Building Your Query Step by Step:
A Query Wizard for the MY DGS – ANNIS Portal of the DGS Corpus

Amy Isard
Institute of German Sign Language and Communication of the Deaf
and House of Computing and Data Science
University of Hamburg, Germany
amy.isard@uni-hamburg.de

Abstract

MY DGS – ANNIS makes the Public DGS Corpus available through the corpus query and visualization tool ANNIS. Due to the complex nature of the corpus, composing queries for advanced research questions can quickly become increasingly complicated. We present a Query Wizard which assists users in building valid queries for *MY DGS – ANNIS*. Complex queries are built up from smaller blocks, which can be linked to each other through context-sensitive connections. Blocks provide options specific to a given annotation tier and dynamically lead users through their construction while preventing the creation of invalid queries. Once completed, queries can be opened directly in *MY DGS – ANNIS*.

Keywords: German Sign Language (DGS), corpus query tool, ANNIS, query wizard

1. Introduction

In 2022 the DGS-Korpus project introduced *MY DGS – ANNIS* (Isard and Konrad, 2022), a third portal to provide access to release 3 of the Public DGS Corpus (Hanke et al., 2020). ANNIS (Krause and Zeldes, 2016) is a corpus query and visualization tool which allows corpus queries written in the ANNIS Query Language AQL[1] to be performed over multiple annotation tiers and corpus metadata. Our interface allows researchers to search either the German or the English version of the Public DGS Corpus.

MY DGS – ANNIS has enabled sign language researchers to make complicated queries over the Public DGS Corpus, but the combination of multiple annotation and metadata tiers with complex glossing conventions on the one hand, and AQL syntax on the other, means that novice users have not always found it easy to create valid queries which exactly match what they were searching for. The ANNIS interface contains a general-purpose Query Builder tool, but the dyadic and sign-based nature of our data (see Section 2.1) necessitates a more customised approach.

Several corpus projects which make their data available through ANNIS have provided "simple search" interfaces for their ANNIS instances (Dipper, 2015), including the Reference Corpus Middle Low German/Low Rhenish (1200–1650)[2] and the Reference Corpus of Middle High German (1050–1350)[3]. Inspired by these we decided to create a "simple search" interface for the Public DGS Corpus: the *MY DGS – ANNIS* Query Wizard.

Our Query Wizard allows users to create complex queries out of smaller building blocks by creating connections between them, and uses visual elements to make the connections between the blocks and the resulting AQL query easier to understand. We hope that this will help users to learn about the structure of AQL queries, so that if their needs surpass the scope of the Query Wizard, they will be ready to manually refine queries in *MY DGS – ANNIS*.

The Query Wizard ensures that only valid queries are generated and makes query building easier in a number of ways:

- Users select annotation and metadata tiers from a comprehensive list, ensuring only valid tiers are involved and avoiding issues like spelling errors.
- Instead of writing complex regular expressions to refine search to only certain tokens, users can compose these expressions using context-sensitive check boxes.
- Connections within tiers can be added without knowledge of the exact syntax necessary.

This article introduces the Query Wizard and explains how it integrates with *MY DGS – ANNIS*. In Section 2 we describe the Public DGS Corpus data available through *MY DGS – ANNIS*, with the annotations and metadata in Section 2.1, the ANNIS Query Language (AQL) in Section 2.2 and the *MY*

[1] http://korpling.github.io/ANNIS/4.11/user-guide/aql/index.html

[2] https://www.slm.uni-hamburg.de/en/ren/korpus/datenzugang/einfache-suche.html

[3] https://www.linguistics.rub.de/rem/access/simplesearch.en.html

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 131–139
25 May 2024.

DGS – ANNIS interface in Section 2.3. In Section 3 we describe the interface and usage of the Query Wizard, and in Section 4 we show how some example queries can be built up. Section 5 contains conclusions and describes further features which we intend to add to the Query Wizard in future.

2. MY DGS – ANNIS

2.1. Annotations and Metadata

MY DGS – ANNIS provides datasets representing both release 3 and release 4 of the Public DGS Corpus, with two versions of each dataset, one each for the English and German versions of its annotations[4].

Table 1 lists the main annotation tiers with a brief description of their content. Each element in a tier contains text which can be searched; in the case of translations and mouthings the text is fairly simple, and the HamNoSys tier can be searched by inputting HamNoSys characters directly, which can be done using the HamNoSys editor.[5]

For Gloss and GlossType tiers a special syntax is used which makes search more complex. In these tiers, each token is represented by a type gloss. Each type gloss contains a gloss word, one or two digits which denote different lexical variants, and an optional letter denoting phonological variants. Types that denote form without specifying meaning (i.e. they are supertypes rather than subtypes) are indicated by the caret character (^). In the Gloss tier, an asterisk (*) indicates that a token gloss diverges in some way from the citation form of the type[6]. We provide one gloss tier for each participant to enable collocation searches within a tier (for an example see Section 4.3). Signs in DGS may be one- or two-handed, and it is possible for each hand to articulate a different sign, so when these complex signs occur, we combine the two glosses into a single token, separated by "||". For example, the token $INDEX1* || CAT1B* indicates that the participant simultaneously signed $INDEX1* with their right hand and CAT1B* with their left hand.

Table 2 shows the eight types of metadata included in *MY DGS – ANNIS*, six of which are available for each transcript, and three for each individual participant.

2.2. Annis Query Language (AQL)

Using AQL it is possible to query just one annotation tier or to make arbitrarily complex queries which refer to multiple annotation and metadata tiers. *MY DGS – ANNIS* provides a number of simple examples which users can use as a basis for creating their own queries. However, these cannot cover all possible combinations, and users have not always found it easy to work from these examples to create the queries which they need for their research. The main query types used in *MY DGS – ANNIS* are:

- regular expression search of the text associated with an element
- links between items in different tiers
- collocation distances between items in the same tier
- metadata

Each item in an AQL query must be linked to at least one other item. To facilitate this, each query item is automatically assigned a sequential number which can be used to refer to it later in the query. For example in Query 1, Gloss and English are connected using identifiers automatically assigned to them: #1 refers to the Gloss item and #2 to English. The identifiers can also be explicitly assigned, and we describe this process in section Section 4.1.

(1) Gloss=/CAT.*/ & English=/.* [Cc]at .*/
& #1 ->ident #2

Collocation distances are expressed using the dot (.) or caret (^) operators, followed by the tier name, then optionally by two numbers which specify the minimum and maximum distances. An example can be seen in Query 8.

Some examples of AQL searches in *MY DGS – ANNIS* can be found in Isard and Konrad (2022), and the full AQL manual is available online[7]. In Section 4 we show how the Query Wizard allows users to build up complex queries from smaller building blocks without the need to know the details of AQL syntax.

2.3. ANNIS Interface

Queries created in the Query Wizard can be opened directly in *MY DGS – ANNIS* (see Section 3). Figure 1 shows the *MY DGS – ANNIS* interface with the AQL query input window on the top left and the query results on the right for Gloss tokens with the gloss name CAT (see Section 4 for a discussion of this query). Each result contains

[4] Mouthings are provided in German for both versions, as they relate directly to articulation of German words and are therefore not suited for translation.

[5] https://www.sign-lang.uni-hamburg.de/hamnosys/input/

[6] Further details of the Public DGS Corpus annotation conventions can be found in Konrad et al. (2022).

[7] https://korpling.github.io/ANNIS/4.11/user-guide/aql

Annotation	Description
Gloss	subtypes or types used to lemmatize tokens
GlossType	parent types
HamNoSys	HamNoSys notations of type citation forms
Mouth	mouthings or mouth gestures
Translation	for each utterance

Table 1: Annotation tiers in *MY DGS – ANNIS*

Metadata	Description	Refers to
TranscriptId	the unique identifier for the transcript	Whole Transcript
Region	where the transcript was recorded	
RegionCode	a shorter code for the region	
Date	date of the recording	
Theme	the task given to the participants for this transcript	
Keywords	a list of the topics discussed in this transcript	
Name	the unique (anonymous) identifier for each participant	Participant
AgeGroup	one of a set of four age categories	
Gender	the gender declared by each participant at the time of recording	

Table 2: Metadata included in *MY DGS – ANNIS*

three tabs which can be independently opened or closed; the first shows the five visible annotation tiers, with query results highlighted in red, the second the video for the transcript which can be played by clicking on any token or on the play button, and the third shows clickable links to another corpus portal, *MY DGS – annotated* (Konrad et al., 2024).

Figure 2 shows a view where we have zoomed in on the query result window, showing one match for the gloss CAT1A*, highlighted in red. GlossType and HamNoSys are also highlighted as they are directly linked to Gloss as alternative representations of the same gloss, while Mouth and English are independent tiers whose tokens can have different durations from the Gloss tokens. Unlike in Figure 1, the links tab has been opened in addition to the annotation and video tabs, providing links to the *MY DGS – annotated* Viewer and list of sign types.

3. The Query Wizard

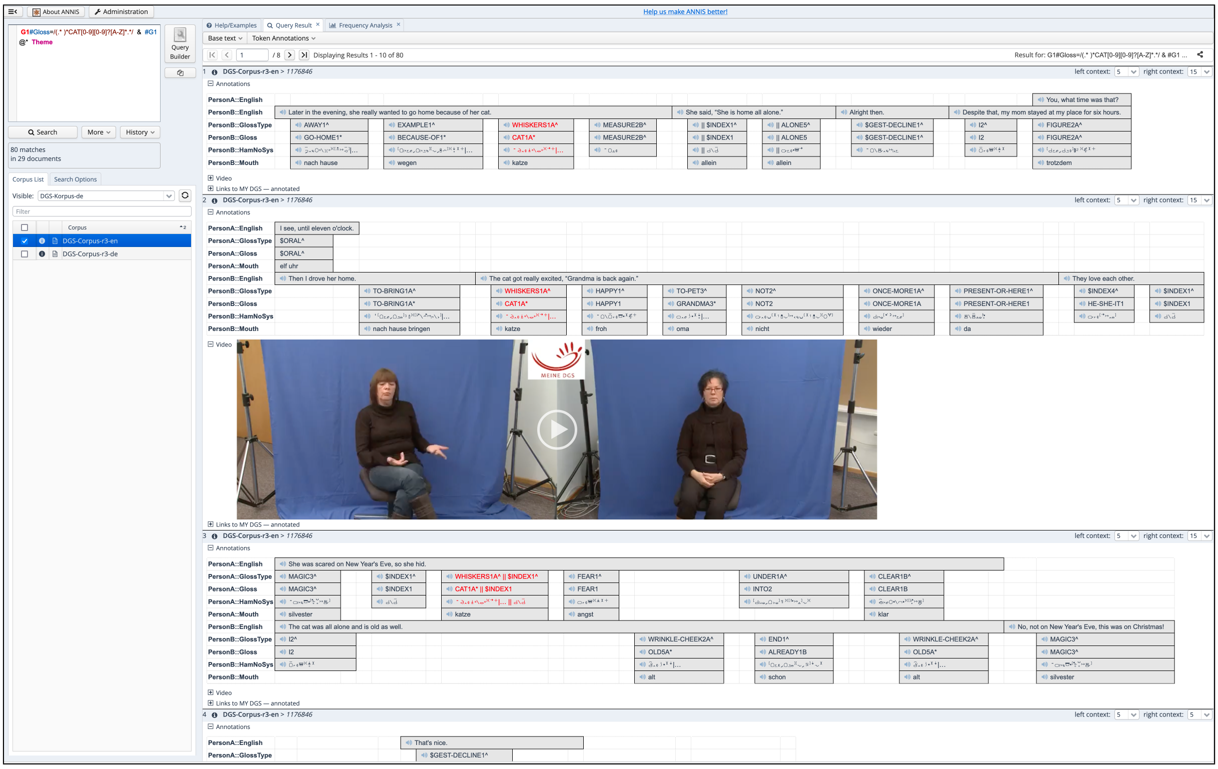
The Query Wizard interface is a web application developed by us and written in JavaScript, that allows users to create a query by creating and linking smaller building blocks. It is available in English for creating queries for the English version of *MY DGS – ANNIS* and in German for the German version of the corpus. All examples in this article are shown for the English interface and corpus.

A user can create a block for any of the annotation tiers, and the search can then be refined by the addition of search text. The options available for the text search depend on the tier selected, with the Gloss and GlossType tiers having the most additional options due to their more complex syntax as described in Section 2.1. Once a query has been generated, the user can click a button to open the query directly in *MY DGS – ANNIS*.

Figure 3 shows the initial state of the Query Wizard interface. The options available are to add a new block for an annotation tier or for a chosen metadata type. At this point, there is nothing displayed in the AQL query box at the top, and the button for opening the query in *MY DGS – ANNIS* is therefore greyed out. There is a button to create connections between elements and a display for the list of connections between annotation elements, but both are empty, as no elements have yet been created.

When the user has selected a tier and clicked the add button, a new block appears, where they can refine the search as shown in Figure 4. This can be done by adding text in the search box, and if desired, constraining the search with further options. If they want to find glosses with a particular gloss name, they can enter free text in the search box, and constrain whether the gloss name should exactly match the text entered, start with the text, or contain the text.

Each annotation or metadata block can be temporarily excluded from the query by unchecking the



G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & #G1 @* Theme

80 matches in 29 documents

Visible: DGS-Korpus-de

	Corpus
☑	DGS-Corpus-r3-en
□	DGS-Corpus-r3-de

Displaying Results 1 - 10 of 80 — Result for: G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & #G1 ...

1 DGS-Corpus-r3-en > *1176846* — left context: 5 right context: 15

PersonA::English									You, what time was that?
PersonB::English	Later in the evening, she really wanted to go home because of her cat.				She said, "She is home all alone."		Alright then.	Despite that, my mom stayed at my place for six hours.	
PersonB::GlossType	AWAY1^	EXAMPLE1^	WHISKERS1A^	MEASURE2B^	|| $INDEX1^	|| ALONE5^	$GEST-DECLINE1^	I2^	FIGURE2A^
PersonB::Gloss	GO-HOME1*	BECAUSE-OF1*	CAT1A*	MEASURE2B^	|| $INDEX1	|| ALONE5	$GEST-DECLINE1^	I2	FIGURE2A^
PersonB::HamNoSys	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PersonB::Mouth	nach hause	wegen	katze		allein	allein			trotzdem

2 DGS-Corpus-r3-en > *1176846* — left context: 5 right context: 15

PersonA::English	I see, until eleven o'clock.								
PersonA::GlossType	$ORAL^								
PersonA::Gloss	$ORAL^								
PersonA::Mouth	elf uhr								
PersonB::English	Then I drove her home.	The cat got really excited, "Grandma is back again."						They love each other.	
PersonB::GlossType	TO-BRING1A^	WHISKERS1A^	HAPPY1^	TO-PET3^	NOT2^	ONCE-MORE1A^	PRESENT-OR-HERE1^	$INDEX4^	$INDEX1^
PersonB::Gloss	TO-BRING1A*	CAT1A*	HAPPY1	GRANDMA3*	NOT2	ONCE-MORE1A	PRESENT-OR-HERE1	HE-SHE-IT1	$INDEX1
PersonB::HamNoSys	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PersonB::Mouth	nach hause bringen	katze	froh	oma	nicht	wieder	da		

Video

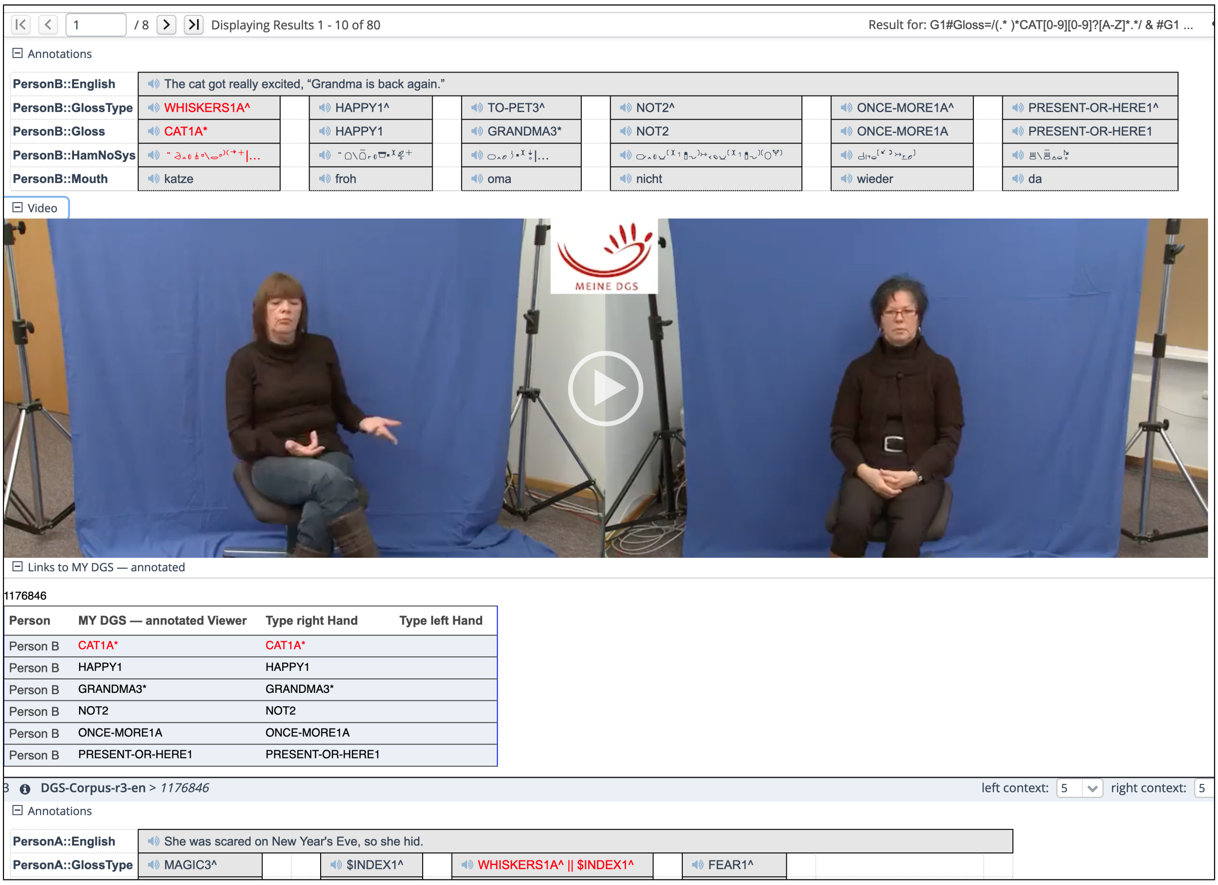
3 DGS-Corpus-r3-en > *1176846* — left context: 5 right context: 15

PersonA::English	She was scared on New Year's Eve, so she hid.						
PersonA::GlossType	MAGIC3^	$INDEX1^	WHISKERS1A^ || $INDEX1^	FEAR1^	UNDER1A^	CLEAR1B^	
PersonA::Gloss	MAGIC3^	$INDEX1	CAT1A* || $INDEX1	FEAR1	INTO2	CLEAR1B	
PersonA::HamNoSys	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
PersonA::Mouth	silvester		katze	angst		klar	
PersonB::English	The cat was all alone and is old as well.					No, not on New Year's Eve, this was on Christmas!	
PersonB::GlossType	I2^	WRINKLE-CHEEK2A^	END1^	WRINKLE-CHEEK2A^	MAGIC3^		
PersonB::Gloss	I2	OLD5A*	ALREADY1B	OLD5A*	MAGIC3^		
PersonB::HamNoSys	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		
PersonB::Mouth		alt	schon	alt	silvester		

4 DGS-Corpus-r3-en > *1176846* — left context: 5 right context: 5

PersonA::English	That's nice.
PersonA::GlossType	$GEST-DECLINE1^

Figure 1: *MY DGS – ANNIS* showing the query results for tokens with gloss name CAT and the metadata "Theme".



/ 8 Displaying Results 1 - 10 of 80 — Result for: G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & #G1 ...

PersonB::English	The cat got really excited, "Grandma is back again."					
PersonB::GlossType	WHISKERS1A^	HAPPY1^	TO-PET3^	NOT2^	ONCE-MORE1A^	PRESENT-OR-HERE1^
PersonB::Gloss	CAT1A*	HAPPY1	GRANDMA3*	NOT2	ONCE-MORE1A	PRESENT-OR-HERE1
PersonB::HamNoSys	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PersonB::Mouth	katze	froh	oma	nicht	wieder	da

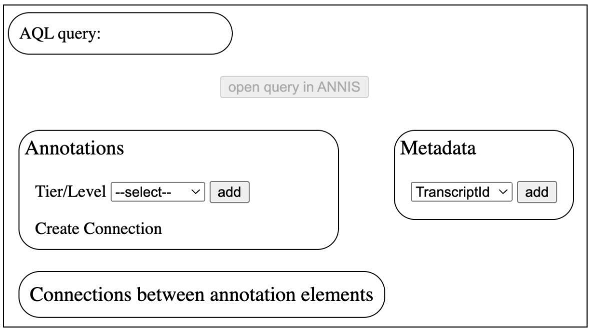
1176846

Person	MY DGS — annotated Viewer	Type right Hand	Type left Hand
Person B	CAT1A*	CAT1A*	
Person B	HAPPY1	HAPPY1	
Person B	GRANDMA3*	GRANDMA3*	
Person B	NOT2	NOT2	
Person B	ONCE-MORE1A	ONCE-MORE1A	
Person B	PRESENT-OR-HERE1	PRESENT-OR-HERE1	

3 DGS-Corpus-r3-en > *1176846* — left context: 5 right context: 5

PersonA::English	She was scared on New Year's Eve, so she hid.			
PersonA::GlossType	MAGIC3^	$INDEX1^	WHISKERS1A^ || $INDEX1^	FEAR1^

Figure 2: Zoomed-in view of Figure 1 showing one specific result with all tabs expanded.



AQL query:

open query in ANNIS

Annotations

Tier/Level --select-- add

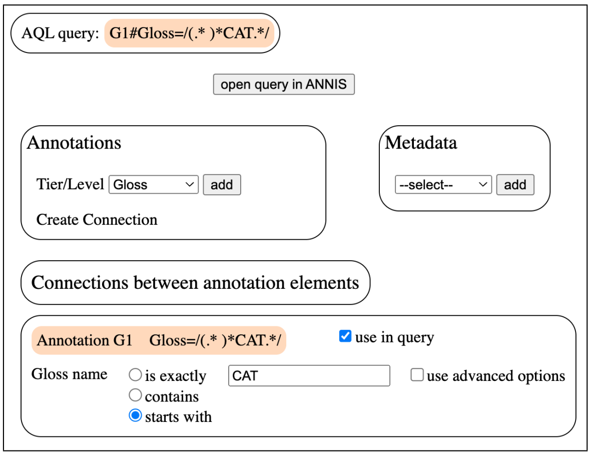
Create Connection

Metadata

TranscriptId add

Connections between annotation elements

Figure 3: The start interface for the Query Wizard



AQL query: G1#Gloss=/(.*)*CAT.*/

open query in ANNIS

Annotations

Tier/Level Gloss add

Create Connection

Metadata

--select-- add

Connections between annotation elements

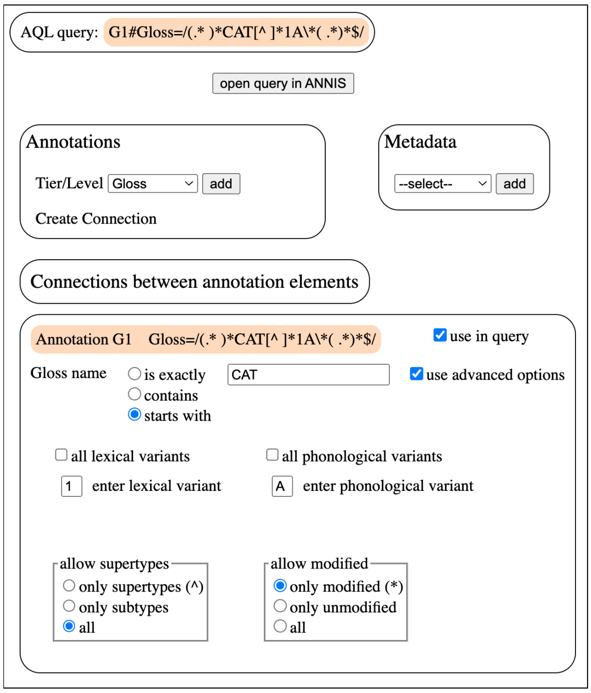
Annotation G1 Gloss=/(.*)*CAT.*/

☑ use in query

Gloss name ○ is exactly ○ contains ● starts with CAT

□ use advanced options

Figure 4: Interface with a Gloss block where the gloss name starts with the string CAT



AQL query: G1#Gloss=/(.*)*CAT[^]*1A*(.*)*$/

open query in ANNIS

Annotations

Tier/Level Gloss add

Create Connection

Metadata

--select-- add

Connections between annotation elements

Annotation G1 Gloss=/(.*)*CAT[^]*1A*(.*)*$/

☑ use in query

Gloss name ○ is exactly ○ contains ● starts with CAT

☑ use advanced options

□ all lexical variants

1 enter lexical variant

□ all phonological variants

A enter phonological variant

allow supertypes

○ only supertypes (^)

○ only subtypes

● all

allow modified

● only modified (*)

○ only unmodified

○ all

Figure 5: Interface with a Gloss block where the gloss name starts with the string CAT, with advanced options selected

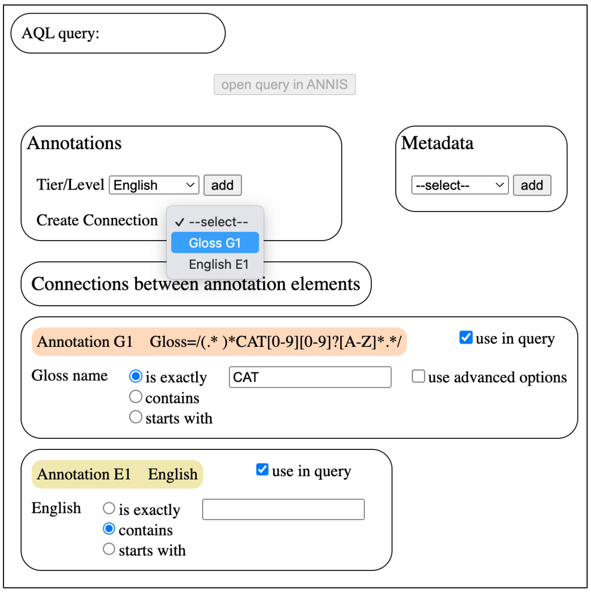
"use in query" checkbox; if the box is later checked again, all search parameters previously entered are still active. As edits are carried out in a block, the AQL query display at the top of the interface changes accordingly.

We mentioned in Section 2.2 that each item in an AQL query is automatically assigned a sequential number which can be used to refer to it later in the query. It is also possible to explicitly assign an identifier to each item in a query, and the Query Wizard does this, giving each item a code which starts with a letter representing the annotation tier (G for Gloss, GT for GlossType and so on) and a number which is incremented every time an item from the same tier is created. In Figure 4 there is a single Gloss item so it receives the identifier G1. This is used to refer to it in the AQL query, but also every time the item is referenced elsewhere in the interface.

Figure 5 shows the search block from Figure 4 with the "use advanced options" checkbox selected. When this checkbox is first selected, the checkboxes for "all lexical variants" and "all phonological variants" are selected, and the "allow supertypes" and "allow modified" checkboxes are set to "all". To avoid the creation of invalid searches, the lexical variants box only becomes active when the user has entered some text into the search box, and the phonological variants box only becomes active when a lexical variant has been entered. In Figure 5 the search is restricted to lexical variant 1, phonological variant A and only tokens which diverge from the citation form.

When a second annotation block is created, it is assigned a different colour in order to help the user to identify which part of the AQL query comes from which block. The main AQL query display temporarily becomes blank, because all annotation items in a valid AQL query must be linked to one another in some way. If there are annotation items which are not yet connected, a dropdown list of the items becomes available, using the identifiers which have been assigned by the Query Wizard. There are two kinds of connections: links between tokens which occur at the *same time* on different tiers, and collocation distances between tokens on the same tier. Collocation searches can find tokens *before* or *after* a token on the same tier, or permit both directions with the option *near*. In addition, it is possible to constrain the collocation distance with minimum and maximum values.

Figure 6 shows how the creation of a Gloss block and an English block has made available a dropdown menu to connect the two. Once a connection has been configured, it can be added and then appears in the list of connections and is integrated into the main AQL query, as can be seen in Figure 7.



AQL query:

open query in ANNIS

Annotations

Tier/Level English ▾ add

Create Connection ✓ --select-- | Gloss G1 | English E1

Metadata

--select-- ▾ add

Connections between annotation elements

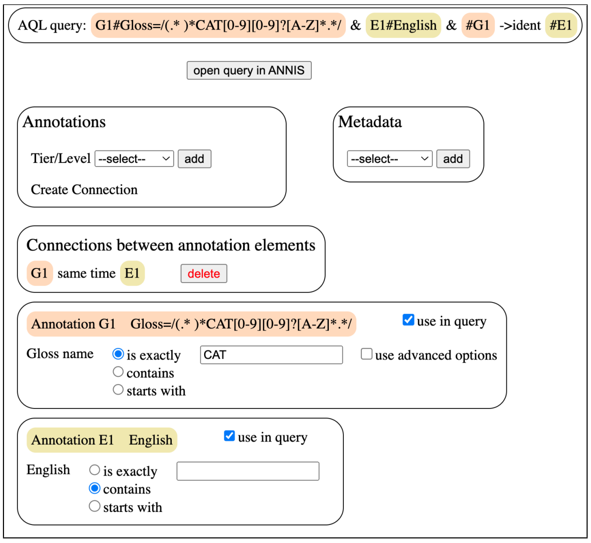
Annotation G1 Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ ☑ use in query

Gloss name ● is exactly CAT ☐ use advanced options
○ contains
○ starts with

Annotation E1 English ☑ use in query

English ○ is exactly
● contains
○ starts with

Figure 6: Interface for connecting two blocks.



AQL query: G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & E1#English & #G1 ->ident #E1

open query in ANNIS

Annotations

Tier/Level --select-- ▾ add

Create Connection

Metadata

--select-- ▾ add

Connections between annotation elements

G1 same time E1 delete

Annotation G1 Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ ☑ use in query

Gloss name ● is exactly CAT ☐ use advanced options
○ contains
○ starts with

Annotation E1 English ☑ use in query

English ○ is exactly
● contains
○ starts with

Figure 7: Interface with a Gloss G1, English G2 and a *same time* connection between them.

4. Examples

4.1. AQL Regular Expressions and Gloss Syntax

A simple query might be, for example, to search the corpus for all translations which contain the word "cat", which can be expressed in AQL as shown in Query 2 and for which we find 194 matches.

(2) English=/.* [Cc]at .*/

In this case, the user needs basic knowledge of the Public DGS Corpus annotations, plus simple AQL and regular expression syntax:

- English translations are in a tier named "English"
- AQL regular expression search is denoted by a query between two forward slashes ("/")
- In a regular expression, ".*" matches 0 or more characters of any kind
- In a regular expression, [Cc] matches either "C" or "c"

Now, if we want to instead search for tokens with the gloss word CAT, we could try the query in Query 3. As before, this requires some knowledge of the DGS Corpus annotations, AQL, and regular expression syntax.

(3) Gloss=/CAT.*/

Query 3 gives us 119 matches, which seems plausible given the 194 matches for Query 2, but if we examine them, we discover that only 75 are actually matches for varieties of CAT, while 25 are varieties of CATHOLIC, 13 CATHEDRAL, 4 CATASTROPHE and 2 CATTLE.

As described in Section 2.1, lexical variants of a sign are indicated with different digits after the gloss word. A next attempt would therefore be Query 4, which does indeed give 75 results – success!

(4) Gloss=/CAT[0-9].*/

However, we then remember that signs performed with the left hand are prefixed with "|| " (as explained in 2.1), and with Query 5 we indeed discover 4 instances of CAT1A* performed with the left hand, and one of CAT1B* performed with the left hand co-articulated with $INDEX1* performed with the right hand.

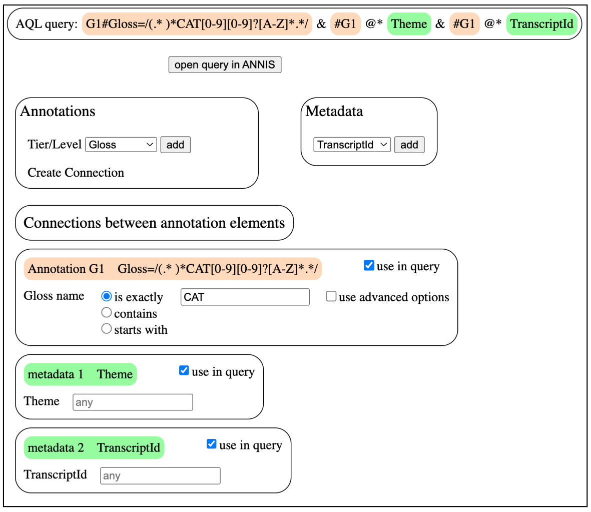
(5) Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/

By this point, the regular expression has already become fairly complicated, and if we wanted to further restrict this query to only supertypes or subtypes, or to only modified or unmodified glosses (see Section 2.1 for explanations), it would become significantly more so.

In the Query Wizard we only need to write the gloss name CAT, and select the button "is exactly", and the regular expression is created automatically, as shown in Figures 6 and 7.

4.2. Corpus Metadata

After finding all the tokens with gloss name CAT, a user may be curious in which corpus themes and in which transcripts these tokens most frequently occurred. In order to find this out, they need to add two metadata items to the query. To build this query manually, the user would have to know not only the type structures described above, but also the exact names of the two metadata types and the syntax for linking them. Each metadata item must be linked from an annotation item using the string "@*", as shown in Query 6.



AQL query: G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & #G1 @* Theme & #G1 @* TranscriptId

Annotations

Tier/Level: Gloss

Create Connection

Metadata

TranscriptId

Connections between annotation elements

Annotation G1 Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ ☑ use in query

Gloss name: ● is exactly ○ contains ○ starts with — CAT ☐ use advanced options

metadata 1 Theme ☑ use in query

Theme: any

metadata 2 TranscriptId ☑ use in query

TranscriptId: any

Figure 8: Interface with a Gloss block with gloss name CAT plus metadata Theme and TranscriptId

(6) G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & #G1 @* Theme & #G1 @* TranscriptId

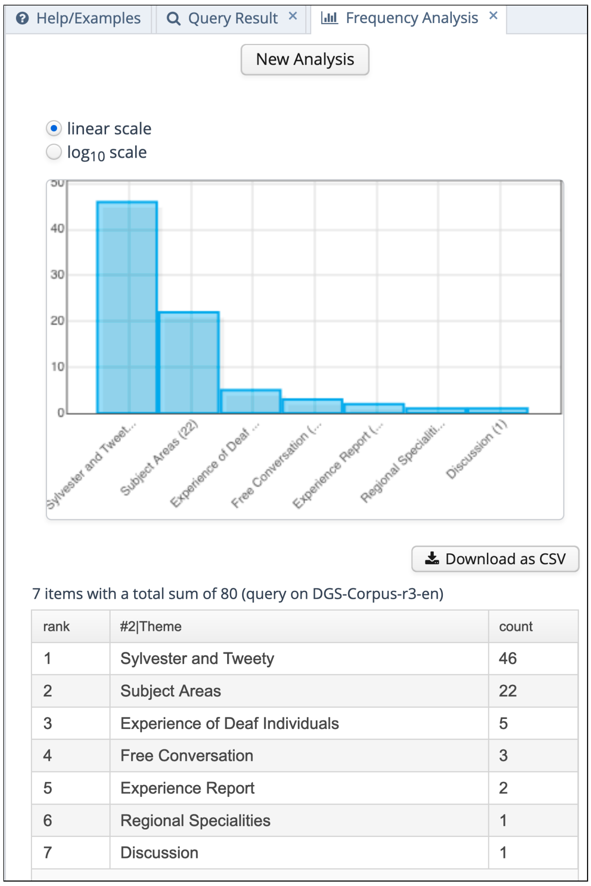
In Query Wizard this query can be created simply by creating the annotation block for the Gloss CAT as shown before, and two metadata blocks, one for Theme and one for TranscriptId, as in Figure 8. The results of opening this query in *MY DGS – ANNIS* are shown in Figures 1 and 2. The *frequency analysis* tab of *MY DGS – ANNIS*, shown in Figure 9, can then be used to discover that the CAT glosses occur most frequently in the "Sylvester and Tweety" task, where participants are asked to retell the popular cartoon story, but also often in discussions on specific "Subject Areas" and occasionally in 5 other themes, including "Experience of Deaf Indivuals".

Alternatively, a user might want to search only for results from participants from the age group "18–30". This is more complicated because of the way that the metadata is stored internally in the ANNIS database. In order to search metadata specific to one participant it is necessary to create a query which explicitly links each person's tokens to their metadata, as shown in Query 7. Again this query can be simply created in the Query Wizard by creating a gloss block for CAT and a metadata block for age group, as shown in Figure 10.

(7) (G1#PersonA:Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/
@* PersonA:AgeGroup="18-30") |
(G1#PersonB:Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/
@* PersonB:AgeGroup="18-30")

4.3. Collocation Distances

In the final example, shown in Figure 11, the user has created two Gloss blocks. The first, G1, searches as before for all tokens with gloss name CAT. The second, G2, does not specify any search text, and has a collocation distance from G1 of 1 to

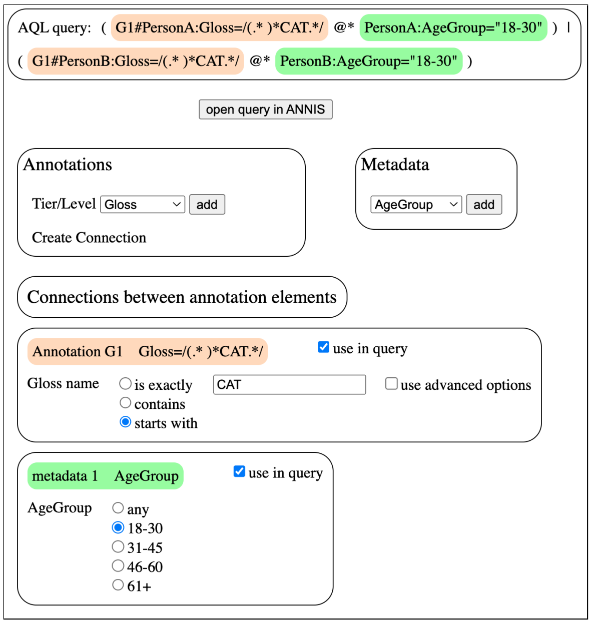


● linear scale
○ log10 scale

7 items with a total sum of 80 (query on DGS-Corpus-r3-en)

rank	#2|Theme	count
1	Sylvester and Tweety	46
2	Subject Areas	22
3	Experience of Deaf Individuals	5
4	Free Conversation	3
5	Experience Report	2
6	Regional Specialities	1
7	Discussion	1

Figure 9: MY DGS – ANNIS frequency analysis showing in which themes the tokens with gloss name CAT appear most frequently



AQL query: (G1#PersonA:Gloss=/(.*)*CAT.*/ @* PersonA:AgeGroup="18-30") | (G1#PersonB:Gloss=/(.*)*CAT.*/ @* PersonB:AgeGroup="18-30")

Annotations

Tier/Level: Gloss

Create Connection

Metadata

AgeGroup

Connections between annotation elements

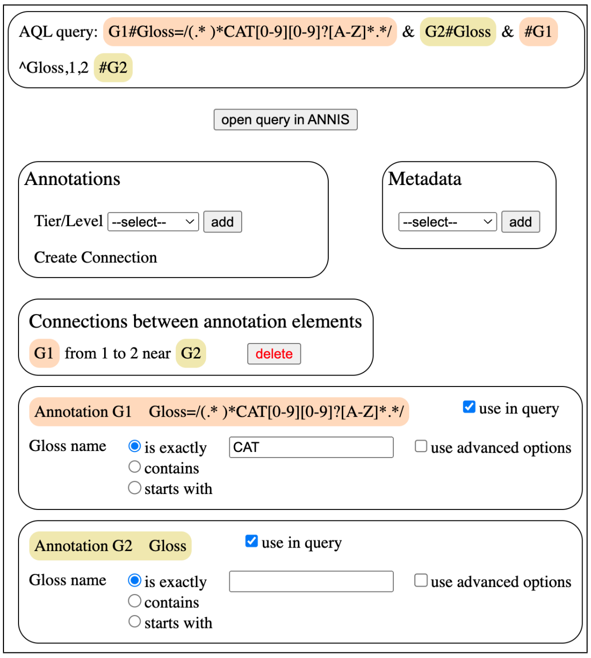
Annotation G1 Gloss=/(.*)*CAT.*/ ☑ use in query

Gloss name: ○ is exactly ○ contains ● starts with — CAT ☐ use advanced options

metadata 1 AgeGroup ☑ use in query

AgeGroup: ○ any ● 18-30 ○ 31-45 ○ 46-60 ○ 61+

Figure 10: Interface with a Gloss block with gloss name CAT plus metadata AgeGroup=18-30



AQL query: G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & G2#Gloss & #G1 ^Gloss,1,2 #G2

open query in ANNIS

Annotations

Tier/Level --select-- add

Create Connection

Metadata

--select-- add

Connections between annotation elements

G1 from 1 to 2 near G2 delete

Annotation G1 Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ ☑ use in query

Gloss name ● is exactly ○ contains ○ starts with CAT □ use advanced options

Annotation G2 Gloss ☑ use in query

Gloss name ● is exactly ○ contains ○ starts with □ use advanced options

Figure 11: Interface with two Gloss blocks with a collocation distance of 1 to 2.

2 in either direction, and the AQL query is shown in Query 8.

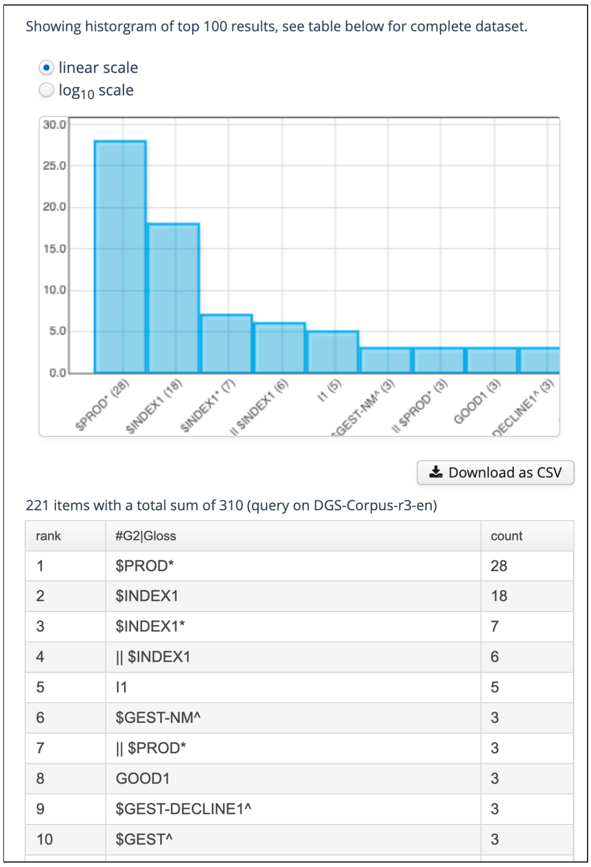
(8) G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & G2#Gloss & #G1 ^Gloss,1,2 #G2

When this query is opened in the frequency analysis tab of ANNIS (see Figure 12), we can see that the most frequent collocations are special signs, including productive signs and pointing gestures. The most frequent lexical sign is GOOD1, which leads us to a tentative and humorous conclusion that the corpus participants are well-disposed towards cats.

5. Conclusions and Future Work

We have introduced the new Query Wizard for *MY DGS – ANNIS* and shown how it simplifies the process of building queries in the ANNIS AQL query language for the DGS Corpus. It allows users to build queries out of small building blocks, and helps them to understand how the queries are built up. It removes the burden of regular expression building from users, and means that they do not have to remember the spellings of annotation tier and metadata names. It also allows users to select from the valid sets of metadata options. While user studies are still ongoing, initial feedback has been very favourable.

New corpus releases will be published as separate datasets in *MY DGS – ANNIS*. These new datasets may introduce new tiers or change structural aspects to account for new corpus (meta)data and improvements to the ANNIS software. The Query Wizard will allow users to choose the desired corpus release and will adjust its query outputs accordingly.



Showing historgram of top 100 results, see table below for complete dataset.

● linear scale
○ $\log_{10}$ scale

Download as CSV

221 items with a total sum of 310 (query on DGS-Corpus-r3-en)

rank	#G2|Gloss	count
1	$PROD*	28
2	$INDEX1	18
3	$INDEX1*	7
4	|| $INDEX1	6
5	I1	5
6	$GEST-NM^	3
7	|| $PROD*	3
8	GOOD1	3
9	$GEST-DECLINE1^	3
10	$GEST^	3

Figure 12: *MY DGS – ANNIS* frequency analysis of signs with a collocation distance of 1 to 2 from signs with gloss word CAT

There are also a number of features yet to be added to the Query Wizard. These include negated searches and fine-grained control over handedness of sign execution. Entering HamNoSys may be further improved by integrating the HamNoSys Builder interface more directly into the Query Wizard or supporting the use of HamNoSys character names, such as *hampinch12open*. Options could also be included to allow users to search special classes of signs such as numbers and fingerspellings.

6. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

7. Bibliographical References

Stefanie Dipper. 2015. Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch. *Zeitschrift für germanistische Linguistik*, 43(3):516–563.

Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).

Amy Isard and Reiner Konrad. 2022. MY DGS – ANNIS: ANNIS and the Public DGS Corpus. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).

Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken, and Oliver Böse. 2024. Corpus à la carte – improving access to the Public DGS Corpus. In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy. European Language Resources Association (ELRA).

Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.