				
		0.16				
			ለተሰማሩ ' for those working in..'	2	7.37	
			ቢባል ' if [it is] said so'	3	6.15	
ጊዜ 'time' This is an adverbial, not a subject	2		የተሰማሩ 'those working in ..'	2	5.31	
			የሚሰሩ 'those who work'	2	5.16	
			ቢያንስ 'at least'	2	4.45	
			አይደሉም 'they are not'	2	3.72	
			ማለት 'means'	3	3.56	
			የሚለው 'means'	2	1.93	
			ተብሎ 'alled, referred to'	2	1.73	

Word sketch of an adjective:

አዲስ 'new'

Amharic WaC [2013 + 2015 + 2016] freq = 20,015 (986.58 per million)

modifiers of "አዲስ"	**393**	**0.10**
በሸራተን compound noun with አዲስ meaning Sheraton Addis	11	9.69
ትናንት 'yesterday', adverbial	21	7.94
ዘንድሮ 'this year', adverbial	7	7.30
ሰሞኑን 'these days', adverbial	19	7.28
በተለይም 'specially'	24	6.68
በተለይ 'specially'	31	6.57
ዛሬ 'today', adverbial	52	6.53
ገና 'not yet'	18	6.51
የ I should be attached to a word, cannot stand alone.	17	6.22
ነገ 'tomorrow'	9	6.22
ሁልጊዜ 'every time', 'always'	5	5.84
ዛሬም 'even today', adverbial	5	5.59
አሁን 'now' , adverbial	29	5.49
እንደገና	6	5.29
ለምን	5	3.84
አቶ 'Mr.' title of a person. አዲስ can be the name of a person.	8	3.67
nouns modified by "አዲስ"	**12,916**	**4.70**
አበባ 'compound noun with አዲስ meaning Addis Ababa	4,285	12.46
ዘመን 'time'	648	9.68
አበባን "compound noun with አዲስ meaning Addis Ababa	264	9.34
ነገር 'thing'	479	8.83
ዓለም 'compound noun with አዲስ meaning Addis Alem	288	8.78
ምዕራፍ 'chapter'	122	8.08
ጋዜጣ 'newspaper'	100	7.67
ትርጉም 'interpretation'	101	7.59
ከተማ 'town', also 'compound noun with አዲስ meaning Addis keteme, one of the districts of Addis Ababa	133	7.54
ኪዳን 'testament' , 'compound		

Thesaurus

http://corpora.fi.muni.cz/habit/run.cgi/thes_form?corpname=amwac16

Though there are good similar words in the thesaurus, the majority of them are just modifiers or collocations with the word presented.

Example መንግስት 'government' has the following similar words:

መንግሥት 'government;
ህዝብ 'people'
ፓርቲ 'party'
ኢህአዴግ 'EPRDF'

But the following are collocations

አንድ 'one', የኢትዮጵያ 'that of Ethiopia', ጥያቄ 'question'.

There are punctuation marks and symbols that come as words in the thesaurus as well. They need to be taken care of.

Example of the thesaurus results for *government*:

መንግስት 'government'

Amharic WaC [2013 + 2015 + 2016] freq = 17,661 (870.54 per million)

Lemma	Score	Freq
መንግሥት 'government;	0.490	16,300
“ punctuation	0.333	102,859
ህዝብ 'people’	0.331	15,175
ፓርቲ 'party’	0.327	12,220
ኢህአዴግ 'EPRDF’	0.319	6,503
አንድ 'one’	0.315	50,442
ጊዜ 'time’	0.312	64,423
ቤት 'house’	0.310	37,620
ሁኔታ 'situation’	0.303	26,643
የኢትዮጵያ 'that of Ethiopia’	0.301	24,067
ሕዝብ 'people’	0.299	11,645
ጉዳይ 'affair’, ‘issue’	0.299	23,272
” punctuation	0.298	108,069
መንገድ 'road’	0.295	24,393
ኢትዮጵያ 'Ethiopia’	0.294	20,687
በኢትዮጵያ 'That of Ethiopia’	0.294	12,532
ድርጅት 'organization’	0.294	9,368
ወቅት 'season’	0.293	19,982
ሰው 'man’	0.293	41,861
ወያኔ 'derogatory name of TPLF’	0.292	7,085
አገር 'country’	0.290	15,184

5. Somali

Evaluated by Ahmed Yusuf

Word frequency list

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=sowac16&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

Almost all words on the frequency list are correct Somali words. There is the word 'home' in the top list, which might be the 'home' button from the home page source. When in addition inspecting the Somali WaC [2016], a few errors were found, for example nouns modified by 'dariiq'; however, except for *nin, waana, meel,* and *mid,* all the rest would definitely be verbs together with 'dariiq'. The word *dariiq* ('way, road, path') is itself a noun, from Arabic طريق. In the list there is only one noun: *nin* ('man'). This needs to be further checked.

Concordances

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=sowac16

An example of a Somali concordance is given below.

Query **dariiq** **709** (8.89 per million)

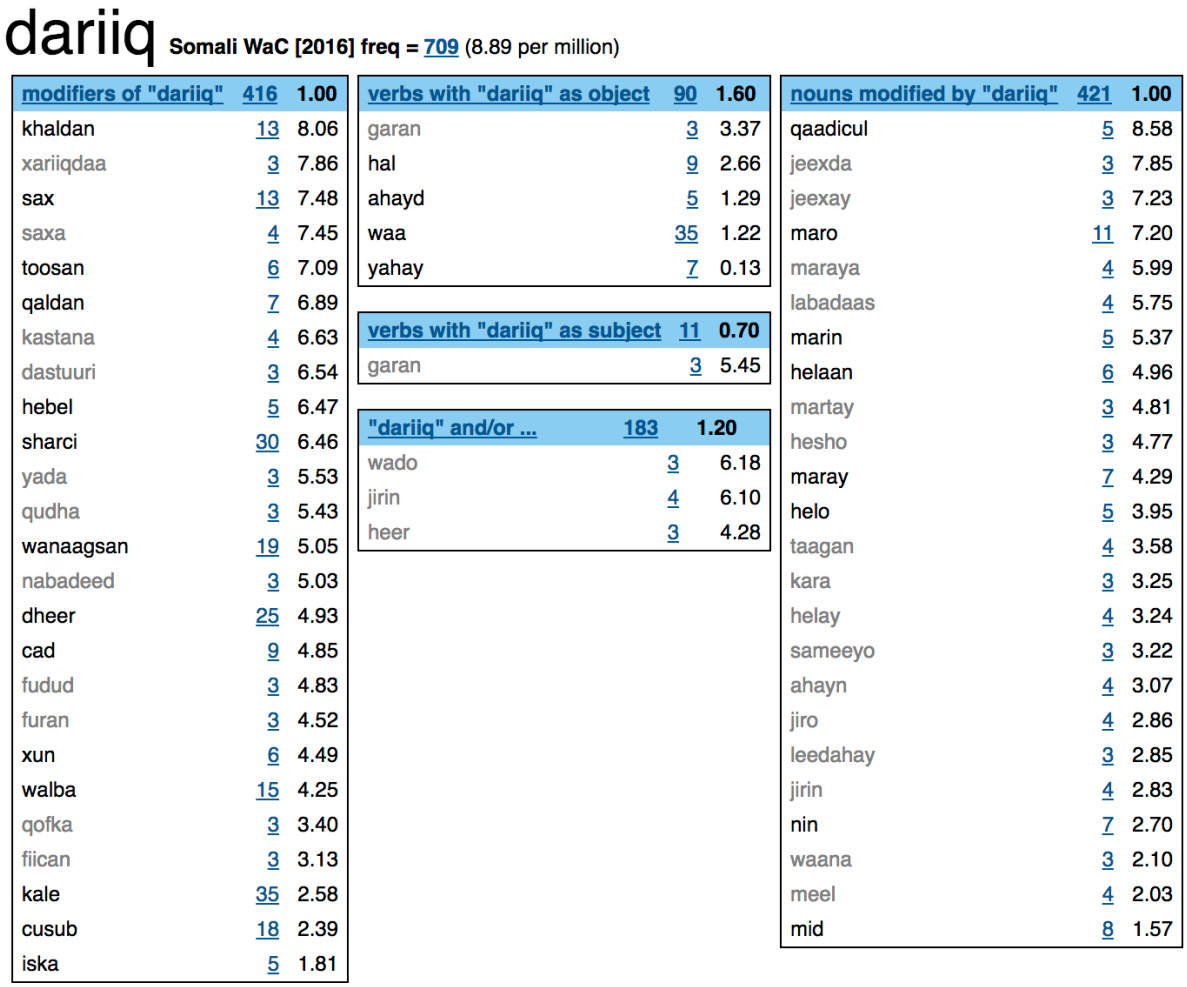
Page 1 of 36 Go Next | Last

infopankki.fi	labaatankii saac. Hosteelku caadi ahaan waa	dariiq	ka jaban hoteelka, lakiin heerka tayada
haatuf.net	siyaasada taasina ay shirka ka dhigtay mid	dariiq	khaldan ku socda sidaa darteed ayay yidhaahdeen
shabellenews.com	ay dib ugu soo noqon karto ay doonayaan,	dariiq	kasta haloo mare,taasi oo keeni karta in
hablaha.com	gurigiisa, ooridiisa dhabta ah ee haddii aad	dariiq	wada martaan aanad ka fikiraynin in lagu
haatuf.net	umada reer Somaliland, waxay usoo mareen	dariiq	dheer, waxaynu umad ahaan ku difaacnaa
hoygasuugaanta.com	gabyeen rag kaloo gabyaa waaweyn ahaa oo uu	dariiq	cad marsiiyey. Gabayaagii Cabdi Gahayr
allgedo.com	lasiiyo wey wada gadgadn waana qaadicul	dariiq	aan wadini aheyn majeerteen muxuu kadali
puntlandtop.net	baxaa, dhawr bilood. Waxaa jira dhawr	dariiq	oo dhakhtarku u baari karo hadaad sonkor
halbeegnews.com	waxa Ummada reer Somaliland ay usoo martay	Dariiq	aad u dheer oo dhib badnaa iyo dedaal xooggan
gabiley.net	siyaasiyan-na u fiicneyn in uu waddan qudhi	dariiq	wax usoo maraan u noqdo. Wuxu lagama-maarmaan
gabiley.net	gaajoonaya loo waddo hellaan marin amma	dariiq	ay maraan oo si dhakhso ah ku gaadhi karaan
gabiley.net	Meles oo waxa Berbera Corridor lagu daray	dariiq	-yada muhiimada dhaqaale u leh Afrika-da
gabiley.net	1961 inqilaabkii uu sababay midowgii aan	dariiq	caddaaladeed iyo wax wada lahaanshaha loo
allsomali24.org	3- Beeshu waxay is ticmaali doontaa	dariiq	kasta oo ay xaqeeda ku heleyso. 1-B
tooshnews.net	aaminsan in khilaafaadka Dawliga ah lagu xalilo	dariiq	Nabadeed iyo Wadahadal, waxaanay aaminsantahay
somaliland.org	Mustafe Cabdi Ciise (Mustafe Shiine) "Nin	dariiq	xariiqdaa garan-kara dayoobay, dhidarkii
somaliland.org	waxa ay cadeynaysaa in qofka leh ama jeexda	dariiq	uu doonayo in uu meel […] Maayarka
somaliland.org	dibna reer Somaliland taas mid la mid ah	dariiq	loogu jeexayaa, in lagu sameeyo tartan
somalilandpost.net	Ismaaciil (Saylici) waxa uu sheegay in saddex	dariiq	ay sahlayaan in ardaydu ay waxbarasho helaan
tooshnews.net	weeyi inaynu hal u jeedo yeelanyo inaynu hal	dariiq	u wada marno aayaha mustaqbalka caruurteena

Page 1 of 36 Go Next | Last

Word sketches

An example of a Somali word sketch is given below.



dariiq **Somali WaC [2016] freq = 709** (8.89 per million)

modifiers of "dariiq"	416	1.00
khaldan	13	8.06
xariiqdaa	3	7.86
sax	13	7.48
saxa	4	7.45
toosan	6	7.09
qaldan	7	6.89
kastana	4	6.63
dastuuri	3	6.54
hebel	5	6.47
sharci	30	6.46
yada	3	5.53
qudha	3	5.43
wanaagsan	19	5.05
nabadeed	3	5.03
dheer	25	4.93
cad	9	4.85
fudud	3	4.83
furan	3	4.52
xun	6	4.49
walba	15	4.25
qofka	3	3.40
fiican	3	3.13
kale	35	2.58
cusub	18	2.39
iska	5	1.81

verbs with "dariiq" as object	90	1.60
garan	3	3.37
hal	9	2.66
ahayd	5	1.29
waa	35	1.22
yahay	7	0.13

verbs with "dariiq" as subject	11	0.70
garan	3	5.45

"dariiq" and/or ...	183	1.20
wado	3	6.18
jirin	4	6.10
heer	3	4.28

nouns modified by "dariiq"	421	1.00
qaadicul	5	8.58
jeexda	3	7.85
jeexay	3	7.23
maro	11	7.20
maraya	4	5.99
labadaas	4	5.75
marin	5	5.37
helaan	6	4.96
martay	3	4.81
hesho	3	4.77
maray	7	4.29
helo	5	3.95
taagan	4	3.58
kara	3	3.25
helay	4	3.24
sameeyo	3	3.22
ahayn	4	3.07
jiro	4	2.86
leedahay	3	2.85
jirin	4	2.83
nin	7	2.70
waana	3	2.10
meel	4	2.03
mid	8	1.57

6. Tigrinya

Evaluated by Shimelis Mazengia and Tsegay (Tigrinya Team, AAU)

Word frequency list

http://corpora.fi.muni.cz/habit/run.cgi/wordlist?corpname=tiwac16&refs=%3Dword&wlmaxitems=200&wlsort=f&wlattr=word&wlminfreq=5&wlmaxfreq=0&include_nonwords=1&wlnums=frq&wltype=simple

1. Stop words and punctuations are presented as the most frequent words. This shows that the text needs cleaning and normalization. For instance, the full stop in Tingrinya has not been normalized. At times, it appears as a single punctuation mark (። 116,579 times). At other times, it appears as a combination of two two-dots (:: 545 times).

2. The most frequent nouns in the first list of 200 words are:

ኣምላኽ	9,089
ሓደ	8,791
ቅዱስ	8,330
ሰብ	7,862
ህዝቢ	4,896

3. The most frequent verbs in the first list of 200 words are auxiliary verb as seen below:

ዘሎ	7212
ኣሎ	4303
ኢሉ	4228
ማለት	4145
ዝነበረ	4098

4. The first main verb was encountered in the third list of 200 words which is ጀማሩ (632 times).

5. In the word list, there are numerals included in the category of the most frequent words.

6. There are some dependent morphemes and also letters which appear as frequent words (ን ይ ዶ ሞ).

7. The first element of some compound nouns (which is modified for compounding) is presented as independent frequent word (ስነ ደቀ). These were found in the third list of 200 words.

Concordances

http://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=tiwac16

The five nouns in 2 and five verbs in 3 above were checked in the concordance. They were found to be in the correct Tigrinya contexts. They all make sense in their collocations.

An example of a concordance in Tigrinya is given below:

Query ዘሎ **7,212** (2,848.97 per million)

Page 1 of 361 Go Next | Last

zaratbebat.com	ብዛዕባ መፈጠር ወዲ ሰብ ናይ'ዚ ግዜ'ዚ ዓውደ ፍልጠት ኣቕሪቡዎ ከም	ዘሎ	መረዳእታ'ኳ ... ምሉእ ትሕዝቶ 12/16/2015 | Comments
nselam.com	ተቓውሞ ክስጉም ኣይኮነን ምንቕ ንኽብል እውን ከልኪልዎን ዓረትዎን	ዘሎ	ሓላኺ ፍጹም ዘየቀራርብን ዘየረዳድእን በታኒ ሓሳብ ገለ ውድባት
nselam.com	ዝመርሑን ዝዝውሮን ጉጅለ ህግደፍ ኣብ ናይ ቀቢጸ ተስፋ ኩነታት ኢዩ	ዘሎ	። ልዕሊ እዚ በሲሉ ፈጋዕጋዕ ዝበለ እዋን ንደምበ ተቓውሞ ዝበለጸን
nselam.com	ብመትከልን ዕላማን ተኣማሚንካ ኣብ ምስራሕ ስግኣትን ትዕዝብትን ከም	ዘሎ	ዝከሓድ ኣይኮነን። ስለዚ ከምቲ ናይ ዝሓለፈ ንስራሕ ዘይኮነ ንስለ
nselam.com	ሕቶ ንስለ ሕቶ ኣይኮንኩን ኣልዒለዮ ዘለኹ። ኣብ ደምበ ተቓውሞ	ዘሎ	ከፍወስ ዘይተኻእለ ክሮኒክ ሕማም ስለ ዝኾነ ኢየ። ጉጅለ ህግደፍ
nselam.com	ስለ ዝኾነ ኢየ። ጉጅለ ህግደፍ ነቲ ፍትሒ ይንገስ ኢሉ ዝቃወሞ	ዘሎ	ኤርትራዊ ንገለ ምስ ሃይማኖት ንገለ ምስ ወገን እና ጠቀነ ዕድመ
nselam.com	ማእሰርትን ግልያነትን መሪርዎ እግሩ ናብ ዝመረሑ ፋሕ ብትን ኣብ ዝብለሉ	ዘሎ	እዋን፡ ንፍትሕን ዲሞክራስን ንቃልስ ኣለና ካብ ዝብሉ መራሕቲ
gov.et	የብሉን። መ.ግዝያዊ ኮሚቴ ብሓባር ብምኺን ኣብ ዕዳጋ ካብ	ዘሎ	ቀረብ ዘሓሸ ፅሬትን ዋጋን ዘለዎ ወይ ዘለዎም መቕረብቲ ብምምራፅ
nselam.com	2003ዓ.ም ጀሚሩ ዝፀንዐ ይኸውን። ነዚ ንምግባር ዚኋፍኣና	ዘሎ	ምድራዊ ወይ ሓላፊ ነገር ንምዕታር ወይ ከብርን ዝናን ተመኒና
nselam.com	ዬ። ስለ'ዚ ሃገር ናይ ሰባት ድኻ ትኸውን ከምዘላ ብግልጺ ዚረአ	ዘሎ	ኩሉ፡ ኣብ ዝኾነማሳበራዊ ጕዳይ ከም እንዕዘቦ፡ እቲ መንእሰይ
nselam.com	ኣይከውንን፡ ሃገር ከኣ ሃገር ኣይትኸውንን። እዚ ዘብለና	ዘሎ	ድማ፡ ኣብ ቅድሚ ዓይንና ይፍጸም ዘሎ፡ ናይ ሰባት ቀጻሊ ዋሕድን
nselam.com	ኣይትኸውንን። እዚ ዘብለና ዘሎ ድማ፡ ኣብ ቅድሚ ዓይንና ይፍጸም	ዘሎ	፡ ናይ ሰባት ቀጻሊ ዋሕድን ዋሕዝን ብሕቱ ዘይኮነ ንጽባሕ ከጋጥም
nselam.com	ዘይኮነ ንጽባሕ ከጋጥም ዘለዎ ጻንታ እዩ እቲ ዚረኣየናን ዘፍረሓናን	ዘሎ	። ሓደ ስደተይና ኣብ ዝኸደ ኸይዱ።ናይ ሃገሩ ናፍቆትን ዝኽርን
nselam.com	ሕጂ ንኡስን ማእከላይን ዝበልናዮ ወለዶ ጥራሕ ኣይኮነን ዚጠፍእ	ዘሎ	፡ እቲ ኪትከኦ ዝነበሮ፡ ማለት እቶም ብሕፃናቶም ዝኸዱን፡ ኣብ
nselam.com	ትርጕም ምርካብ ይከኣል...”17። ነዚ ንምግባር “ዚኋፍኣና	ዘሎ	ምድራዊ ወይ ሓላፊ ነገር ንምዕታር ወይ ከብርን ዝናን ተመኒና
nselam.com	? ዕላማኡኸ? ናብ'ቲ ሓቀይና ታሕጓስ ዚወስድ መንገጺ ኣበይ እዩ	ዘሎ	? ብዛዕባ ሞት/ፍርዲ፡ከንየው ሞት ብዛዕባ ዘሎ ዓስቢ ተግባራትካ
nselam.com	መንገጺ ኣበይ እዩ ዘሎ? ብዛዕባ ሞት/ፍርዲ፡ከንየው ሞት ብዛዕባ	ዘሎ	ዓስቢ ተግባራትካ ብዚርኢኸ እቲ ሓቁ ኣየናይ እዩ? ንህላዌና ከቢብዎ
nselam.com	ምኺኑ በሓቂ “ዜሕንኸ ተቐባልነት ዘይብሉ፣ ምስ'ቲ ወድሰብ በጺሕዎ	ዘሎ	ናይ ምዕባሌን ሥልጣኔን ዷረጃ ዘይኸይድ” ኪብሉ ርእሰ ሊቓነጻጻሳት
nselam.com	ከኣ፡ ሀልው መነባብሮ ዘይጽወር ኮይንዎ፣ በቲ ኣብ ስግር-ማዕዶ	ዘሎ	ንብረትን፣ ዘይረኣዮ ሕይወትን ተሳሒቡ፡ እቲ ዷላይ ለውጢ መንእሰይ
nselam.com	ዘይረኣዮ ሕይወትን ተሳሒቡ፡ እቲ ዷላይ ለውጢ መንእሰይ ዚገብሮ	ዘሎ	ጕዕዞ፣ ኪጅምር ከሎ ብሕጋዊ መንገድን ኣገባብን ዝነበረ፡ ሕጂ

Page 1 of 361 Go Next | Last

Word sketches

A word sketch is given below:

ዘሎ **Tigrinya WaC [2016] freq = 7,212** (2,848.96 per million)

modifiers of "ዘሎ"	336	0.40
ሕጂ	123	9.95
ሎሚ	82	9.49
ቅድሚ	37	8.35
ገና	10	8.19
ድሕሪ	25	8.02
ደው	13	7.96
ልዕሊ	22	7.75
ምሳይ	4	7.73
ብሓባር	4	7.18
ሰለስተ	4	6.88

objects of "ዘሎ"	5,342	2.90
ስርዓት	101	8.82
ህዝቢ	78	8.21
ኩነታት	52	8.00
ፍልልይ	42	7.94
ሰማይ	45	7.87
እዋን	57	7.81
ፖለቲካዊ	41	7.72
ሰማያት	37	7.70
፡	61	7.64
ህዝብና	35	7.62
ኣቦይ	33	7.57
ቃልሲ	36	7.53
ሃገርና	33	7.51
ዓለም	44	7.45
ግዜ	49	7.42
ጉዳይ	33	7.37
ሰብ	62	7.35
ጸገማት	27	7.26
ሓቂ	31	7.17
ጉጅለ	25	7.09
ጽሑፍ	25	7.08
ዝምድና	23	7.06
ሕጊ	27	6.98
ጸገም	23	6.98
ሽግር	22	6.98

subjects of "ዘሎ"	553	1.30
መንነ	52	10.95
ኤርትራን	12	8.98
ዝበገነ	7	8.62
ሓጢኣት	7	8.51
ጽፍሪ	5	8.19
መንጮ	5	8.15
ክርስቶስ	10	8.08
ዓይኒ	5	8.03
ደንበ	5	7.97
ደምበ	5	7.92
ምእንታኹም	4	7.85
ተራ	4	7.66
ሰማይን	4	7.64
ስድራ	4	7.43
ዓዲ	4	7.42
ጉጅለ	4	7.37
ስርዓት	6	7.34

pronominal subjects of "ዘሎ"	25	0.80
ንሱ	14	7.23
ኣነ	5	6.22

7. Conclusion

There is no doubt that the four Ethiopian corpora with their web based search and results interfaces are a major contribution to the possibilities for linguistic research into these languages in the future. The same is true for the Norwegian resources. The table below summarizes the results.

	Norwegian Bokmål	Norwegian Nynorsk	Amharic	Oromo	Somali	Tigrinya
No. of Words in Corpus	1,178,357,993	54,511,854	17,320,000	4,249,953	71,871,585	2,087,613
No. of Documents	3,443,807	214,379	33,542	8,851	385,338	1,907
Grammatically Tagged	tagged	no	yes	no	no	yes
Concordance	yes	yes	yes	yes	yes	yes
Word List	yes	yes	yes	yes	yes	yes
Word Sketches	yes	no	yes	no	yes	yes
Thesaurus and Word Cloud	yes	no	yes	no	no	yes

References

Guevara, Emiliano Raul (2010). NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, Association for Computational Linguistics page 1 - 7.