Archimob Corpus on Sketch Engine and Annis
Getting Started

Noëmi Aepli, Fatima Stadler, Tanja Samardžić

March 23, 2017

1 Searching the Archimob Corpus

This is a brief tutorial describing how to access and search the Archimob Corpus with two online search tools namely Sketch Engine and Annis. The purpose is to show how to get started with these tools and to provide a few examples of possible queries. For further information and advanced search techniques please refer to the respective websites and documentations.

2 Annis

Annis is a web browser-based tool to search linguistic corpora. We will explain the most important queries to get started. For a more detailed explanation on corpus queries, we recommend going through the Annis tutorial or reading the Annis user guide: http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.4.3.pdf

2.1 Access to Annis

To access the Archimob corpus on Annis visit the website https://annis.linguistik.uzh.ch/annis-gui/archimob, click the “Login” button in the upper right corner and enter the credentials:

username: archimob password: archimob@UZH4All

Annis comes with a built-in tutorial that is shown when opening the website (see figure 1). It can always be re-accessed by choosing the tutorial tab. After logging in, you will find the Annis query builder on the left side of the window.
and below it the corpus list, as shown in figure 2. In order to perform queries, you first need to activate the Archimob corpus by clicking on its name in the list, which will then be highlighted in blue.

2.2 Queries

Within Annis you can perform simple word queries (see section 2.2.1), use regular expressions (see section 2.2.2), search for co-occurrences (see section 2.2.3) and define the context for the search (see section 2.2.4). Moreover, you can for instance search for different dialectal variants of a Standard German word by
executing a query on the normalised layer or you can use the part of speech information on the tag layer for your queries. The following annotation levels are available:

- **id**: identification number of a specific document, utterance or word
- **tok**: transcribed words
- **normalised**: normalisation of the transcribed word
- **tag**: part of speech tags
- **type**: pauses, para- and nonverbal contents, comments

After performing a query, you have the option to do a frequency analysis on the result (see section 3.4) and export (see section 3.5) the query results or the frequency analysis of the query.

### 2.2.1 Simple Word Query

To find an exact variant of a word, e.g. *miuch* (eng. ‘milk’ according to a specific pronunciation), you can perform a query on the token level by specifying the annotation level *tok* along with the exact word in quotation marks:

```
tok="miuch"
```

If you are interested in all dialectal variants you can execute the query on the normalisations. In order to do so, set the annotation level to *normalised* and wrap the normalised form in quotation marks.

```
normalised="milch"
```

The above query on the normalisation layer will yield dialectal variants appearing in the “Query Result” as shown in figure 3.

### 2.2.2 Using Regular Expressions

With the use of regular expressions you can specify more flexible and more specific queries. Instead of quotation marks as in the “normal” search, wrap

```
annotation_level="/xyz/"
```

4Normalisations are labels that group dialectal variants together. In most cases they are similar to Standard German forms. More information on normalisation is included in the corpus documentation.

5For those who are not familiar with regular expressions, here are some basic operations: `.` matches any character, `*` quantifies the preceding character, meaning it occurs zero or more times whereas with a `+` it must occur at least once and `?` states that the preceding character is optional.
your queries in slashes as shown above. If you are interested in particular pronunciations like for example the merged form of ‘have’ 1. P. Pl. and the corresponding personal pronoun, you can perform a RegEx search on the tok-level: \texttt{tok=/h\[\textae\]mm?er/}

### 2.2.3 Co-occurrences

Annis can also be used to find co-occurring tokens, the abbreviated Annis syntax for the corresponding query is the dot operator. This makes it possible to find interesting examples of Swiss German syntax like for example word order phenomena. To search for the verb 
\textit{lassen} preceding any other verb you can search a range of variants on the token level (\texttt{tok}) and combine it with any subsequent verb on the part of speech level (\texttt{tag}):

\texttt{tok=/l\[\textao\]+/ .\ tag=/VV.*/}

With this expression we find sentences like the following:

(1) \textit{hend mi\cch deet l\textoe ligge} (Sempach, LU)
If you want to specify a span of tokens within which to search, you can indicate the minimal and the maximal number of positions directly attached to the dot operator, separated by a comma.

\texttt{tok=/l[^ao` o]+/ .1,3 tok=/l[^ao` o]+/}

The results of the query expression above contains cases of a dialectal reduplication of the verb 'lassen' with the particle \textit{la} and the infinitive \textit{laoa} in combination with \textit{z} (dialectal variant for 'zu').

\begin{itemize}
\item \textit{dâ jung pûrschtel da la furt z laa} (Bern, BE)
\end{itemize}

\section*{2.2.4 Defining the Range of Word Context}

Depending on your research question, the span of the context you are interested in may vary. You can adjust the range around the keyword by selecting the “Search Options” tab below the query window on the left and defining the number of positions around your query match (see figure 4).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure4.png}
\caption{Options to define the query span}
\end{figure}

\section*{2.3 Frequency Analysis}

After performing a query, you can do a frequency analysis on the results and export the search results and rankings. To do so, click the “More” button below the query field and choose the desired action (see figure 5).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure5.png}
\caption{Frequency analysis and export options on the performed query}
\end{figure}

Choosing the “Frequency Analysis” option will result in a new tab where the analysis can be specified as shown in figure 6. In this step you can choose the
level on which the frequency analysis shall be performed by changing “Selected annotation of node” before confirming with “Perform frequency analysis”.

Figure 6: Define options for the frequency analysis

The frequency analysis can take a while. Once finished, the result is displayed as a histogram. Figure 7 shows the frequency analysis result for the query normalised="uns", i.e. a query on the normalised layer to get all the available variants of the first person plural pronoun ‘uns’. Furthermore, a table with frequency counts (as shown in figure 8) is generated in which you can sort and rank the result by selecting the corresponding table header. Note that the displayed plot is interactive: by selecting a specific bar in the histogram, the corresponding variant is activated in the table. The analysis can be downloaded by clicking “Download as CSV”.

Figure 7: Histogram for the query normalised="uns"
Choosing “Export” in the menu “More” (see figure 5), the results of queries can be exported in different formats. According to your needs you can choose between a WEKA, CSV, Token, Grid or SimpleText format (see figure 9). Exports from Annis are time consuming, therefore it can be advantageous to narrow the query scope to sub-queries and then perform several exports. A description of the formats in Annis is visible on the right side when you choose a format in the menu. The following extracts give you a quick impression of the export formats.

Figure 8: Interactive table with variants and corresponding counts

![Interactive table with variants and corresponding counts](image)

**Figure 8: Interactive table with variants and corresponding counts**

**2.4 Export**

Choosing “Export” in the menu “More” (see figure 5), the results of queries can be exported in different formats. According to your needs you can choose between a WEKA, CSV, Token, Grid or SimpleText format (see figure 9). Exports from Annis are time consuming, therefore it can be advantageous to narrow the query scope to sub-queries and then perform several exports. A description of the formats in Annis is visible on the right side when you choose a format in the menu. The following extracts give you a quick impression of the export formats.

![Export function in Annis](image)

**Figure 9: Export function in Annis**
2.4.1 SimpleText

0. echli blöüi [m’üuch] und geschwelt härdepf dus hemmer
1. bi der [milch] det isch eh dän äu
2. pro litter [milch] und då hånd sich d

2.4.2 WEKA

@relation name
@attribute #1_id string
@attribute #1_span string
@attribute #1_anno_default_ns:id string
@attribute #1_anno_default_ns:normalised string
@attribute #1_anno_default_ns:tag string
@data

'9825126','m’üuch','d1007-u778-w3','milch','NN'
'9842020','milch','d1048-u92-w11','milch','NN'
'9842040','milch','d1048-u94-w14','milch','NN'

2.4.3 CSV

1_id 1_span 1_anno_default_ns:id 1_anno_default_ns:normalised 1_anno_default_ns:tag
9825126 m’üuch d1007-u778-w3 milch NN
9842020 milch d1048-u92-w11 milch NN
9842040 milch d1048-u94-w14 milch NN

2.4.4 Token

0. echli/d1007-u778-w1/ein klein/PIAT blöüi/d1007-u778-w2/blaue/ADJA [m’üuch]/d1007-u778-w3/milch/NN
1. bi/d1048-u92-w9/bei/APPR der/d1048-u92-w10/der/ART [milch]/d1048-u92-w11/milch/NN
2. pro/d1048-u94-w12/pro/APPR litter/d1048-u94-w13/liter/NN [milch]/d1048-u94-w14/milch/NN

2.4.5 Grid

0. tok echli blöüi m’üuch
start media_pointers#d1007-T1328[1-6]
who person_db#EJos1007[1-6]
1. tok  bi der milch
who  person_db#WRos1048[1-3]

...
3 Sketch Engine

The second search tool that can be used for searching the Archimob corpus is Sketch Engine\textsuperscript{2}. This tutorial captures only a fraction of queries possible with Sketch Engine, for more detailed information please refer to the software user guide at \url{https://www.sketchengine.co.uk/user-guide/} where you can also find video tutorials \url{https://www.sketchengine.co.uk/user-guide/videos/sketch-engine-in-videos/}.

3.1 Access to Sketch Engine

To access the Archimob corpus on Sketch Engine, login on the website \url{https://www.sketchengine.co.uk/}. Please contact us\textsuperscript{6} in order to get the login credentials. After logging in, you can select the corpus you want to search; choose Archimob\_Release1. You will see a query window as in figure 10 where simple queries can be performed. Clicking “Query types”, “Context” and “Text types” will extend the window to provide the respective query options (see figure 11).

![Figure 10: Query window for the simple search](image)

3.2 Queries

Figure 11 shows the extended query window for a corpus search. The first third of the window (“Query types”) displays the different types of queries which can be executed (see section 3.2.1). The other two sections provide options to limit the context or the text types (see section 3.2.3) on which the search is executed.

After each query, further options will be available in the left menu in order to sort (see section 3.3), analyse (see section 3.4) or save (see section 3.5) the result. In figures 12 and 13 you can see the left menu options.

\textsuperscript{6}tanja.samardzic@uzh.ch
3.2.1 Query Types

- **simple**: By default, the query is performed as flexible search. This means, if we search for *gäl* (eng. ‘money’), which is normalised to *geld*, the result contains all kinds of variants such as *gäud* or *gült*, etc. which have the same normalisation. However, potentially different words not relevant to the search might be included in the result. In order eliminate unwanted examples, choose a different query type such as **word** or **CQL**.

- **word** finds the exact word form as it is typed (without orthographic flexibility).

- **character** finds a sequence of characters inside a token. *men* for instance corresponds to the RegEx .*men.*, meaning that any characters can surround *men*. The result will include words like *men*, *amen*, *bimene*, *umened* etc.

- **phrase** finds examples of a sequence of words exactly as typed.

- **CQL** allows more complex searches. The Corpus Query Language is explained in more detail in section 3.2.2. In this setting, the layer to be searched can be specified: **word** (as described above), **tag** (part of speech tags) or **normalised**. This option is to be used in combination with simple query in cases where we are not sure how a word should be written. In this case, we first perform a simple query typing the word in any of Swiss Ger-
man variants. If the word is recognised, the system will return instances of the desired word, but probably some other words too. The resulting concordances will show (in grey) the normalised form of each returned instance, which we can then use to search for the desired normalisation using the “CQL” query type.

3.2.2 Corpus Query Language

[CQL]

CQL allows to set more complex criteria in a search. The basic format is an attribute value pair as shown above. In order to search for a phrase, each token has to be surrounded by square brackets. The attribute can be word, normalised or tag as described above.

RegEx can be used with values, e.g. words ending with -zion: [word=".*zion"]:

(3)  oder irgend än informazion chönd si sich

... or different variations of haben wir: [word="h[a|e]mm?er"]

(4)  häröpfu hemer sowisoo sälber ghaa
e schtund hämmer müse lauffe

RegEx can also be useful when searching the part of speech tag layer. For example, to search for any auxiliary verb: [tag = "VA.*"], any conjunction [tag = "KO.*"] or any particle [tag = "PTK.*"]. Especially interesting in the case of the Archimob corpus might be to search for merged words, which are labelled with a part of speech tag ending with +. To do so, the plus sign has to be escaped with a backslash (as otherwise it has a special meaning in regular expressions): [tag = ".*\+"] The result are words like wos, hets, wemme, dasch, etc.

It is also possible to specify a part of speech tag for a specific word: [word = "äss[eâ"] & tag = "NN"]

(5)  mi händ das ässe ghaa wommer

By placing a question mark after the square bracket, the token is made optional, i.e. it can appear once or not at all:

\[^\]

The part of speech tags used in the Archimob corpus are an extended version of the Stuttgart Tübingen Tagset, please refer to the STTS guidelines [http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf] and the documentation of the Archimob corpus.

For more information about the use of the plus sign in the Archimob corpus, please refer to the documentation of the corpus.
Parentheses can be used for grouping. For example \([(\text{word} = "z" \mid \text{word} = "zum"))\] means there must be either the word \(z\) or the word \(zum\).

\[
(6) \quad \text{tischli inschtalirt f"ur z "asse draf aagfrogt woorde f"ur cho z singe sii het "pper haa f"ur der z h"alf fe gw"uss geng}
\]

Curly brackets \(\{m,n\}\) allow \(m\) to \(n\) repetitions of the preceding token:

\[
(7) \quad \text{d"a jung p"urschtel da la furc z laa i d"are ziit}
\]

### 3.2.3 Text Types

The query can be limited to a specific document ID\(^9\) or to a specific location which is available in the corpus (e.g. NW \((\text{Stans})\)). The information about the document ID or location can also be accessed by clicking the blue number on the left in any search result (see figure 12).

### 3.3 Sort

The results of the performed query can be sorted. Check the sort options in the menu on the left (see figure 13). By clicking “Left” under the point “Sort”, the resulting sentences will be sorted alphabetically according to the word appearing on the left side of the queried word. “Node” will sort the according to the searched word itself. By clicking “Sort” itself, more complex sorting options can be specified.

### 3.4 Analysis

In order to compute frequencies of the performed search, click “Frequency” in the menu on the left (see figure 13). A window like in figure 14 will pop up to specify the criteria for the frequency analysis. Figure 15 shows an example

---

\(^9\)Please refer to the Archimob corpus documentation.
frequency analysis for the simple query "füf" (eng. ‘five’). In the left menu you can choose further options for processing the frequency analysis or saving it.

3.5 Export

The result of the query can be exported as a text file or in XML format. In order to do so, click “Save” in the menu on the left and specify the export options shown in the window as in figure 16.

3.6 Word Sketch: Collocations

Along with the above described “Search” options, Sketch Engine also offers the option “Word sketch”. A word sketch shows the word’s collocates categorised by grammatical relations including statistical analyses. Figure 17 shows an example word sketch for "ässe" (eng. ‘eat’). The blue number shows the frequency of the collocation. The source concordances can be accessed by clicking on the number. Grey phrases represent the most typical use of the collocation.
Figure 13: Right-sorted result result of the query [word = "1[aoö]+"]

In the left menu you can choose options for further processing the word sketch or saving it.
Figure 14: Options to be set for the frequency analysis

Figure 15: Frequency analysis for the query füf
Figure 16: Save the result of a query

Figure 17: Word sketch for ässe