# Kaydka Af Soomaaliga
# *Somali Corpus*

Hagaha Adeegsiga
*User manual*

Jama Musse Jama
2016

Ku dhawaad 3 milyan oo erey oo suntan ● 1100 dhigane oo baadhan ● 52 kun oo erey saleed ● Ku dhawaad 6 milion oo erey farcame ● 10 kun oo turjumaadood

*Almost 3 million words tagged ● 1100 documents parsed and indexed ● Over 52000 head words ● Almost 6 million inflected forms generated ● Over 10 thousand translation in four languages*

REDSEA
CULTURAL
FOUNDATION
www.redsea-online.org

**RCF Somali Corpus – User Manual**
Ver. 1.0 – June 2016
www.somalicorpus.com

SIGILLVM INSTITVTI ORIENTALIS NEAPOLITANI

# TABLE OF CONTENTS

# 1. Introduction

This User Manual aims to introduce different functionalities of the Redsea Cultural Foundation Somali Corpus management platform and to allow linguists, researchers and other potential users of the corpus to access it easily, search and retrieve data from both the corpus and its auxiliary tables (dictionary, synonyms, etc.).

The RCF Somali Corpus incorporates a “Word Behaviour” (WB) page, a seminal work of corpus-based analysis of a word in  Somali discourse. It is based on a powerful querying tool to access the grammatically tagged Somali Corpus to summarize the syntactic context of the searched word;for instance, to show the relationship that the searched word has with other words, which words are mostly associated with the searched word; the frequency of the searched word within specific sub-corpora; the etymology; synonyms and antonyms; spelling variants of the searched word; and finally the definitions from a list of reference dictionaries as well as  translations to English, Italian, French and Swedish languages.

This document covers the following major sections:

1. Access to the system, basic search and data retrieval;
2. Major components of the WB view page: definitions, synonyms, etymology, concordance, translations, grammatical behaviour of the word in pre-defined structures and collocations;
3. *SomlSearch* - an intelligent search engine, incorporated in the WB. It can search for words in all their inflected forms. It can also search through translations in English, French, Italian and Swedish, or through synonyms in Somali: for instance, one can search the word *gobannimo*, which means ‘freedom’, and the engine will also search for *xornimo*, which means ‘independence’, a near synonym of *gobannimo.* Finally *SomlSearch* develops an Advanced Query Language for the RCF Somali Corpus (see chapter 5 for more details).
4. Exporting data and printing.

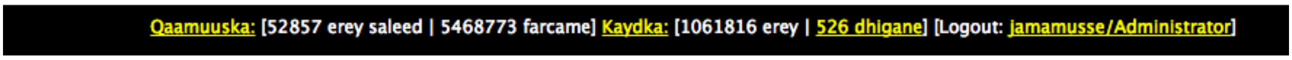
This document is written in English as an integral part to the work Jama Musse Jama, “A Syntactically Annotated Corpus of Somali Literature”, 2016 that was initially written in English.

## 2. Access to the system, basic search and data retrieval

### *Access*

RCF Somali Corpus is a web-based online platform and it needs a valid user id and password in order to access it and to research data. The system administrator, who defines also a specific role for each approved user, provides the user id and the password upon request. If you do not yet have a valid user id and password, please write to info@somalicorpus.com and cc to somalicorpus@gmail.com.

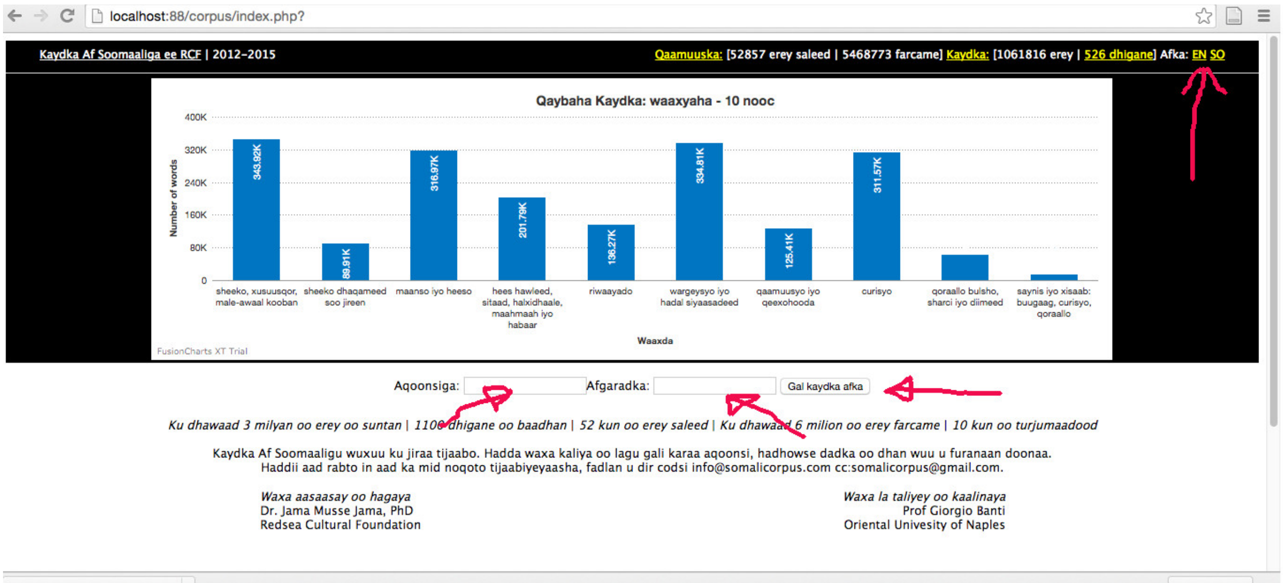
Whether you are connected or not, the system shows you on the top bar of the page some statistics: number of headwords in the dictionary, number of inflected forms automatically generated by the system, number of words tagged in the corpus and number of documents indexed. If you are not connected, at the end of the top bar, you have language options (Somali, English). When you are connected, at the end of the top bar, you see your name, role and the possibility to logout, as shown below.



Qaamuuska: [52857 erey saleed | 5468773 farcame] Kaydka: [1061816 erey | 526 dhigane] [Logout: jamamusse/Administrator]

### *Login*

Once you have obtained a user id and password, you can connect to the system as indicated in the first page (see the image below). Put your id in the "*Aqoonsiga"* (login) field and your password in the "*Afgaradka"* (password) field, then click on the "*Gal kaydka"* (enter the corpus) button for authentication. This opening page is the only part of the platform available bilingually (Somali and English), the rest of the user interface and its functionalities, when connected, is presented only in Somali.



localhost:88/corpus/index.php?
Kaydka Af Soomaaliga ee RCF | 2012–2015
Qaamuuska: [52857 erey saleed | 5468773 farcame] Kaydka: [1061816 erey | 526 dhigane] Afka: EN SO
Qaybaha Kaydka: waaxyaha - 10 nooc
400K
320K
240K
160K
80K
0
Number of words
343.92K
89.91K
316.97K
201.79K
136.27K
334.81K
125.41K
311.57K
sheeko, xusuusqor, male-awaal kooban
sheeko dhaqameed soo jireen
maanso iyo heeso
hees hawleed, sitaad, halxidhaale, maahmaah iyo habaar
riwaayado
wargeysyo iyo hadal siyaasadeed
qaamuusyo iyo qeexohooda
curisyo
qoraallo bulsho, sharci iyo diimeed
saynis iyo xisaab: buugaag, curisyo, qoraallo
Waaxda
FusionCharts XT Trial
Aqoonsiga:
Afgaradka:
Gal kaydka afka

*Ku dhawaad 3 milyan oo erey oo suntan | 1100 dhigane oo baadhan | 52 kun oo erey saleed | Ku dhawaad 6 milion oo erey farcame | 10 kun oo turjumaadood*

Kaydka Af Soomaaligu wuxuu ku jiraa tijaabo. Hadda waxa kaliya oo lagu gali karaa aqoonsi, hadhowse dadka oo dhan wuu u furanaan doonaa.
Haddii aad rabto in aad ka mid noqoto tijaabiyeyaasha, fadlan u dir codsi info@somalicorpus.com cc:somalicorpus@gmail.com.

*Waxa aasaasay oo hagaya*
Dr. Jama Musse Jama, PhD
Redsea Cultural Foundation

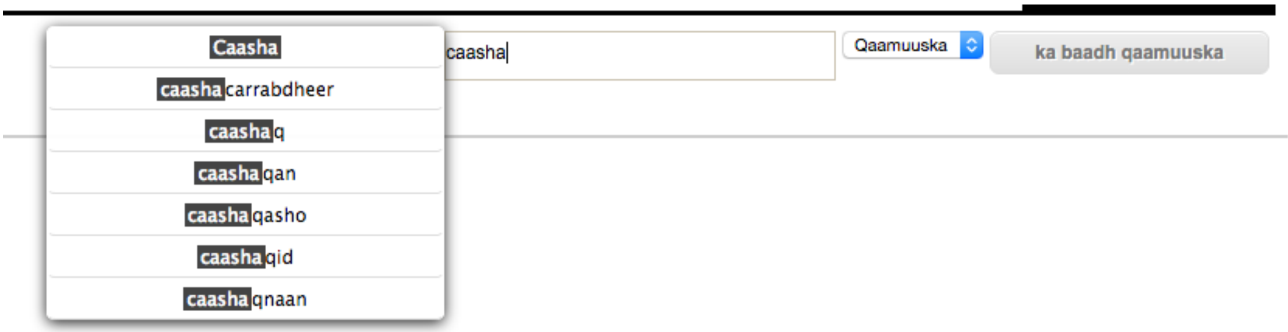
*Waxa la taliyey oo kaalinaya*
Prof Giorgio Banti
Oriental Univesity of Naples

## Browse

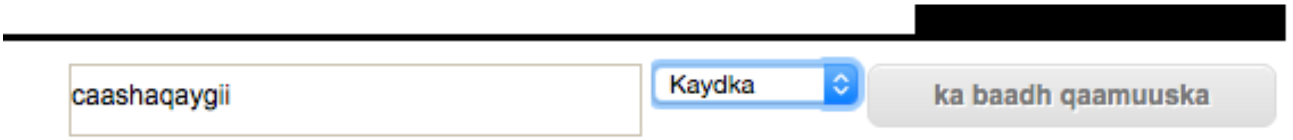
Statistics given by the system include the distribution of the dictionary in Parts of Speech; distribution of the corpus in eras; and distribution of the corpus in sub corpora. All these distributions are seen in the form of graphs. Links to *Qaamuuska* (Dictionary), *Kaydka* (Corpus) and *Dhigane* (Documents) are in the top bar. When you click on "*Dhigane* (Documents)", in particular, you can browse the corpus by author: a list of authors is shown ordered by number of words in the contribution to the corpus by that author. When an author's name is clicked, then you see the list of titles written by that author, and each link will take you to the analysis of the single document within the corpus: for instance frequency of used words, list of successfully tagged words, list of unrecognized words or spelling errors found in the document, etc.

## Search

When you log into the system, you will see the distribution of the dictionary in Parts of Speech. You can perform three types of simple search and the Advanced Query Language search type: First, to search a root word, choose "*Qaamuuska"* (dictionary) at the search drop-down and insert characters in the field "search". The system will perform a word prediction process while you are typing and, as long as a match is found in headwords, the word prediction system will guide you to the right spelling. You can also put here an inflected form of the word you are looking for, and again, as long as what you are typing matches to a headword in the dictionary, the system will retrieve it and show the headword definition.



The second type of simple search is on the corpus. Choose "Kaydka" (corpus) in the search drop-down, and insert the word. The system will not perform a prediction and will search the inserted word in the corpus and not in the dictionary of headwords.

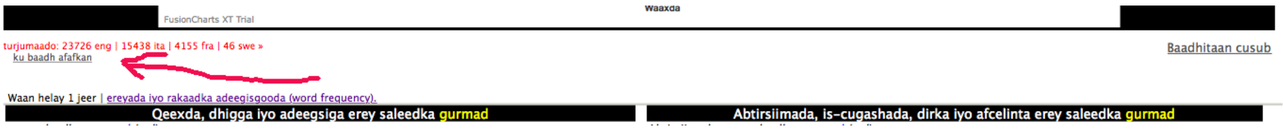


caashaqaygii

Kaydka

If you are not using the Advanced Query Language (AQL) (see chapter 3), the query system expects one single word search (it can be a headword or its inflected form). Make sure not to have space in the text even if you are looking for a compound word.  Similarly the use of wildcards and special characters are not to be used, these are only available within AQL (see chapter 3).

The third type of simple search function is through its translations in different languages (at present English, Italian, French and Swedish are available for some words). To perform this search, follow the link *"ku baadh afafkan"* at the left side of the screen. It will open a new popup window.



FusionCharts XT Trial
Waaxda

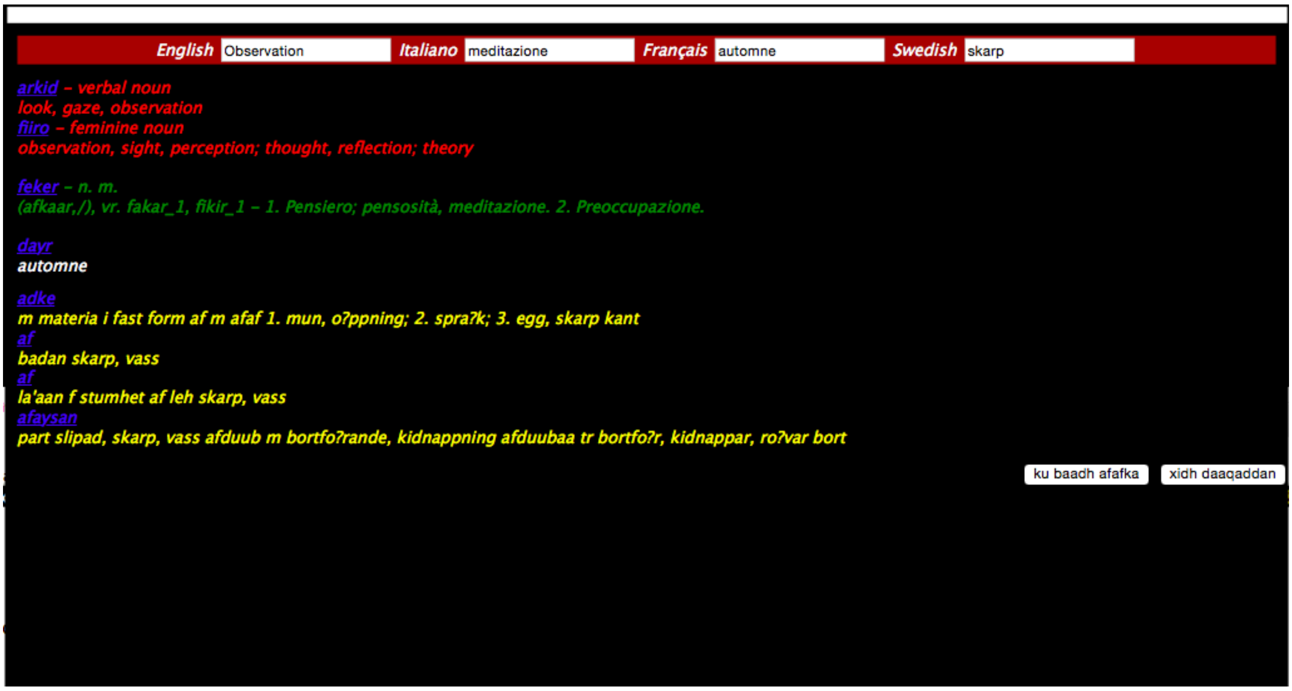
turjumaado: 23726 eng | 15438 ita | 4155 fra | 46 swe »

ku baadh afafkan
Baadhitaan cusub

Waan helay 1 jeer | ereyada iyo rakaadka adeegisgooda (word frequency).

Qeexda, dhigga iyo adeegsiga erey saleedka gurmad
Abtirsiimada, is-cugashada, dirka iyo afcelinta erey saleedka gurmad

Insert one or all your English, Italian, France and Swedish words in the corresponding fields and click *"ku baadh afafka"*.



English Observation  Italiano meditazione  Français automne  Swedish skarp

arkid – verbal noun
look, gaze, observation
fiiro – feminine noun
observation, sight, perception; thought, reflection; theory

feker – n. m.
(afkaar,/), vr. fakar_1, fikir_1 – 1. Pensiero; pensosità, meditazione. 2. Preoccupazione.

dayr
automne

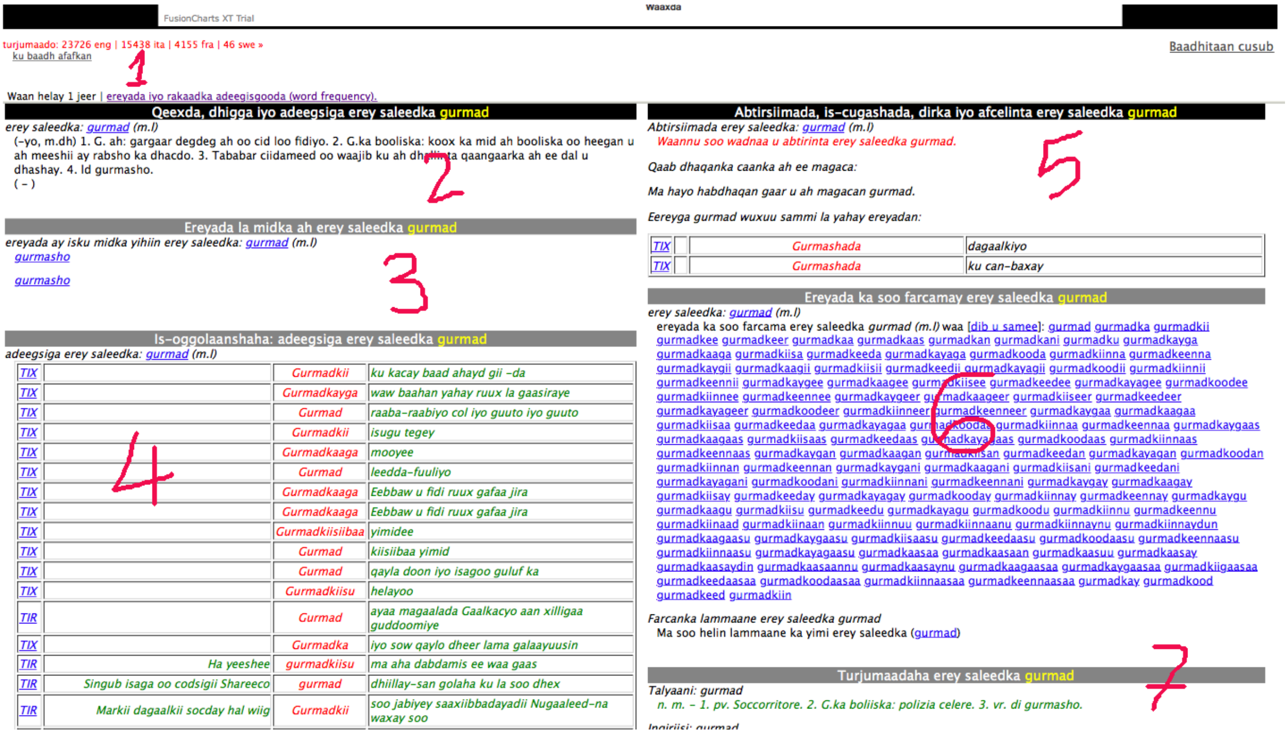
adke
m materia i fast form af m afaf 1. mun, o?ppning; 2. spra?k; 3. egg, skarp kant
af
badan skarp, vass
af
la'aan f stumhet af leh skarp, vass
afaysan
part slipad, skarp, vass afduub m bortfo?rande, kidnappning afduubaa tr bortfo?r, kidnappar, ro?var bort

You do not need the words to have the same meaning as the system will perform separate and parallel searches. The results will be shown in different colours as shown below (red for English, green for Italian, white for French and yellow for Swedish). The corresponding Somali word is shown through an active link which, when clicked, will open a new search in the dictionary.

To close and return back to the main page, click “xidh daaqaddan” (close this window) button.

The Advanced Query Language (AQL) (see Chapter 5) applies only to the search on “*Kaydka”* (Corpus) and you do not have word prediction hints here while entering the search string.

Whatever service you use (browse, search, translation, AQL), once the word is identified, the system will bring you to the Word Behaviour view structure. See Chapter 3 for more details.



turjumaado: 23726 eng | 15438 ita | 4155 fra | 46 swe »
ku baadh afafkan

Baadhitaan cusub

Waan helay 1 jeer | ereyada iyo rakaadka adeegisgooda (word frequency).

**Qeexda, dhigga iyo adeegsiga erey saleedka gurmad**

*erey saleedka: gurmad (m.l)*
(-yo, m.dh) 1. G. ah: gargaar degdeg ah oo cid loo fidiyo. 2. G.ka booliska: koox ka mid ah booliska oo heegan u ah meeshii ay rabsho ka dhacdo. 3. Tababar ciidameed oo waajib ku ah dh[illegible]ta qaangaarka ah ee dal u dhashay. 4. Id gurmasho.
( - )

**Ereyada la midka ah erey saleedka gurmad**

*ereyada ay isku midka yihiin erey saleedka: gurmad (m.l)*
gurmasho

gurmasho

**Is-oggolaanshaha: adeegsiga erey saleedka gurmad**

*adeegsiga erey saleedka: gurmad (m.l)*

| | | | |
|---|---|---|---|
| TIX | | Gurmadkii | ku kacay baad ahayd gii -da |
| TIX | | Gurmadkayga | waw baahan yahay ruux la gaasiraye |
| TIX | | Gurmad | raaba-raabiyo col iyo guuto iyo guuto |
| TIX | | Gurmadkii | isugu tegey |
| TIX | | Gurmadkaaga | mooyee |
| TIX | | Gurmad | leedda-fuuliyo |
| TIX | | Gurmadkaaga | Eebbaw u fidi ruux gafaa jira |
| TIX | | Gurmadkaaga | Eebbaw u fidi ruux gafaa jira |
| TIX | | Gurmadkiisiibaa | yimidee |
| TIX | | Gurmad | kiisiibaa yimid |
| TIX | | Gurmad | qayla doon iyo isagoo guluf ka |
| TIX | | Gurmadkiisu | helayoo |
| TIR | | Gurmad | ayaa magaalada Gaalkacyo aan xilligaa guddoomiye |
| TIX | | Gurmadka | iyo sow qaylo dheer lama galaayuusin |
| TIR | Ha yeeshee | gurmadkiisu | ma aha dabdamis ee waa gaas |
| TIR | Singub isaga oo codsigii Shareeco | gurmad | dhiillay-san golaha ku la soo dhex |
| TIR | Markii dagaalkii socday hal wiig | Gurmadkii | soo jabiyey saaxiibbadayadii Nugaaleed-na waxay soo |

**Abtirsiimada, is-cugashada, dirka iyo afcelinta erey saleedka gurmad**

*Abtirsiimada erey saleedka: gurmad (m.l)*
*Waannu soo wadnaa u abtirinta erey saleedka gurmad.*

*Qaab dhaqanka caanka ah ee magaca:*

*Ma hayo habdhaqan gaar u ah magacan gurmad.*

*Eereyga gurmad wuxuu sammi la yahay ereyadan:*

| | | | |
|---|---|---|---|
| TIX | | Gurmashada | dagaalkiyo |
| TIX | | Gurmashada | ku can-baxay |

**Ereyada ka soo farcamay erey saleedka gurmad**

*erey saleedka: gurmad (m.l)*
ereyada ka soo farcama erey saleedka *gurmad (m.l)* waa [dib u samee]: gurmad gurmadka gurmadkii gurmadkee gurmadkeer gurmadkaa gurmadkaas gurmadkan gurmadkani gurmadku gurmadkayga gurmadkaaga gurmadkiisa gurmadkeeda gurmadkayaga gurmadkooda gurmadkiinna gurmadkeenna gurmadkaygii gurmadkaagii gurmadkiisii gurmadkeedii gurmadkayagii gurmadkoodii gurmadkiinnii gurmadkeennii gurmadkaygee gurmadkaagee gurmadkiisee gurmadkeedee gurmadkayagee gurmadkoodee gurmadkiinnee gurmadkeennee gurmadkaygeer gurmadkaageer gurmadkiiseer gurmadkeedeer gurmadkayageer gurmadkoodeer gurmadkiinneer gurmadkeenneer gurmadkaygaa gurmadkaagaa gurmadkiisaa gurmadkeedaa gurmadkayagaa gurmadkoodaa gurmadkiinnaa gurmadkeennaa gurmadkaygaas gurmadkaagaas gurmadkiisaas gurmadkeedaas gurmadkayagaas gurmadkoodaas gurmadkiinnaas gurmadkeennaas gurmadkaygan gurmadkaagan gurmadkiisan gurmadkeedan gurmadkayagan gurmadkoodan gurmadkiinnan gurmadkeennan gurmadkaygani gurmadkaagani gurmadkiisani gurmadkeedani gurmadkayagani gurmadkoodani gurmadkiinnani gurmadkeennani gurmadkaygay gurmadkaagay gurmadkiisay gurmadkeeday gurmadkayagay gurmadkooday gurmadkiinnay gurmadkeennay gurmadkaygu gurmadkaagu gurmadkiisu gurmadkeedu gurmadkayagu gurmadkoodu gurmadkiinnu gurmadkeennu gurmadkiinaad gurmadkiinaan gurmadkiinnuu gurmadkiinnaanu gurmadkiinnaynu gurmadkiinnaydun gurmadkaagaasu gurmadkaygaasu gurmadkiisaasu gurmadkeedaasu gurmadkoodaasu gurmadkeennaasu gurmadkiinnaasu gurmadkayagaasu gurmadkaasaa gurmadkaasaan gurmadkaasuu gurmadkaasay gurmadkaasaydin gurmadkaasaannu gurmadkaasaynu gurmadkaagaasaa gurmadkaygaasaa gurmadkiigaasaa gurmadkeedaasaa gurmadkoodaasaa gurmadkiinnaasaa gurmadkeennaasaa gurmadkay gurmadkood gurmadkeed gurmadkiin

*Farcanka lammaane erey saleedka gurmad*
Ma soo helin lammaane ka yimi erey saleedka (gurmad)

**Turjumaadaha erey saleedka gurmad**

*Talyaani: gurmad*
*n. m. - 1. pv. Soccorritore. 2. G.ka boliiska: polizia celere. 3. vr. di gurmasho.*

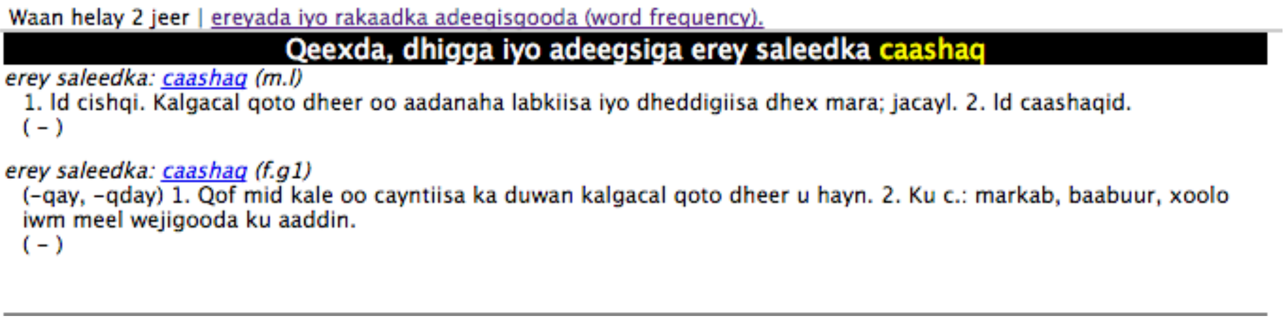
*Ingiriisi: gurmad*

## 3. Word Behaviour view structure

As the above screenshot indicates, the Word Behaviour view has 7 major components: Box (1) indicates the number of times the headword is found in the dictionary (meaning also different Parts of Speech); box (2) contains the definitions of the headword as provided in the different dictionaries; box (3) contains the links to the synonyms of the headword; box (4) contains the concordance of the headword (and all its inflected forms); box (5) contains the etymology of the word (if known) and collocations defined through pre-established configurations (for example if the headword is a Noun then we extract the collocation with the conjunction expression *"iyo"*, etc.(see chapter 4 for pre-configured clusters of rules for each part of speech; box (6) contains all the inflected forms of verbs and defined forms of nouns generated by SomMorph; and box (7) contains the translations of the word in other languages (English, Italian, French and Swedish). Below is a capture screen of each section for the word *caashaq* (love, affection).

### *Dictionary definitions*

Definitions are searched in several monolingual dictionaries and for each entry the Part of Speech is shown. There is also an active link for each PoS entry found in the major dictionary of reference (see Jama Musse Jama, 'A Syntactically Annotated Corpus of Somali Literature', 2016 for more details of the dictionaries) that shows you the inflected forms of the selected item.



Waan helay 2 jeer | ereyada iyo rakaadka adeegisgooda (word frequency).

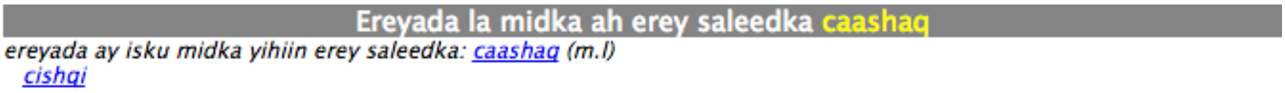
**Qeexda, dhigga iyo adeegsiga erey saleedka caashaq**

*erey saleedka:* caashaq *(m.l)*
1. Id cishqi. Kalgacal qoto dheer oo aadanaha labkiisa iyo dheddigiisa dhex mara; jacayl. 2. Id caashaqid.
( – )

*erey saleedka:* caashaq *(f.g1)*
(–qay, –qday) 1. Qof mid kale oo cayntiisa ka duwan kalgacal qoto dheer u hayn. 2. Ku c.: markab, baabuur, xoolo iwm meel wejigooda ku aaddin.
( – )

### *Synonyms*

Synonyms are collected automatically from the main dictionary and subsequently some of the entries were edited manually to link synonymous words. There is an active link that brings you to the WB view for the new word.

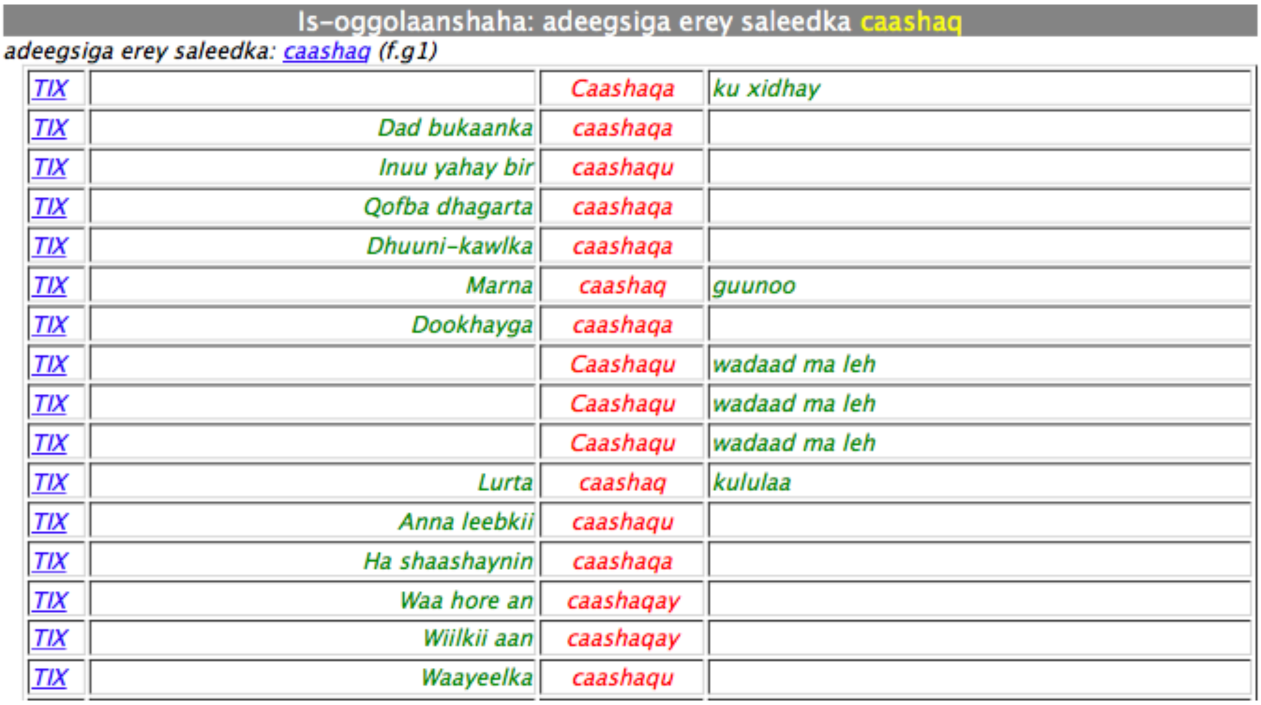


**Ereyada la midka ah erey saleedka caashaq**

*ereyada ay isku midka yihiin erey saleedka:* caashaq *(m.l)*
cishqi

## Concordance

In this box you can see the searched word and all its inflected forms used in the corpus. The first column contains an active link that indicates the sub corpus in which the instance is found. When clicked the link opens a popup notice that tells you the details of the document this sentence belongs to. The rest of each row shows the searched word (or the inflected form of the headword)



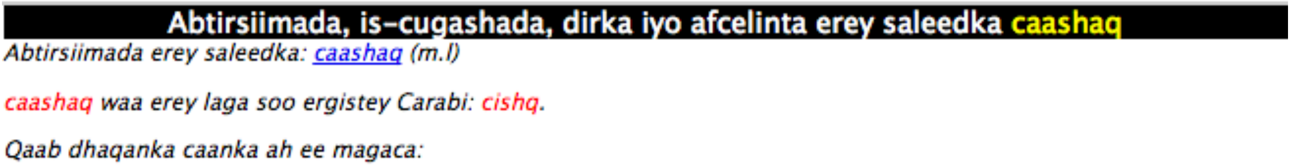
Is-oggolaanshaha: adeegsiga erey saleedka caashaq

*adeegsiga erey saleedka: caashaq (f.g1)*

| | | | |
|---|---|---|---|
| TIX | | Caashaqa | ku xidhay |
| TIX | Dad bukaanka | caashaqa | |
| TIX | Inuu yahay bir | caashaqu | |
| TIX | Qofba dhagarta | caashaqa | |
| TIX | Dhuuni-kawlka | caashaqa | |
| TIX | Marna | caashaq | guunoo |
| TIX | Dookhayga | caashaqa | |
| TIX | | Caashaqu | wadaad ma leh |
| TIX | | Caashaqu | wadaad ma leh |
| TIX | | Caashaqu | wadaad ma leh |
| TIX | Lurta | caashaq | kululaa |
| TIX | Anna leebkii | caashaqu | |
| TIX | Ha shaashaynin | caashaqa | |
| TIX | Waa hore an | caashaqay | |
| TIX | Wiilkii aan | caashaqay | |
| TIX | Waayeelka | caashaqu | |

## Etymology

Etymology is an area that has received very little attention in the academic research on Somali language. In this box of WB, you will find a small number of Somali words of which we do know the etymology, and all loanwords from other languages (Arabic, Italian, English, French, etc.). An active link indicates the language of origin and the word, and if you click on the language, you will find a list of Somali words loaned from that specific language. For Arabic there are some font issues to sort out.



Abtirsiimada, is-cugashada, dirka iyo afcelinta erey saleedka caashaq

*Abtirsiimada erey saleedka: caashaq (m.l)*

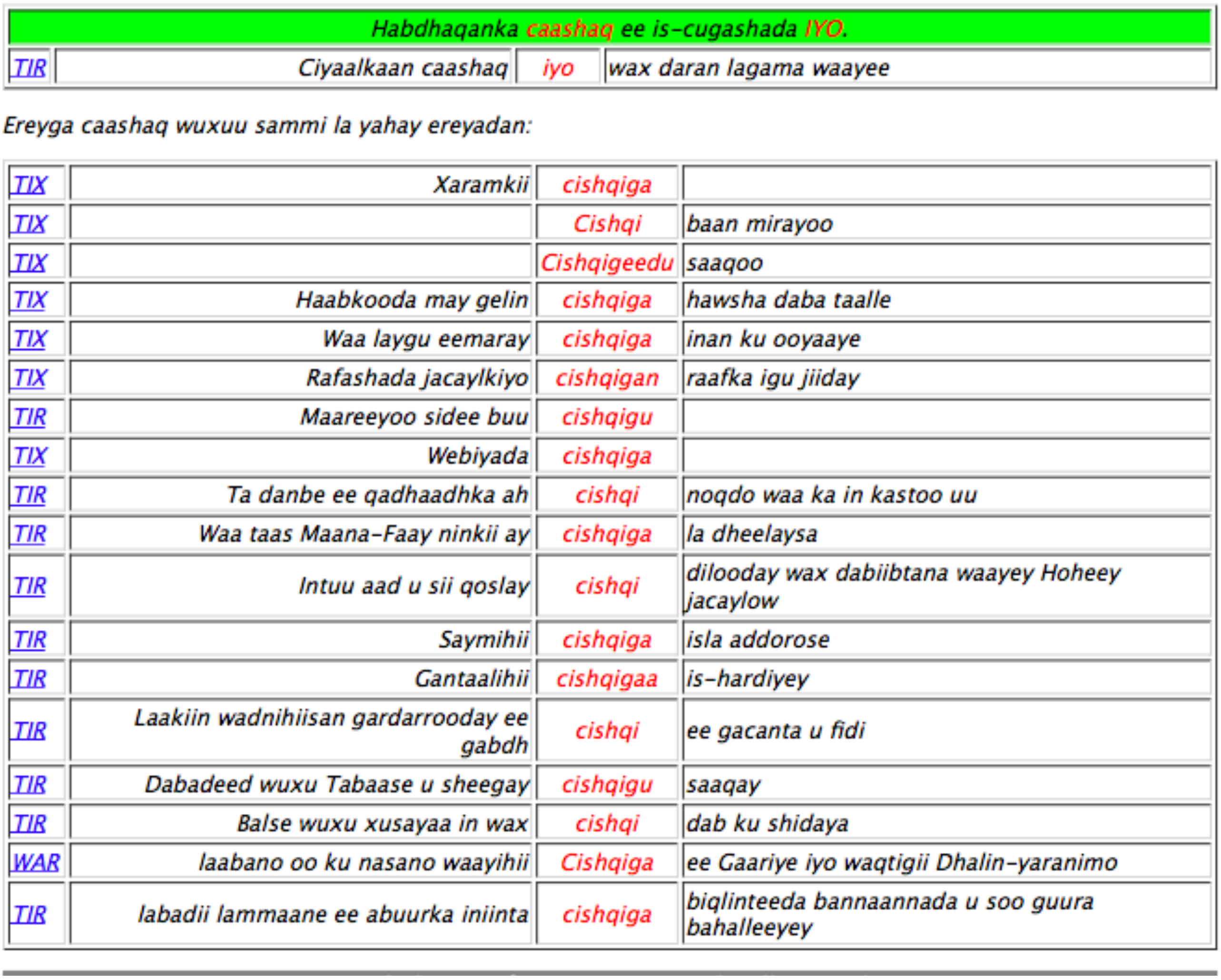
*caashaq waa erey laga soo ergistey Carabi: cishq.*

*Qaab dhaqanka caanka ah ee magaca:*

## Word behaviour in a pre-defined grammatical structures

Immediately after the etymology box, you have several pre-defined grammatical structures. For each type of Part of Speech, the system checks if there are particular collocations of linguistic interest. For instance if it is a noun, as said earlier, the system checks the collocation of the noun with the conjunction expression "*iyo*". Another collocation for nouns of particular interest is with the verb "*leh*" (to have).

In case of the verbs, when looking for verb V, the system checks and displays the collocation of V with "*is*", and the collocation of V with "*baa*", "*uu*", "*ayaa*", "*uu*", etc. And finally if the word has defined synonyms, also shown in this box is the concordance of the synonym word in the corpus.



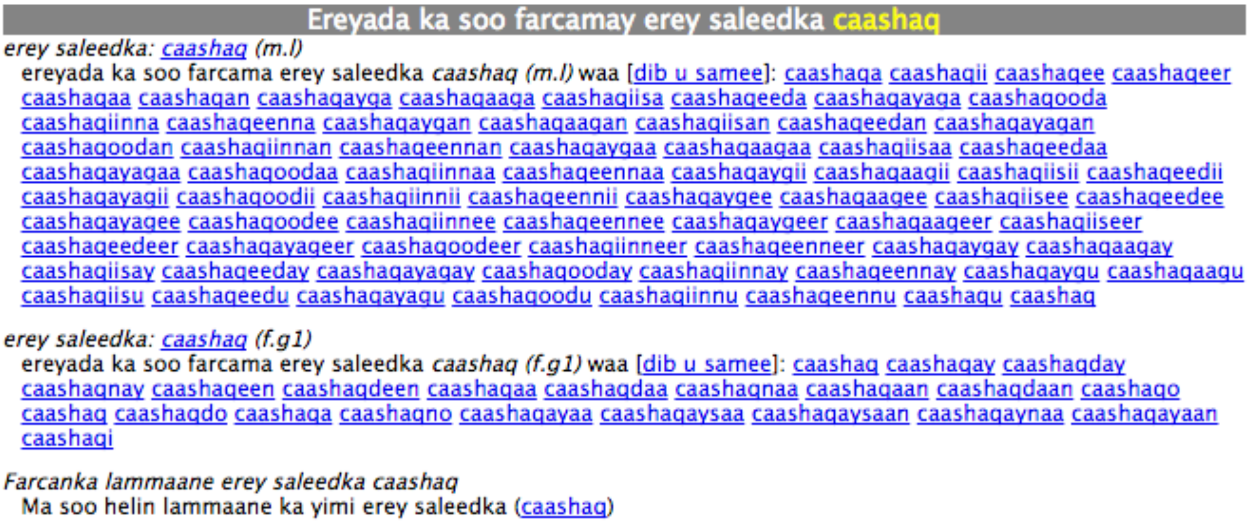
| Habdhaqanka caashaq ee is-cugashada IYO. | | | |
|---|---|---|---|
| TIR | Ciyaalkaan caashaq | iyo | wax daran lagama waayee |

*Ereyga caashaq wuxuu sammi la yahay ereyadan:*

| | | | |
|---|---|---|---|
| TIX | Xaramkii | cishqiga | |
| TIX | | Cishqi | baan mirayoo |
| TIX | | Cishqigeedu | saaqoo |
| TIX | Haabkooda may gelin | cishqiga | hawsha daba taalle |
| TIX | Waa laygu eemaray | cishqiga | inan ku ooyaaye |
| TIX | Rafashada jacaylkiyo | cishqigan | raafka igu jiiday |
| TIR | Maareeyoo sidee buu | cishqigu | |
| TIX | Webiyada | cishqiga | |
| TIR | Ta danbe ee qadhaadhka ah | cishqi | noqdo waa ka in kastoo uu |
| TIR | Waa taas Maana-Faay ninkii ay | cishqiga | la dheelaysa |
| TIR | Intuu aad u sii qoslay | cishqi | dilooday wax dabiibtana waayey Hoheey jacaylow |
| TIR | Saymihii | cishqiga | isla addorose |
| TIR | Gantaalihii | cishqigaa | is-hardiyey |
| TIR | Laakiin wadnihiisan gardarrooday ee gabdh | cishqi | ee gacanta u fidi |
| TIR | Dabadeed wuxu Tabaase u sheegay | cishqigu | saaqay |
| TIR | Balse wuxu xusayaa in wax | cishqi | dab ku shidaya |
| WAR | laabano oo ku nasano waayihii | Cishqiga | ee Gaariye iyo waqtigii Dhalin-yaranimo |
| TIR | labadii lammaane ee abuurka iniinta | cishqiga | biqlinteeda bannaannada u soo guura bahalleeyey |

## *Inflected forms of the word*

*SomMorph* is an application written by the author that develops noun and verb derivatives from head words according to the rules defined in the accessible authorative sources published from the the time the Somali language was being institutionalized as the national language. One major reference is Annarita Puglielli and Abdalla Omar Mansur, *Qaamuuska Af-soomaaliga* (2012), but also other sources including John Saeed, Giorgio Banti, Martin Orwin and others. These rules define different ways of recognizing derivatives of the headword and generate all inflected forms. In this box you see the generated forms in an active link that, when clicked, will show you the concordance of the specific inflected form in the corpus (instead of the lemma). If there are combined words of which the searched word is part of the combination, again here you see the list.



Ereyada ka soo farcamay erey saleedka caashaq

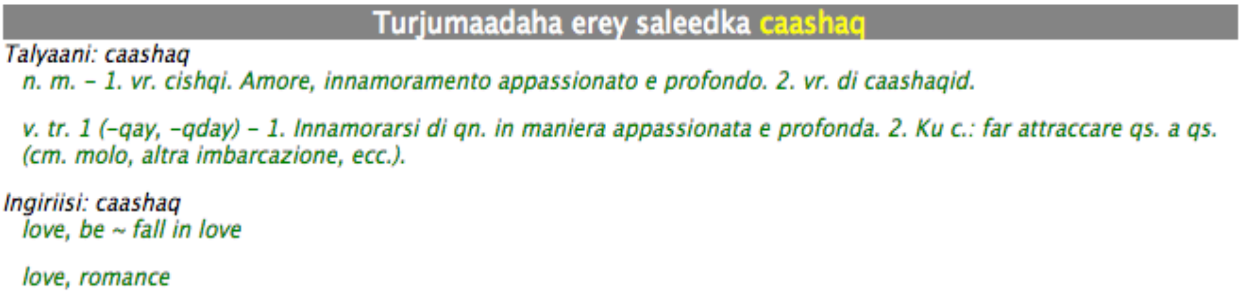
*erey saleedka:* caashaq *(m.l)*
ereyada ka soo farcama erey saleedka *caashaq (m.l)* waa [dib u samee]: caashaqa caashaqii caashaqee caashaqeer caashaqaa caashaqan caashaqayga caashaqaaga caashaqiisa caashaqeeda caashaqayaga caashaqooda caashaqiinna caashaqeenna caashaqaygan caashaqaagan caashaqiisan caashaqeedan caashaqayagan caashaqoodan caashaqiinnan caashaqeennan caashaqaygaa caashaqaagaa caashaqiisaa caashaqeedaa caashaqayagaa caashaqoodaa caashaqiinnaa caashaqeennaa caashaqaygii caashaqaagii caashaqiisii caashaqeedii caashaqayagii caashaqoodii caashaqiinnii caashaqeennii caashaqaygee caashaqaagee caashaqiisee caashaqeedee caashaqayagee caashaqoodee caashaqiinnee caashaqeennee caashaqaygeer caashaqaageer caashaqiiseer caashaqeedeer caashaqayageer caashaqoodeer caashaqiinneer caashaqeenneer caashaqaygay caashaqaagay caashaqiisay caashaqeeday caashaqayagay caashaqooday caashaqiinnay caashaqeennay caashaqaygu caashaqaagu caashaqiisu caashaqeedu caashaqayagu caashaqoodu caashaqiinnu caashaqeennu caashaqu caashaq

*erey saleedka:* caashaq *(f.g1)*
ereyada ka soo farcama erey saleedka *caashaq (f.g1)* waa [dib u samee]: caashaq caashaqay caashaqday caashaqnay caashaqeen caashaqdeen caashaqaa caashaqdaa caashaqnaa caashaqaan caashaqdaan caashaqo caashaq caashaqdo caashaqa caashaqno caashaqayaa caashaqaysaa caashaqaysaan caashaqaynaa caashaqayaan caashaqi

*Farcanka lammaane erey saleedka caashaq*
Ma soo helin lammaane ka yimi erey saleedka (caashaq)

## *Translations*

Finally the translation box will show you different translations from different bilingual dictionaries (Somali-Italian, Somali-English, Somali-Swedish). Dictionaries of reference are being revised for the English, however the Italian and French are in good quality, while a request for permission of use of the Swedish-Somali dictionary is being processed. Other languages can be added to the platform.



Turjumaadaha erey saleedka caashaq

*Talyaani: caashaq*
*n. m. – 1. vr. cishqi. Amore, innamoramento appassionato e profondo. 2. vr. di caashaqid.*

*v. tr. 1 (–qay, –qday) – 1. Innamorarsi di qn. in maniera appassionata e profonda. 2. Ku c.: far attraccare qs. a qs. (cm. molo, altra imbarcazione, ecc.).*

*Ingiriisi: caashaq*
*love, be ~ fall in love*

*love, romance*

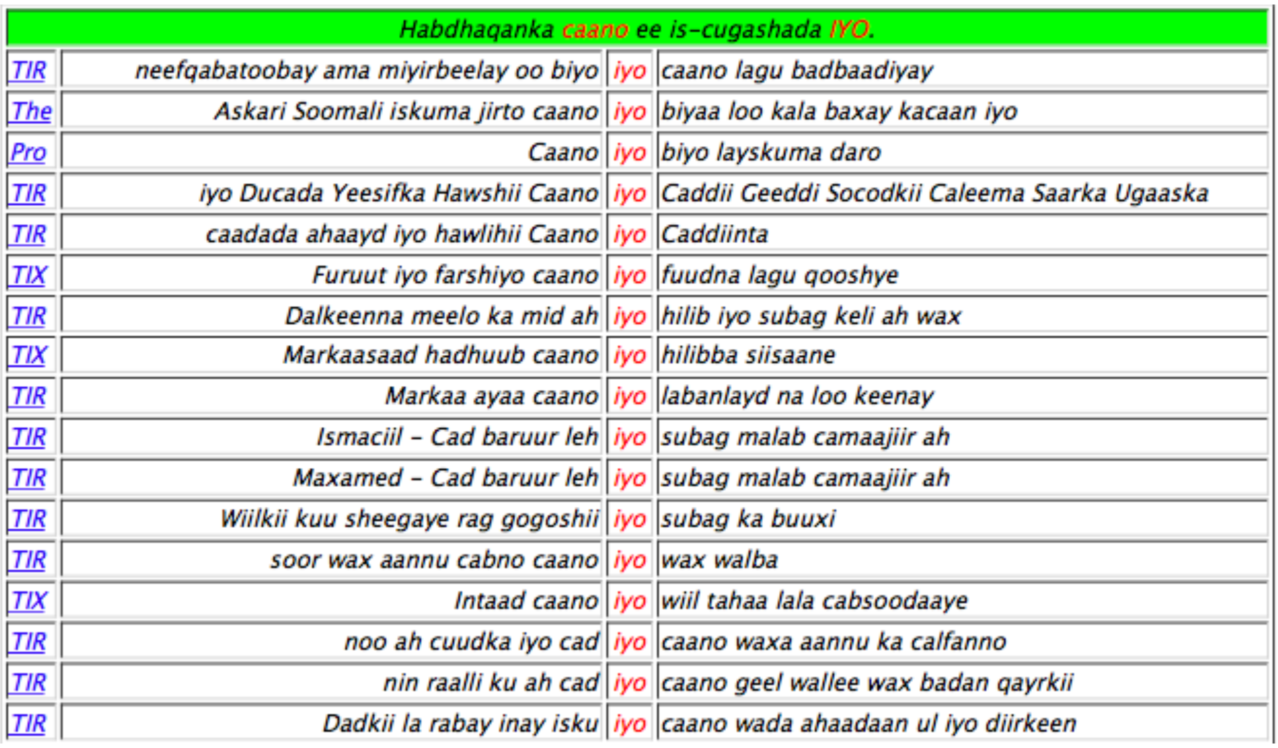
## 4. Automatic identification of linguistic structures in the corpus

There are few configurations defined within the platform for specific types of Parts of Speech. The aim is to show how the selected word behaves within the discourse (for instance we have already seen if the selected word is Noun, we want to check the collocation of the N with the word “*iyo*”, so we can figure out all other nouns (N1) where the configuration “N *iyo* N1” matches. These results are shown in the Word Behaviour view. Different configurations apply to different types of Parts of Speech.

### *Noun behaviour*

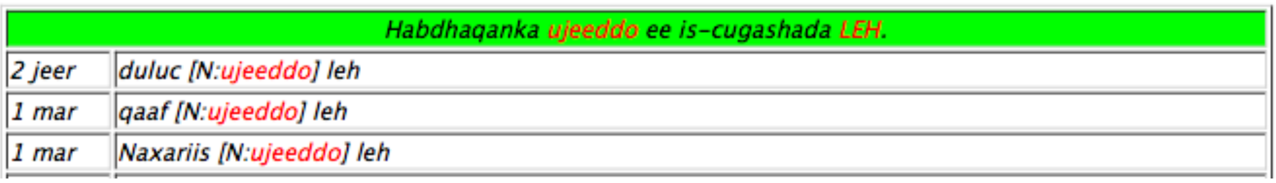
Nouns have different defined configurations: 1) the collocation with the conjunction “*iyo*” (and) as explained above and 2) the collocation with verb *leh* (to have) as mentioned above; 3) for locational nouns (*ag*, *dhex*, *kor*, *hoos*, *dul*, *dhinac*), we enlist the list of verbs immediately following the locational noun.

Example 1: Collocation of the noun *caano* (milk) with “*iyo*” (and).



| Habdhaqanka caano ee is-cugashada IYO. | | | |
|---|---|---|---|
| TIR | neefqabatoobay ama miyirbeelay oo biyo | iyo | caano lagu badbaadiyay |
| The | Askari Soomali iskuma jirto caano | iyo | biyaa loo kala baxay kacaan iyo |
| Pro | Caano | iyo | biyo layskuma daro |
| TIR | iyo Ducada Yeesifka Hawshii Caano | iyo | Caddii Geeddi Socodkii Caleema Saarka Ugaaska |
| TIR | caadada ahaayd iyo hawlihii Caano | iyo | Caddiinta |
| TIX | Furuut iyo farshiyo caano | iyo | fuudna lagu qooshye |
| TIR | Dalkeenna meelo ka mid ah | iyo | hilib iyo subag keli ah wax |
| TIX | Markaasaad hadhuub caano | iyo | hilibba siisaane |
| TIR | Markaa ayaa caano | iyo | labanlayd na loo keenay |
| TIR | Ismaciil - Cad baruur leh | iyo | subag malab camaajiir ah |
| TIR | Maxamed - Cad baruur leh | iyo | subag malab camaajiir ah |
| TIR | Wiilkii kuu sheegaye rag gogoshii | iyo | subag ka buuxi |
| TIR | soor wax aannu cabno caano | iyo | wax walba |
| TIX | Intaad caano | iyo | wiil tahaa lala cabsoodaaye |
| TIR | noo ah cuudka iyo cad | iyo | caano waxa aannu ka calfanno |
| TIR | nin raalli ku ah cad | iyo | caano geel wallee wax badan qayrkii |
| TIR | Dadkii la rabay inay isku | iyo | caano wada ahaadaan ul iyo diirkeen |

Example 2: Collocation of the noun *ujeeddo* (subject) with verb *leh* (to have)

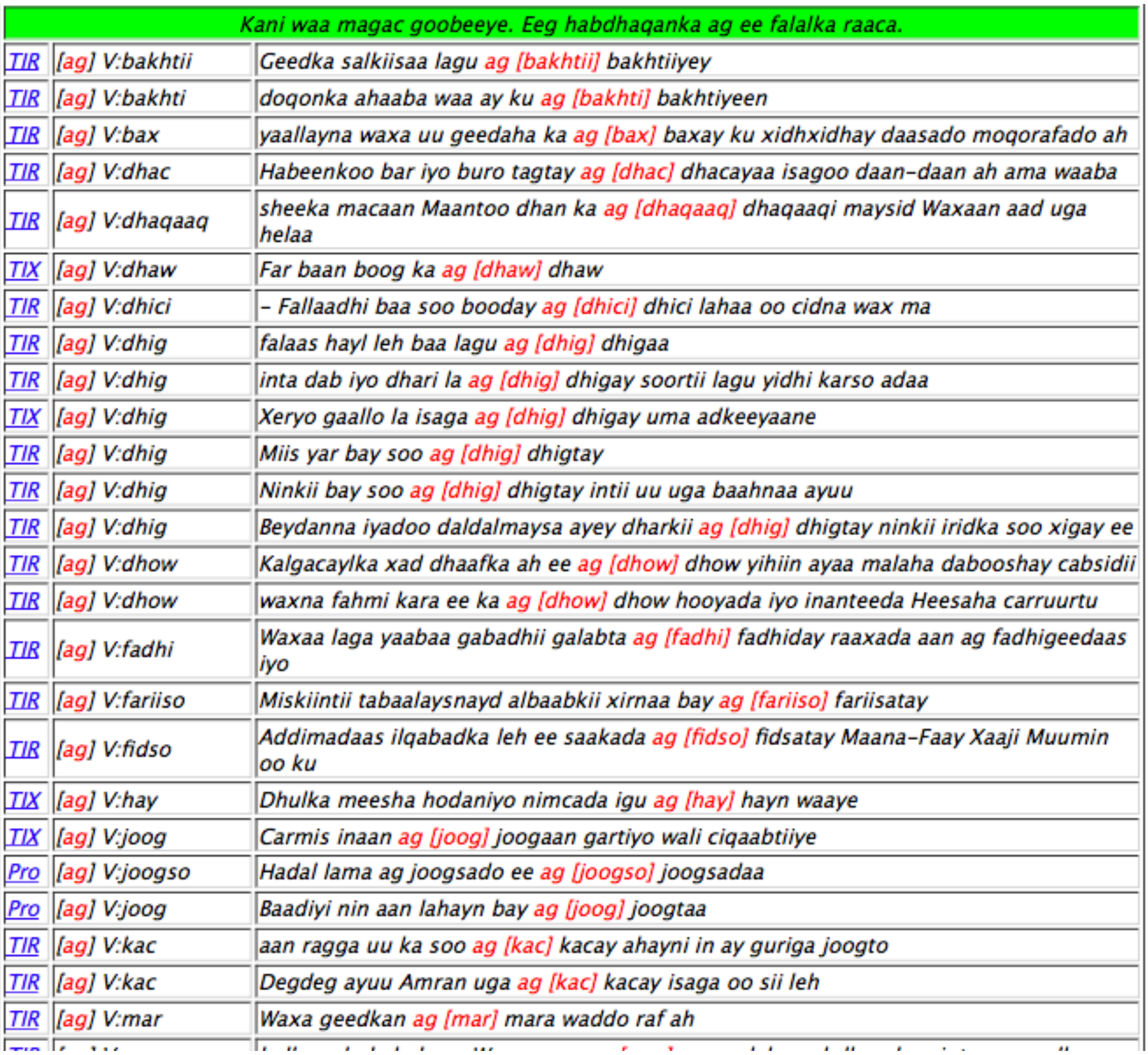


| Habdhaqanka ujeeddo ee is-cugashada LEH. | |
|---|---|
| 2 jeer | duluc [N:ujeeddo] leh |
| 1 mar | qaaf [N:ujeeddo] leh |
| 1 mar | Naxariis [N:ujeeddo] leh |

## Locational nouns

Banti calls locational nouns (magac goobeeye) those nouns that indicate reference to a place: *ag*, *dhex*, *kor*, *hoos*, *dhinac*, *dul*, *daba*. Whenever these nouns are shown in the WP view, the system locates and displays the collocation of Verbs that immediately follow this locational noun. For example shows the result for the locational noun “*ag*”:

Example 3: collocation for locational noun “*ag*”

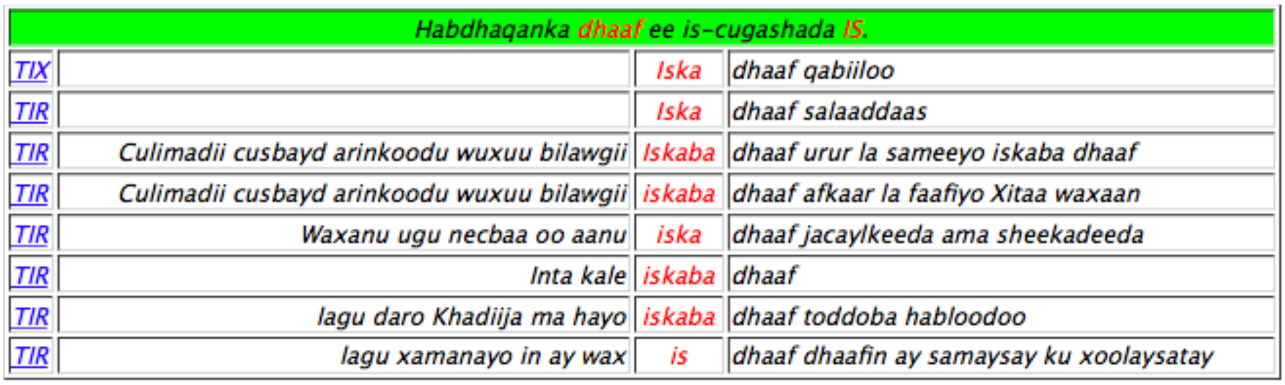


| Kani waa magac goobeeye. Eeg habdhaqanka ag ee falalka raaca. | | |
|---|---|---|
| TIR | [ag] V:bakhtii | Geedka salkiisaa lagu ag [bakhtii] bakhtiiyey |
| TIR | [ag] V:bakhti | doqonka ahaaba waa ay ku ag [bakhti] bakhtiyeen |
| TIR | [ag] V:bax | yaallayna waxa uu geedaha ka ag [bax] baxay ku xidhxidhay daasado moqorafado ah |
| TIR | [ag] V:dhac | Habeenkoo bar iyo buro tagtay ag [dhac] dhacayaa isagoo daan-daan ah ama waaba |
| TIR | [ag] V:dhaqaaq | sheeka macaan Maantoo dhan ka ag [dhaqaaq] dhaqaaqi maysid Waxaan aad uga helaa |
| TIX | [ag] V:dhaw | Far baan boog ka ag [dhaw] dhaw |
| TIR | [ag] V:dhici | – Fallaadhi baa soo booday ag [dhici] dhici lahaa oo cidna wax ma |
| TIR | [ag] V:dhig | falaas hayl leh baa lagu ag [dhig] dhigaa |
| TIR | [ag] V:dhig | inta dab iyo dhari la ag [dhig] dhigay soortii lagu yidhi karso adaa |
| TIX | [ag] V:dhig | Xeryo gaallo la isaga ag [dhig] dhigay uma adkeeyaane |
| TIR | [ag] V:dhig | Miis yar bay soo ag [dhig] dhigtay |
| TIR | [ag] V:dhig | Ninkii bay soo ag [dhig] dhigtay intii uu uga baahnaa ayuu |
| TIR | [ag] V:dhig | Beydanna iyadoo daldalmaysa ayey dharkii ag [dhig] dhigtay ninkii iridka soo xigay ee |
| TIR | [ag] V:dhow | Kalgacaylka xad dhaafka ah ee ag [dhow] dhow yihiin ayaa malaha dabooshay cabsidii |
| TIR | [ag] V:dhow | waxna fahmi kara ee ka ag [dhow] dhow hooyada iyo inanteeda Heesaha carruurtu |
| TIR | [ag] V:fadhi | Waxaa laga yaabaa gabadhii galabta ag [fadhi] fadhiday raaxada aan ag fadhigeedaas iyo |
| TIR | [ag] V:fariiso | Miskiintii tabaalaysnayd albaabkii xirnaa bay ag [fariiso] fariisatay |
| TIR | [ag] V:fidso | Addimadaas ilqabadka leh ee saakada ag [fidso] fidsatay Maana-Faay Xaaji Muumin oo ku |
| TIX | [ag] V:hay | Dhulka meesha hodaniyo nimcada igu ag [hay] hayn waaye |
| TIX | [ag] V:joog | Carmis inaan ag [joog] joogaan gartiyo wali ciqaabtiiye |
| Pro | [ag] V:joogso | Hadal lama ag joogsado ee ag [joogso] joogsadaa |
| Pro | [ag] V:joog | Baadiyi nin aan lahayn bay ag [joog] joogtaa |
| TIR | [ag] V:kac | aan ragga uu ka soo ag [kac] kacay ahayni in ay guriga joogto |
| TIR | [ag] V:kac | Degdeg ayuu Amran uga ag [kac] kacay isaga oo sii leh |
| TIR | [ag] V:mar | Waxa geedkan ag [mar] mara waddo raf ah |

### *Verb behaviour*

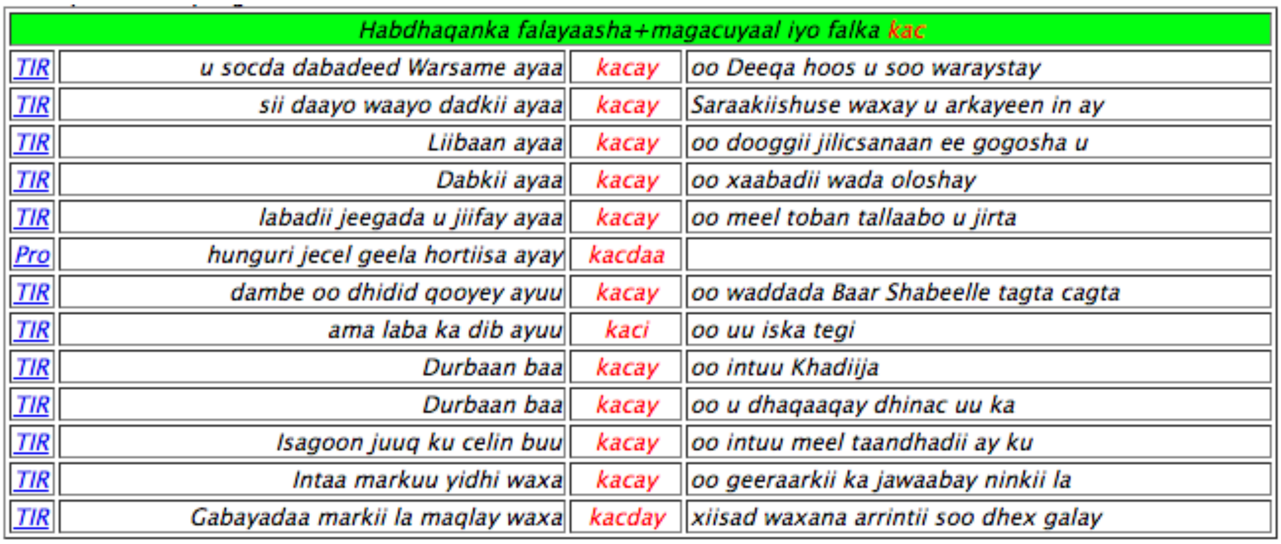
Similar to the nouns, also the verbs have pre-defined collocation configurations in place. For example 1) the collocation of the search verb V with the reflexive pronoun “*is*”, or 2) the collocations with focus markers either with our without pronouns ('*ayaa*', ‘baa’, ‘waxa’, '*ayuu*', '*ayay*', ‘*buu’, ‘bay*‟, '*wuxuu*', '*waxay*').

Example 4: Collocation of the verb "*dhaaf*" (leave, omit) with *"is"*



| Habdhaqanka dhaaf ee is-cugashada IS. | | | |
|---|---|---|---|
| TIX | | Iska | dhaaf qabiiloo |
| TIR | | Iska | dhaaf salaaddaas |
| TIR | Culimadii cusbayd arinkoodu wuxuu bilawgii | Iskaba | dhaaf urur la sameeyo iskaba dhaaf |
| TIR | Culimadii cusbayd arinkoodu wuxuu bilawgii | iskaba | dhaaf afkaar la faafiyo Xitaa waxaan |
| TIR | Waxanu ugu necbaa oo aanu | iska | dhaaf jacaylkeeda ama sheekadeeda |
| TIR | Inta kale | iskaba | dhaaf |
| TIR | lagu daro Khadiija ma hayo | iskaba | dhaaf toddoba habloodoo |
| TIR | lagu xamanayo in ay wax | is | dhaaf dhaafin ay samaysay ku xoolaysatay |

Example 5: Collocation of verb "kac" (up) with focus makers



| Habdhaqanka falayaasha+magacuyaal iyo falka kac | | | |
|---|---|---|---|
| TIR | u socda dabadeed Warsame ayaa | kacay | oo Deeqa hoos u soo waraystay |
| TIR | sii daayo waayo dadkii ayaa | kacay | Saraakiishuse waxay u arkayeen in ay |
| TIR | Liibaan ayaa | kacay | oo dooggii jilicsanaan ee gogosha u |
| TIR | Dabkii ayaa | kacay | oo xaabadii wada oloshay |
| TIR | labadii jeegada u jiifay ayaa | kacay | oo meel toban tallaabo u jirta |
| Pro | hunguri jecel geela hortiisa ayay | kacdaa | |
| TIR | dambe oo dhidid qooyey ayuu | kacay | oo waddada Baar Shabeelle tagta cagta |
| TIR | ama laba ka dib ayuu | kaci | oo uu iska tegi |
| TIR | Durbaan baa | kacay | oo intuu Khadiija |
| TIR | Durbaan baa | kacay | oo u dhaqaaqay dhinac uu ka |
| TIR | Isagoon juuq ku celin buu | kacay | oo intuu meel taandhadii ay ku |
| TIR | Intaa markuu yidhi waxa | kacay | oo geeraarkii ka jawaabay ninkii la |
| TIR | Gabayadaa markii la maqlay waxa | kacday | xiisad waxana arrintii soo dhex galay |

## 5. Advanced Query Language

### *Introduction to Advanced Query Language for RCF Somali Corpus*

The RCF Somali Corpus platform has its own advanced query language. Users can by themselves compose complex query strings and search with these strings against the corpus.[1] The valid keywords for the query formulations consist of indicators and placeholders (see below table 1):

| **Keyword** | **Function** | **Type** |
|---|---|---|
| HORRAAD: | the word I am searching follows immediately […] | Indicator |
| DANBEED: | the word I am searching precedes immediately […] | Indicator |
| F | the word in this place is a verb | Placeholder |
| M | the word in this place is a noun | Placeholder |
| NOOC:*XX* | the word in this place belongs to *XX* Part of Speech[2] | Placeholder |

Table 1: Defined keywords in Advanced Query Language for RCF Somali Corpus.

### Example 6: cases of complex Advanced Query Language

| **Query** | **Description** | **Example results** |
|---|---|---|
| cun HORRAAD:uu | Find all sentences containing any inflected form of the verb *cun* (eat) preceded by the word *uu.* | • wax cunto ahna ma **uu cunin**<br>• mar kale na waa qof **uu cunayba**<br>• qofku si uu ku noolaado waxa **uu cunayaa** waa inay ahaataa...<br>• wax ku baranayey in **uu cunay** hilib doofaar<br>• maandooriyaha **uu cuno** oo keliya ayay wehel wadaag...<br>• […] |
| M DAMBEED:dheh | Find all sentences where the verb *dheh* (say) is preceded by a noun. | • **doobir dheh**<br>• markaasuu ku yidhi **car dheh**<br>• kute **wir dheh**<br>• afartaa **shax dheh** maanso<br>• […] |
| M DAMBEED:NOOC:v.g4 | Find all sentences containing a Noun followed by a verb of the 4th type | • Ciise oo **xaalladiisa fahansan** baa<br>• qobol uga keeni karaa **dad raacsan** oo u buuxiya tirada laga doonayo<br>• markaasaa **dadkii raacsanaa** ka cadhoodeen |

Table 2: Examples of Advanced Query Language for RCF Somali Corpus.

[1] These queries are heavy for the system to manage, so they are partially allowed to all users. The administrator can grant which user can run different types of queries.

[2] The abbreviations for the Parts of Speech types are defined in Jama Musse Jama, "A syntactically annotated corpus for Somali literature", 2016, and are based on the abbreviations used in Annarita Puglielli and Cabdalla Cumar Mansoor, "Qaamuuska Af-soomaaliga", 2012. They can be also accessed through the "Qaamuuska/Dictionary" link at the top bar of the user interface.

### *Using wildcard, multi word search and special characters*

RCF Somali Corpus supports three wildcard characters: *, ? and %. 1) * matches zero or more words of any part of speech; 2) ? matches exactly one word of any part of speech in that place; and finally 3) % allows to partial search of the single word in the sentence.

The system supports also special characters “” matching exactly the sequence of words within quotes.

Example 7: use of wildcard and special characters in AQL

| **Query** | **Description** | **Example results** |
|---|---|---|
| NOOC:m.dh*heli* | Find all sentences containing any inflected form of feminine noun, followed by zero or more words followed by the word *heli* followed by any word. | • **Dawadeeda** waan **heli** lahaa dayaxaba haddaan tago.<br>• Adoon **baahi** qaba baad **heli** kartaa booska uu yahay e.<br>• **Tirada** xarummaha oo kooban darteed marmar waxaad **heli** kartaa<br>• **Labadaas** buug ka **heli** maynno oo keligood meel yaalla.<br>• […] |
| M?NOOC:f.g1garee | Find a sentence where a noun is followed an inflected form of 1st conjugation verb followed by the word “*garee*” | • u midowdo oo la dhiso oo **ummadda** diyaar **gareeya** waajib<br>• **dadkii** caasi **gareeya**<br>• […] |
| cuna% | Find sentences containing a word starting with *cuna* | • bahalku waxa uu **cunayaa** hilib qeedhin<br>• iyagu way **cunaan** geedaha<br>• […] |
| qalin cas | Sentence containing both words *cas* and *qalin* | • **qalin** madow iyo mid **cas** buu lahaa<br>• shimbir cas iyo balanbaalis **qalin** leh ayaan arkay<br>• macallinku **qalin cas** ayuu haystaa.<br>• [..] |
| “qalin cas” | | • macallinku **qalin cas** ayuu haystaa. |

Table 3: Examples of Advanced Query Language for RCF Somali Corpus: wildcards.

Note that if at least a AQL keyword (ie. HORRAAD: NOOC: DANBEED:) is found in the query string, or if the searched word is a single word, then the system will perform the search on the structure and the grammar of the sentence. Otherwise (if there are multiple words search with spaces, or special characters %, “”), the query will be done as free text string.