ANNIS and the reference corpora of historical German

partly based on Perlitz & Odebrecht, ANNIS Einführung RIDGES Herbology 4.1. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/lehre/folien-material/2015_historische_korpora (CC-BY)

Historical reference corpora

REA (750-1050)

Reference Corpus of Old High (and Low) German
Referenzkorous Altdeutsch
https://www.deutschdiachrondigital.de/rea/

REM (1050-1350)

Reference Corpus of Middle High German
Referenzkorpus Mittelhochdeutsch
https://www.linguistics.rub.de/rem/

REN (1200–1650)

Reference Corpus of Middle Low German / Low Rhenish - *Referenzkorpus Mittelniederdeutsch / Niederrheinisch*
https://www.fdr.uni-hamburg.de/record/9195

REF (1350-1650)

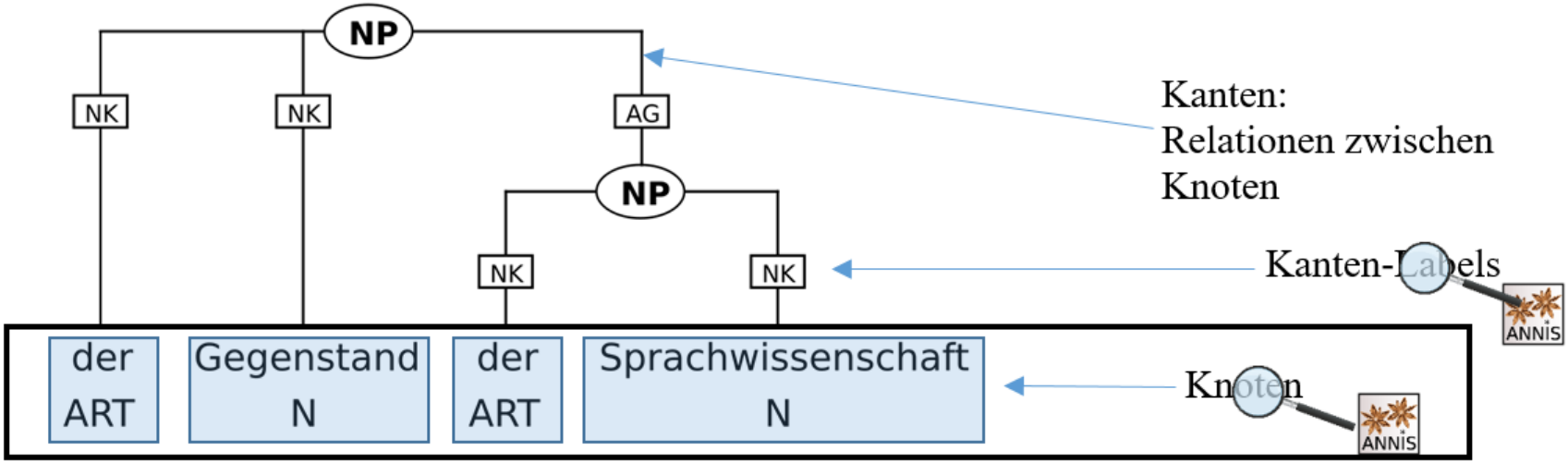
Reference Corpus of Early New High German
Referenzkorpus Frühneuhochdeutsch
https://www.linguistics.rub.de/ref/

ANNIS

- Another corpus query system…
- developed at HU Berlin
- for searching existing, annotated corpora (no inbuilt annotation function)
- some basic visualization possibilities

ANNIS Query Language

- can search for **nodes** and **edge labels**
- The latter are only available in **treebanks**, i.e. syntactically annotated corpora



ANNIS Query Language

- How do I tell ANNIS what I'm looking for?
- Basic principle: variable-value pairs

variable = value

e.g. pos="NN"

- and **relations** between variable-value pairs

pos= " ART " . pos="NN" 'followed by'

lemma= "Tanzen" _=_ pos="NN" 'congruent'

ANNIS
Search & visualization tool

- web-based search tool for corpora
  - project website http://annis-tools.org/
  - Access to web-based ANNIS tool https://korpling.german.hu-berlin.de/annis3

➢ What can ANNIS do?
  - ➢ finds annotations in a corpus
  - ➢ finds multiple annotations in a specific relation to one another
  - ➢ can visualize the annotations of the hits in multiple ways
  - ➢ exports the hits
  - ➢ etc.

ANNIS

Examples for corpora available via ANNIS:

- Referenzkorpus Altdeutsch (= Deutsch Diachron Digital; REA)
- Referenzkorpora Mittel- und Frühneuhochdeutsch (REM / REF)
- Refernzkorpus Mittelniederdeutsch /Niederrheinisch (REN)
- RIDGES (historical herbology texts)

ANNIS

Examples for synchronic corpora in ANNIS:

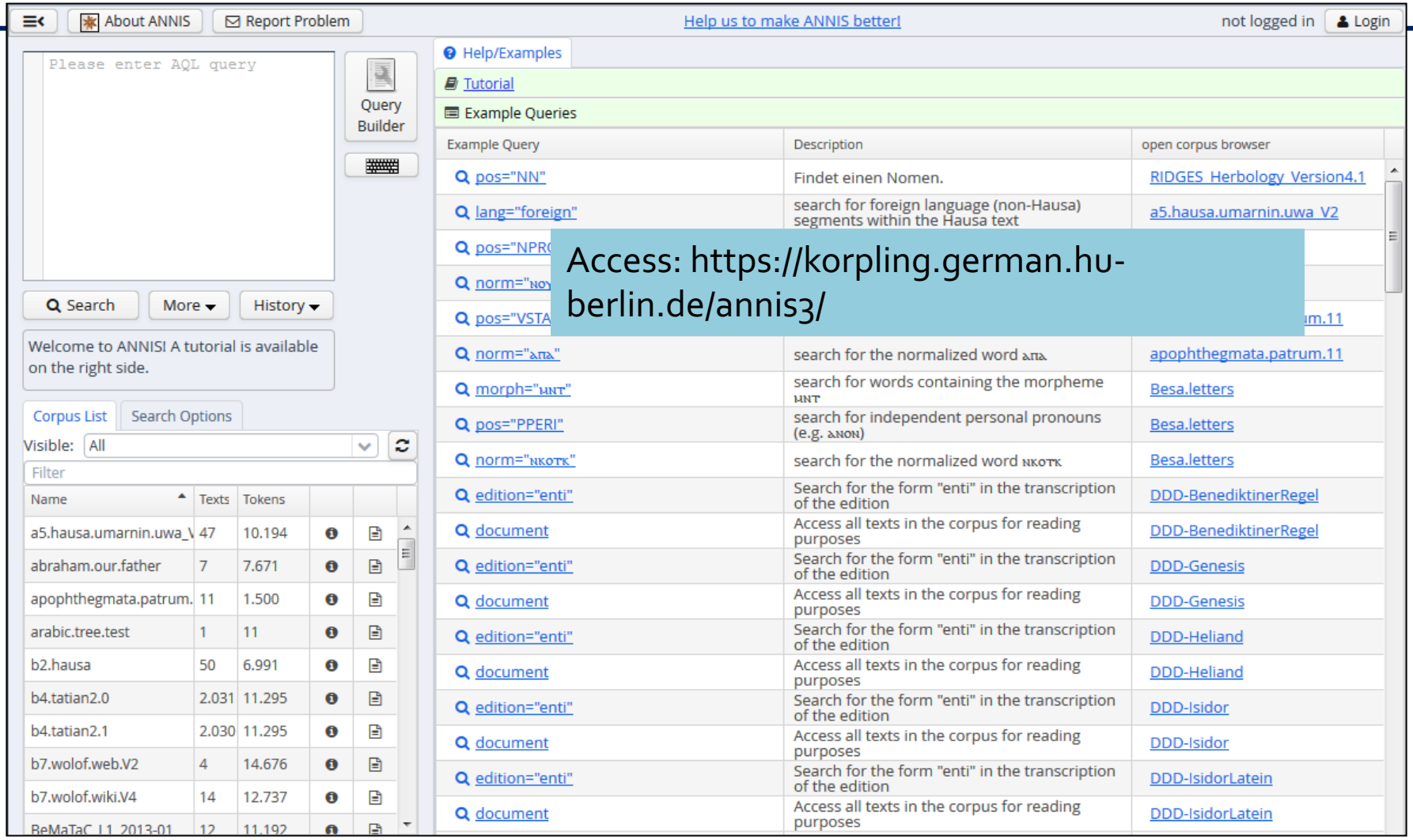
- Potsdamer Kommentarkorpus (Zetungskommentare)
- KiezDeutschKorpus
- Corpus d'Études pour le Français Contemporain (CEFC)
- Georgetown University Multilayer Corpus (GUM)

Example corpus: RIDGES Herbology

Ridges
Register in Diachronic German Science

- Herbology corpus
  - 15th-19th century
  - different dialects from the German-speaking area
  - annotated by ~50 students from three B.A./M.A. seminars
  - project website: http://korpling.german.hu-berlin.de/ridges/documentation_v4.1_en.html
- Annotation guidelines
  - http://korpling.german.hu-berlin.de/ridges/download/pubs/annotationGuidelines_v4.1.pdf

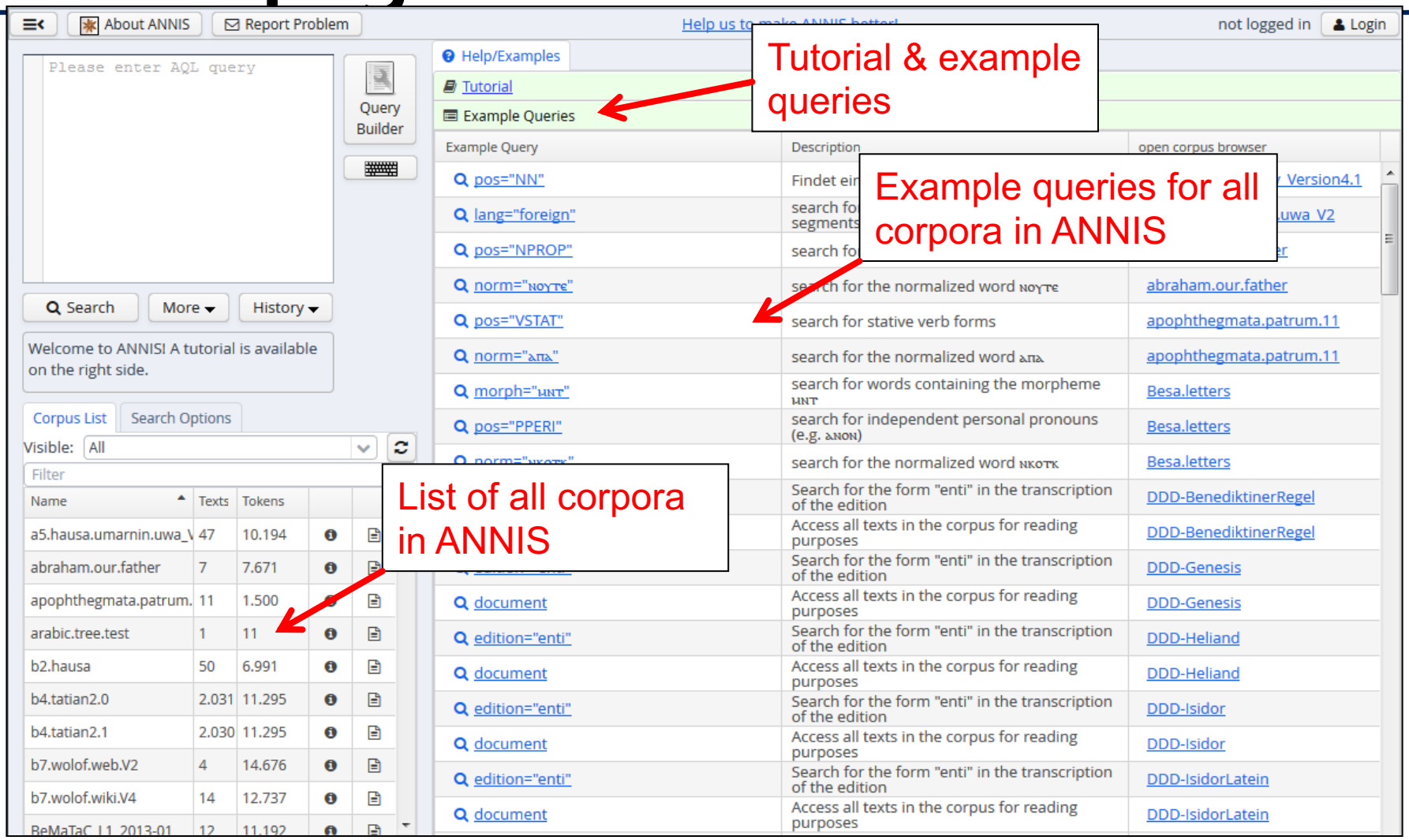
https://korpling.german.hu-berlin.de/annis3-snapshot/



Interface

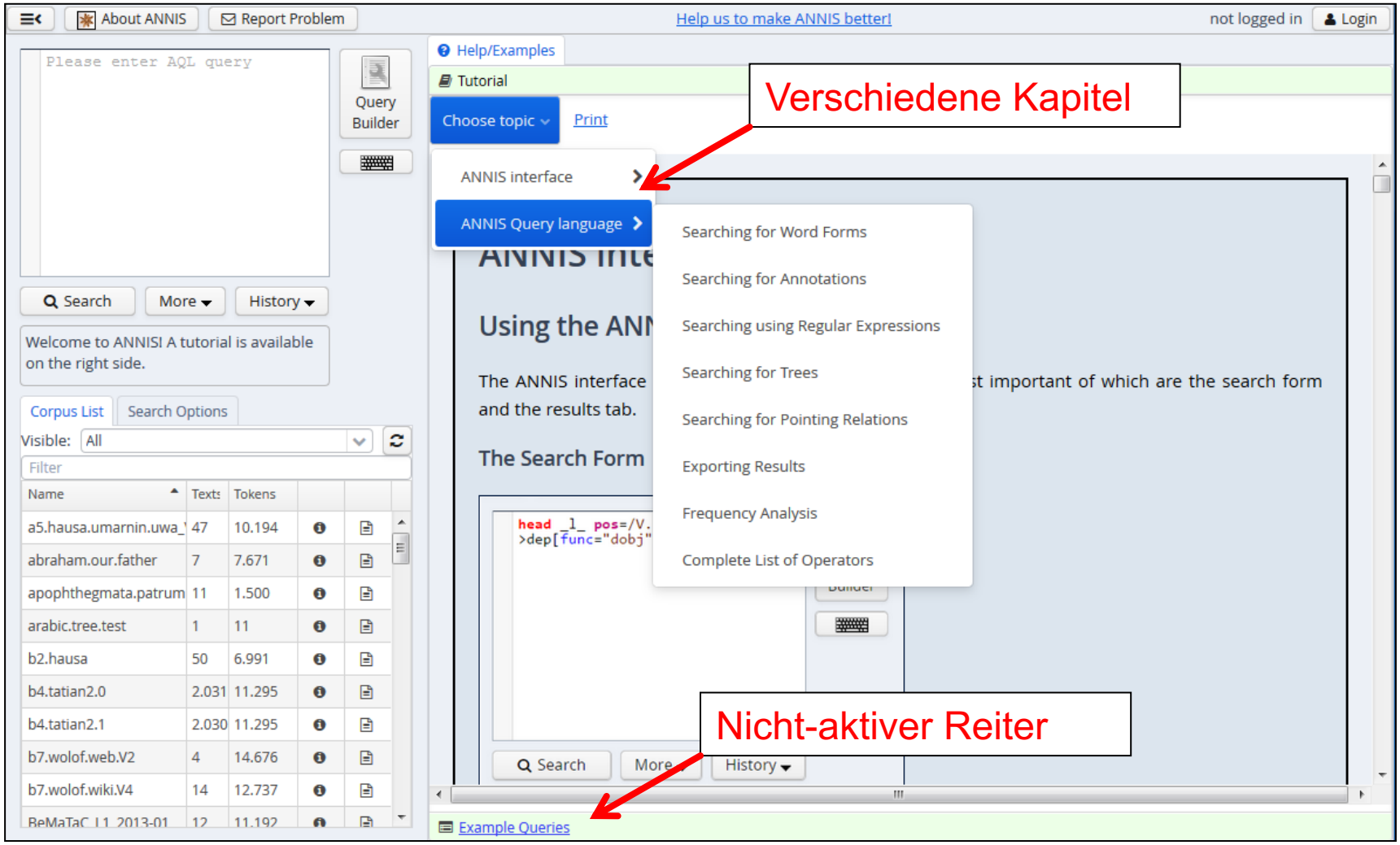
Start page, help, query window

Interface
Start page



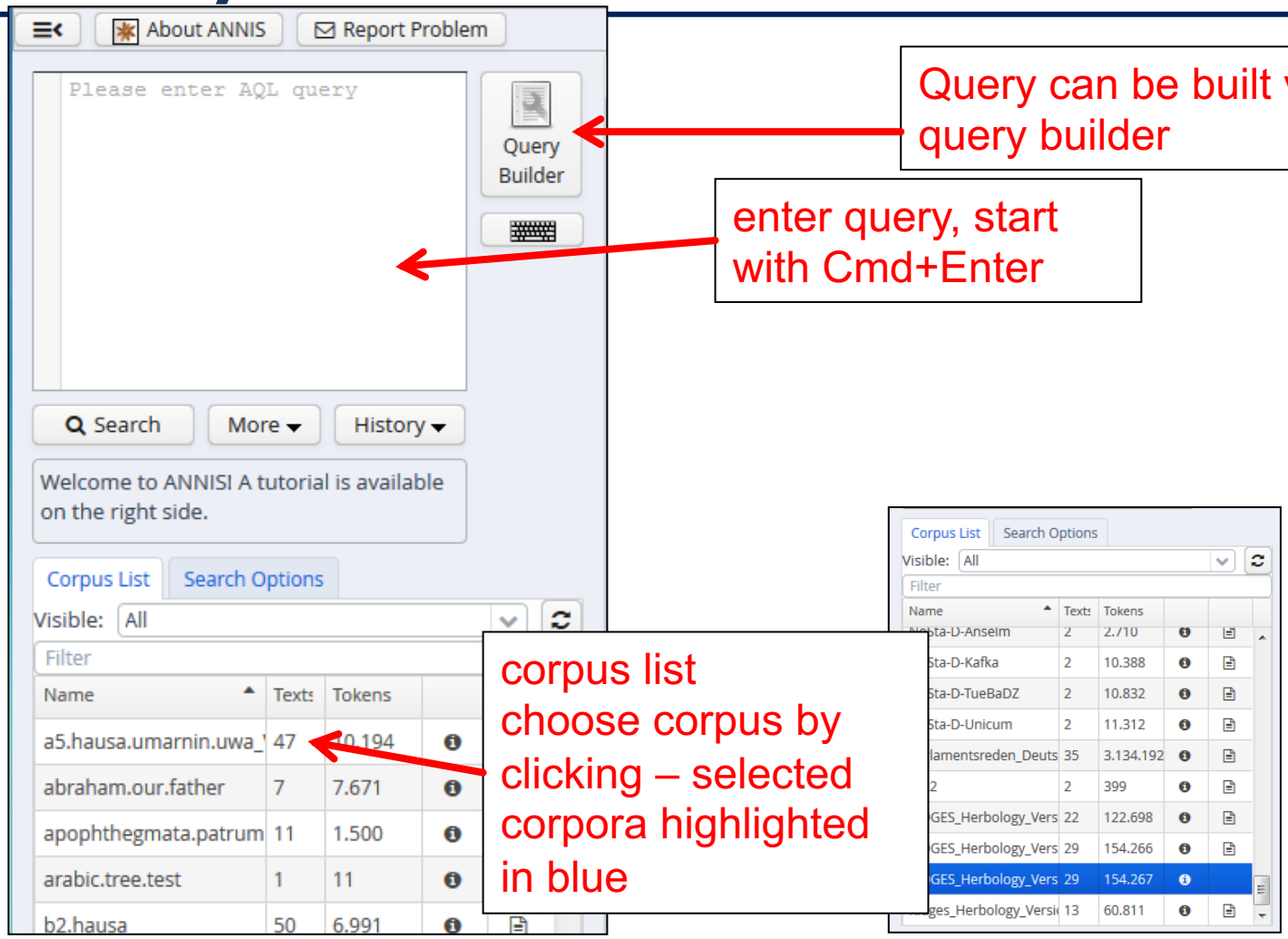
Slide based on the ANNIS tutorial by L. Perlitz & C. Odebrecht

Interface Tutorial



Slide based on the ANNIS tutorial by L. Perlitz & C. Odebrecht

Interface
Query window



Query can be built via the query builder

enter query, start with Cmd+Enter

corpus list
choose corpus by clicking – selected corpora highlighted in blue

ANNIS Query Language

- oder: How do I tell ANNIS what I'm looking for?
- two basic principles:
  - Variable-value pairs
  - Relations between variable-value pairs

Slide based on the ANNIS tutorial by L. Perlitz & C. Odebrecht

Tokens

1) Tokens are the smallest (technical) unit in a corpus.
2) A token usually, but not always, corresponsd to one orthographic word or punctuation mark!
3) You can search for this items in ANNIS.

Token Token Token Token Token Token ...

dipl	wider	auff	ein	friſches	zerſtoſſenes	kraut	/	thue	es	wider	in
clean	wider	auff	ein	frisches	zerstossenes	kraut	/	thue	es	wider	in
norm	wieder	auf	ein	frisches	zerstoßenes	Kraut	/	tue	es	wieder	in

Principle 1: Variable-value pair

- AQL
  - Name of annotation level →Variable
  - annotated category →value

dipl = /kraut/

- We have to know which variables are available!
  - Corpus metadata
  - annotation guidelines of the corpus (e.g. for RIDGES http://korpling.german.hu-berlin.de/ridges/download/pubs/annotationGuidelines_v4.1.pdf)

Principle I: Variable-value pairs

1) Prerequisite: Existence of a level called „dipl“. (Metadata!)
2) Expected result: finding all attestations of this character string on the level „dipl“ in the selected corpus (or corpora)

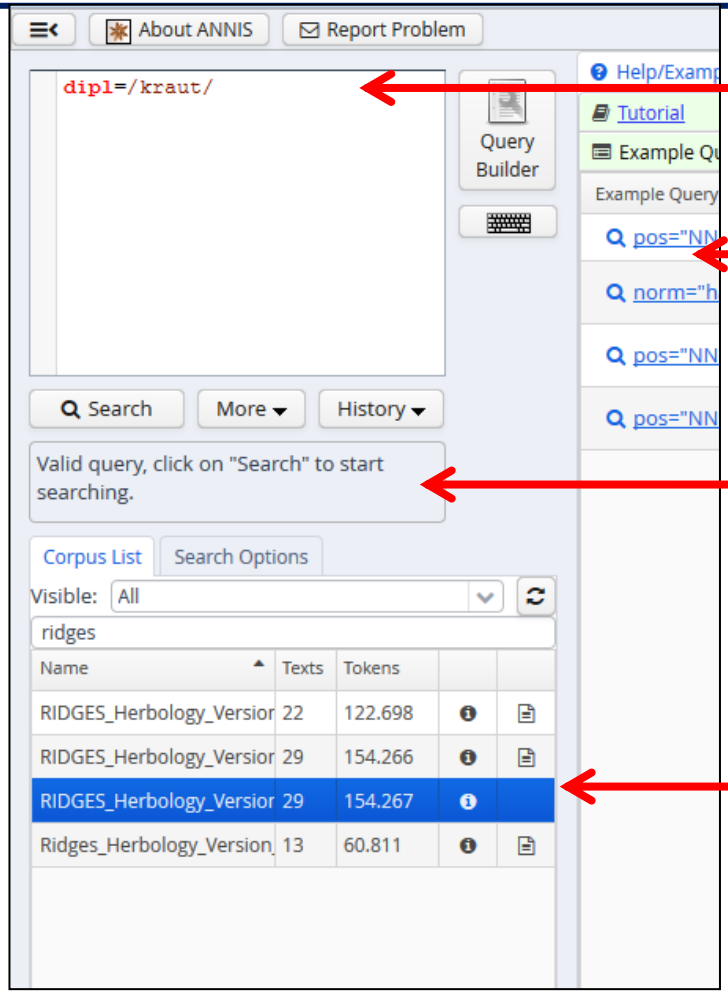
dipl= /kraut/

Variable (Layer, Tier, Ebene ...)

Wert (Wort, Lemma, Satz, Wortart ...)

dipl	wider	auff	ein	friſches	zerſtoſſenes	kraut	/	thue	es	wider	in
clean	wider	auff	ein	frisches	zerstossenes	kraut	/	thue	es	wider	in
norm	wieder	auf	ein	frisches	zerstoßenes	Kraut	/	tue	es	wieder	in

Interface Query



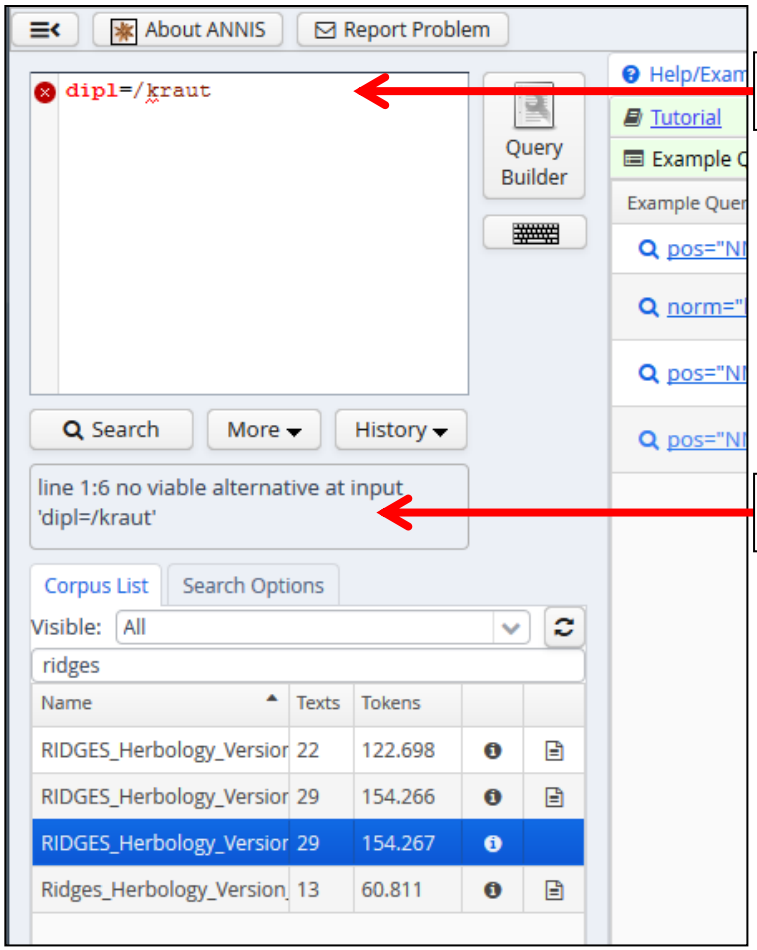
query language AQL

example queries for RIDGES corpus

shows if query is valid

for every query at least one corpus has to be selected

Interface
Errors in the query

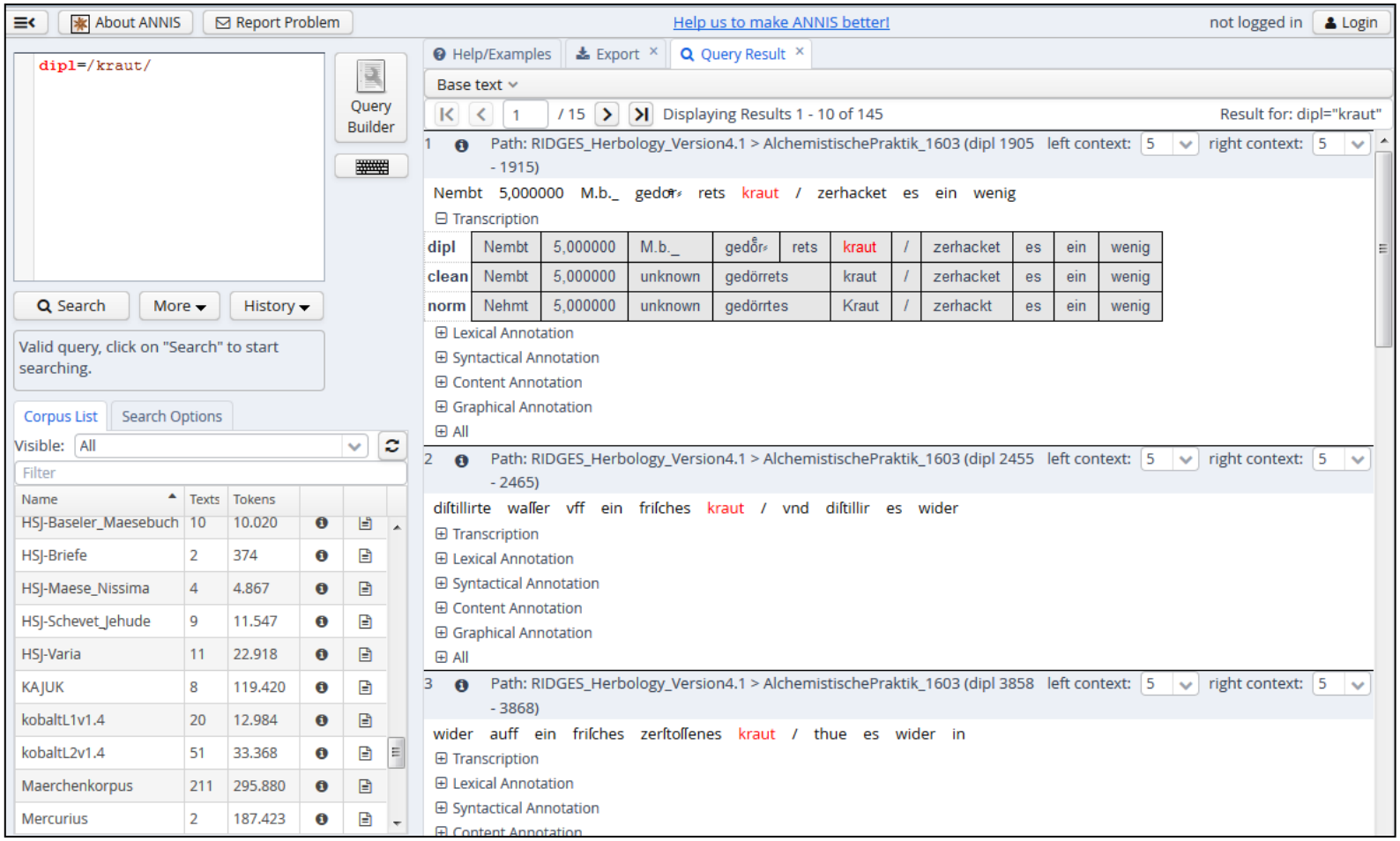


AQL with errors!

Feedback on the error

Variable-value pair

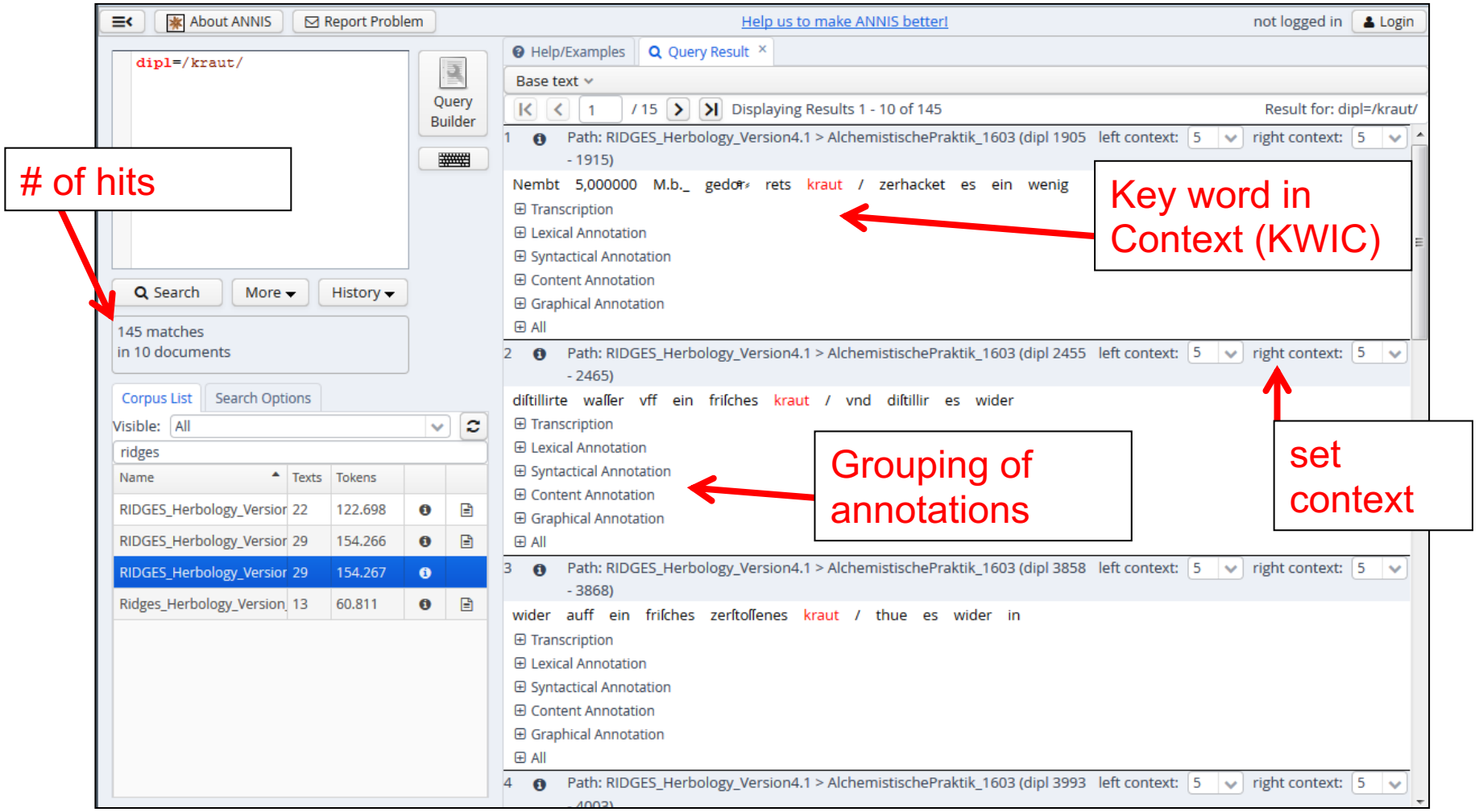
dipl= /kraut/



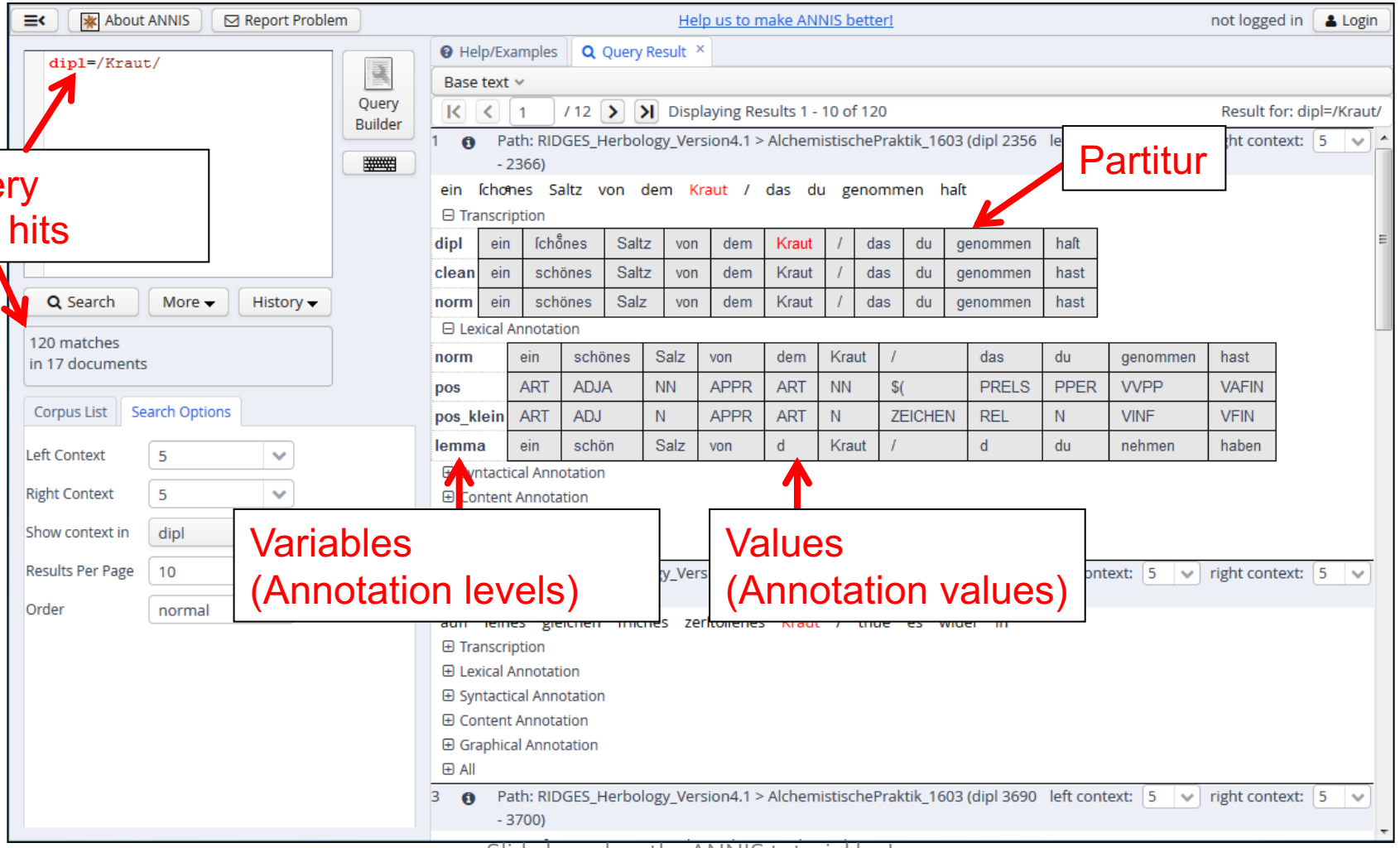
Interface

Hits, visualization, metadata, search history

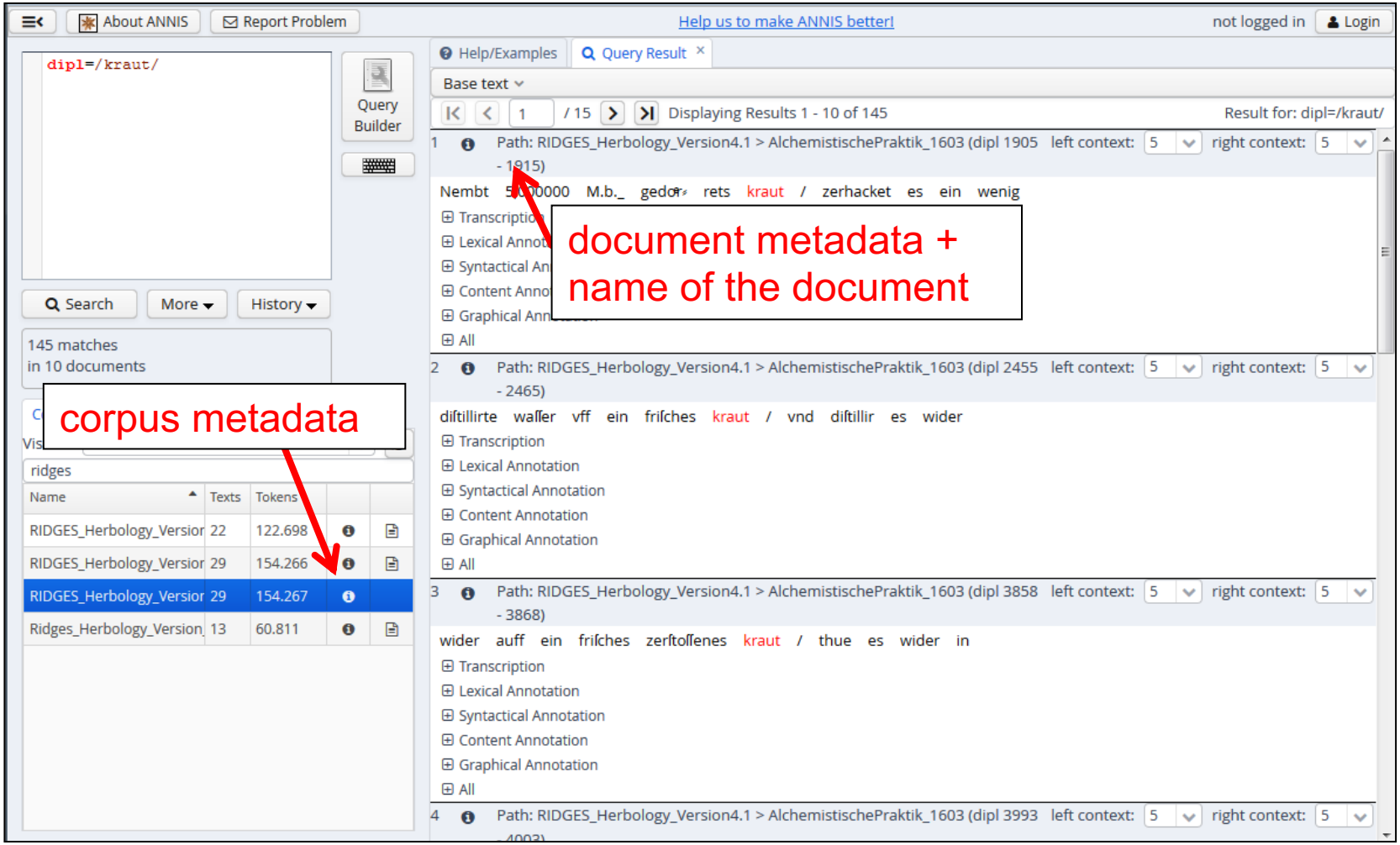
Interface hits



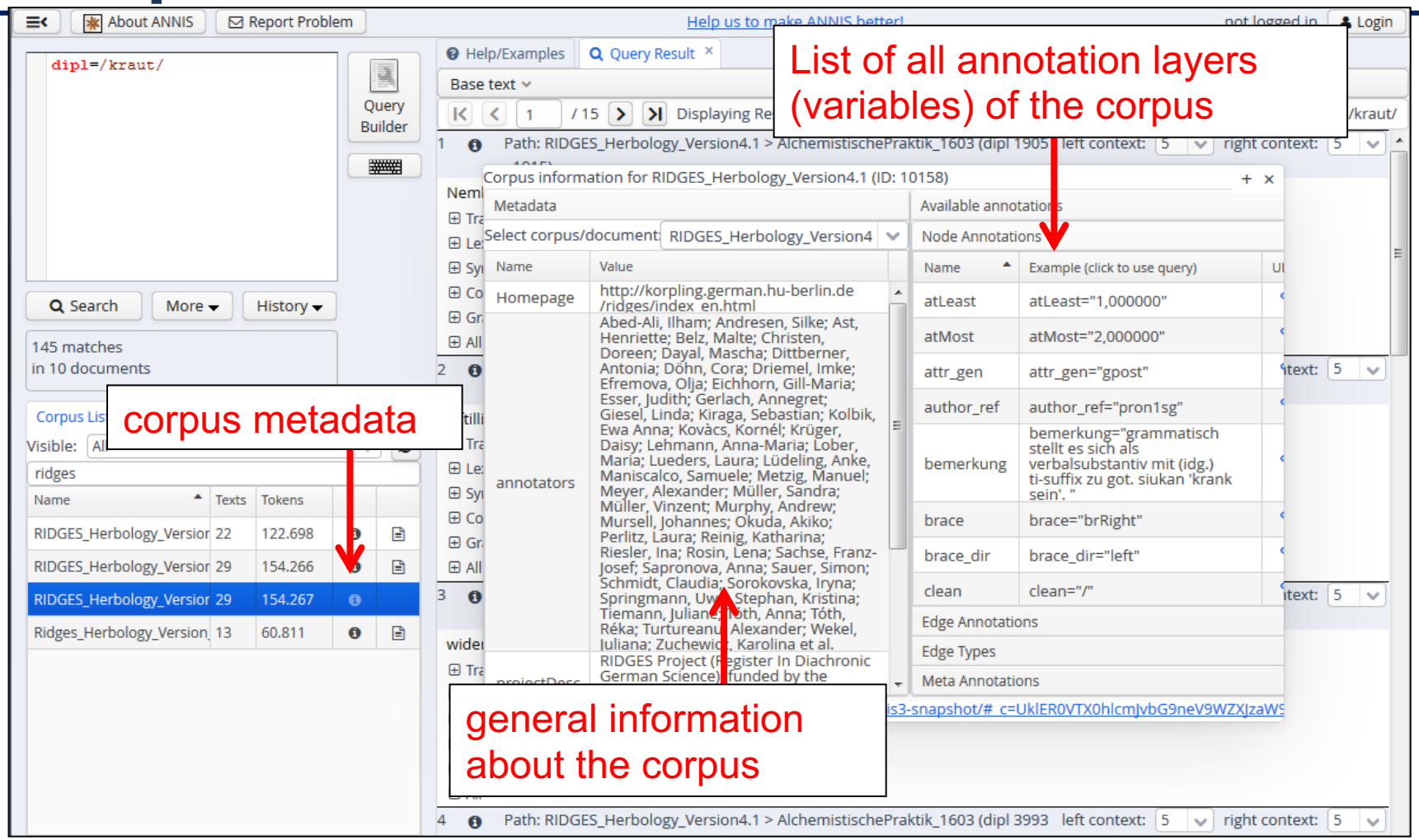
Interface Visualization



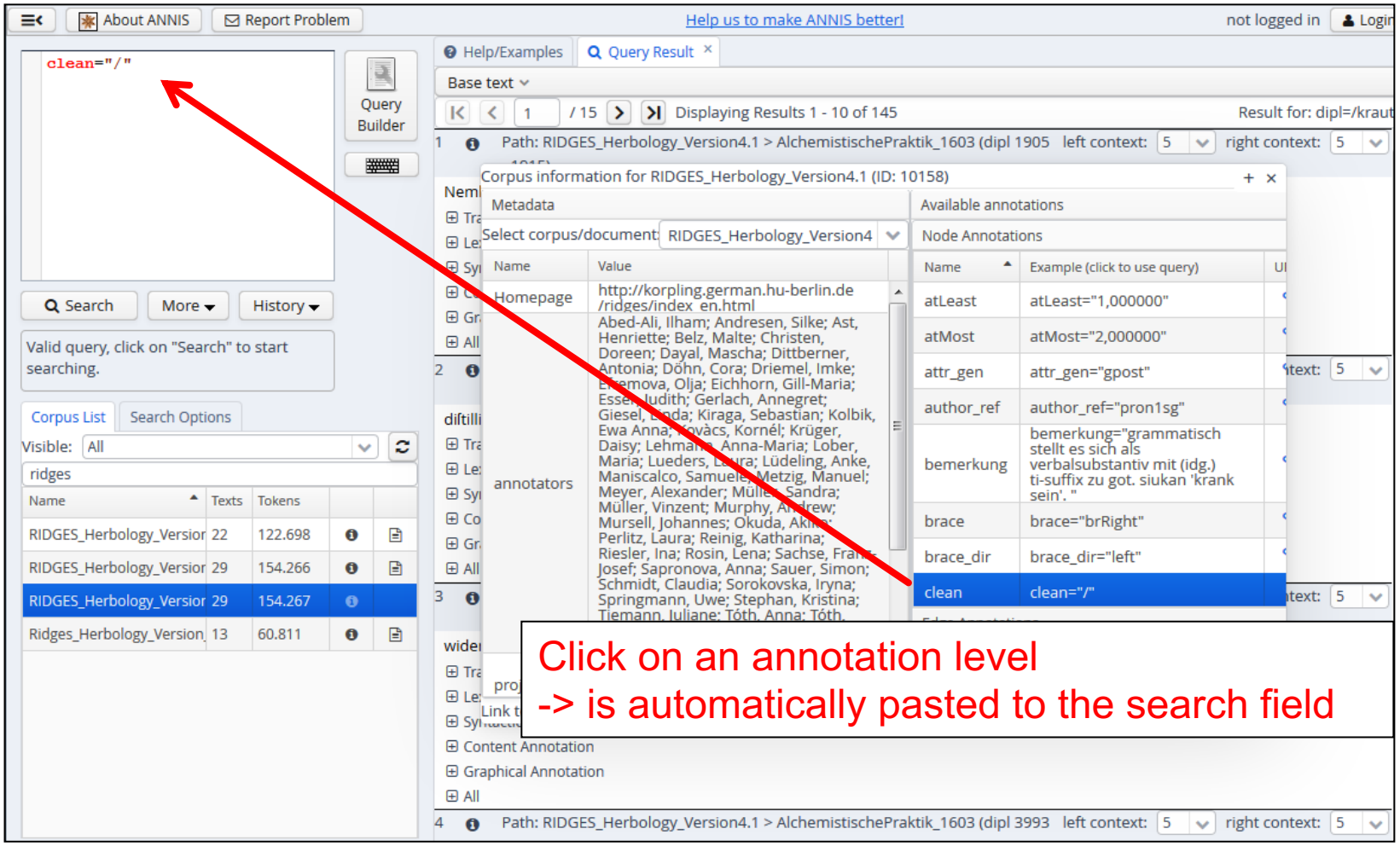
Interface Metadata



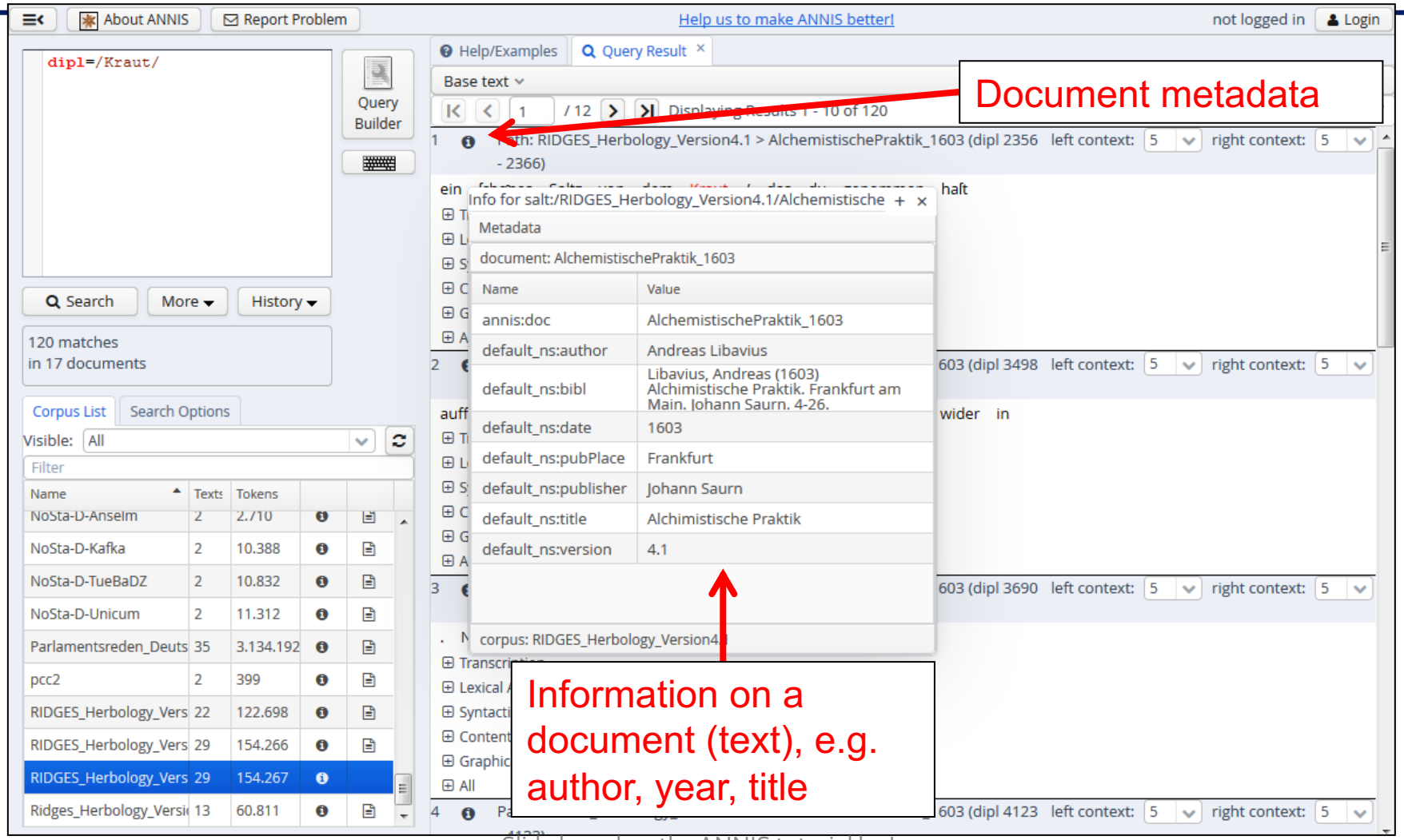
Interface
Corpus metadata



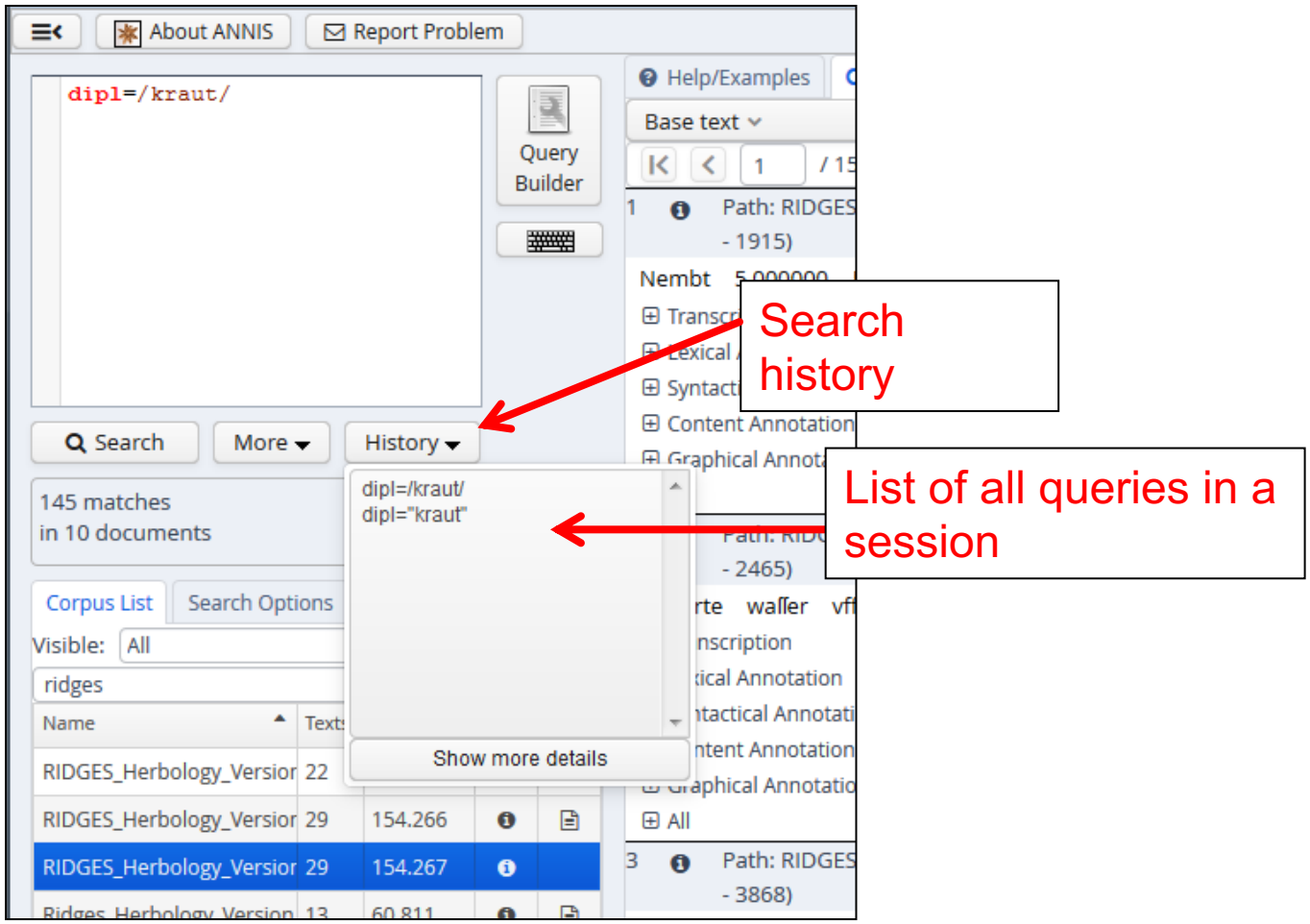
Interface
Corpus metadata II



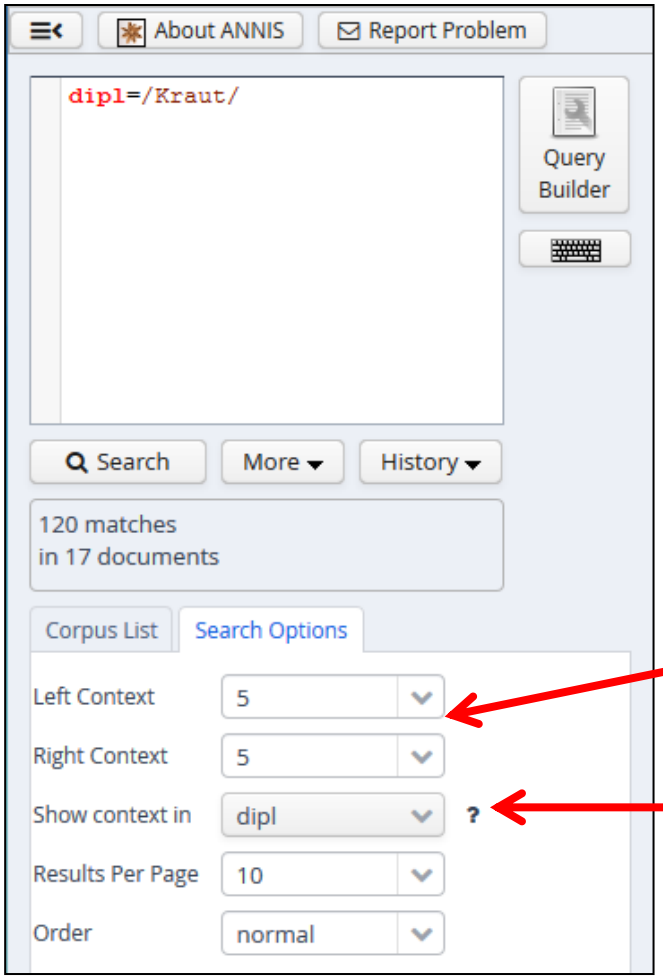
Interface
Document metadata



Interface
Search history



Interface
Query options



set size of context

Which transcription level should be shown in the list of hits?

Historical texts

Ridges
Register in Diachronic German Science

- Variance in orthography and composition
- some things are quite "unpredictable"
- graphemic / (text)-structural information
  - line break, page break, page composition

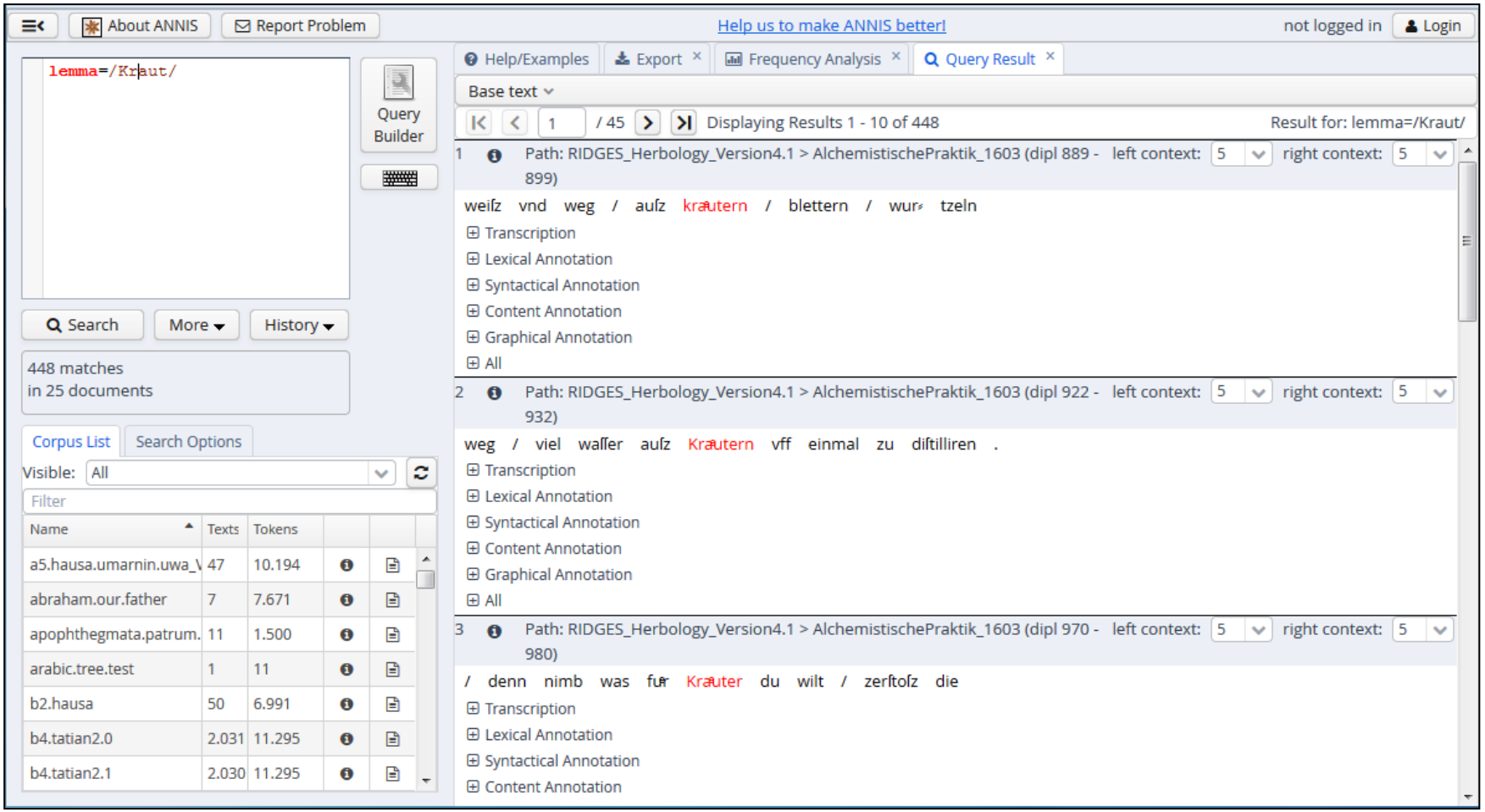
	Krautern	Alchimistische Praktik 1603
	Kraut	Alchimistische Praktik 1603
	kraut	Alchimistische Praktik 1603
	Kreutern	Alchimistische Praktik 1603
	Kreutter	Alchimistische Praktik
	kreüter	Fuchs New Kreüterbuch 1543
	Krauteren	Pflantz.Gart. 1639
	Kreuter	Alchmistische Praktik 1603
	Kräuter	Deutsche Pflanzennamen 1870

Example
Graphemic variants

- Try to find all graphemic and inflectional variants of *Kraut* 'herb'!
  - use the fitting annotation layer (variable) for the query
    - most suitable variable would be **lemma**
  - and enter the value **/Kraut/**

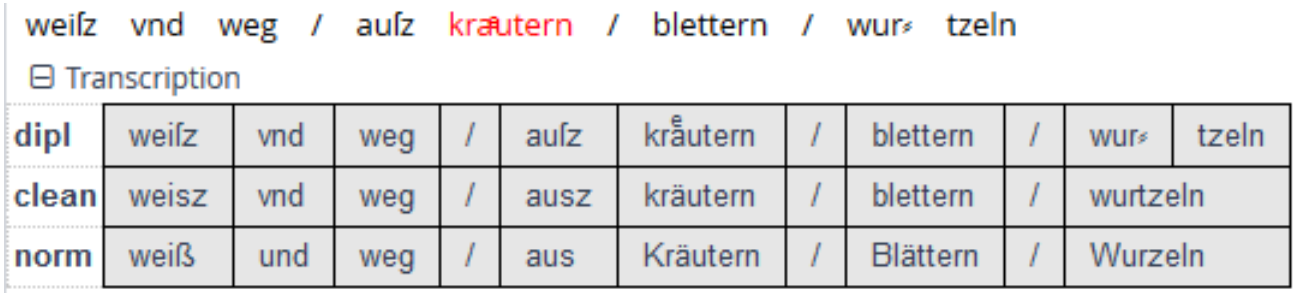
➢ lemma=/Kraut/

Example
Graphemic variants



Normalization

- Take a closer look at the hits for lemma=/Kraut/



  - Our query finds all historical word forms of *Kraut* attested in the corpus.

Example: Normalization

- Search for the lemma of *zusammensetzen*!
  - We find very different historical spelling variants…
  - ➢ lemma=/zusammensetzen/

  - Hit 1

    erſten grade / vnd ſey zůſammen geſetzt auſz widerwertiger ſubſtantz . das

  - Hit 2

    , die aus drei Worten zuſammen geſetzt ſind , wie die oben

  - Hit 3

    muſz aus zwei einfachen Worten zuſammengeſetzt ſein , von denen der

Slide based on the ANNIS tutorial by L. Perlitz & C. Odebrecht

Example: Normalization

- Stepwise normalization (dipl>clean>norm)
  - dipl: Unification/merging of historical spelling variants
  - POS and lemma annotation on the normalized level (norm)

3 Path: RIDGES_Herbology_Version4.1 > DeutschePflanzennamen_1870 (dipl 5444 - 5455) left context: 5 right co

muſz aus zwei einfachen Worten zuſammengeſetzt ſein , von denen der

⊟ Transcription

dipl	muſz	aus	zwei	einfachen	Worten	zuſ	ſammengeſetzt	ſein	,	von	denen	der
clean	musz	aus	zwei	einfachen	Worten	zusammengesetzt		sein	,	von	denen	der
norm	muss	aus	zwei	einfachen	Worten	zusammengesetzt		sein	,	von	denen	der

⊟ Lexical Annotation

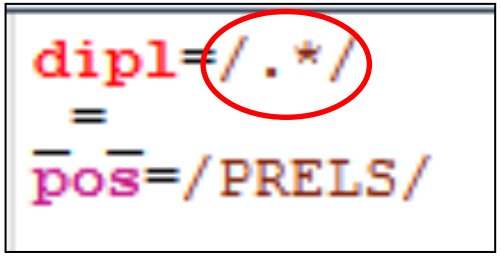
norm	muss	aus	zwei	einfachen	Worten	zusammengesetzt	sein	,	von
pos	VMFIN	APPR	CARD	ADJA	NN	VVPP	VAINF	$,	APPR
pos_klein	VFIN	APPR	CARD	ADJ	N	VINF	VINF	ZEICHEN	APPR
lemma	müssen	aus	zwei	einfach	Wort	zusammensetzen	sein	,	von

Operators for pattern search

.	any character
?	0 or 1 repetition (of preceding element)
*	0-infinitely many rep. (of preceding element)
+	1-infinitely many rep. (of preceding element)
\\	verbatim (following element)
!	not
(a|b)	a or b (also: [ab])

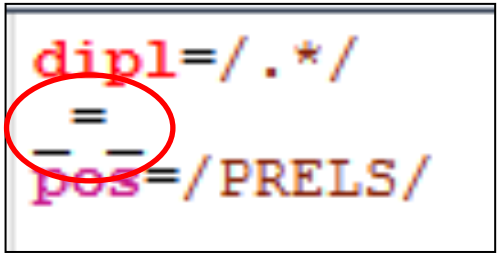
Important operators for pattern search

.	any character
?	0 or 1 repetition (of preceding element)
*	0-infinitely many rep. (of preceding element)
+	1-infinitely many rep. (of preceding element)
\\	verbatim (following element)
!	not
(a|b)	a or b (also: [ab])
[abc]	set (or [^abc]=all except abc)
a{2,3}	a 2-3times



important relations between VV pairs

.	direct precedence
.*	indirect precedence
_=_	identical coverage
_i_	inclusion
_o_	overlap
_l_	left overlap
_r_	right overlap



Example: Operators

- Which results do you expect for the following pattern searchers? norm=/g.b./
  - ➢ *gebe*, *gibt* (for RIDGES, others are theoretically possible)
  - – dipl=/r(a|o)t/
  - ➢ *rot*, *rat* (for RIDGES)
  - – dipl=/meint?/
  - ➢ *mein*, *meint* (for RIDGES)

Example
Part of speech

- How can we find all common nouns (not proper nouns) in RIDGES?
  - relevant variable is **pos**
  - ➢ pos=/NN/

Example
Part of speech

- How can we find all adjectives in RIDGES?
  - What do the guidelines say (STTS)?
  - relevant variable here **pos**
  - STTS: difference between ADJA and ADJD
    - ADJA: attributive adjective, *die schöne Katze* 'the beautiful cat'
    - ADJD: predicative or adverbial adjectives, e.g. *er fährt schnell* 'he drives fast', *er ist schnell* ' he is fast'

➢ pos=/ADJ./

Example: Querying annotation layers

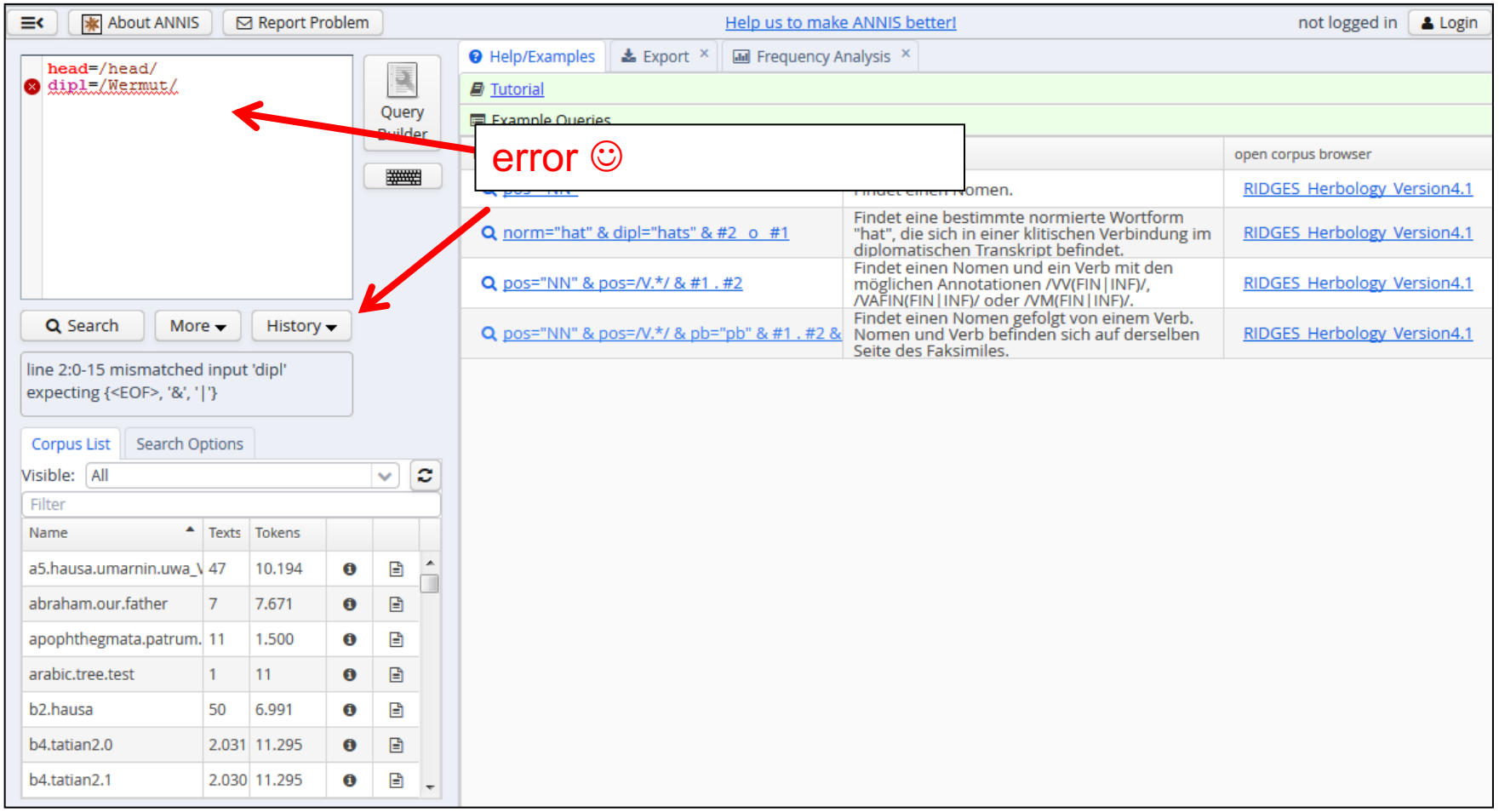
- Find out if there is a dedicated layer for headers in RIDGES!
  - Can we build a query that searchers for headers?
  - relevant variable is **head**
  - ➢head=/head/

Example: Combining searches

- Find the diplomatic word form *Wermut* occurring in headers!
  - relevant variables **head** and **dipl**
  - ➢ head=/head/
  - ➢ dipl=/Wermut/

  What's happening?

Aufgabe
Kombinierte Suche



Principle II: Relations

head=/head/ **VV pair1**

_o_ **Relation**

dipl=/Wermut/ **VV pair2**

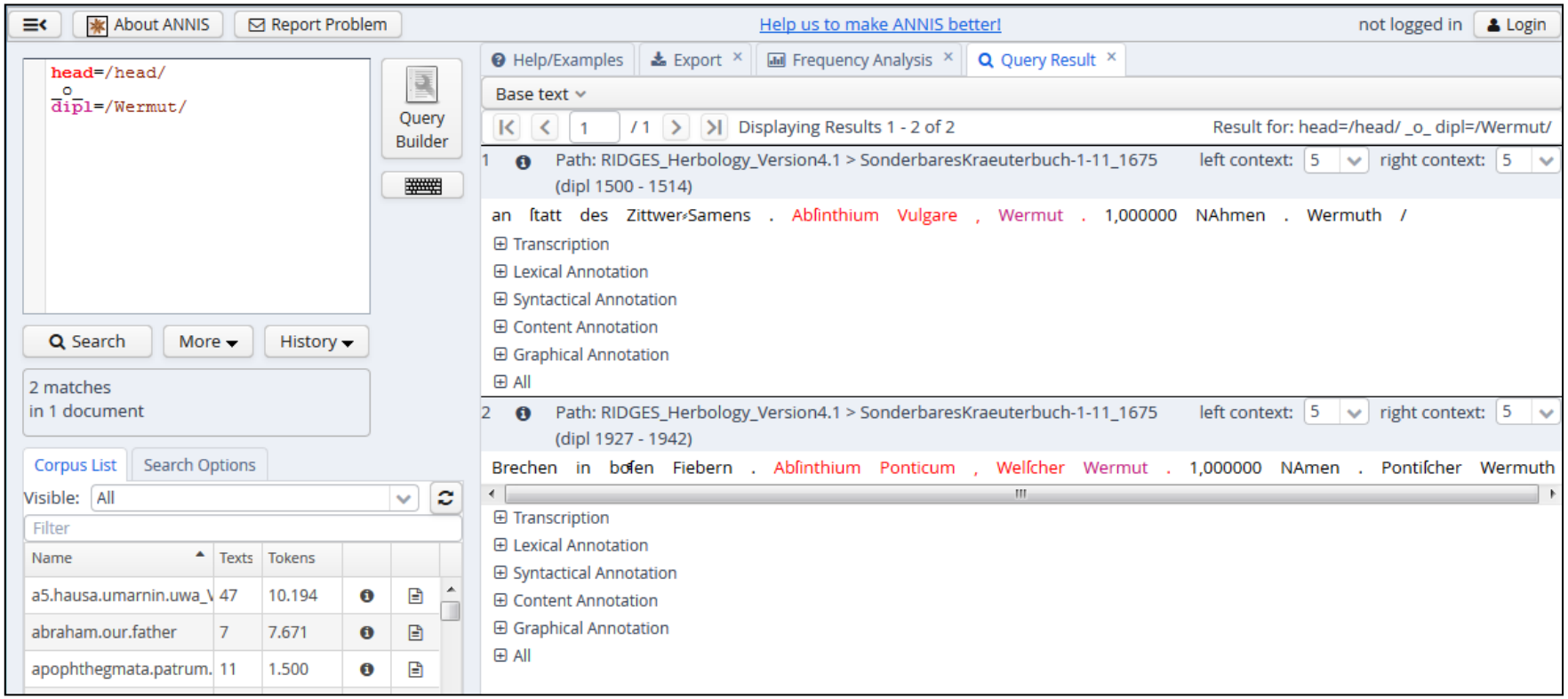
1) There are several types of annotations!
2) We have to know which relation these annotations can have to one another.
3) Spans (head) overlap with token annotations (dipl)!

XVIII. Saturniniſche Kræuter . Von ħ Kræutern ihrer Natur

⊟ Transcription

dipl	XVIII.	Saturniniſche	Kr̊äuter	.	Von	ħ	Kr̊äutern	ihrer	Natur
clean	XVIII.	Saturninische	Kräuter	.	Von	ħ	Kräutern	ihrer	Natur
norm	XVIII.	Saturninische	Kräuter	.	Von	?	Kräutern	ihrer	Natur
head	head								

Aufgabe
Kombinierte Suche



Syntax-Highlighting

- one color per VV pair - some examples:
  - Red for all values of the variable **head**
  - Purple for all values of the variable **dipl**

```
head=/head/
  _o_
dipl=/Wermut/
```

- The hits are marked in the same colors:
  - multiple tokens are red, as they occur in a header together
  - one token is purple because it is the dipl value we were searching for: Abſinthium Vulgare , Wermut .

Example: Overlaps

- Search for a verb that occurs in a header.
  - relevant variables: **pos** and **head**
  - Operator **_o_**
  - ➢ pos=/V.*/ _o_ head=/head/

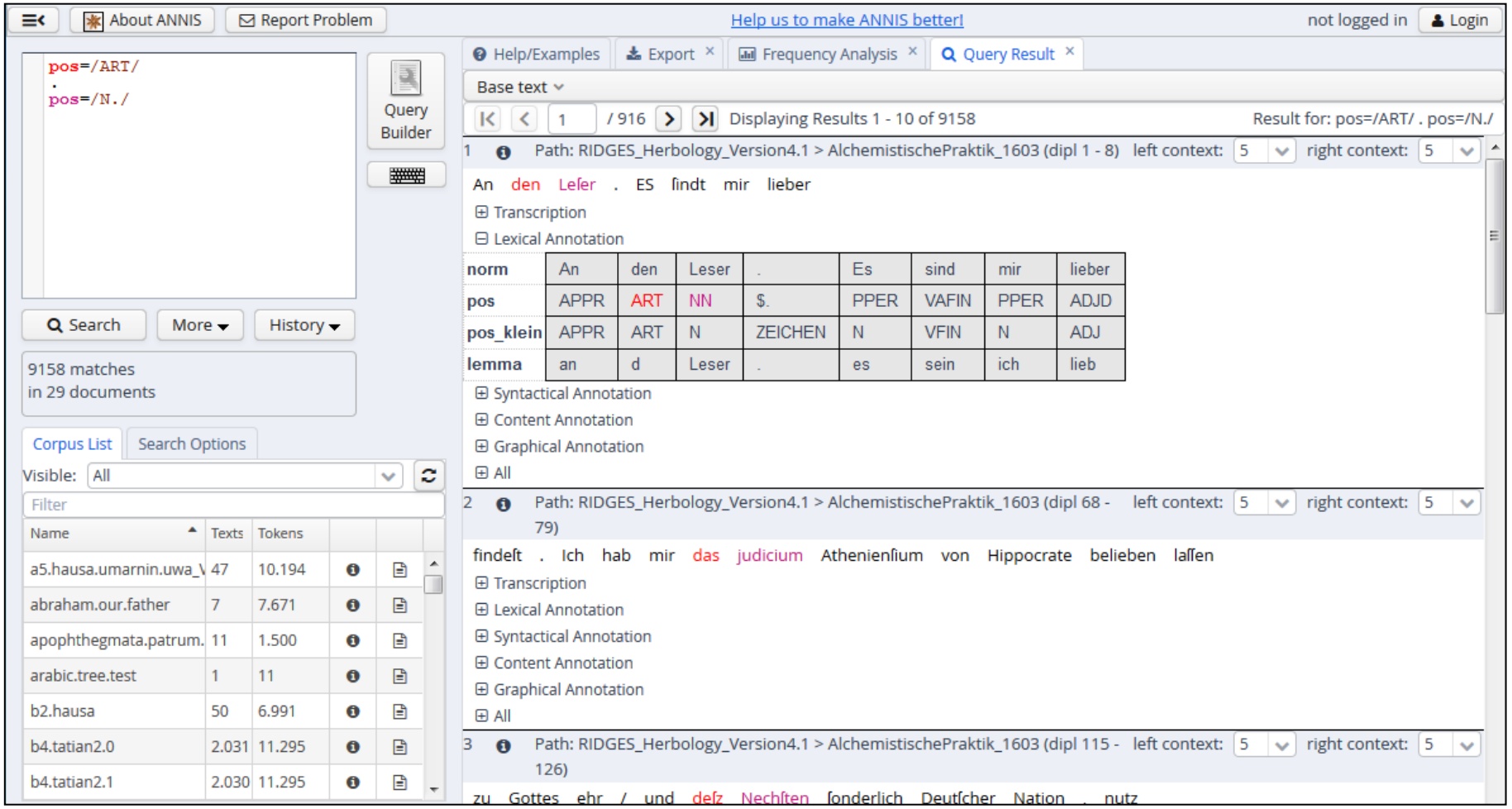
Example: Identity

- Look for all diplomatic word forms tagged as substituting relative pronouns.
  - relevant variables: **dipl** and **pos**
  - Operator _=_
  - ➢ dipl=/.*/ _=_ pos=/PRELS/

Example: Direct precedence

- Search for an article directly preceding a noun.
  - relevant variable: **pos**
  - Operator **.**
  - ➢ pos=/ART/ . pos=/N./

Example: Direct precedence



Example: indirect precedence

- Search for an article (**A**) that **indirectly precedes** a noun (**B**). Why indirectly? You also want to catch a potential prenominal modifier (**C**) in your results: *die schöne Katze.*
  - relevant variable: **pos**
  - Operator **.** and distance 1,2 (read: I search **A** and **B** either in direct precedence or one token **C** can be in-between)

➢ pos=/ART/ .1,2 pos=/N./

Example: Sequences

- Search for a noun directly followed by a preposition which in turn is directly followed by an adjective.
  - relevant variable: **pos**
  - Operator **.**
  - ➢ pos=/N./ . pos=/APPR/ . pos=/ADJ./

Example: Sequences

- Search for a noun directly followed by a preposition directly followed by an adjective.
  - relevant variable **pos**
  - Operator **.**

➢ pos=/N./ . pos=/APPR/ . pos=/ADJ./

/ zwey feine Alchy⸗ miſtiſche Tractätlein von guten freundten zukom̃en / wiewol vnbewuſt

⊞ Transcription

⊟ Lexical Annotation

norm	/	zwei	feine	Alchimistische	Traktätlein	von	guten	Freunden	zukommen
pos	$(	CARD	ADJA	ADJA	NN	APPR	ADJA	NN	VVINF

Examle: Identity

- Search for particle verbs containing the particle *auf* (e.g. *aufwachen* 'wake up')
  - relevant variables: **lemma** and **pos**
  - Operator _=_
  - ➢ pos=/VV.*/ _=_ lemma=/auf.*/

Examle: Identity

- Search for particle verbs containing the particle *auf* (e.g. *aufwachen* 'wake up')
  - relevant variables: **lemma** and **pos**
  - Operator _=_
  - ➢ pos=/VV.*/ _=_ lemma=/auf.*/

➢ What do we **not** find using this query?

Example: Metadata

- Find out if the normnalized word form *Blume* can be found in a text from the year 1543.
- If yes, how often?
  - relevant variables **norm** and **meta::date**
  - Verknüpfungszeichen **&**
  - ➢ norm=/Blume/ & meta::date=/1543/
  - ➢ 2 matches

Hints

- The list of hits only shows the annotation layers (variables) that have been annotated in the span that is shown. In other words, there might be more annotation layers that are not displayed because there are no annotations for the span currently displayed.
- Take a look at corpus metadata and annotation guidelines to know which annotation layers are actually in the corpus.
- Not all documents in a corpus necessarily have the same (number and type of) annotation layers.

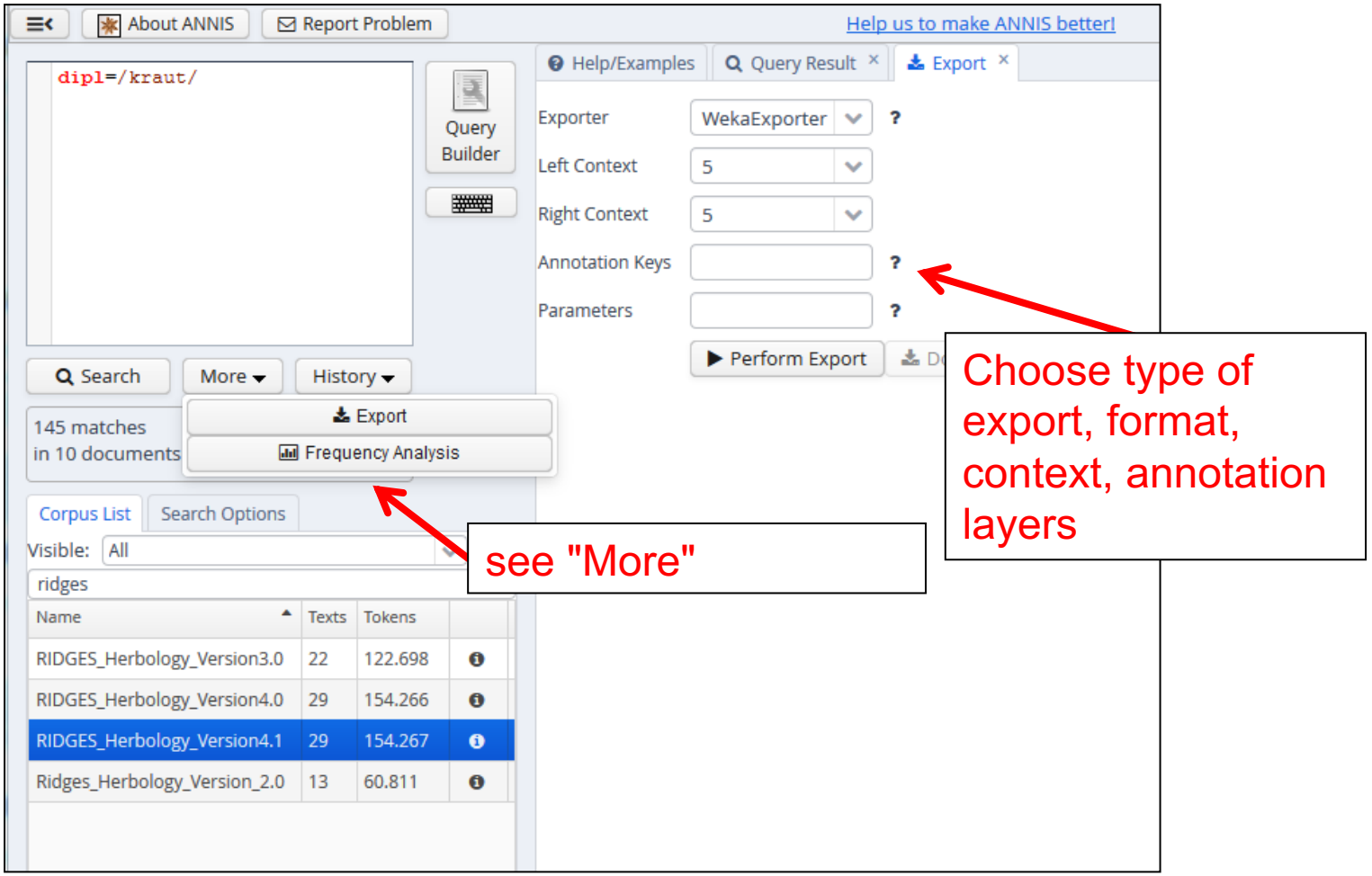
Summing up

- Querying ANNIS is based on
  - Variables (Annotation layers) and values (categories in the annotation layers), e.g.
    - Search for **exact values,** e.g. pos=/ADJA/
    - Search for **patterns**, e.g. pos=/ADJ./
  - variable-value pairs can be searched for in relation to one another, e.g.
    - Annotations can **overlap**
    - Annotations cover an **identical area**
  - Variable-value pairs can be searched for together with metadata, e.g.:
    - pos=/ADJA/ & meta::date=/1870/

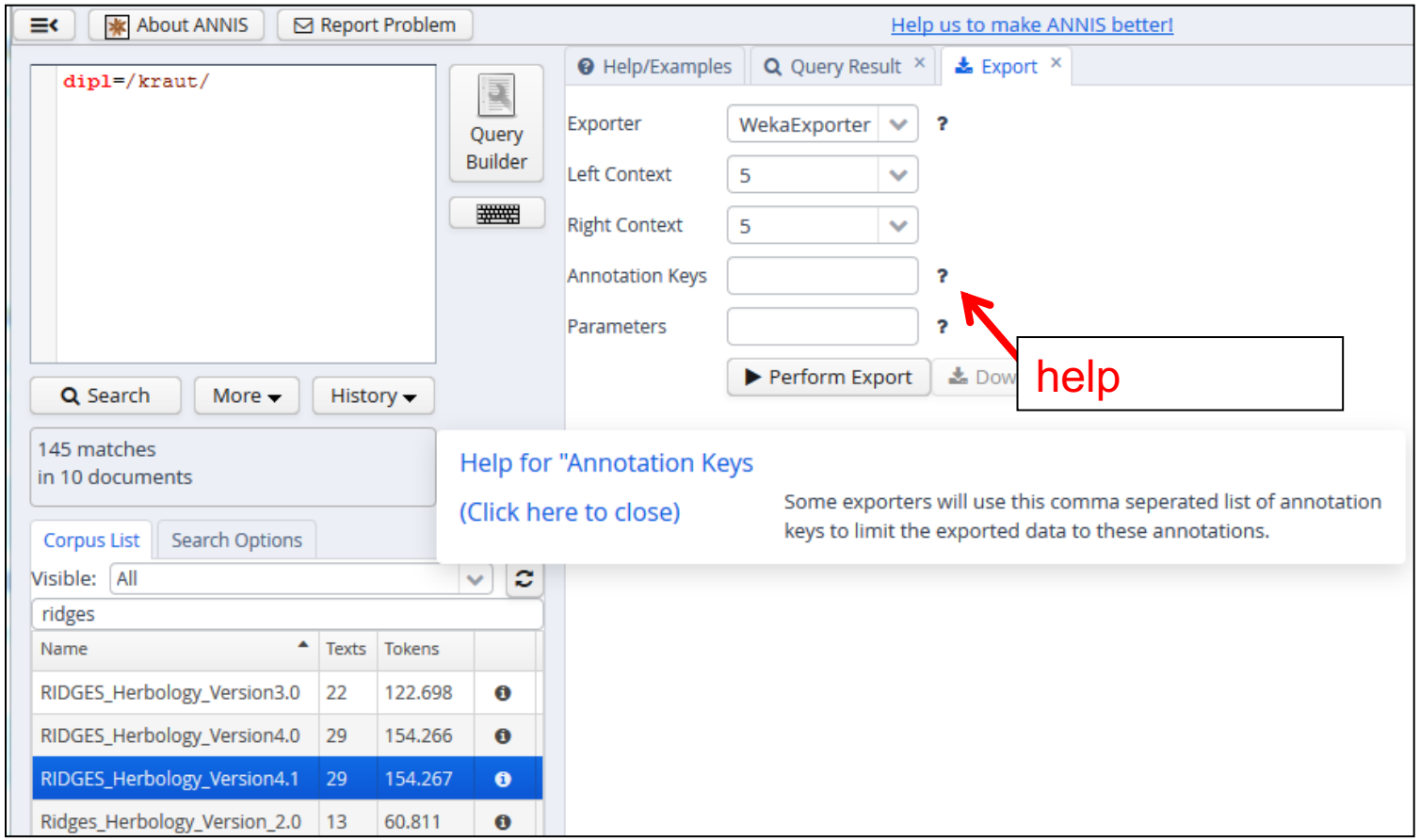
Interface

Export, Frequency analysis

Interface
Exporting hits



Interface
Exporting hits



Interface
Exporting hits

Text-Exporter

```
0. of the International Brotherhood of [Magicians] Wednesday , October 9
1. Magic Month in the United [States] . Wikinews spoke with William
2. of the International Brotherhood of [Magicians] , about the current st
3. - " Scarne on Card [Tricks] " and " Scarne on
4. and " Scarne on Magic [Tricks] " . That started me
```

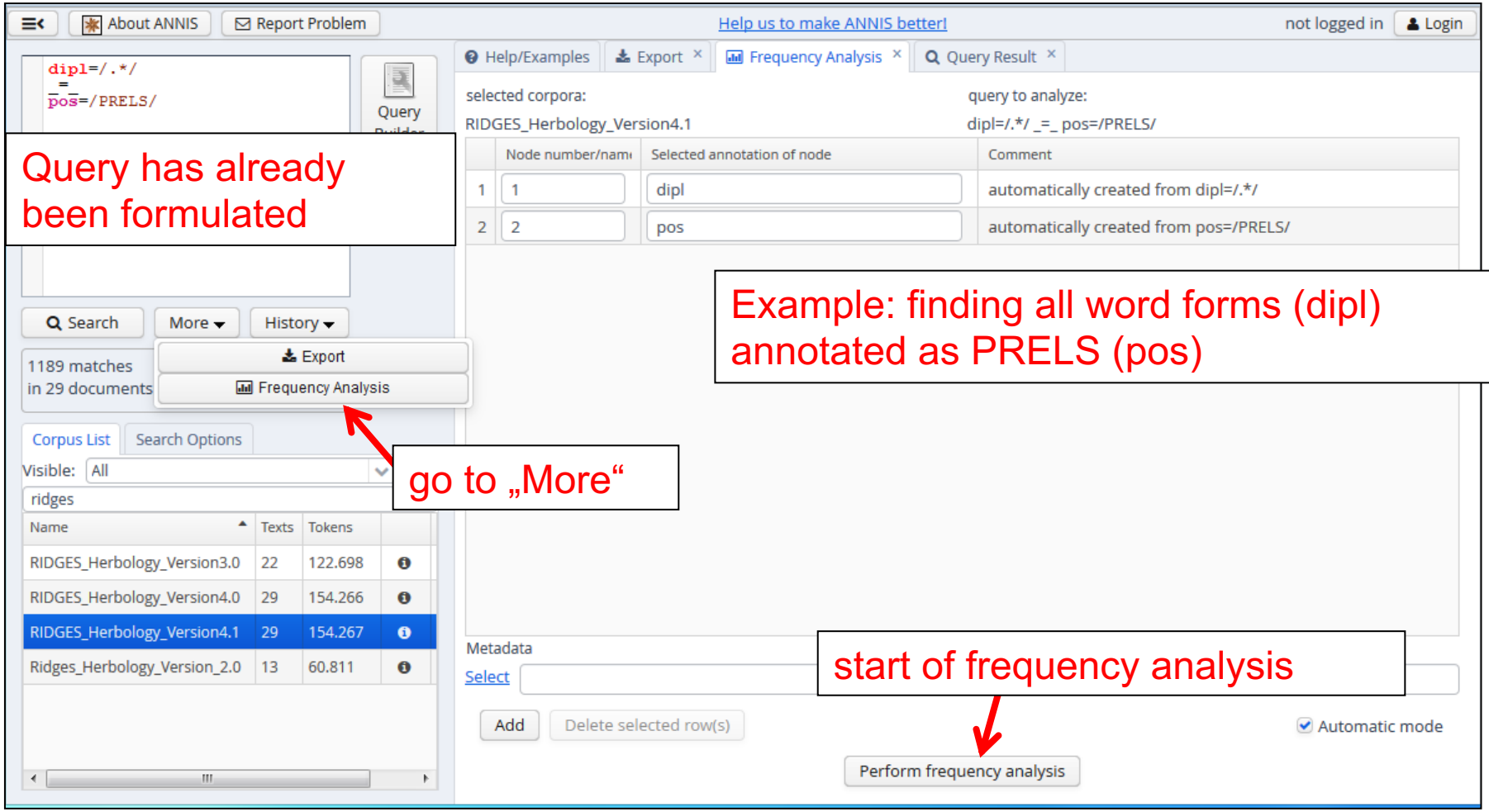
Grid-Exporter

```
0.      tok     of the International Brotherhood of Magicians Wednesday
        pos     IN[1-1] DT[2-2] NP[3-3] NP[4-4] IN[5-5] NPS[6-6] NP[7-7]
        cat     S[1-6] VP[1-6] NP[1-6] PP[1-6] NP[2-4] PP[5-6] NP[6-6] NP[7-12]
```

CVS-Exporter

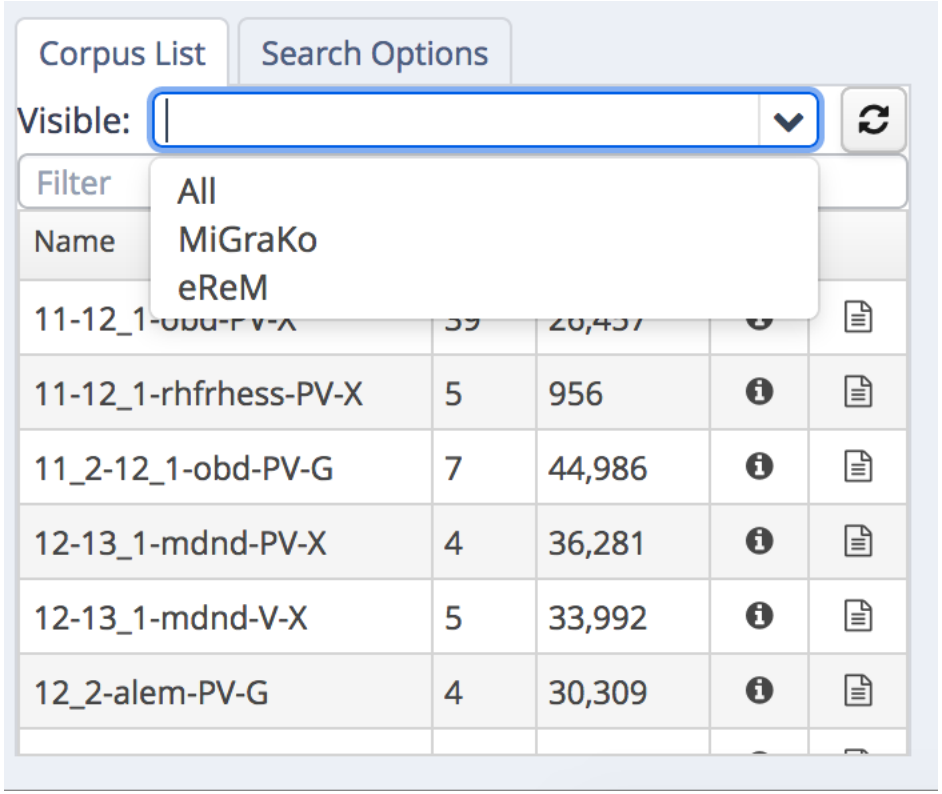
```
'11318611','the current state','NP','11318616','current','AJ0','current','JJ'
'11318686','magic','NP','11318688','magic','AJ0','magic','JJ'
'11318757','some basic tricks','NP','11318760','basic','AJ0','basic','JJ'
```

Interface
Frequency Analysis



REM

REM: Choosing a corpus



- MiGraKo: Corpus of MHD Grammar
- eReM: Additional texts

Find individual tokens

- very easy – but not always helpful, see

"so"

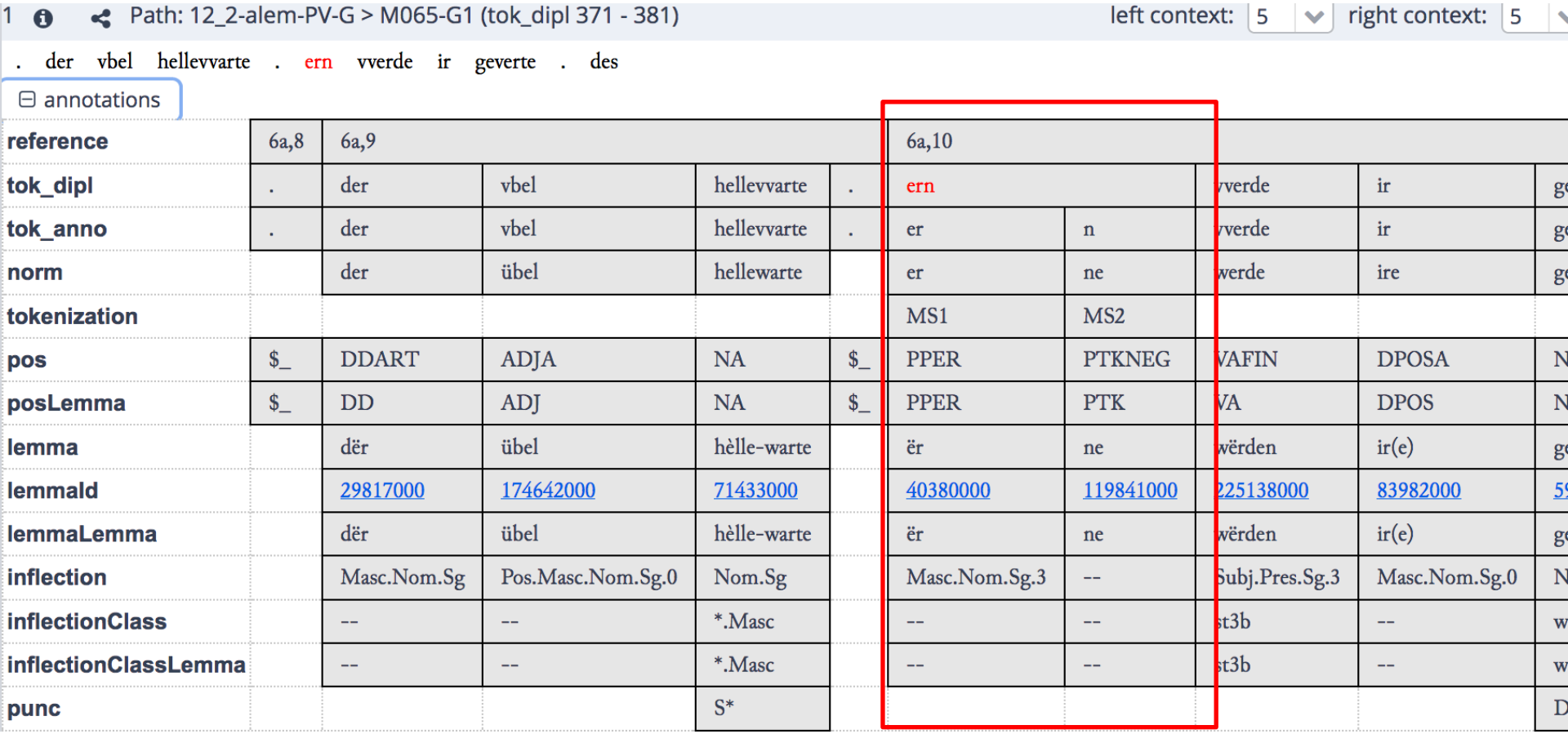
"mann"

"vrouwe" – keine Treffer

Lemma search

- Lemmas in REM are MHG lemmas (as in MHD dictionary)
- in case of doubt, check mhdwb.de to search for the right lemma (or use the query builder)

Layers in REM



reference	6a,8	6a,9				6a,10				
tok_dipl	.	der	vbel	hellevvarte	.	ern		vverde	ir	ge
tok_anno	.	der	vbel	hellevvarte	.	er	n	vverde	ir	ge
norm		der	übel	hellewarte		er	ne	werde	ire	ge
tokenization						MS1	MS2			
pos	$_	DDART	ADJA	NA	$_	PPER	PTKNEG	VAFIN	DPOSA	N
posLemma	$_	DD	ADJ	NA	$_	PPER	PTK	VA	DPOS	N
lemma		dër	übel	hèlle-warte		ër	ne	wërden	ir(e)	ge
lemmaId		29817000	174642000	71433000		40380000	119841000	225138000	83982000	59
lemmaLemma		dër	übel	hèlle-warte		ër	ne	wërden	ir(e)	ge
inflection		Masc.Nom.Sg	Pos.Masc.Nom.Sg.0	Nom.Sg		Masc.Nom.Sg.3	--	Subj.Pres.Sg.3	Masc.Nom.Sg.0	N
inflectionClass		--	--	*.Masc		--	--	st3b	--	w
inflectionClassLemma		--	--	*.Masc		--	--	st3b	--	w
punc				S*						D

Example: Negation in MHG

- We want to investigate the development of negation in MHG:

- "double" negation with *en a*nd *niht* vs. simple negation
- What do we have to look for?

Example: Conversion in MHD

- We want to find all nominalized infinitives in MHG (e.g. NHG *das Singen*, *das Tanzen*)
- We can use the pos and posLemma annotation for that purpose.

Finding individual tokens

- Pay attention for spelling variants (upper-/lowercase!)
- *so* only finds lowercase attestations, *So* only uppercase ones
- For finding all variants, we use regular expressions

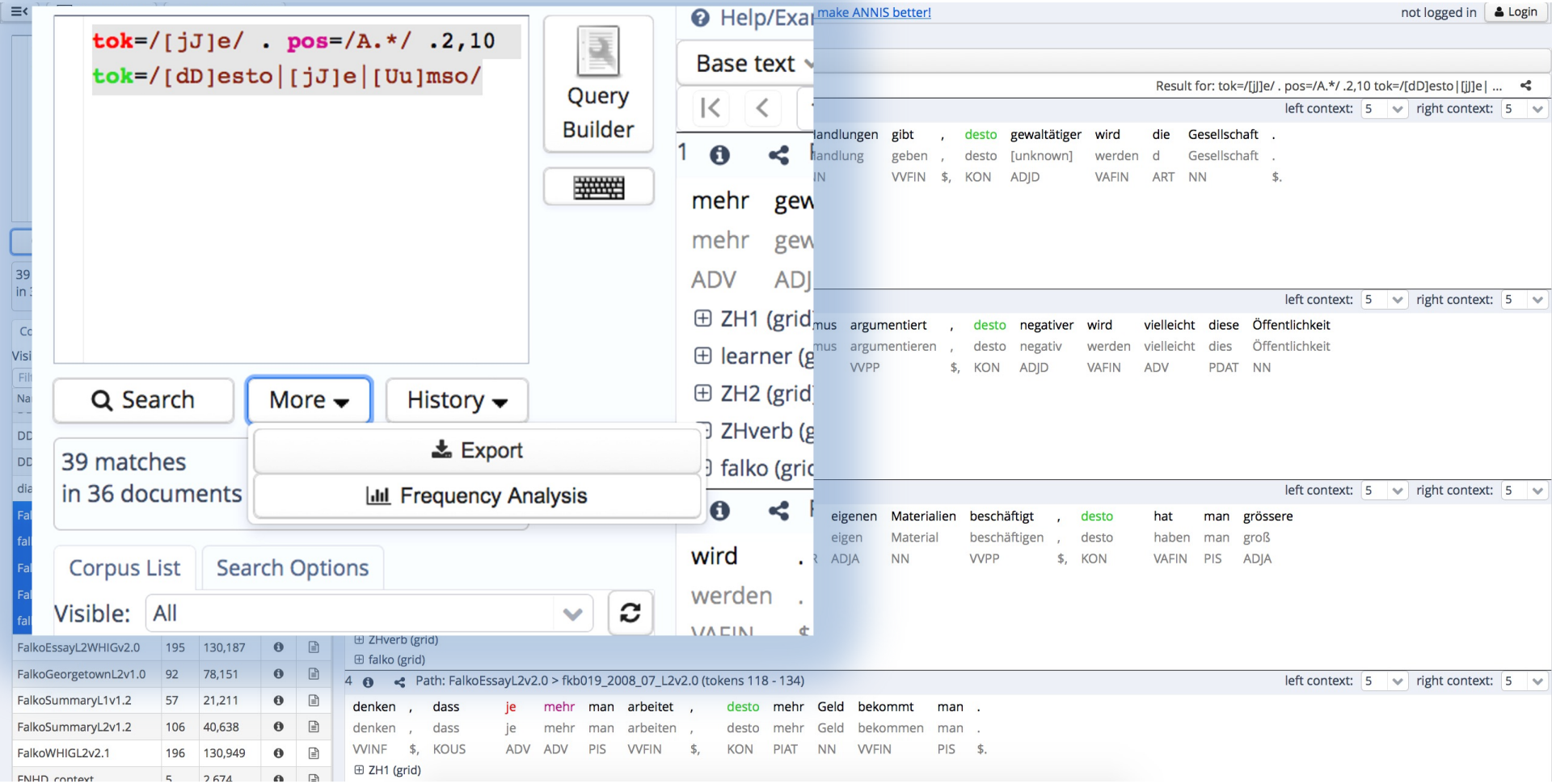
Finding individual tokens

"so" | "So" | "SO"

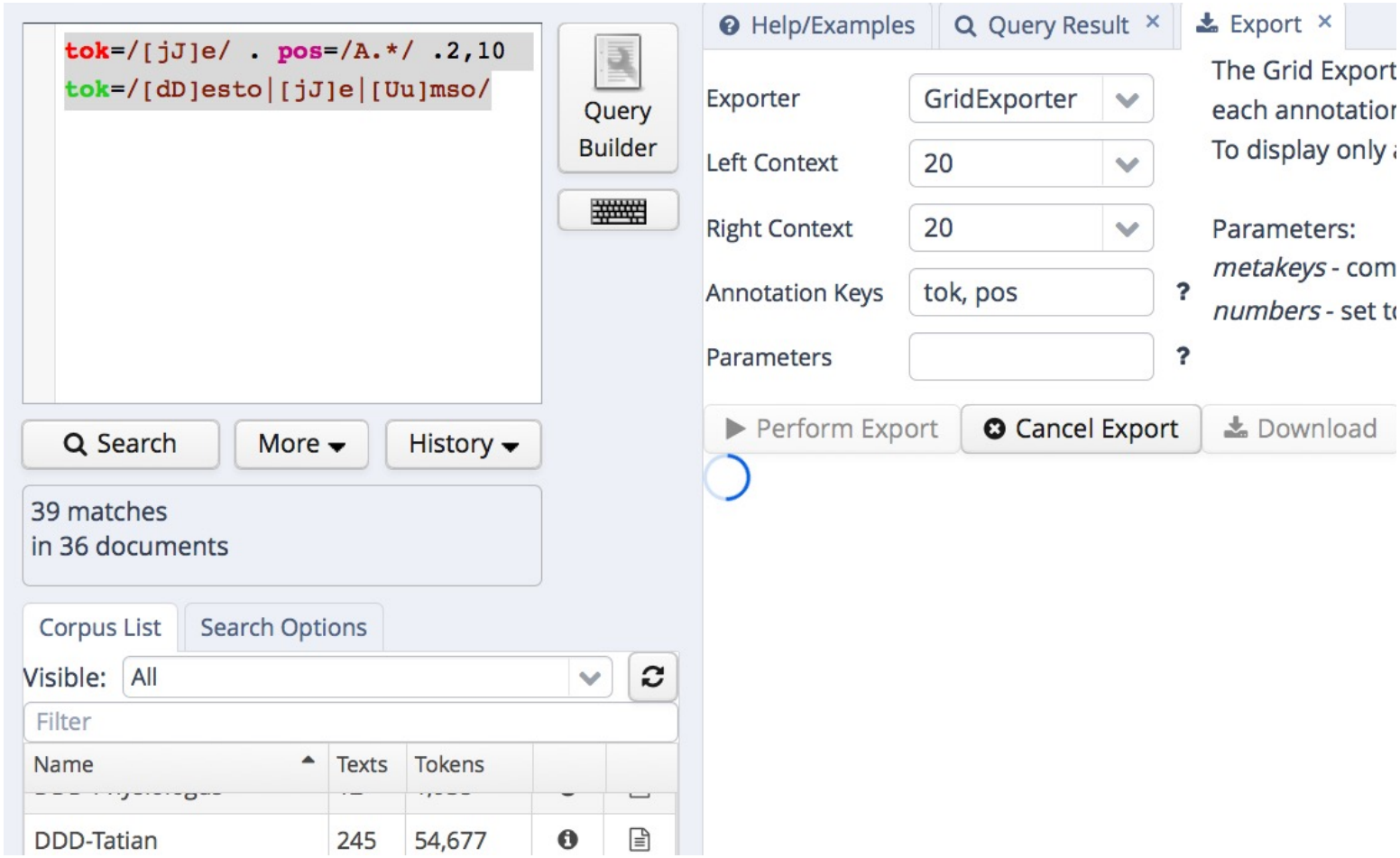
Finding individual lemmas

- Important: Attributes can have different names!
- Example: The layer that is called "lemma" in virtually all other corpora is called "hw" in the British National Corpus.

From result to export file

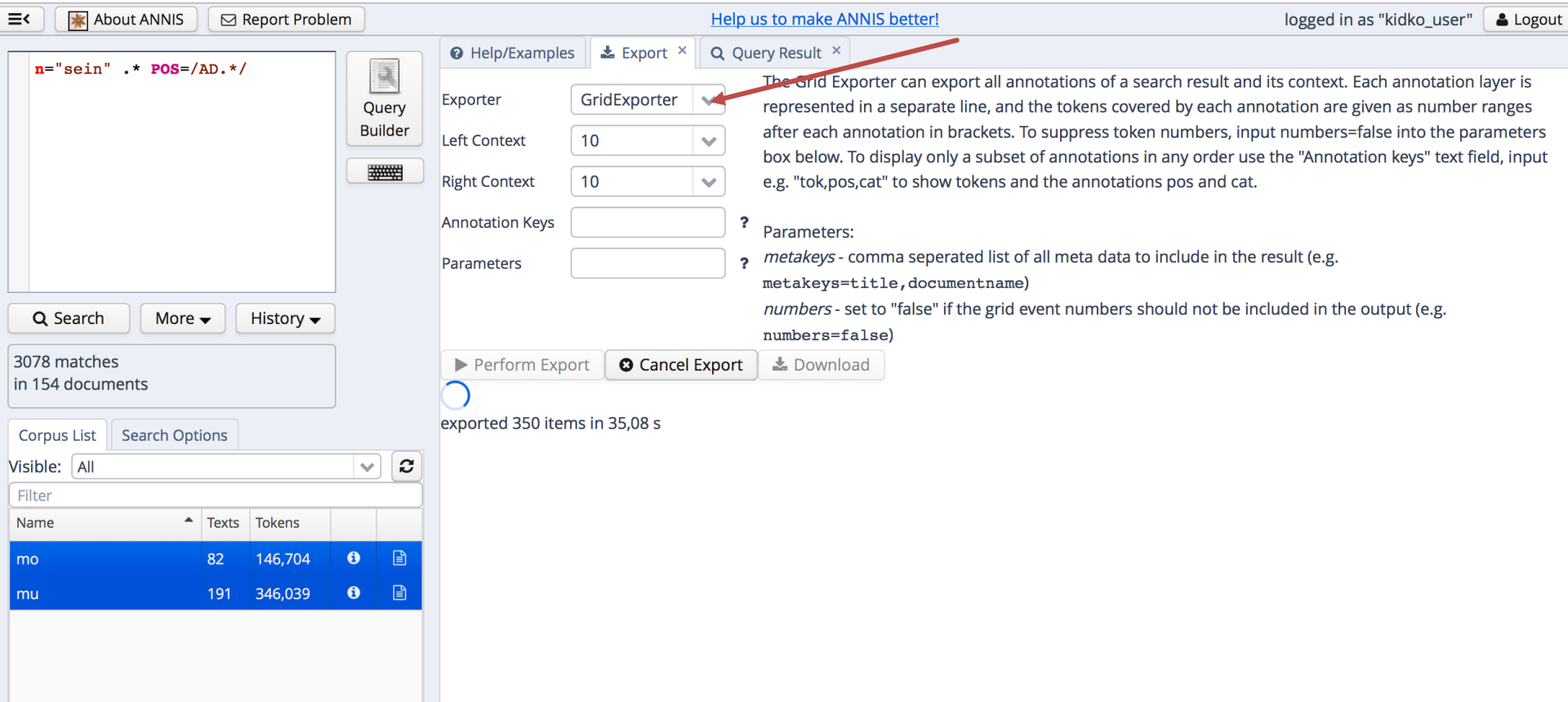


From result to export file

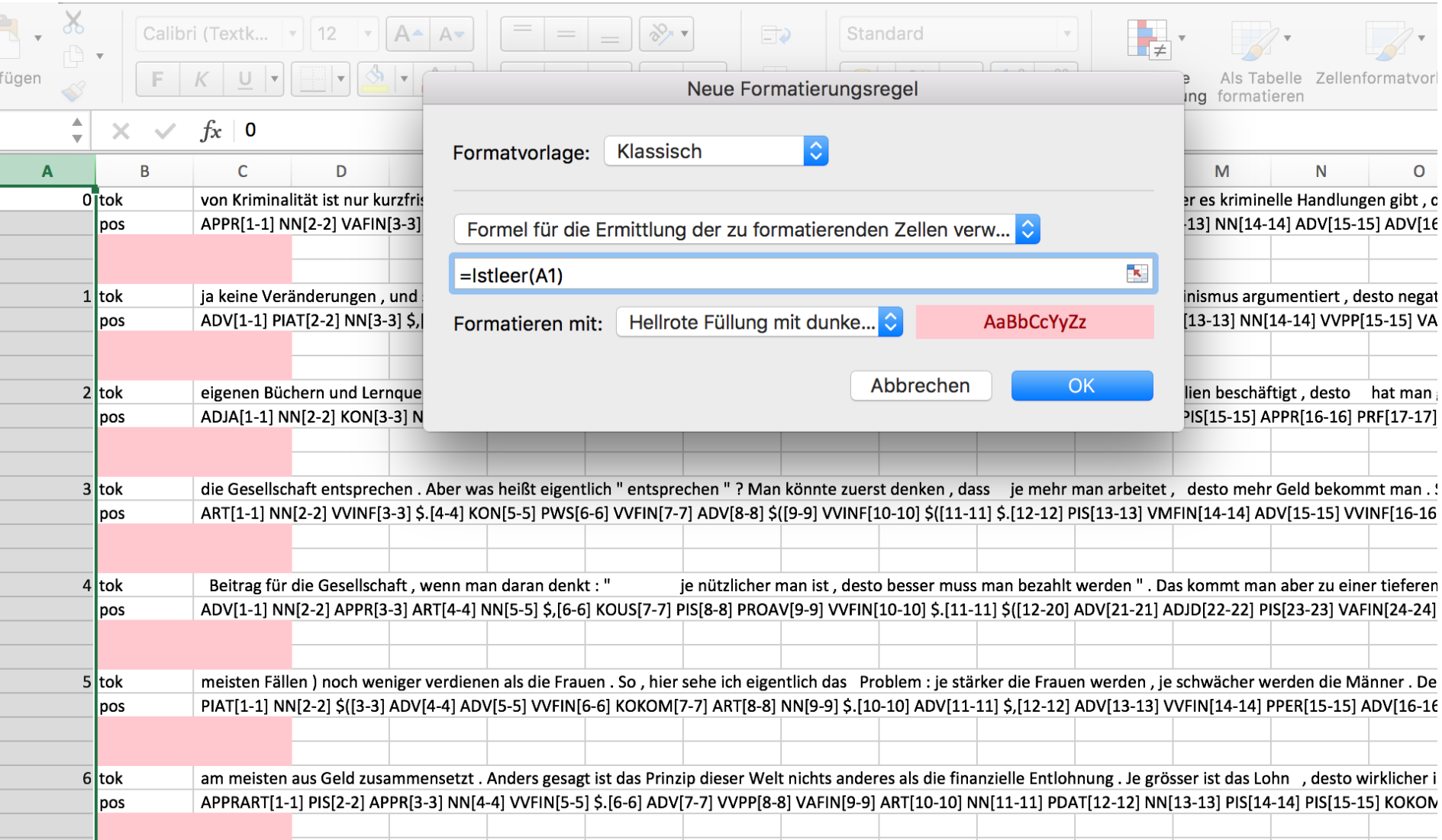


Export

- not all exporters work for all ANNIS corpora ☹



Using conditional formatting in Excel

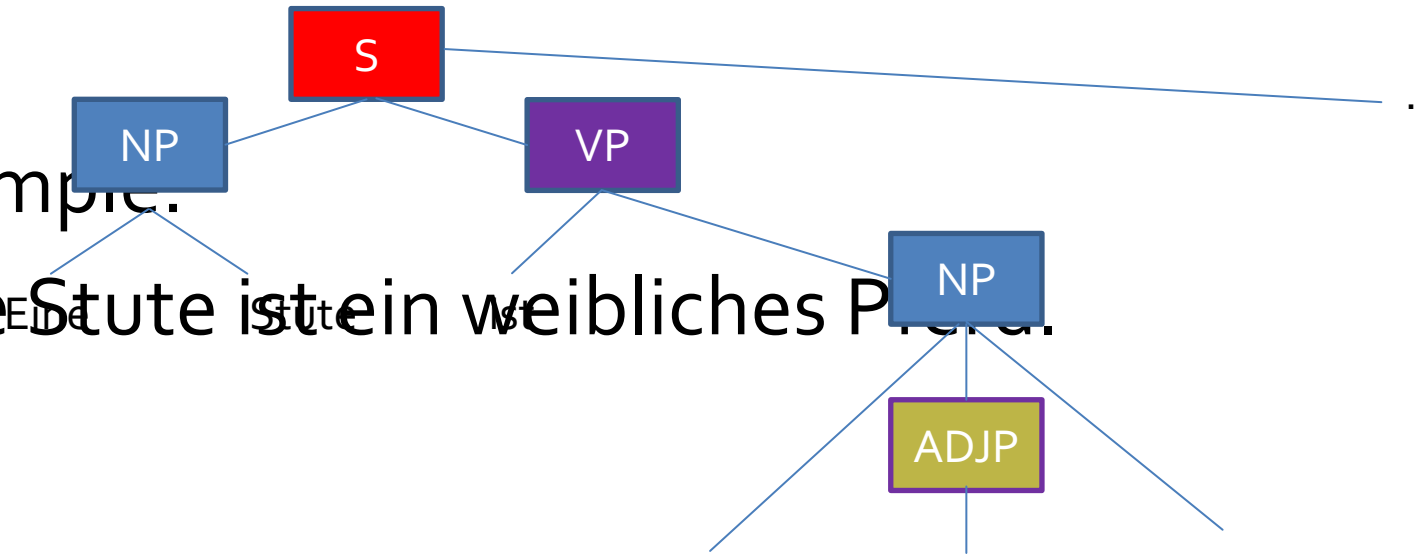


Treebanks

- syntactically parsed corpora
- syntax represented via syntactic trees
- a syntactic tree consists of **edges** and **nodes**
- widespread tagset for German: Tagset of the TIGER treebank (Albert et al. 2004)

Constituent vs. dependency structure

- two key syntactic formalisms
- Constituent structures: Words are grouped into **phrases**, with one word functioning as **head** in each phrase.



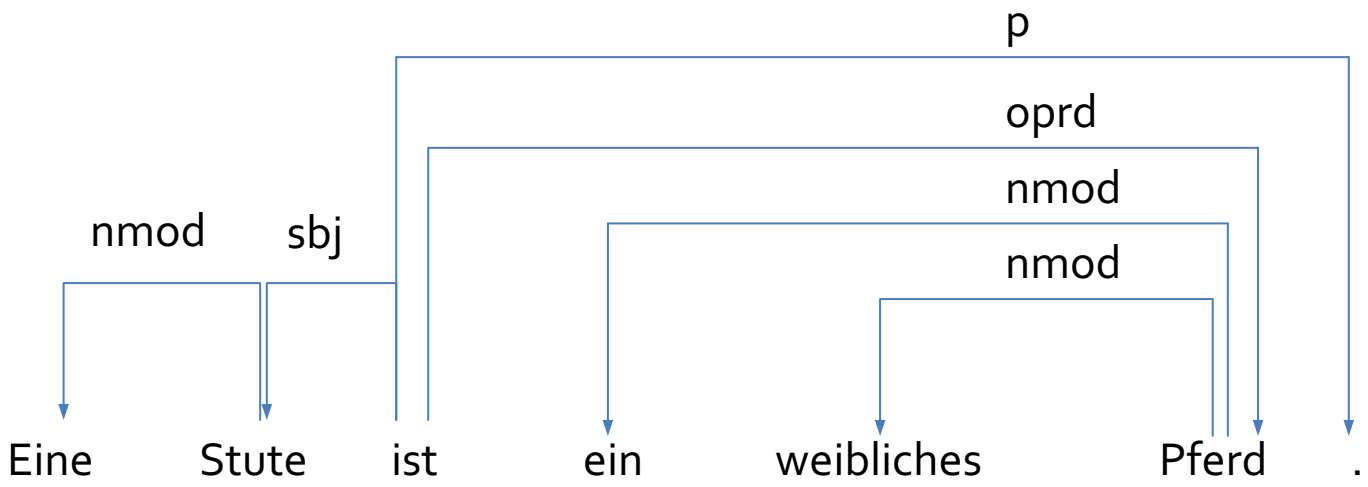
Example:

Eine Stute ist ein weibliches Pferd.

(based on Kübler & Zinsmeister 2015: 57)

Constituent vs. dependency structure

- Dependency structure: Relationships between word pairs (**head** and **dependent**)



(based on Kübler & Zinsmeister 2015: 58)

Constituent vs. dependency structure

- both analyses assume hierarchical structuring of sentences
- but: in constituent structure analysis constituents (i.e. abstract entities) are ordered hierarchically
- dependency analysis, by contrast, is constrained to the words themselves.
- Constituent structure represents syntactic categories, dependency structure takes syntactic functions into account.

(see Lemnitzer & Zinsmeister 2015)

Hybrid models

- some projects use a mixture of both variants
- advantage: being able to take both syntactic categories and syntactic functions into account.

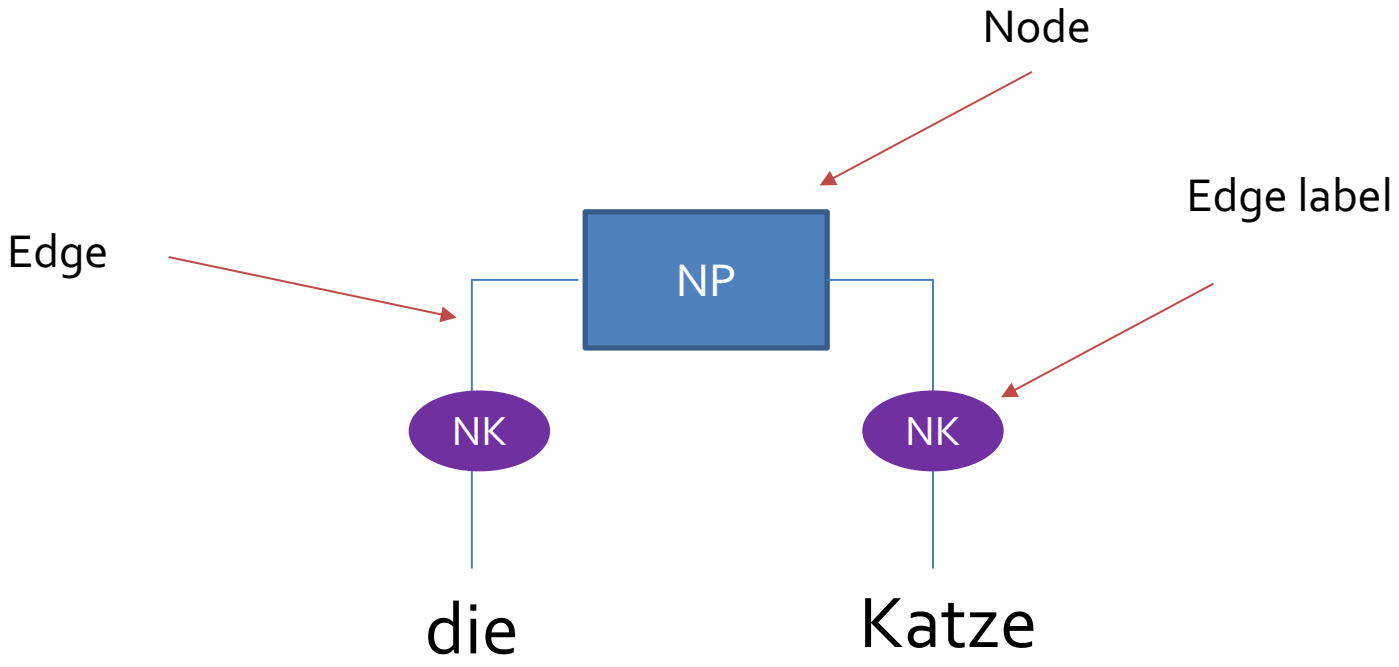
Terminology

- "Constituent": category-neutral description of a phrase
- A phrase is always assigned a specific category, e.g. VP, NP
- Apart from phrases, syntactic analysis has started to work increasingly with chunks
- Chunks are prosodic units of speech (recognizable by small pauses in spoken language) or multi-word units that belong together.

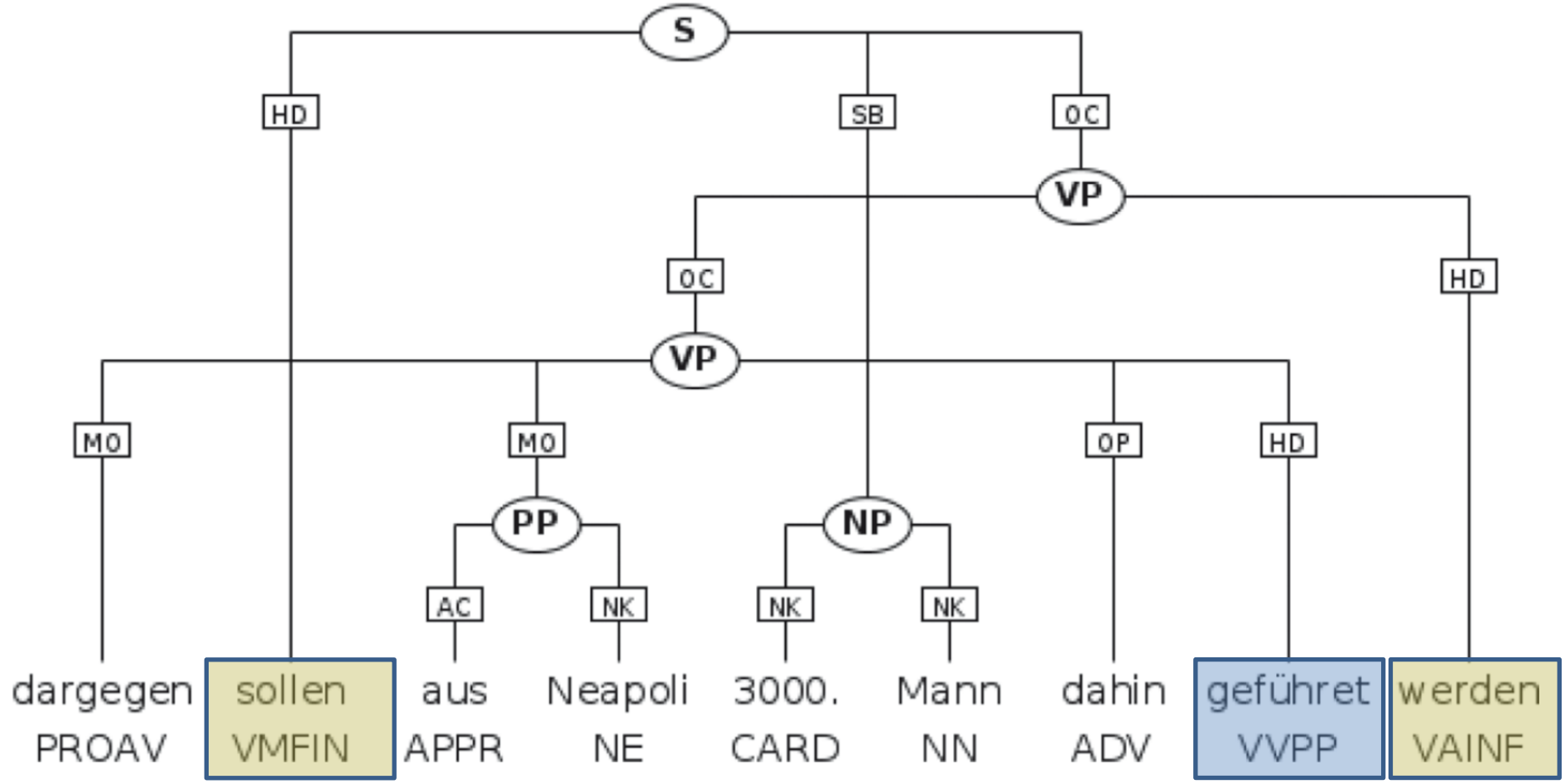
Historical treebanks of German

- Mercurius corpus
  - Newspaper texts from two different years: *Annus Christi* 1597, *Mercurius* 1667
- Deutsche Diachrone Baumbank
  - c. 2500 Tokens from two texts each for OHD, MHG, ENHG

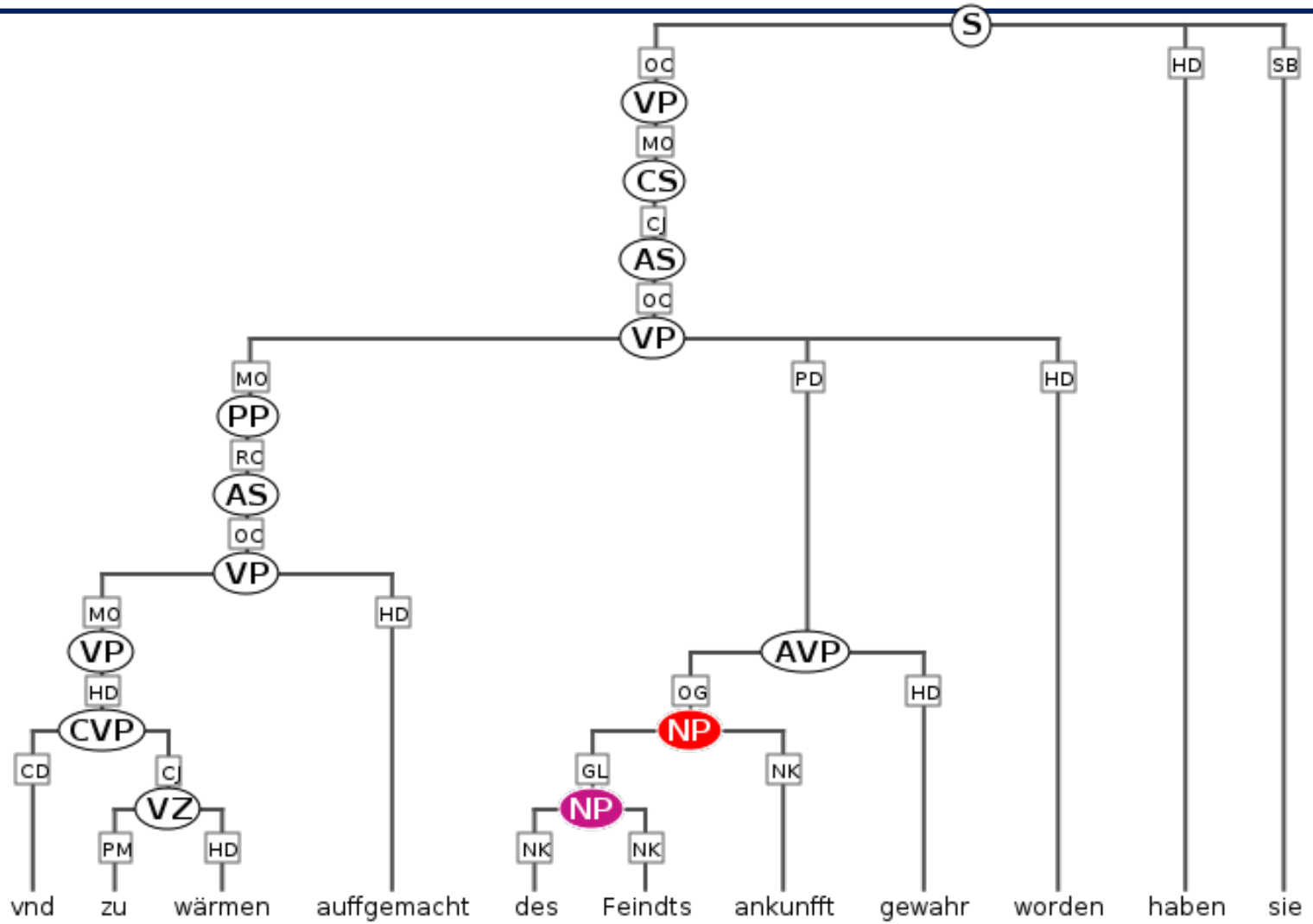
Syntax tree: Example



Crossing edges



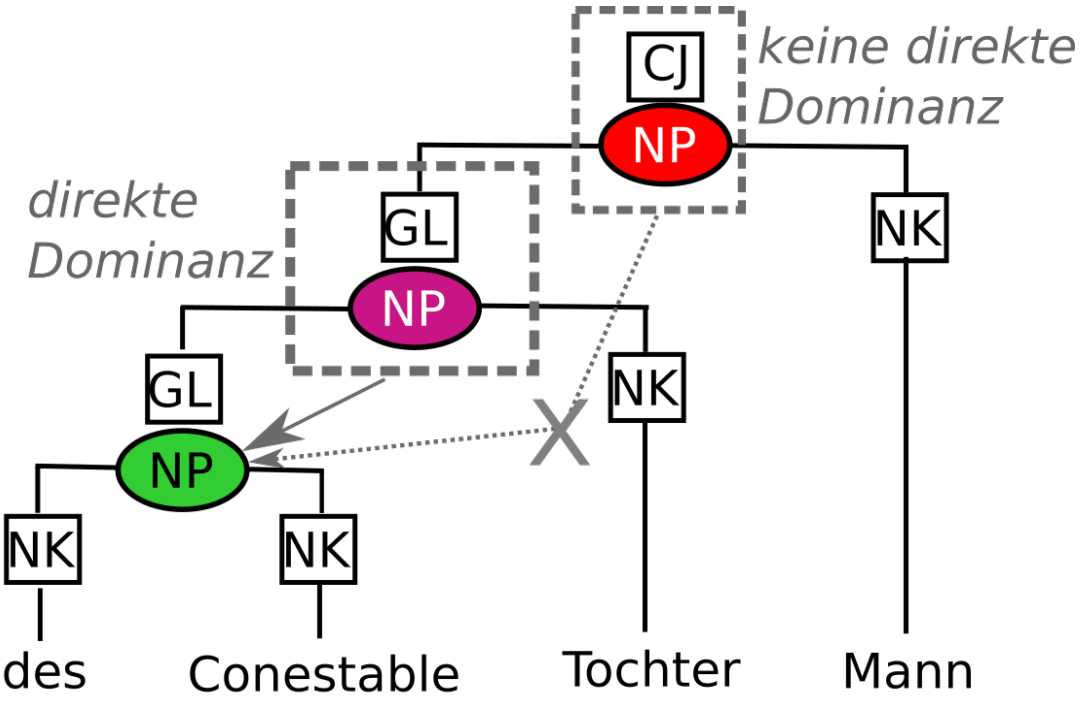
Example: Mercurius



Example: Mercurius

- node > [label="GL"] cat="NP"

any node — direct dominance — edge label — noun phrase



Example: Mercuris

What can we find using the following queries?

- node > cat="NP",
- node >[label="GL"] cat="NP" & meta::doc="Mercurius-1667"
- node >* cat="NP"