Documentation:
GeWiss Corpus Web services

Feb 2015
authors: Daniel Jettka, David Stoppel

Contents

I. Introduction

In the course of the CLARIN Curation Project CLARIN-KP-GeWiss, several RESTful Java Web services have been developed which can be used for basic processing of the GeWiss Corpus data. These are server-side applications which may be accessed by users in a standardized, parametric way to provide pre-defined data.

The GeWiss Web services allow for the retrieval of complete lists of

- the existing sub-corpora, filtered by base language, academic context, or discourse genre,
- the communications a certain sub-corpus comprises of,
- speakers involved in a certain communication,
- transcripts of the communications,
- audio recordings of the communications,
- the different types of metadata and annotations,

moreover, they can be used to

- display complete metadata for any single communication or speaker,
- search transcripts and annotations within transcripts,
- retrieve concordances, i.e. to search for text in the whole corpus or a set of sub-corpora, and to retrieve and filter for the according metadata.

The output takes different formats such as XML, which enhances post-processing and analysis through the users' Web services or applications, or HTML, which allows for displaying the results in a web browser.

For background information concerning the GeWiss project, communications and speakers as well as transcription and annotation methods, please refer to the GeWiss Corpus manual in German (Gräfe et al. (2015). Additional information on metadata can be found at https://gewiss.uni-leipzig.de/index.php?id=metadata, and at https://gewiss.uni-leipzig.de/index.php?id=annotations on annotations (in German).

II. Prerequisites

Some of the Web services access corpus data containing sensitive information on speakers, and thus require password authentication. Registered users of the GeWiss interface may get access to the restricted Web services upon request (please notify us at gewiss-korpus@uni-leipzig.de). The Web service PlayRecording, which permits access to audio recordings of whole communications, is subject to further data protection requirements. Access will be granted only upon individual agreement.

Accessibility restrictions for Web services are indicated in the following sections, where all GeWiss Web services are outlined in detail. These can be executed via wget or cURL, e.g., (CLI programs for downloading data from the WWW via HTTP- and FTP, available for different operating systems) or using the browser plugin RESTClient. You can also call the Web services directly from your web browser.

III. Using the Web services

The data output format is given for each Web service in the following list. If only one output format is indicated, this format is returned automatically when the respective Web service is called.

By appending parameters to the URL String, the results can be specified (in general, this means specifying metadata for more selective retrieval). A complete list of parameters is given for each Web service. While most of the parameters are optional, some are obligatory as indicated.

The parameters appear in the query string of the URL; i.e., after the URL of the Web service a question mark "?" indicates the start of the query string, and the parameters following take the form "parameter name=parameter value":

> *Example: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora?acad-context=UK*

Web services taking multiple parameters allow for freely combining these, if not indicated otherwise. The single parameters are separated by ampersands "&":

> *Example: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora?acad-context=UK&lang=DEU&lang-competence=L2&disc-genre=EV*

Each parameter may appear only once within a given query string, if not indicated otherwise.

> *not possible: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora?lang=DEU&lang=ENG*

For some optional parameters, a default value is given. If the parameter is not appended to the query string manually, the parameter will be set to the default value automatically by the Web service.

> *Example:* ListMetadata; parameter *codes* is optional, default is *"true"*. Calling *http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListMetadata* corresponds to *http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListMetadata?codes=true*

Note on password restricted Web services: When calling the Web services via wget, include your user name and password as follows:

> *Example: –http-user="USERNAME" –http-password="PASSWORD" http://….*

When calling the Web service in your web browser, a pop-up window appears that asks you to submit your user credentials.

IV. List of Web services

1. List Subcorpora (open access)

Description:
Returns a list of all GeWiss subcorpora in the form of their codes. The subcorpora may be further specified by base language, academic context etc. The codes serve as parameter values for other Web services.
For a complete list of the subcorpora, see appendix.

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
acad-context	academic context	yes	UK, PL, BG
lang	language	yes	DEU, ENG, POL, ITA, BUL
lang-competence	language competence	yes	L1, L2
disc-genre	discourse genre	yes	EV, PG, SV (expert talk, oral examination, student talk)

Example:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSubcorpora?acad-context=UK&lang=DEU&lang-competence=L2&disc-genre=EV

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<result-list>
   <result>
      <data name="corpus-name">DEU_L2_UK_EV</data>
   </result>
</result-list>
```

2. List Communications (open access)

Description:
Returns a list of all communications of a subcorpus in form of their codes. The codes serve as parameter values for other Web services.

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListCommunications
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListCommunications/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
cor-name	subcorpus code	no	Ex.: DEU_L1_EV, DEU_L2_BG_SV for complete list, see appendix

Example:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListCommunications?cor-name=DEU_L1_EV

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<result-list>
   <result>
      <data name="communication-name">EV_DE_004</data>
   </result>
   <result>
      <data name="communication-name">EV_DE_005</data>
   </result>

   <!-- ... abbr. ... -->

   <result>
      <data name="communication-name">EV_DE_105</data>
   </result>
   <result>
      <data name="communication-name">EV_DE_106</data>
   </result>
</result-list>
```

3. List Speakers (open access)

Description:
Returns a list of all speakers involved in a certain communication in form of their codes. The codes serve as parameter values for other Web services.

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSpeakers
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSpeakers/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
com-name	communication code	no	Ex.: EV_DE_004, PG_DE_047

Example:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ListSpeakers?com-name=EV_DE_004

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<result-list>
   <result>
      <data name="speaker-name">LV_0295</data>
   </result>
   <result>
      <data name="speaker-name">JS_0215</data>
   </result>

   <!-- ... abbr. ... -->

</result-list>
```

4. View Metadata (open access)

Description:
Returns the metadata of a subcorpus, communication, or speaker.

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewMetadata
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewMetadata/html
Method: GET
Output: text/html

Parameters:

name	description	optional*	possible values
cor-name	subcorpus code	yes*	Ex.: DEU_L1_EV, DEU_L2_BG_SV for complete list, see appendix
com-name	communication code	yes*	Ex.: EV_DE_004, PG_PL_090
spk-name	speaker code	yes*	Ex.: MH_0226, RO_1505

*set one parameter

Example:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewMetadata?com-name=EV_IT_002

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<Corpus Name="GeWiss" Id="ID7C223B9A-3273-F969-8822-4E006B5D93B8-X6">
  <DBNode/>
  <CorpusData>
    <Communication Id="ID56B4CC45-81ED-AD08-4318-3F064E48BF6E-ITA_L1_EV"
      Name="EV_IT_002">

      <!-- ... abbr. ... -->

    </Communication>
  </CorpusData>
</Corpus>
```

5. View Transcript (access restricted)

Description:
Returns single communications as transcripts.

URL: http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewTranscript
Method: GET
Output: text/xml

URL:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewTranscript/html

Method: GET
Output: text/html

URL:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewTranscript/pdf
Method: GET
Output: application/pdf

Parameters:

name	description	optional	possible values
tra-name	communication code	no	Ex..: EV_DE_004, PG_PL_090

Example:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewTranscript?tra-name=EV_DE_004

via wget:
wget –http-user="USERNAME" –http-password="PASSWORD" http://gewiss.uni-leipzig.de:8282/gewiss/webresources/ViewTranscript?tra-name=EV_DE_004

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- (c) http://www.rrz.uni-hamburg.de/exmaralda -->
<basic-transcription Id="CIDID2AB43F30-3B69-F4E6-5CCF-20345564498C">
  <head>
    <meta-information>
      <project-name>GeWiss</project-name>
      <transcription-name>EV_DE_004</transcription-name>
      <referenced-file url="../Audio/EV_DE_004.wav" />

      <!-- ... abbr. ... -->

    </speakertable>
  </head>
  <basic-body>
    <common-timeline>
      <tli id="T0" time="0.0" />
      <tli id="T2" time="0.43333277368063916" />
      <tli id="T4" time="5.693326576545131" />

      <!-- ... abbr. ... -->

    </common-timeline>
    <tier id="TIE0" speaker="SPK0" category="v" type="t" display-name="MOD
      [v]">
      <event start="T0" end="T2">(0.4) </event>
      <event start="T2" end="T4">ich freue mich die nächste vortragende
        vorstellen zu dürfen </event>

      <!-- ... abbr. ... -->

    </tier>
  </basic-body>
</basic-transcription>
```

6. Play Audio Recording (access restricted)

Description:
Returns single communications in form of audio recordings.

Note:
Due to data protection restrictions, access to this Web service requires a separate registration and will be granted only in individual cases. Feel free to contact us for further information.

URL:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/PlayRecording/wav
Method: GET
Output: audio/wav

URL:
http://gewiss.uni-leipzig.de:8282/gewiss/webresources/PlayRecording/mp3
Method: GET
Output: audio/mpeg3

Parameters:

name	description	optional	possible values
rec-name	code of recording	no	Ex.: EV_DE_004, PG_PL_090
seg-name	code of segment in transcription	yes	Ex.: T169
start	start of audio segment (seconds)	yes (only in connection with end)	Ex.: 15
end	end of audio segment (seconds)	yes (only in connection with start)	Ex.: 25

Note:
The parameters seg-name, start, and end are available for audio/wav only. audio/mp3 returns the whole communication.

Example (wget):
wget –http-user="USERNAME" –http-password="PASSWORD" –header="Accept: audio/wav" http://gewiss.uni-leipzig.de:8282/gewiss/webresources/PlayRecording/wav?rec-name=EV_DE_004&seg-name=T169

7. List Metadata Types (open access)

Description:
Returns a list of all metadata types and their respective values. In addition, codes of metadata types and values (see below) can be retrieved; with "C" = communication, "S"= speaker. Some metadata types do not have fixed values (e.g. C4).
For a complete list of metadata types, see appendix.

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListMetadata
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListMetadata/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
codes	codes of metadata types & values	yes	true – display codes (default) false – do not display codes

Example:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListMetadata?codes=true

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<metadata-list>
  <meta type="C" code="C1" value="1" name="Teilkorpus">Deutschland</meta>
  <meta type="C" code="C1" value="2"
    name="Teilkorpus">Großbritannien</meta>
  <meta type="C" code="C1" value="3" name="Teilkorpus">Polen</meta>
  <meta type="C" code="C1" value="4" name="Teilkorpus">Bulgarien</meta>
  <meta type="C" code="C1" value="5" name="Teilkorpus">Italien</meta>
  <meta type="C" code="C2" value="1" name="Muttersprachliche
    Kommunikation">ja</meta>
  <meta type="C" code="C2" value="2" name="Muttersprachliche
    Kommunikation">nein</meta>

  <!-- ... abbr. ... -->

  <meta type="C" code="C4" name="Art"/>
  <meta type="C" code="C5" name="Kurzbezeichnung"/>

  <!-- ... abbr. ... -->

</metadata-list>
```

8. List of Annotation Types (open access)

Description:
Returns a list of all annotation types (incl. subtypes for D1, D2, Verweis and Zitat) and resp. codes for the annotation search (see below). For additional information on the annotations, please refer to https://gewiss.uni-leipzig.de/index.php?id=annotations.

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListAnnotations
Method: GET
Output: text/xml

Parameters: none

Example:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/ListAnnotations

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
   <annotation-list>
       <anno code="0" type="Wechsel"/>
       <anno code="1" type="D1">Daten</anno>
       <anno code="1" type="D1">Ende</anno>
       <anno code="1" type="D1">Fazit</anno>
       <anno code="1" type="D1">Makrostruktur</anno>
       <anno code="1" type="D1">Performanz</anno>
       <anno code="1" type="D1">Rueckbezug</anno>
       <anno code="1" type="D1">Sprechhandlung-A</anno>
       <anno code="1" type="D1">Sprechhandlung-F</anno>
       <anno code="1" type="D1">Thema</anno>
       <anno code="1" type="D1">Zeit</anno>
       <anno code="2" type="D2">Anfang</anno>
       <anno code="2" type="D2">Diskussion</anno>
       <anno code="2" type="D2">Rederecht</anno>
       <anno code="2" type="D2">Vorstellung</anno>
       <anno code="3" type="D3"/>
      <anno code="4" type="Situation"/>
      <anno code="5" type="Verweis">Konzept</anno>
      <anno code="5" type="Verweis">Studie</anno>
      <anno code="5" type="Verweis">Publikation</anno>
      <anno code="5" type="Verweis">unspezifisch</anno>
      <anno code="6" type="Zitat">sinn</anno>
      <anno code="6" type="Zitat">woertl</anno>
</annotation-list>
```

9. Simple Search - Text (open access)

Description:
Returns a list of communications containing the search text, and the total number of matches.

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Text
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Text/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
search	search text	no	lower case letters,_,-, and * as placeholder for any sequence of characters within a word (letters,_,-): test test* -> teste,tester they_re
acad-context	academic context	yes	D, UK, PL, BG
lang	language	yes	DEU, ENG, POL, ITA, BUL
lang-competence	language competence	yes	L1, L2
disc-genre	discourse genre	yes	EV, PG, SV

Example:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Text?search=inhaltlich&lang-competence=L1

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<result-list total-occurrences="7">
  <result>
    <data name="communication-name">PG_DE_132</data>
  </result>
  <result>
    <data name="communication-name">PG_DE_136</data>
  </result>

  <!-- ... abbr. ... -->

</result-list>
```

10. Simple Search - Annotations (open access)

Description:
Returns a list of all communications containing a particular annotation and subtype (see also section 8).
The search for annotations can be combined with the text search, by adding the parameter *search* to the url. This returns the respective annotations containing the searched text.

URL:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Anno
Method: GET
Output: text/xml

URL:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Anno/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
anno-type	annotation type	yes	0(default),1,2,3,4 – see appendix
search	search text	yes	lower case letters,_,-, and * as placeholder for any sequence of characters within a word (letters,_,-): test test* -> teste,tester they_re
subtype	subtype	yes	only with annotation types having predefined subtypes, e.g. subtype "Data" of annotation type D1 – see appendix
acad-context	academic context	yes	D, UK, PL, BG
lang	language	yes	DEU*
lang-competence	language competence	yes	L1, L2
disc-genre	discourse genre	yes	EV, PG, SV

*Only the subcorpora in German contain annotations. All these subcorpora are annotated for code-switching (annotation type 0), while only subcorpus DEU_L1_EV is annotated for discourse commentaries (annotation types 1-4) and their respective subtypes (cf. Baur et al. 2014, among others). Subcorpora DEU_L1_EV and DEU_L1_SV are annotated for references and citations (Verweise und Zitate) (cf. Maier et al. 2015).

Example:
http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/SimpleSearch/Anno?anno-type=2&subtype=Diskussion&lang-competence=L1

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<result-list total-occurrences="7">
  <result>
    <data name="communication-name">EV_DE_093</data>
  </result>
  <result>
    <data name="communication-name">PG_UK_063</data>
  </result>

  <!-- ... abbr. ... -->

</result-list>
```

11. Extended Search – Text (access restricted)

Description:
Returns all occurrences of the search text, and additional information on communication and speakers, as well as the total number of matches. Optional parameters allow for retrieval of context and metadata, and filtering and sorting the results.

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text/html
Method: GET
Output: text/html

Parameters:

name	description	optional	possible values
search	search text	no	lower case letters,_,-, and * as placeholder for any sequence of characters within a word (letters,_,-): test test* -> teste,tester they_re
acad-context	academic context	yes	D, UK, PL, BG
lang	language	yes	DEU, ENG, POL, ITA, BUL
lang-competence	language competence	yes	L1, L2
disc-genre	discourse genre	yes	EV, PG, SV
first	first result returned	yes	1-10000, Default:1
last	last result returned	yes	1-10000, Default:10000
context	context returned, no. of characters	yes	integers, Default: 100
show	metadata returned	yes*	codes of metadata types (complete list, see appendix): {C1-C17,S1-S8}
fltr	filtering for specific values of single metadata types	yes*	code of metadata type + IN | EX + value: C1IN1,C1EX1,C1IN2,…;
sort	sorting of results	yes*	Ca,Cd,Sa,Sd; C1a,C1d,C2a,C2d…code of metadata type + "a"|"d"

*each of these parameters may be used repeatedly to combine different values of the same parameter. E.g.,
"PARAMETER1=PARAMETERVALUE1&PARAMETER1=PARAMETERVALUE2&PARAMETER1=PARAMETERVALUE3…"

Search
Search for the text "kurz" using the parameter **search** in the German subcorpora (**lang**=DEU) of discourse genre "oral examination" (**disc-genre**=PG).

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text?search=kurz&lang=DEU&disc-gerne=PG

Specifiy number of results

To return a subset of the total matches, use the parameters **first** and **last** for the first and the last match to be returned, respectively (e.g, first=100, last=199 for returning matches 100 through 199).

Context

Using the parameter **context**, you may retrieve the context of each match. The value of the parameter refers to the numbers of characters to the left and to the right of the actual match (e,g, context=80 for 80 characters to the left and 80 characters to the right of the match).

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text?search=kurz&first=100&last=199&context=80

Show metadata

Use the parameter **show** to display additional metadata on the communication or speaker. For retrieving metadata of types "subcorpus" (C1) and "sex" of the speaker (S2) add: show=C1&show=S2 – for codes and values, see appendix.

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text?search=kurz&show=C1&show=S2

Filtering

The parameter **fltr** lets you filter the results for specific metadata values. The parameter values take the form *Metadata type+IN/EX+Metadata value*. The filter IN returns all matches containing the given metadata value, the command EX returns all matches which do **NOT** contain the given value.

For example, to retrieve all results for which the following holds: "native language communication"="yes" and "subcorpus"≠"Poland", set fltr=C2IN1&fltr=C1EX3.

- fltr=C2IN1 returns all matches subject to the following condition: metadata type C2 ("native language communication") has value 1 ("yes") – for codes and values, see appendix
- fltr=C1EX3 returns all matches subject to the following condition: metadata type C1 ("subcorpus") does not have value 3 ("Poland") – for codes and values, see appendix

Note: C1EX3 returns only results for which C1 has a value (other than "Polen"). Results without a value C1 are neither retrieved by C1EX3 nor by C1IN3.

In other words: Total Results (without filtering) = Results C1IN3 + Results C1EX3 + Results (without a C1 value)

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text?search=kurz&show=C1&show=C2&fltr=C2IN1&fltr=C1EX3

Notes:

- It is possible, but not necessary, to display the metadata types you are filtering for (C1 and C2).
- For metadata types without fixed values (see appendix), a value must be set manually. E.g. all matches with "courses abroad"(S6)="Poland": fltr=S6INPolen.

Sorting

The parameter **sort** lets you sort the results. The parameter values take the form *Sorting condition + Order*; e.g. Ca, Sa, C1d, S3a etc.:

- Sorting conditions are
  - C: sorting for communication codes in alphabetic order (EV_DE_004, PG_DE_047 etc.)
  - S: sorting for speaker codes in alphabetic order (MH_0226, RO_1505 etc.)
  - C1,C2,..S1…: sorting for the values of the respective metadata type in alphabetic order (e.g. C2 sorts alphabetically for values "gemischt"/"ja"/"nein" of type „native language communication" – for codes and values, see appendix)
- Orders are
  - a for ascending
  - d for descending

Thus, sort=Ca sorts for communications codes in ascending order, sort=Sa sorts for speaker codes in ascending order, sort=C1d sorts for "subcorpus" in descending order (C1) etc.
Multiple sorting parameters can be combined and will be processed in sequence from left to right. E.g. sort=Ca&sort=Sa&sort=C1d sorts the results for communication in ascending order, thereafter, results with the same communication code are sorted for speaker code in ascending order; results with the same communication and speaker codes are sorted for "subcorpus" in descending order.

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Text?search=kurz&sort=Ca&sort=Sa&sort=C1d

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<search-result-list total-occurrences="10">
<search-result selected="true" communication="EV_DE_002" speaker="SY_0644">
<locator file="" xpath=""/>
<left-context>depeschen (.) also aus korrespondenzbriefen (.) wo (.) sehr </left-
context>
<match original-match-start="">kurz</match>
<right-context> (.) und sehr knapp (.) über eine schlacht (.) oder ein (.) </right-
context>
<data name="tier">TIE2</data>
</search-result> <meta type="C1" name="Teilkorpus">Deutschland</meta>
<meta type="S2" name="Geschlecht">weiblich</meta>
</search-result>
<meta type="S6" name="Auslandsstudium">USA</meta>

<!-- ... abbr. ... -->

</search-result-list>
```

12. Extended Search – Annotations (access restricted)

Description:
Returns all occurrences of a given annotation, and additional information on communication and speakers, as well as their overall count. Additional parameters allow for retrieval of context and metadata, and filtering and sorting of the results.
The search for annotations can be combined with the text search, by adding the parameter *search* to the url. This returns the respective annotations containing the searched text.

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno
Method: GET
Output: text/xml

URL: http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno/html
Method: GET
Output: text/html
Parameters:

name	description	optional	possible values
anno-type	annotation type	yes	0(default),1,2,3,4 – see appendix
search	search text	yes	lower case letters,_,-, and * as placeholder for any sequence of characters within a word (letters,_,-): test test* -> teste,tester they_re
subtype	subtype	yes	only with annotation types having predefined subtypes, e.g. subtype "Data" of annotation type D1 – see appendix
acad-context	academic context	yes	D, UK, PL, BG
lang	language	yes	DEU*
lang-competence	language competence	yes	L1, L2
disc-genre	discourse genre	yes	EV, PG, SV
first	first result returned	yes	1-10000, Default:1
last	last result returned	yes	1-10000, Default:10000
context	context returned, no. of characters	yes	integers, Default: 60
show	metadata returned	yes**	codes of metadata types (complete list, see appendix): {C1-C17,S1-S8}
fltr	filtering for specific values of single metadata types	yes**	code of metadata type + IN | EX + value: C1IN1,C1EX1,C1IN2,…;
sort	sorting of result	yes**	Ca,Cd,Sa,Sd; C1a,C1d,C2a,C2d…code of metadata type + "a"|"d"

*Only the subcorpora in German contain annotations. All these subcorpora are annotated for code-switching (annotation type 0), while only subcorpus DEU_L1_EV is annotated for discourse commentaries

(annotation types 1-4) and their respective subtypes (cf. Baur et al. 2014, among others). Subcorpora DEU_L1_EV and DEU_L1_SV are annotated for references and citations (Verweise und Zitate) (cf. Maier et al. 2015).
**each of these parameters may be used repeatedly to combine different values of the same parameter. E.g.,
"PARAMETER1=PARAMETERVALUE1&PARAMETER1=PARAMETERVALUE2&PARAMETER1=PARAMETERVALUE3…"

Search
Search for annotation type D2-Discussiom (**anno-type**=2, **subtype**=Diskussion; for codes, see appendix) in German language subcorpora (**lang**=DEU) of discourse genre "oral examination" (**disc-genre**=PG):

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno?anno-type=2&subtype=Diskussion&lang=DEU&disc-gerne=PG

Specifiy number of results
To return a subset of the total matches, use the parameters **first** and **last** for the first and the last match to be returned (e.g, first=100, last=199 for returning matches 100 through 199).

Context
Using the parameter **context**, you may retrieve the context of each match. The value of the parameter refers to the numbers of characters to the left and to the right of the actual match (e,g, context=80 for 80 characters to the left and 80 characters to the right of the match).

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno?anno-type=0&first=100&last=199&context=80

Show metadata
Use the parameter **show** to display additional metadata on the communication or speaker. For retrieving metadata of types "subcorpus" (C1) and "sex" of the speaker (S2) add: show=C1&show=S2 – for codes and values, see appendix.

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno?anno-type=0&show=C1&show=S2

Filtering
The parameter **fltr** lets you filter the results for specific metadata values. The parameter values take the form *Metadata type+IN/EX+Metadata value*. The filter IN returns all matche, containing the given metadata value, the command EX returns all matches which do **NOT** contain the given value.
For example, to retrieve all results for which the following holds: "native language communication"="yes" and "subcorpus"≠"Poland", set fltr=C2IN1&fltr=C1EX3.

- fltr=C2IN1 returns all matches subject to the following condition: metadata type C2 ("native language communication") has value 1 ("yes") – for codes and values, see appendix
- fltr=C1EX3 returns all matches subject to the following condition: metadata type C1 ("subcorpus") does not have value 3 ("Poland") – for codes and values, see appendix

Note: C1EX3 returns only results for which C1 has a value (other than "Polen"). Results without a value C1 are neither retrieved by C1EX3 nor by C1IN3.
In other words: Total Results (without filtering) = Results C1IN3 + Results C1EX3 + Results (without a C1 value)

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno?anno-type=0&show=C1&show=C2&fltr=C2IN1&fltr=C1EX3

Notes:

- It is possible, but not necessary, to display the metadata types you are filtering for (C1 and C2).
- For metadata types without fixed values (see appendix), a value must be set manually. E.g. all matches with "courses abroad"(S6)="Poland": fltr=S6INPolen.

Sorting

The parameter **sort** lets you sort the results. The parameter values take the form *Sorting condition + Order*; e.g. Ca, Sa, C1d, S3a etc.:

- Sorting conditions are
  - C: sorting for communication codes in alphabetic order (EV_DE_004, PG_DE_047 etc.)
  - S: sorting for speaker codes in alphabetic order (MH_0226, RO_1505 etc.)
  - C1,C2,..S1…: sorting for the values of the respective metadata type in alphabetic order (e.g. C2 sorts alphabetically for values "gemischt"/"ja"/"nein" of type „native language communication" – for codes and values, see appendix)
- Orders are
  - a for ascending
  - d for descending

Thus, sort=Ca sorts for communications codes in ascending order, sort=Sa sorts for speaker codes in ascending order, sort=C1d sorts for "subcorpus" in descending order (C1) etc.
Multiple sorting parameters can be combined and will be processed in sequence from left to right. E.g. sort=Ca&sort=Sa&sort=C1d sorts the results for communication in ascending order, thereafter, results with the same communication code are sorted for speaker code in ascending order; results with the same communication and speaker codes are sorted for "subcorpus" in descending order.

http://gewiss.uni-leipzig.de:8282/CorpusQuery/webresources/Search/Anno?anno-type=0&sort=Ca&sort=Sa&sort=C1d

Result:

```
<?xml version="1.0" encoding="UTF-8"?>
<search-result-list total-occurrences="10">
<search-result selected="true" communication="EV_DE_002" speaker="SY_0644">
<locator file="" xpath=""/>
<left-context>depeschen (.) also aus korrespondenzbriefen (.) wo (.) sehr <match
original-match-start="">kurz</match>
<right-context> (.) und sehr knapp (.) über eine schlacht (.) oder ein (.) </right-
context>
<data name="tier">TIE2</data>
```

```
</search-result> <meta type="C1" name="Teilkorpus">Deutschland</meta>
<meta type="S2" name="Geschlecht">weiblich</meta>
</search-result>
<meta type="S6" name="Auslandsstudium">USA</meta>


<!-- ... abbr. ... -->

</search-result-list>
```

Remark:
Due to technical reasons, some matches may contain only the start of the respective annotations. In this case, the annotation extends into the right context of the match.

References

Baur, Benedikt/Gräfe, Karen/Lange, Daisy/Schmidt, Julia (2014): *Dokumentation zur Annotation der Diskurskommentierungen im GeWiss-Projekt*, available at: https://gewiss.uni-leipzig.de/index.php?id=annotations_discourse.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012). "The GeWiss Corpus: Comparing Spoken Academic German, English and Polish", in: Schmidt, Thomas/Wörner, Kai (Hg.): *Multilingual corpora and multilingual corpus analysis*. Amsterdam: Benjamins. (= Hamburg Studies in Multilingualism).

Gräfe, Karen/Lange, Daisy/Sieradz, Magda/Meißner, Cordula/Slavcheva, Adriana/Stoppel, David (2015): *Handbuch zum Korpus*, available at: https://gewiss.uni-leipzig.de/index.php?id=help

Lange, Daisy/Slavcheva, Adriana/Rogozińska, Marta/Morton, Ralph (2014): "GAT 2 als Transkriptionssystem für multilinguale Sprachdaten? Zur Adaption der Notationskonventionen im Rahmen des Projekts GeWiss", in: Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (Hgg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation)*, 39-55.

Maier, Elisabeth/Sadowski, Sabrina/Schmidt, Julia (2015): *Dokumentation zur Annotation der Verweise und Zitate*, abrufbar unter: https://gewiss.uni-leipzig.de/index.php?id=annotations_citations

Meißner, Cordula/Jettka, Daniel/Fandrych, Christian (2013): CLARIN-KP-GeWiss: Das zweite Kurationsprojekt der F-AG 1 Deutsche Philologie, in: *CLARIN-D-Newsletter* 4: 3-8, available at: http://de.clarin.eu/images/newsletter/CLARIN-D-Newsletter-2013-4.pdf

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian/Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stuckenbrock, Anja/Uhmann, Susanne (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2), in: *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion [Online]*, 10, 353–402, available at: http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf

Appendix

List of Subcorpora

```
DEU_L1_EV
DEU_L1_SV
DEU_L2_PL_PG
DEU_L2_UK_SV
ENG_L1_SV
ENG_L2_SV
POL_PG
DEU_L1_PG
DEU_L2_D_PG
DEU_L2_PL_SV
DEU_L2_UK_EV
ENG_L1_EV
ENG_L2_EV
DEU_L2_D_SV
DEU_L2_PL_EV
DEU_L2_UK_PG
ENG_L1_PG
ENG_L2_PG
POL_EV
POL_SV
ITA_L1_EV
DEU_L2_BG_SV
```

List of Annotations with Codes

```
<anno code="0" type="Wechsel"/>                    - code-switching
<anno code="1" type="D1">Daten</anno>              - data
<anno code="1" type="D1">Ende</anno>               - ending
<anno code="1" type="D1">Fazit</anno>              - conclusion
<anno code="1" type="D1">Makrostruktur</anno>      - macro structure
<anno code="1" type="D1">Performanz</anno>         - performance
<anno code="1" type="D1">Rueckbezug</anno>         - back reference
<anno code="1" type="D1">Sprechhandlung-A</anno>   - speech act-A
<anno code="1" type="D1">Sprechhandlung-F</anno>   - speech act-F
<anno code="1" type="D1">Thema</anno>              - topic
<anno code="1" type="D1">Zeit</anno>               - time
<anno code="2" type="D2">Anfang</anno>             - beginning
<anno code="2" type="D2">Diskussion</anno>         - discussion
<anno code="2" type="D2">Rederecht</anno>          - right to debate
<anno code="2" type="D2">Vorstellung</anno>        - introduction
<anno code="3" type="D3"/>
<anno code="4" type="Situation"/>                  - situation
<anno code="5" type="Verweis">Konzept</anno>       - concept
<anno code="5" type="Verweis">Studie</anno>        - study
<anno code="5" type="Verweis">Publikation</anno>   - publication
<anno code="5" type="Verweis">unspezifisch</anno>  - non-specific
<anno code="6" type="Zitat">sinn</anno>            - verbal
<anno code="6" type="Zitat">woertl</anno>          - paraphrased
```

List of Metadata Types with Codes

```
<meta type="C" code="C1" value="1" name="Teilkorpus">Deutschland</meta> - subcorpus/Germany
<meta type="C" code="C1" value="2" name="Teilkorpus">Großbritannien</meta> - UK
<meta type="C" code="C1" value="3" name="Teilkorpus">Polen</meta> - Poland
<meta type="C" code="C1" value="4" name="Teilkorpus">Bulgarien</meta> - Bulgaria
<meta type="C" code="C1" value="5" name="Teilkorpus">Italien</meta> - Italy
<meta type="C" code="C2" value="1" name="Muttersprachliche Kommunikation">ja</meta> - native
language communication/yes
<meta type="C" code="C2" value="2" name="Muttersprachliche Kommunikation">nein</meta> - no
<meta type="C" code="C2" value="3" name="Muttersprachliche Kommunikation">gemischt</meta> -
mixed
<meta type="C" code="C3" value="1" name="Genre">Expertenvortrag</meta> - genre/expert talk
<meta type="C" code="C3" value="2" name="Genre">Studentischer Vortrag</meta> - student talk
<meta type="C" code="C3" value="3" name="Genre">Prüfungsgespräch</meta> - oral examination
```

<meta type="C" code="C4" name="Art"/> - type
<meta type="C" code="C5" name="Kurzbezeichnung"/> - short description
<meta type="C" code="C6" name="Zusammenfassung"/> - résumé
<meta type="C" code="C7" name="Jahr"/> - year
<meta type="C" code="C8" name="Land"/> - country
<meta type="C" code="C9" name="Institution"/> - institution
<meta type="C" code="C10" name="Dauer"/> - duration
<meta type="C" code="C11" value="1" name="Raum">Computerpool</meta> - location/computer lab
<meta type="C" code="C11" value="2" name="Raum">Seminarraum</meta> - lecture room
<meta type="C" code="C11" value="3" name="Raum">Hörsaal</meta> - lecture hall
<meta type="C" code="C11" value="4" name="Raum">Konferenzraum</meta> - conference room
<meta type="C" code="C11" value="5" name="Raum">Büro</meta> - office
<meta type="C" code="C12" value="1" name="Basissprache">deu</meta> - base language
<meta type="C" code="C12" value="2" name="Basissprache">eng</meta>
<meta type="C" code="C12" value="3" name="Basissprache">pol</meta>
<meta type="C" code="C12" value="4" name="Basissprache">ita</meta>
<meta type="C" code="C13" value="1" name="Grad der Mündlichkeit">frei gesprochen</meta> - degree of orality - total
<meta type="C" code="C13" value="2" name="Grad der Mündlichkeit">zum Teil abgelesen</meta> - read in parts
<meta type="C" code="C13" value="3" name="Grad der Mündlichkeit">vollständig abgelesen</meta> - read completely
<meta type="C" code="C13" value="4" name="Grad der Mündlichkeit">scheint vorformuliert und auswendig gelernt</meta> - pre-formulated and learned by heart
<meta type="C" code="C14" value="1" name="Wechsel in andere Sprache(n)">eng</meta> - code switching to language:
<meta type="C" code="C14" value="2" name="Wechsel in andere Sprache(n)">pol</meta>
<meta type="C" code="C14" value="3" name="Wechsel in andere Sprache(n)">deu</meta>
<meta type="C" code="C14" value="4" name="Wechsel in andere Sprache(n)">bul</meta>
<meta type="C" code="C15" value="1" name="Verwendete Medien">Präsentation</meta> - media used/presentation
<meta type="C" code="C15" value="2" name="Verwendete Medien">Handout</meta> - handout
<meta type="C" code="C15" value="3" name="Verwendete Medien">Skript</meta> - lecture notes
<meta type="C" code="C15" value="4" name="Verwendete Medien">Thesenpapier</meta> - discussion paper
<meta type="C" code="C15" value="5" name="Verwendete Medien">Prüfungsfragen</meta> - exam questions
<meta type="C" code="C15" value="6" name="Verwendete Medien">OHP-Folien</meta> - OHP slides
<meta type="C" code="C15" value="7" name="Verwendete Medien">Audiobeispiel</meta> - audio
<meta type="C" code="C15" value="8" name="Verwendete Medien">Videobeispiel</meta> - video
<meta type="C" code="C15" value="9" name="Verwendete Medien">Analysematerial</meta> - analysed materials
<meta type="C" code="C15" value="10" name="Verwendete Medien">Mindmaps</meta> - mindmaps
<meta type="C" code="C15" value="11" name="Verwendete Medien">Essay</meta> - essay
<meta type="C" code="C15" value="12" name="Verwendete Medien">Internet</meta> - internet
<meta type="C" code="C15" value="13" name="Verwendete Medien">Plakat</meta> - poster
<meta type="C" code="C15" value="14" name="Verwendete Medien">Tafel</meta> - blackboard
<meta type="C" code="C15" value="15" name="Verwendete Medien">keine</meta> - none
<meta type="C" code="C16" name="Anzahl der Teilnehmer"/> - number of participants
<meta type="C" code="C17" name="Beziehung der Sprecher zueinander"/> - speakers’ relations
<meta type="S" code="S1" name="Alter"/> - age
<meta type="S" code="S2" value="1" name="Geschlecht">weiblich</meta> - sex/female
<meta type="S" code="S2" value="2" name="Geschlecht">männlich</meta> - sex/male
<meta type="S" code="S3" value="1" name="Rollen">Vortragender</meta> - role(s)/speaker
<meta type="S" code="S3" value="2" name="Rollen">Prüfling</meta> - examinee
<meta type="S" code="S3" value="3" name="Rollen">Prüfer</meta> - examiner
<meta type="S" code="S3" value="4" name="Rollen">Seminarleiter</meta> - lecturer
<meta type="S" code="S4" value="1" name="Erstsprache">ara</meta> - L1
<meta type="S" code="S4" value="2" name="Erstsprache">ces</meta>
<meta type="S" code="S4" value="3" name="Erstsprache">chi</meta>
<meta type="S" code="S4" value="4" name="Erstsprache">cmn</meta>
<meta type="S" code="S4" value="5" name="Erstsprache">dan</meta>
<meta type="S" code="S4" value="6" name="Erstsprache">deu</meta>
<meta type="S" code="S4" value="7" name="Erstsprache">eng</meta>
<meta type="S" code="S4" value="8" name="Erstsprache">fon</meta>
<meta type="S" code="S4" value="9" name="Erstsprache">fra</meta>
<meta type="S" code="S4" value="10" name="Erstsprache">idd</meta>
<meta type="S" code="S4" value="11" name="Erstsprache">jpn</meta>
<meta type="S" code="S4" value="12" name="Erstsprache">k.A.</meta>
<meta type="S" code="S4" value="13" name="Erstsprache">ndl</meta>
<meta type="S" code="S4" value="14" name="Erstsprache">pol</meta>
<meta type="S" code="S4" value="15" name="Erstsprache">por</meta>
<meta type="S" code="S4" value="16" name="Erstsprache">ron</meta>
<meta type="S" code="S4" value="17" name="Erstsprache">rus</meta>
<meta type="S" code="S4" value="18" name="Erstsprache">slv</meta>
<meta type="S" code="S4" value="19" name="Erstsprache">spa</meta>
<meta type="S" code="S4" value="20" name="Erstsprache">swe</meta>
<meta type="S" code="S4" value="21" name="Erstsprache">tgl</meta>

<meta type="S" code="S4" value="22" name="Erstsprache">tur</meta>
<meta type="S" code="S4" value="23" name="Erstsprache">ukr</meta>
<meta type="S" code="S4" value="24" name="Erstsprache">zho</meta>
<meta type="S" code="S5" value="1" name="L2">afr</meta> - L2
<meta type="S" code="S5" value="2" name="L2">amh</meta>
<meta type="S" code="S5" value="3" name="L2">ara</meta>
<meta type="S" code="S5" value="4" name="L2">arb</meta>
<meta type="S" code="S5" value="5" name="L2">ben</meta>
<meta type="S" code="S5" value="6" name="L2">ces</meta>
<meta type="S" code="S5" value="7" name="L2">cmn</meta>
<meta type="S" code="S5" value="8" name="L2">dan</meta>
<meta type="S" code="S5" value="9" name="L2">deu</meta>
<meta type="S" code="S5" value="10" name="L2">eng</meta>
<meta type="S" code="S5" value="11" name="L2">est</meta>
<meta type="S" code="S5" value="12" name="L2">eth</meta>
<meta type="S" code="S5" value="13" name="L2">fas</meta>
<meta type="S" code="S5" value="14" name="L2">fra</meta>
<meta type="S" code="S5" value="15" name="L2">hbo</meta>
<meta type="S" code="S5" value="16" name="L2">heb</meta>
<meta type="S" code="S5" value="17" name="L2">hrv</meta>
<meta type="S" code="S5" value="18" name="L2">hun</meta>
<meta type="S" code="S5" value="19" name="L2">ind</meta>
<meta type="S" code="S5" value="20" name="L2">isl</meta>
<meta type="S" code="S5" value="21" name="L2">ita</meta>
<meta type="S" code="S5" value="22" name="L2">jpn</meta>
<meta type="S" code="S5" value="23" name="L2">k.A.</meta>
<meta type="S" code="S5" value="24" name="L2">lat</meta>
<meta type="S" code="S5" value="25" name="L2">lit</meta>
<meta type="S" code="S5" value="26" name="L2">ndl</meta>
<meta type="S" code="S5" value="27" name="L2">nld</meta>
<meta type="S" code="S5" value="28" name="L2">nor</meta>
<meta type="S" code="S5" value="29" name="L2">pol</meta>
<meta type="S" code="S5" value="30" name="L2">por</meta>
<meta type="S" code="S5" value="31" name="L2">pus</meta>
<meta type="S" code="S5" value="32" name="L2">rus</meta>
<meta type="S" code="S5" value="33" name="L2">slk</meta>
<meta type="S" code="S5" value="34" name="L2">spa</meta>
<meta type="S" code="S5" value="35" name="L2">swe</meta>
<meta type="S" code="S5" value="36" name="L2">swh</meta>
<meta type="S" code="S5" value="37" name="L2">urd</meta>
<meta type="S" code="S5" value="38" name="L2">wen</meta>
<meta type="S" code="S5" value="39" name="L2">zho</meta>
<meta type="S" code="S6" name="Schulbildung"/> - education
<meta type="S" code="S7" name="Auslandsstudium"/> - study abroad
<meta type="S" code="S8" name="Auslandsaufenthalt"/> - stay abroad