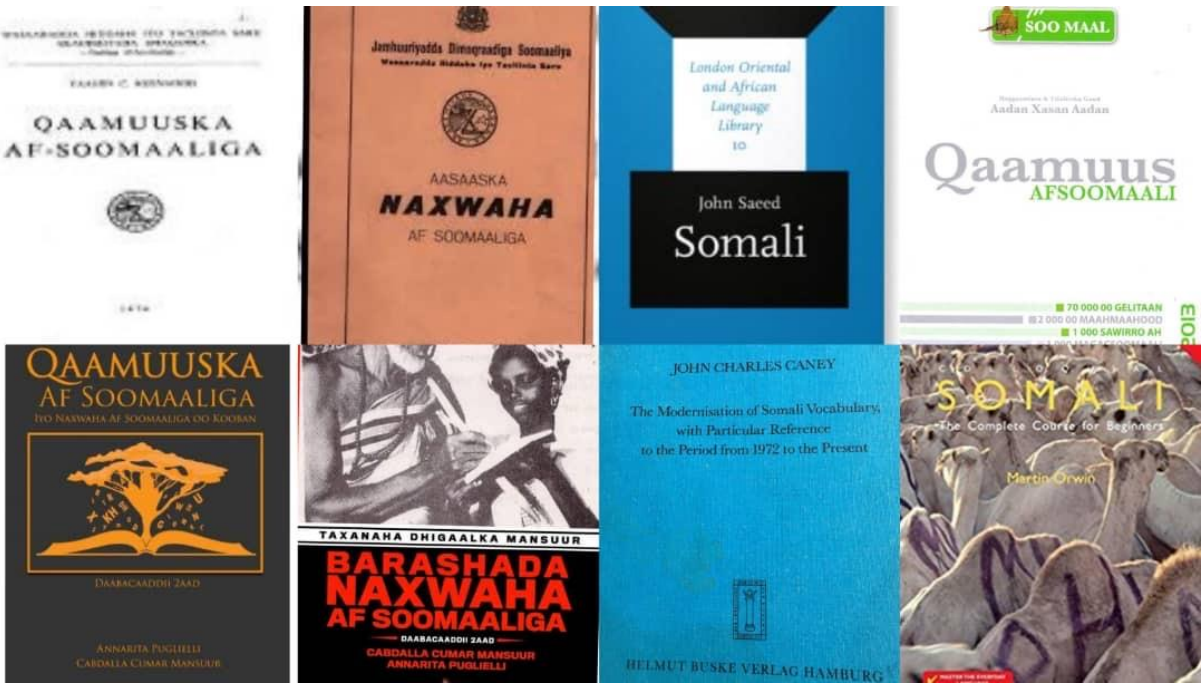
USING THE WEB FOR RESEARCH INTO THE SOMALI LANGUAGE USAGE

By Liban A. Ahmad



Abstract

Collapse of the state in 1991 affected the use of the Somali language for print media due to the destruction of the state printing agency in Mogadishu at which newspapers and books were printed. Digital media has filled the vacuum and proved a fertile ground for research into the Somali language usage. Based on Google searches of Somali words, phrases and sentences with an emphasis on Somali spelling rules and commonly confused Somali words and phrases, this paper takes advantage of the unstructured but continually updated corpus of the World Wide Web to make tentative conclusions from examples discussed in it. The paper raises questions for computational linguists about the reliability of Google Somali spelling suggestions particularly when the search engine suggests an incorrect Somali spelling with fewer hits than a correctly spelt word.

Introduction

"*Language scientists and technologists are increasingly turning to the web as a source of language data, because other resources are not large enough, because they do not contain the types of language the researcher is interested in, or simply because it is free and instantly available. The default means of access to the web is through a search engine such as Google*." – Adam Kilgarriff

The emergence of the Internet was as epoch-making for the Somali language speakers as the decision taken forty eight years ago by the former military regime to use of the Roman script for written Somali. The first Somali language news website was launched in 1999. The number of Somali words in websites far outnumber words in Somali language newspapers and books published between January 1973, when the first daily Somali language newspaper, *Xiddigta Oktoobar*, was published, and January 1991, when the state collapsed.

The collapse of the state did "leave Somalia "out of the loop" of new global technologies, markets, politics, and cultures. Contrary to this expectation, Somalis have not been so excluded" (Issa-Salwe, 2006), p. 54). Since 1999 more than 1300 Somali language news websites have been launched.

For many Somalis the Internet is where they access the news they read, watch or listen to. The amount of Somali language texts online provides language researchers with an opportunity to conduct research into the Somali language. Researchers noted the both the importance and pitfalls of the language data one can access from the Internet for research purposes "Because the Web is a huge, grammatically unanalysed corpus, obtaining information from it raises the same problems that analysing any lexical corpus will raise: the appropriate search tools for the linguistic constructions being studied must be selected…" (Meyer et al., 2003, p. 243). Other researchers have argued that "Google released a trillion-word corpus with frequency counts for all sequences up to five words long… outweighs these drawbacks." (Halevy, Norvig, and Pereira, 2009, p.8).

Methodology

The methodology used for the research concentrates on two aspects of Somali language spelling. The first aspect focuses on the seven doubled consonants in the Somali alphabet (*xarfaha laballaabma*) where the doubled letter is either a part of the word e.g. *ba<u>ll</u>an* (a promise) or a word in which the doubled letter indicates a grammatical function e.g. *saacad<u>d</u>a* (the watch). The second aspect will briefly illuminate the difficulties arising from some commonly confused Somali words and phrases

An equally more important subsidiary question from this paper is: To what extent are Somali spelling suggestions of Google reliable? This is a question that computational linguists can research given reliability issues that this study will bring to the fore.

The specification of the language data reduces the scope of the research. The focus on common spelling mistakes based on specific letters helped the formulation of the language data to be searched online. In each table, lexical items under study are highlighted in bold followed in the row below by the phrase containing the incorrectly spelt lexical items in italics. Each table has separate columns for the number of Google hits for lexical items, translation of lexical items and the date of search. Use of navigational search (Broder, 2002, p. 5) has made the task of searching lexical items in quotation marks easier.

Limitations of the research

The search method used for this paper has some limitations that any researcher should bear in mind. A fraction of Google search results can be accessed. Although Somali is one of the languages in Google Translate, the search engine does not have an advanced search facility in Somali. The number of hits from Google search results keeps changing due to daily updating of Somali language websites. This limitation could be offset by the potential of the web-based research to promote adherence to Somali spelling conventions and raise awareness about the grammatical mistakes caused by commonly confused Somali words and phrases.

Part I: Somali Spelling Conventions about Doubled Somali Letters

How writers in Somali follow spelling conventions has long been a subject of discussions. Before 1991 the now-defunct former national daily, *Xiddigta Oktoobar,* promoted correct spelling of words after the government "initiated a nation-wide literacy campaign and provided language talks, with particular emphasis on punctuation and spelling, which were broadcast on Somali radio." (Caney, 1984, p. 32). "The motivation for language reform was patriotic, and arose from the love and respect the Somali people have for their language". (Andrzejewski, 1979, p. 69).

The Former cartoonist and radio journalist Abdulqadir Mohamed Mursal said that the transformation the Somali language was going through had been severely affected by the state collapse in 1991. "Young Somali journalists and writers have not had the opportunities we had had during the 1970s to benefit from standardisation of the Somali language. The radio stations must set a good example about the use of the Somali language. Although the establishment of the Somali Language Academy is commendable we would like the work of the Academy to go beyond events and ceremonies. We should take the threat to our language seriously" Mursal said. He commended editors and journalists of Hargeisa-based newspapers, and contrasted their role in the promotion of the written Somali language with the radio stations in other parts of Somalia.

After the state collapse in 1991 and before the Internet afforded Somalis an opportunity to read news in their mother tongue wherever they are, the use of written Somali language for print media had declined sharply.

The Internet is where language researchers can find a wealth of Somali language data for different research purposes. The extent to which search engines like Google can play a similar role is now becoming clearer. What Google can help with Somali language researchers is to help linguists to research how rife incorrect Somali spellings in Somali websites are because more online content (articles, news, commentaries, discussions, review essays, comments) generated daily far exceeds the number of words in Somali language books and newspapers published annually.

In this part of the paper we discuss common spelling errors made in written Somali with particular attention to the seven doubled letters (*m,n,l,g,r, d, b*).

Nouns, adjectives and verbs that contain a doubled letter shed more light on how Somali spelling rules are adhered to forty seven after the first daily Somali language newspaper appeared in Mogadishu and Somali was made the medium of instruction at primary and secondary levels. The seven doubled Somali letters are in the sentence *Ma nala garaad baa?* (Are they as wise as we are?).
In the following section we shall discuss how rules of doubling a Somali letter is adhered to by using seven examples based on each of the seven consonants.

Letter M
Many countable Somali words ending in the letter *n*, change into *m* or double *m* in their plural form. The following table contains a sample of nouns ending in *n* and their plural forms.

Table I

Singular noun	**Plural noun**
Ciidan (troops/force)	ciidammo
ballan (appointment)	ballammo
Laan (branch)	laamo
Calan (flag)	calammo
Ashuun (clay water churn)	ashuummo

For this study two the nouns *ciidan* and *ummad* are used to find out if the writers use the correct plural form of each word. First the correct plural form of *ciidan* with the Somali definite article – *ciidammada* – was searched online. The search results showed 232,000 hits (with 83 accessible entries). The suggested spelling of the word, *ciidamada*, is the incorrect spelling of the word which after search, produced 3,060,000 hits (with 70 accessible entries).
The second word is *ummadaha* (the peoples) from *ummad* (peoples) The correct spelling of the word – *ummadaha* – has 41,500 hits (with 110 accessible entries). The incorrectly spelt *umadaha* has 33,100 hits (with 100 accessible entries). Google suggests *muadah*a*,* the name of a city in India, as the correct spelling of when one types *umadaha* in quotation marks.

The Letter N

The two sample words that contain double *n* are *xildhibaanno* (Members of Parliament) and *warbixinno* (reports).

Xildhibaanno, the correct spelling of the plural form of *xildhibaan,* has 171,000 (65 accessible entries), the suggested spelling from Google is the singular form of the word (*xildhibaan*). When the plural form of the word is spelt with one *n* – *xildhibaano* – the search result shows 553,000 hits (with 87 accessible entries).
Warbixinno, the plural form of *warbixin* (a report) has 297,000 hits (with 107 accessible entries). The incorrectly spelt *warbixino* has 198,000 hits (with 100 accessible entries). If the plural form of the word is spelt correctly or incorrectly the suggested spelling from Goole is *warbixin* (a report).

The Letter L

The third letter in the memorable sentence about doubled Somali consonants is *L*. *Ballaarinta* (the expansion) from *ballaarin* (expansion) yields 50,800 hits (with 130 accessible entries) whereas the incorrectly spelt *balaarinta* has 16,300 hits (with 120 accessible entries).
The collective noun *dhallaan* (infants) with a definite article becomes *dhallaank*a. It has 101,000 hits (with 110 accessible entries). When spelt incorrectly with one *l,* as in *dhalaanka,* it has 83,700 hits (with 100 accessible entries).

The Letter G

The two words *gaashaandhigga* (the defence) from *gaashaandhig* (defence) and and *bandhigga* (the exhibition) from *bandhig* (exhibition) contain double *g*. The correctly spelt *gaashaandhigga* has 418,000 hits (with 70 accessible entries). *Gaashaandhiga* (without an extra *g*) is the suggested spelling from Google with 163,000 hits (with 70 accessible entries.) The second word *bandhigga* (the exhibition) has 324,000 hits (with 80 entries). The incorrectly spelt *bandhiga* has 588,000 hits (with 90 accessible entries).

The Letter R

Carruurta (the children) from the undefined word *carruur* (children) has 754,000 hits (with 175 accessible entries) compared to the *caruurta*, which Google suggests as the correct spelling, with 761,000 hits (with 120 accessible entries). The second example *qorraxda* (the sun) from the undefined *qorrax* (a sun), has 235,000 hits (with 110 accessible entries). Google suggests the incorrectly spelt *qoraxda,* with 111,000 hits (100 with accessible entries) despite the correctly spelt *qorraxda* having more hits than *qoraxda*.

The Letter D

The definite article for singular feminine Somali noun ending in the letter d necessitates the addition the letter *d* to the noun. The two widely used words to be used as

examples in the paper are the defined forms of *jaamacad* (university) and *wasaarad (* ministry). *Wasaaradda* (the ministry) has 3,270,000 hits (with 90 accessible entries) When the extra *d* is dropped *wasaarada* has 1,440,000 hits (with 82 accessible entries). *Jaamacadda* (the university) has 981,000 hits (with 80 accessible hits) whereas the incorrectly spelt *jaamacada* has 634,000 hits (with 94 accessible entries).

The Letter B

The last two examples for the letter ***b*** involve two words that separately contain a possessive suffix and a definite article for a masculine noun. The dictionary form of both words contain double *b* in the second syllable.

The word *aabbahay* (my father) has 34,500 hits (with 97 accessible entries). The suggested spelling of the word is *aabahay* with 72,600 hits (with 100 accessible entries). The second word, *cabbirka* (the measurement/ size), has 152,000 hits (with 105 accessible entries). Google suggests the incorrectly spelt *cabirka*, with 136,000 hits (with 145 accessible entries) as the correct spelling of the word.

Table II

Lexical items	**Search hits from google.com**	**Word class**	**Translation of lexical items**	**Date of search**
ciidammada	232,000	Noun, plural	The troops	31/08/2020
Ciidamada	3,060,000	Noun, plural	The troops	31/08/2020
Ummadaha	41,500	Noun, plural	The people	31/08/2020
Umadaha	33,100	Noun, plural	The people	31/08/2020
Xildhibaanno	171,000	Noun, plural	Members of Parliament	31/08/2020
Xildhibaano	553,000	Noun, plural	Members of Parliament	31/08/2020
Warbixinno	297,000	Noun, plural	Reports	31/08/2020
Warbixino	198,000	N, plural	Reports	31/08/2020
Ballaarinta	50,800	Noun, singular	The expansion	31/08/2020
Balaarinta	15,100	Noun, singular	The expansion	31/08/2020
Dhallaanka	**101,000**	Collective noun	Infants	31/08/2020
Dhalaanka	**83,700**	Collective noun	Infants	31//08/2020
Gaashaandhigga	418,000	Noun, singular	The defence	31/08/2020
Gaashaandhiga	163,000	Noun, singular	The defence	31/08/2020
Bandhigga	324,000	Noun, singular	The exhibition	31/08/2020

Bandhiga	588,000	Noun, singular	The exhibition	31/08/2020
Carruurta	754,000	Noun, singular	The children	31/08/2020
Caruurta	761,000	Noun, singular	The children	31/08/2020
Qorraxda	235,000	Noun, singular	The sun	31/08/2020
Qoraxda	111,000	Noun, singular	The sun	31/08/2020
Wasaaradda	3,270,000	Noun, singular	The university	31/08/2020
Wasaarada	1,440,000	Noun, singular	The university	31/08/2020
Jaamacadda	981,000	Noun, singular	The Ministry	31/08/2020
Jaamacada	634,000	Noun, singular	The Ministry	31/08/2020
Aabbahay	34,000	Noun, singular	My father	31/08/2020
Aabahay	72,600	Noun, singular	My father	31/08/20
Cabbirka	152,000	Noun, singular	The measure	31/08/2020
Cabirka	139,000	Noun, singular	The measure	31/08/2020

Part II

In this part of the paper three different examples about Somali language usage will aim to shed light on how a search engine like Google can facilitate a research into awareness about correct use of set phrases, commonly confused Somali verbs and standard and non-standard use of in-clause negative particles.

Commonly Confused Somali Phrases

Somali has set phrases whose usage requires familiarity with the standard Somali grammar. One of those phrases are *waxaa jira* (there is /there are) and *waxaa jirta* (there is, used with a feminine singular noun). We searched the two phrases *waxaa jira qorshe* and *waxaa jirta qorshe* to compare the data on usage from Google searches.

The grammatically correct phrase *waxaa jira qorshe* (there is a plan) has 2,690 (with 95 accessible entries) whereas the grammatically incorrect phrase *waxaa jirta qorshe* (there is a plan) has 87 hits (with 20 accessible entries). The preliminary conclusion one can infer from this search is that writers have an awareness about the standard Somali grammar rules about phrases *waxaa jira* and *waxaa jirta*.

Commonly Confused Somali Verbs

The two verbs *dhaq* (wash) and *dhaqo* (to look after e.g. livestock) are commonly confused verbs. We compared the search results of the two sentences *Sida geela loo dhaqdo* and *Sida geela loo dhaqo*. The first sentence, *Sida geela loo dhaqdo* (how to look after camels) contains a correctly conjugated verb, *dhaq*o, and has 220 hits (with 16 accessible entries) whereas the phrase *Sida geela loo dhaqo* (how to wash camels), meant to convey the meaning 'how to look after camels' in Somali, has 67 results (with 21 accessible entries).

Non-standard Usage of Somali Negative In-clause Particles

The last two examples are about the negative in-clause particle for the third person masculine pronoun (*isaga)*. The standard form the negative in-clause particle is *inuusan.* The non-standard form is *inuunan*. A search of the phrase *Wuxuu sheegay inuusan* (he said that that he did not…) has yielded 8,740 hits (with 84 accessible entries). The phrase containing the colloquial form of the word, *Wuxuu sheegay inuunan,* has 347 hits (with 60 accessible entries).

Table III

Lexical items	**Search hits from google.com**	**Word class**	**Translation of lexical items**	**Date of search**
Waxaa jira qorshe	2,690	Noun	There is a plan	31/08/2020
Waxaa jirta qorshe	87	Noun	There is a plan	31/08/2020
Sida geela loo dhaqdo	22	Noun, plural	How to look after camels	31/08/2020
Sida geela loo dhaqo	67	Noun, plural	How to look after camels	31/08/2020
Wuxuu sheegay inuusan	8740	Negative in-clause particle	He said that he did not	31/08/2020
Wuxuu sheegay inuunan	347	Negative in-clause particle	He said that he did not	31/08/2020

Conclusion

This paper has explored how the World Wide Web can be used to conduct a research into the Somali language usage. The aim is not to make definitive conclusions from examples under discussion but to illustrate the tentative insights into the Somali language usage one can infer from a search engine such as Google to promote

adherence to Somali grammar and spelling rules, and draw lessons that can be incorporated into the Somali language courses for journalists. Questions raised by search results about reliability of Google Somali spelling suggestions are beyond the scope of this paper.

Bibliography

Aadan Xasan Aadan, 2013 Soo Maal: Qaamuus Afsoomaali (A monolingual Somali Dictionary). Djibouti. Naadiga Qalinleyda iyo Hal'abuurka Soomaaliyeed (SSPEN). Andrzejewski, B.W., 1979. Language reform in Somalia and the modernization of the Somali vocabulary. *Northeast African Studies*, pp.59-71.

Broder, A., 2002, September. A taxonomy of web search. In *ACM Sigir forum* (Vol. 36, No. 2, p. 5). New York, NY, USA: ACM.

Caney, J.C., 1984. *The Modernisation of Somali Vocabulary-Particular Reference to the Period from 1972 to the Present*. Helmut Buske Verlag.

Gagliardone, I. and Stremlau, N., 2011. *Digital media, conflict and diasporas in the Horn of Africa*. New York: Open Society Foundations.

Halevy, A., Norvig, P. and Pereira, F., 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), p.8.

Issa-Salwe, A.M., 2006. The Internet and the Somali diaspora: The Web as a new means of expression. *Bildhaan: An International Journal of Somali Studies*, *6*(1).

Jama, J.M., 2017. Somali corpus: state of the art, and tools for linguistic analysis.

Kilgarriff, A., 2003, March. Linguistic search engine. In *Shallow Processing of Large Corpora: Workshop Held in Association with Corpus Linguistics*.

Meyer, C.F., Grabowski, R., Han, H.Y., Mantzouranis, K. and Moses, S., 2003. The world wide web as linguistic corpus. In *Corpus Analysis* (p. 243. Brill Rodopi.

Mursal, Abdulkadir M. (2019) Interviewed by Asha Ibrahim Aden *for VOA Somali Service, 21 January.* Available at: http://iplayer.co.uk/Newsnight/march7 (Accessed: 12 March 2020).

Nilsson, M. and Corpora, S., Corpus driven lexicography opens new horizons for Somali.

Puglielli, A. and Mansuur, C.C., 2016. *Qaamuuska Af-Soomaaliga (G. Diz. Somalo Monolingue)*. Roma TrE-Press.