ANNIS Einführung
RIDGES Herbology 4.1

Ridges
Register in Diachronic German Science

Laura Perlitz und Carolin Odebrecht
Humboldt-Universität zu Berlin

Korpus
RIDGES Herbology

Ridges
Register in Diachronic German Science

- Wiederholung
  - Kräuterkundekorpus
    - 15.-19. Jahrhundert
    - verschiedene Dialekte aus dem deutschsprachigen Raum
    - Arbeit von drei Seminaren aus B.A. und M.A. von ca. 50 Studenten
  - Projekthomepage
    - http://korpling.german.hu-berlin.de/ridges/documentation_v4.1_en.html
  - Annotationsrichtlinie
    - http://korpling.german.hu-berlin.de/ridges/download/pubs/annotationGuidelines_v4.1.pdf

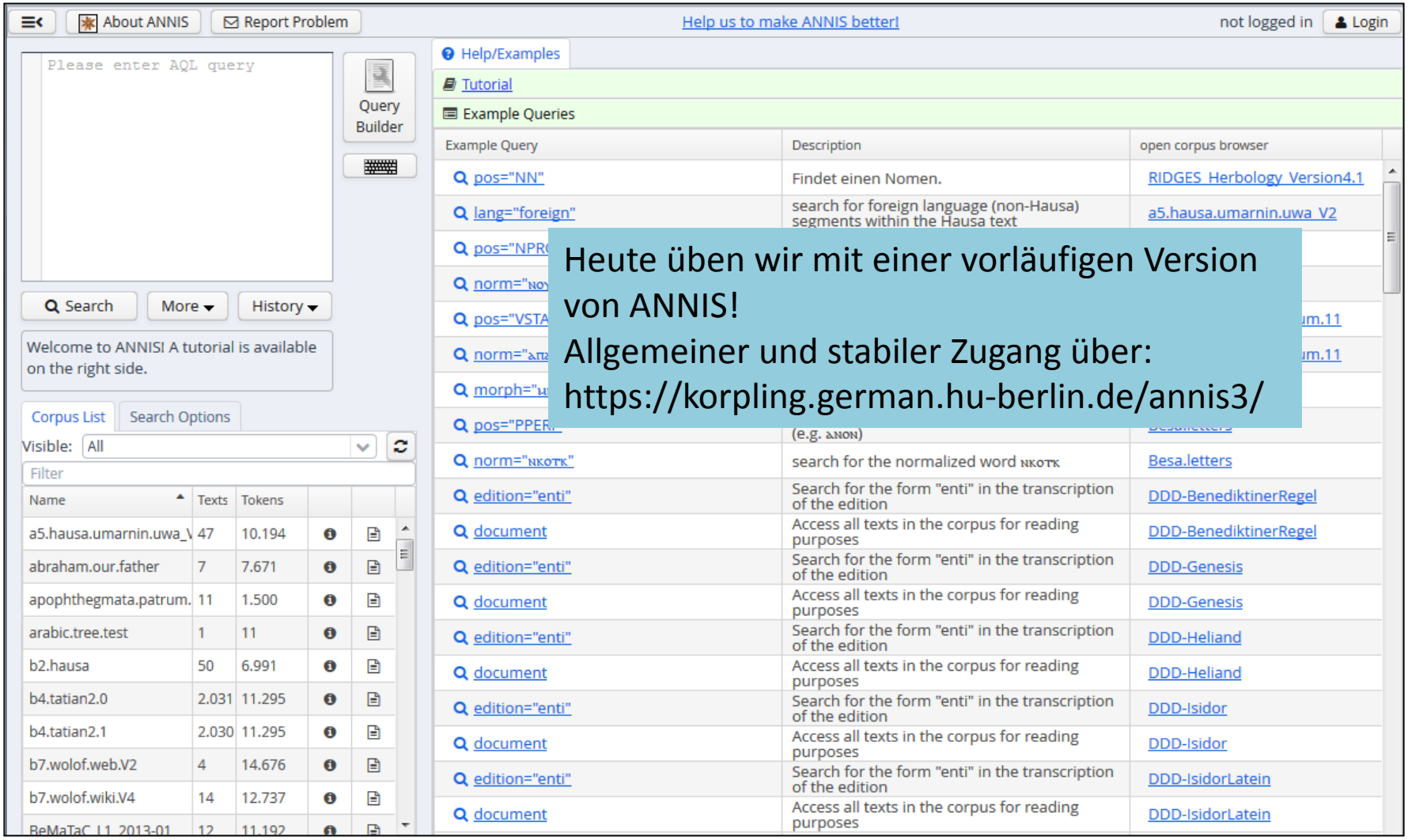
ANNIS
Such- und Visualisierungstool

- webbasiertes Suchtool für Korpora
  - Projekthomepage http://annis-tools.org/
  - Zugang auf das ANNIS-Tool https://korpling.german.hu-berlin.de/annis3

➢ Was kann ANNIS?
  - ➢ findet Annotationen in einem Korpus
  - ➢ findet mehrere Annotationen in bestimmter Relation zu einander
  - ➢ gibt die gesuchten Annotationen als Treffer in vielfältigen Visualisierungen aus
  - ➢ exportiert diese Treffer
  - ➢ u.v.m.

➢ Hier im Seminar lernen wir mit dem Korpus RIDGES Version 4.1, wie man ANNIS benutzt!

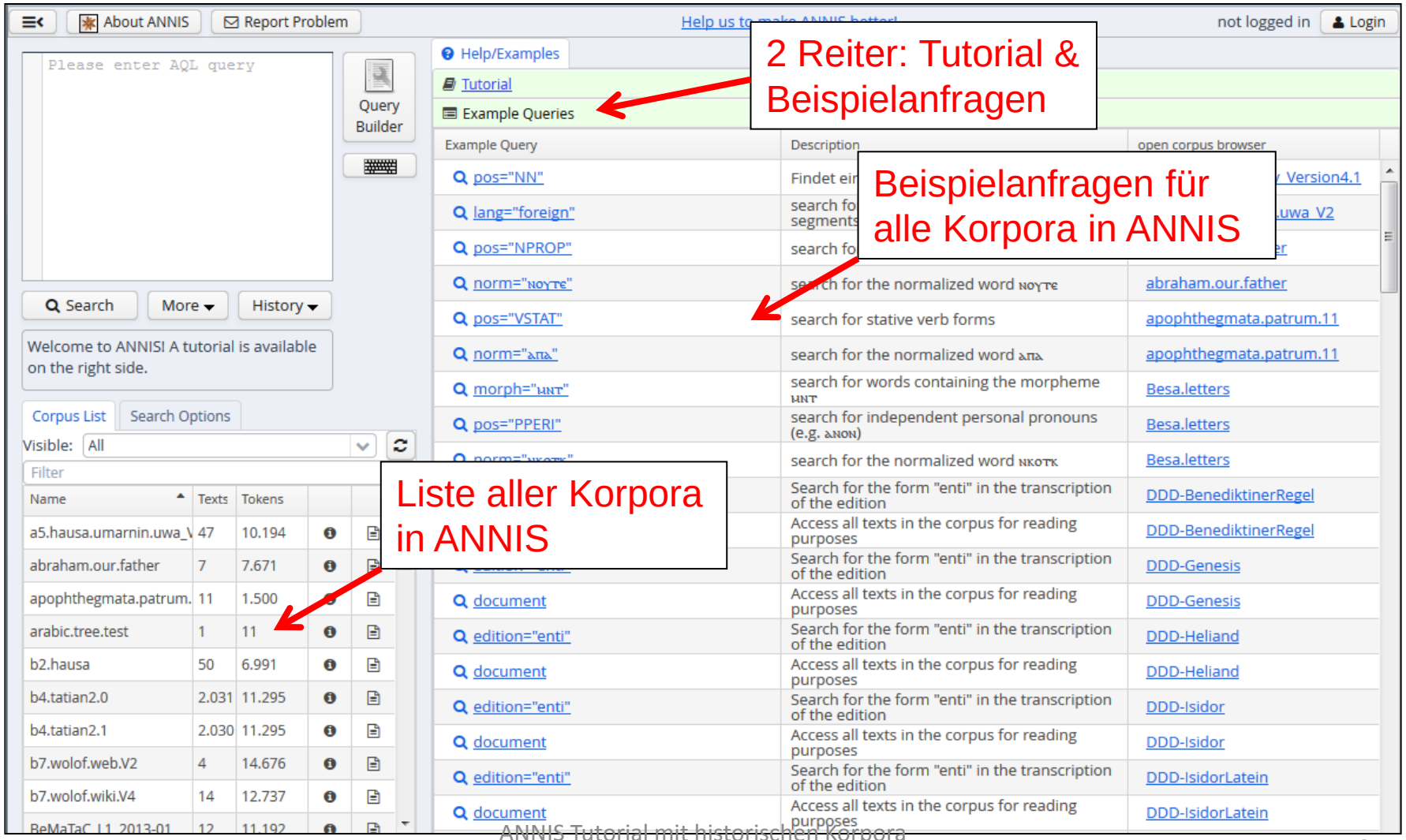
https://korpling.german.hu-berlin.de/annis3-snapshot/



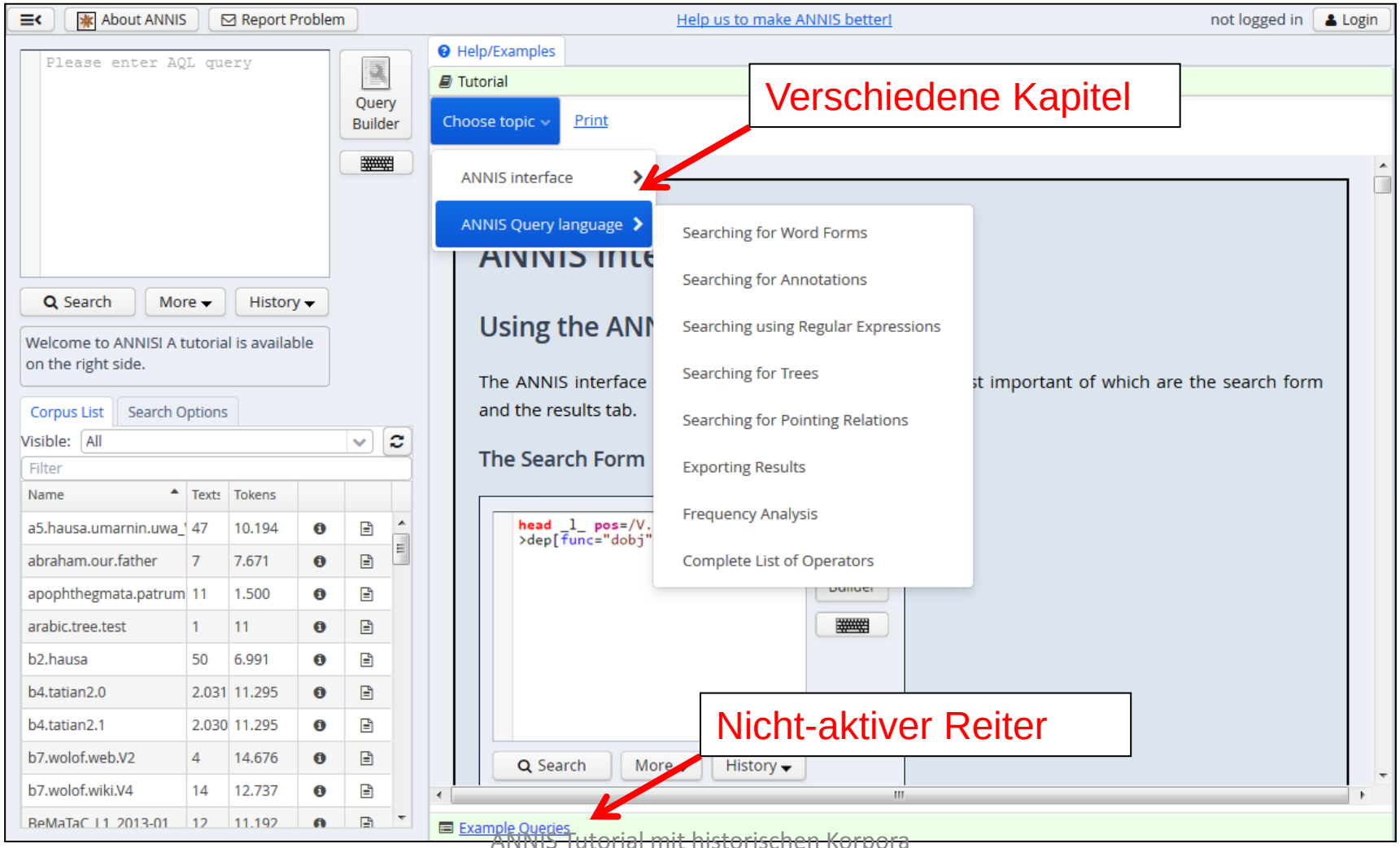
Interface

Startseite, Hilfestellung, Query Fenster

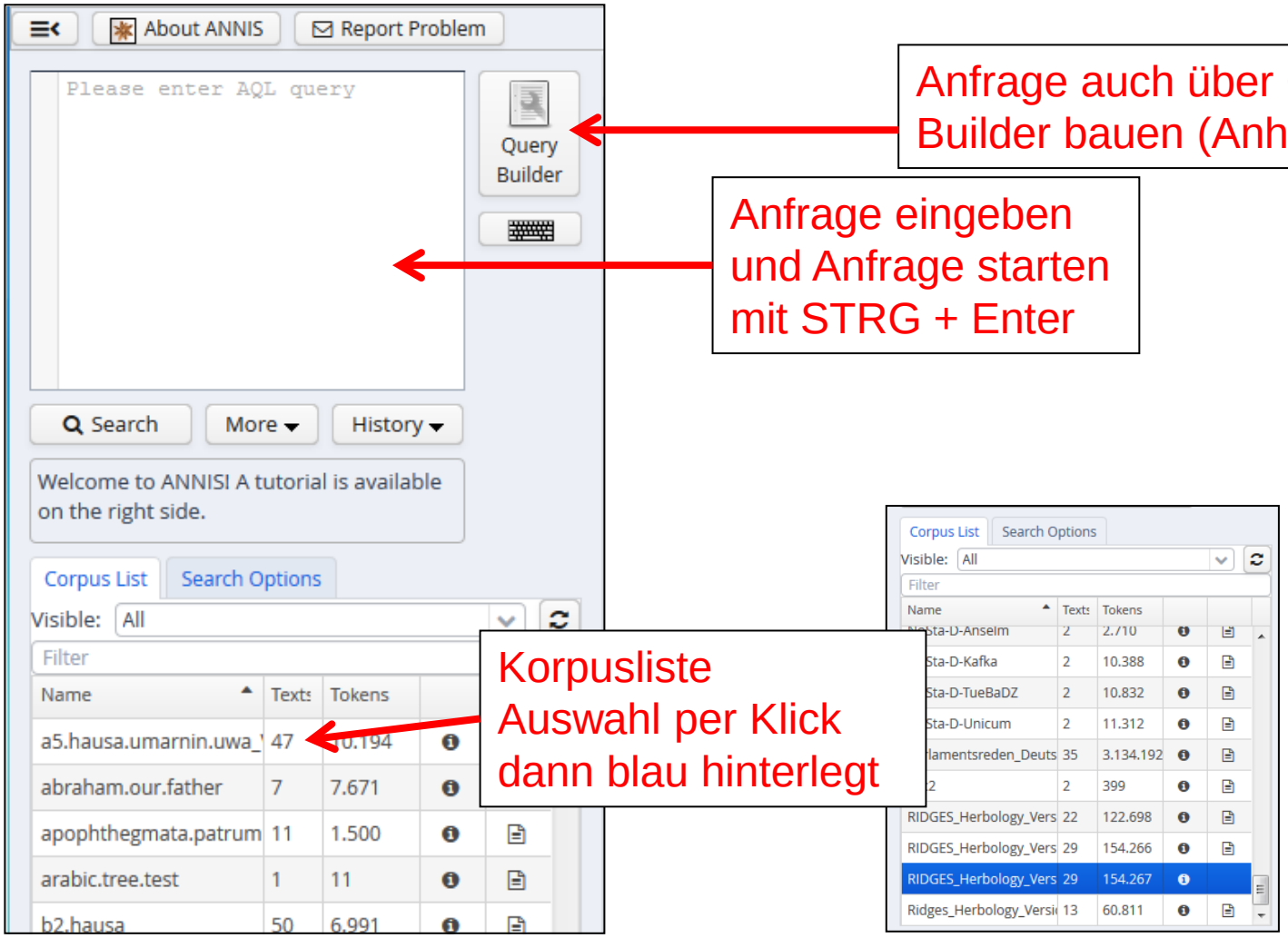
Interface Startseite



Interface Tutorial



Interface
Query-Fenster



Anfrage auch über Query Builder bauen (Anhang)

Anfrage eingeben und Anfrage starten mit STRG + Enter

Korpusliste
Auswahl per Klick
dann blau hinterlegt

ANNIS Query Language

- oder: Wie sage ich ANNIS, was ich suche?
- für die Suche nach Annotationen gelten zwei Prinzipien:
  - Variable-Wert-Paar (VW-Paar)
  - Relationen zwischen VW-Paaren

Token

1) Als Token bezeichnet man häufig die kleinste (technische) Einheit in einem Korpus!
2) Ein Token entspricht oft (aber nicht immer) einem orthographischen Wort oder Satzzeichen!
3) Nach diesen Einheiten kann man in ANNIS suchen.

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

Prinzip I: Variable-Wert-Paar

1) Voraussetzung ist das Vorhandensein einer Ebene namens „dipl“. (Metadaten!)
2) Erwartetes Ergebnis ist es, exakt alle Vorkommen dieser Zeichenkette in „dipl“ im ausgewählten Korpus zu finden.

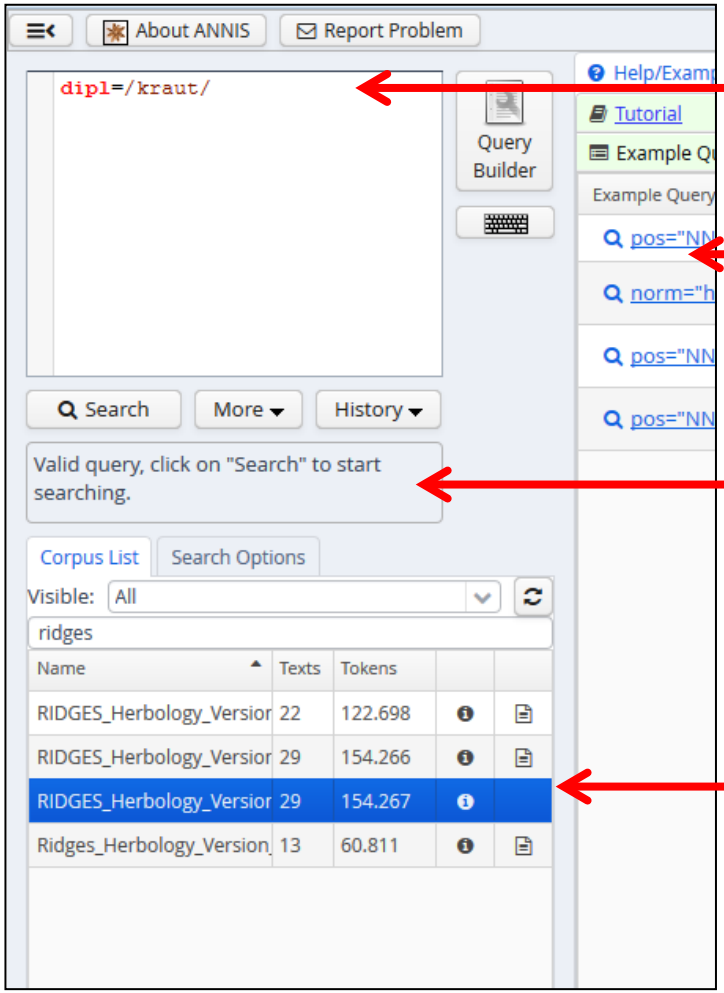
dipl= /kraut/

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 Suchanfrage



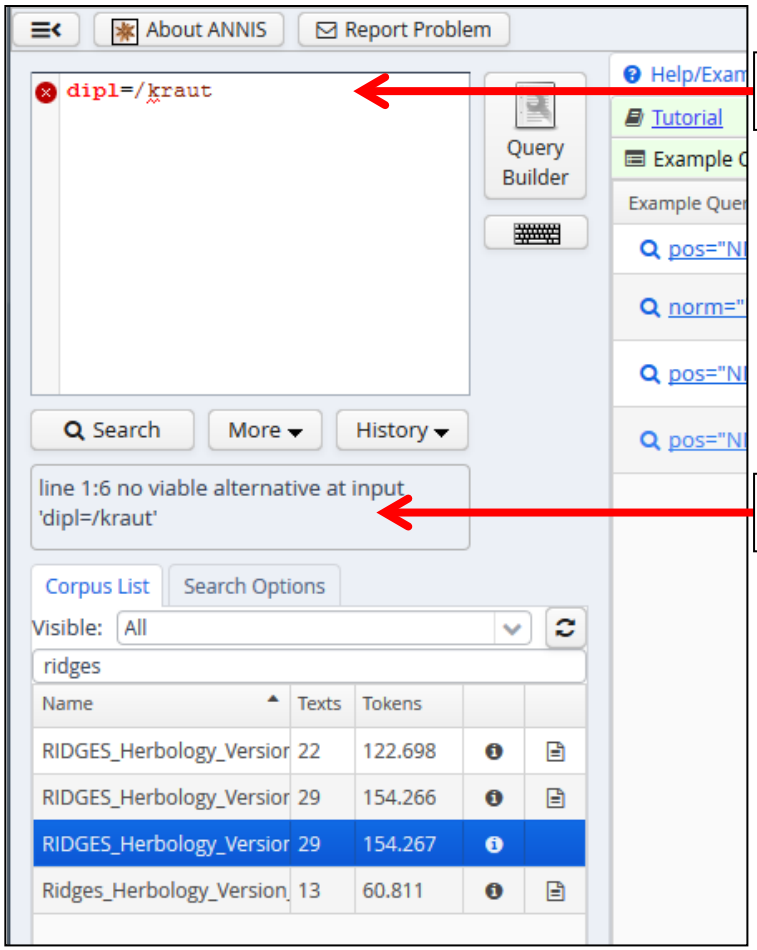
Anfragesprache AQL

Beispielanfragen für RIDGES-Korpus

Zeigt an, ob eine Suchanfrage valide ist (bevor man die Anfrage abschickt)

Für jede Anfrage muss mindestens ein Korpus ausgewählt sein

Interface
Fehler in der Suchanfrage

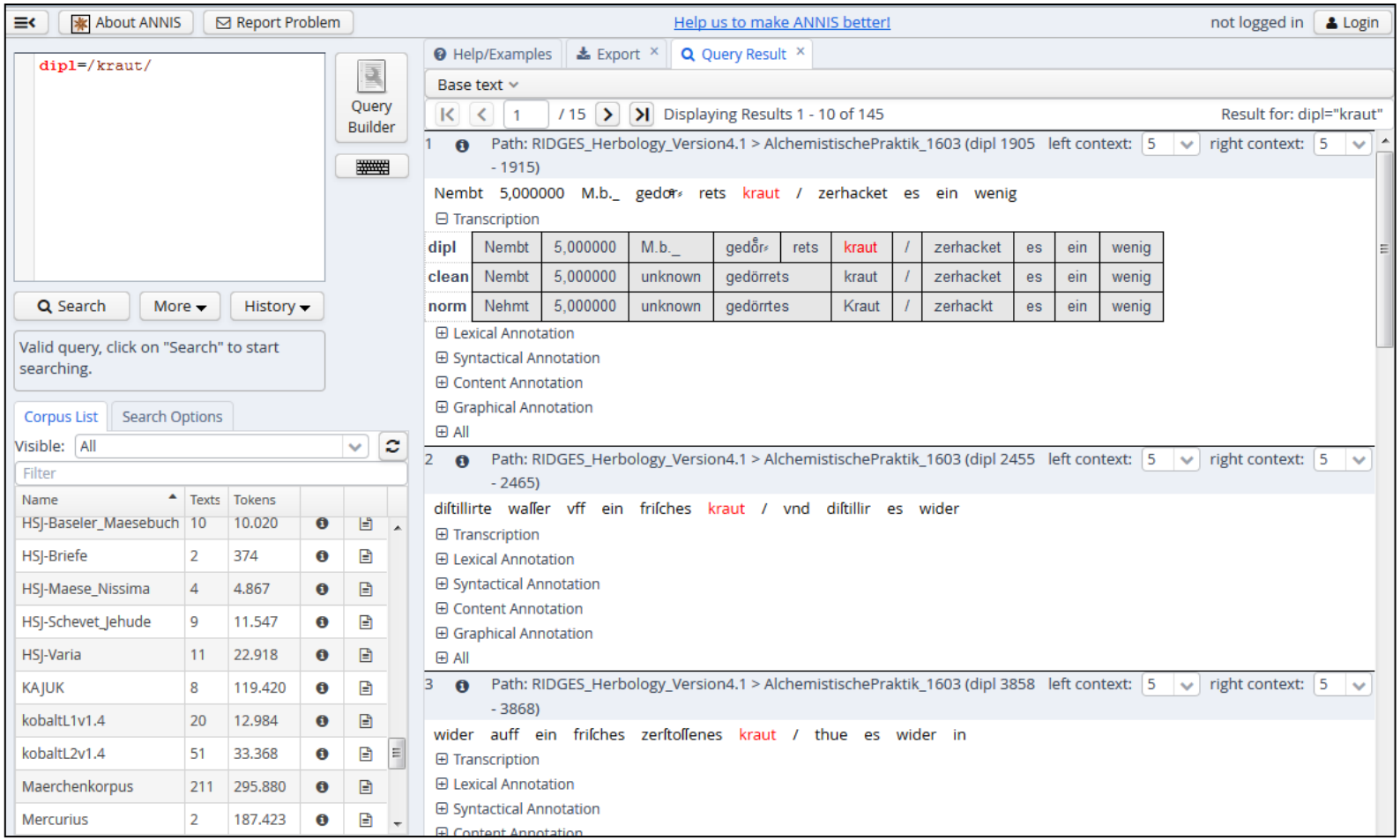


Anfragesprache AQL mit Fehler!

Feedback zum Fehler!

Variable-Wert-Paar

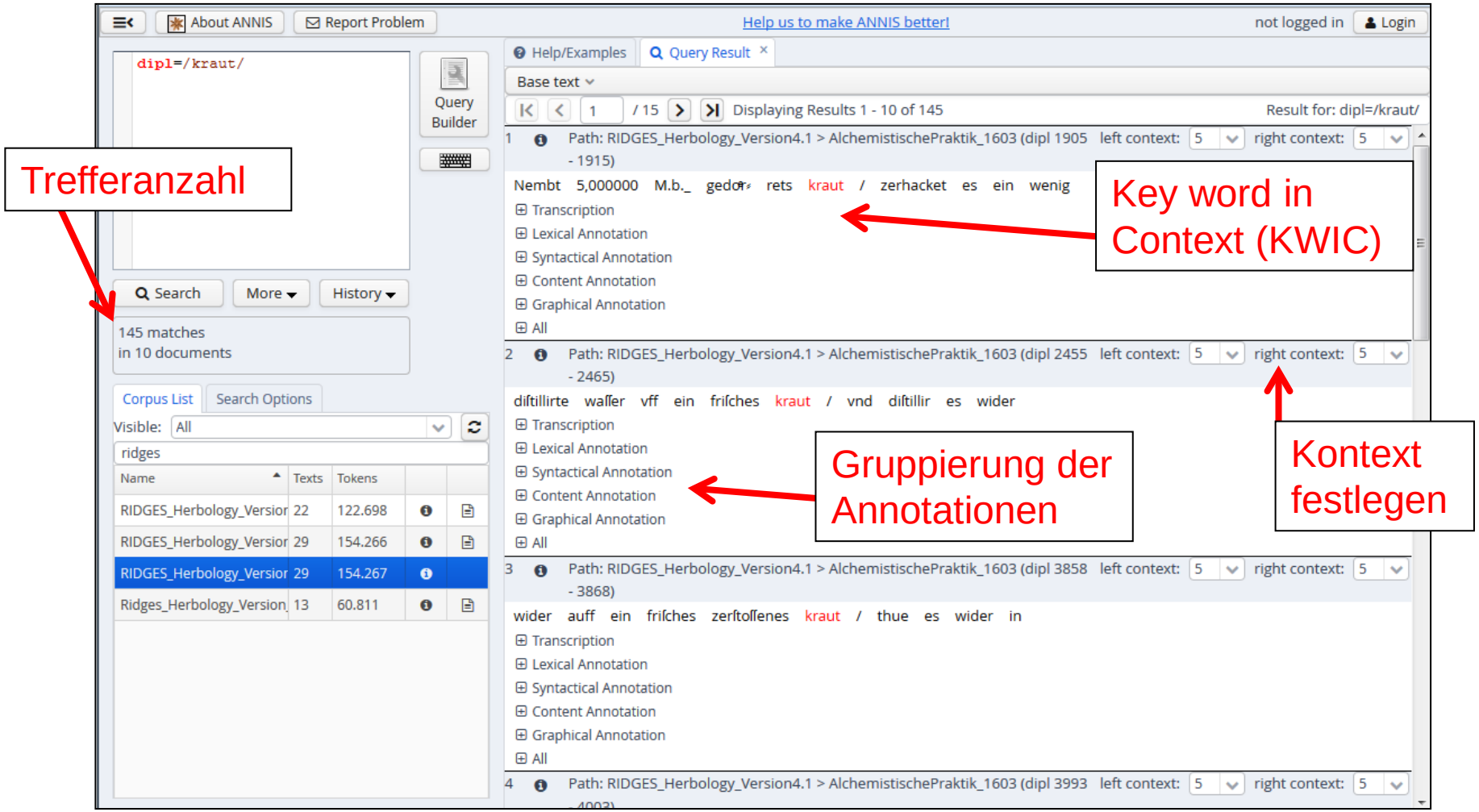
dipl= /kraut/



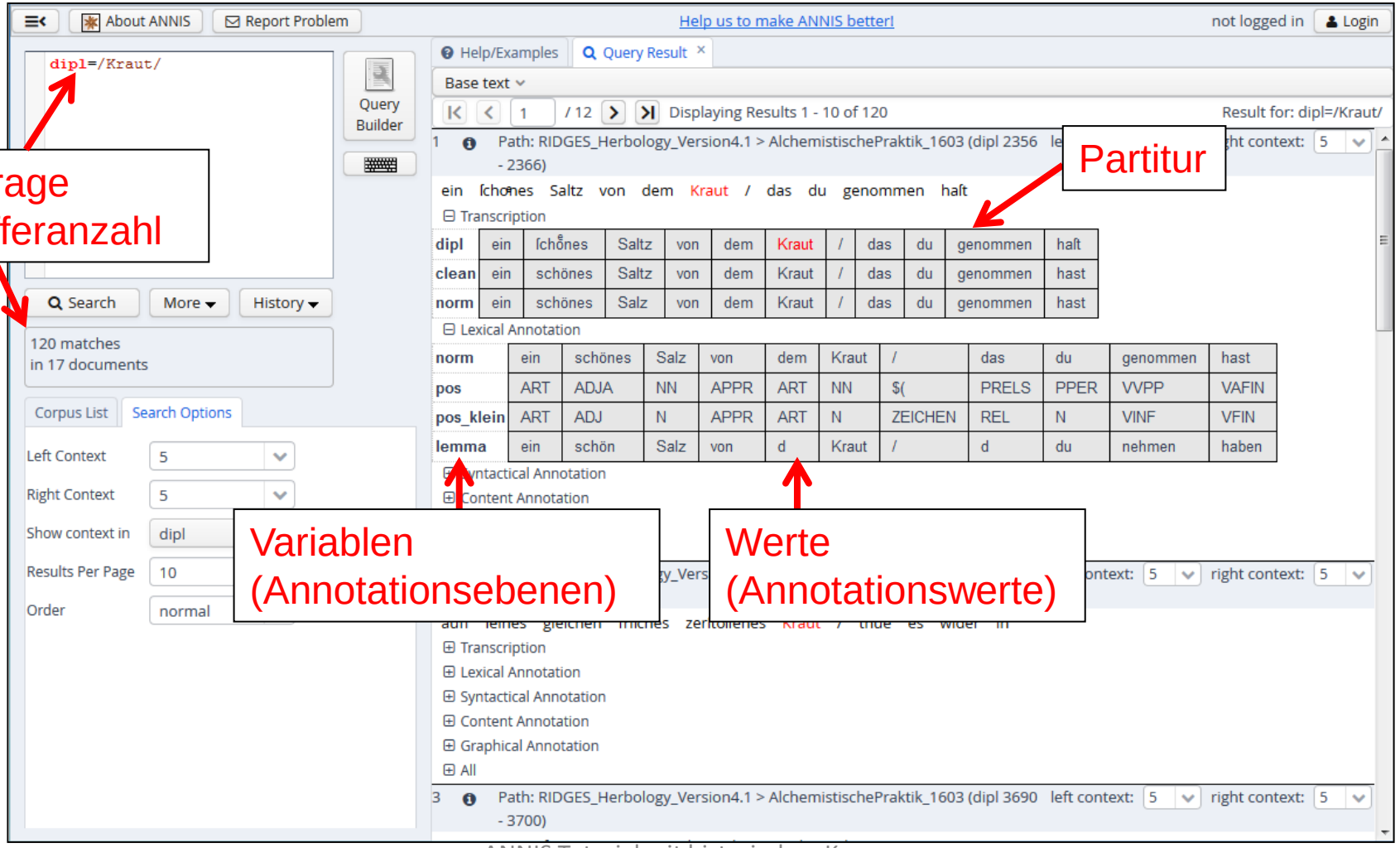
Interface

Treffer, Visualisierung, Metadaten, Suchverlauf

Interface
Treffer



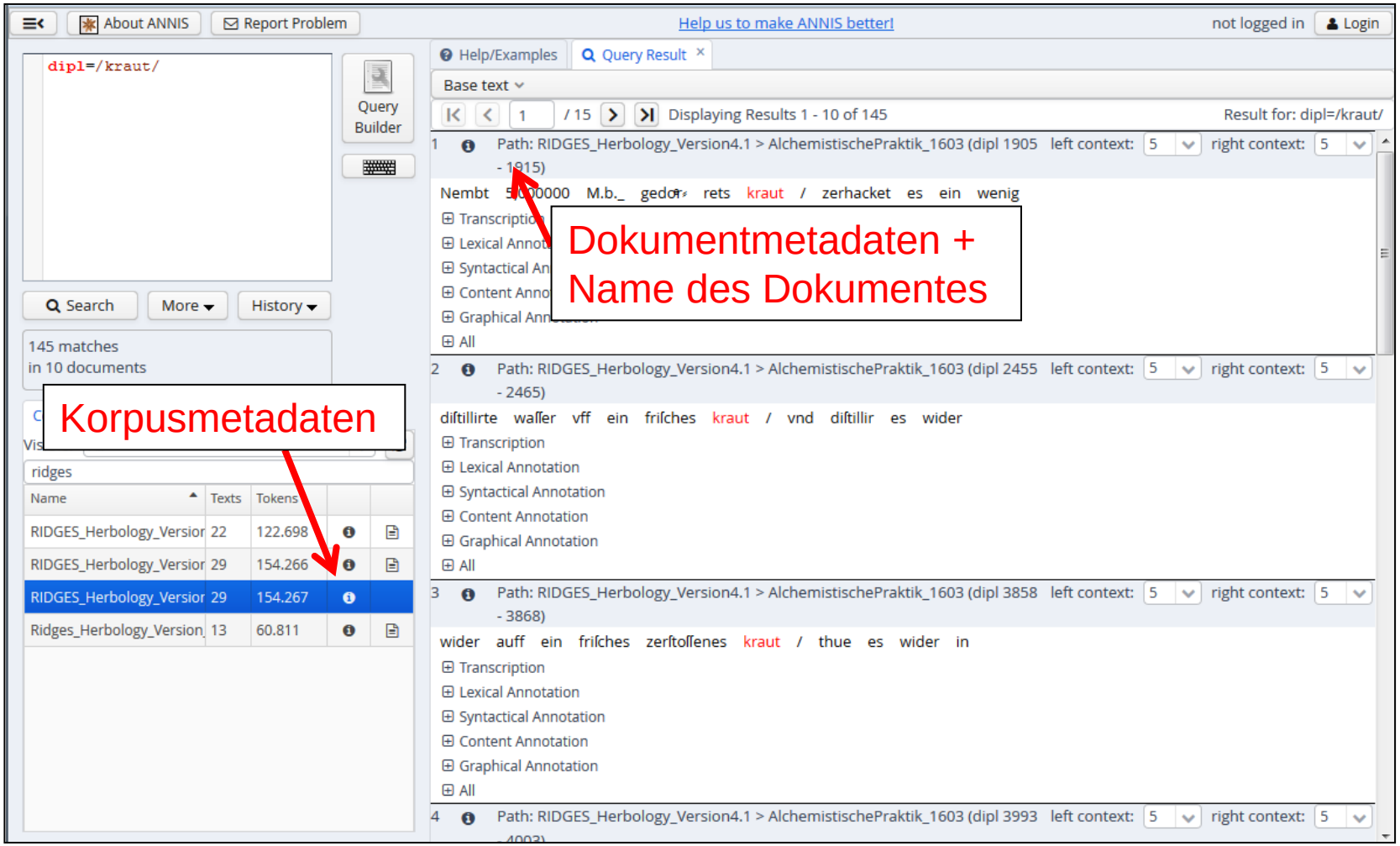
Interface Visualisierung



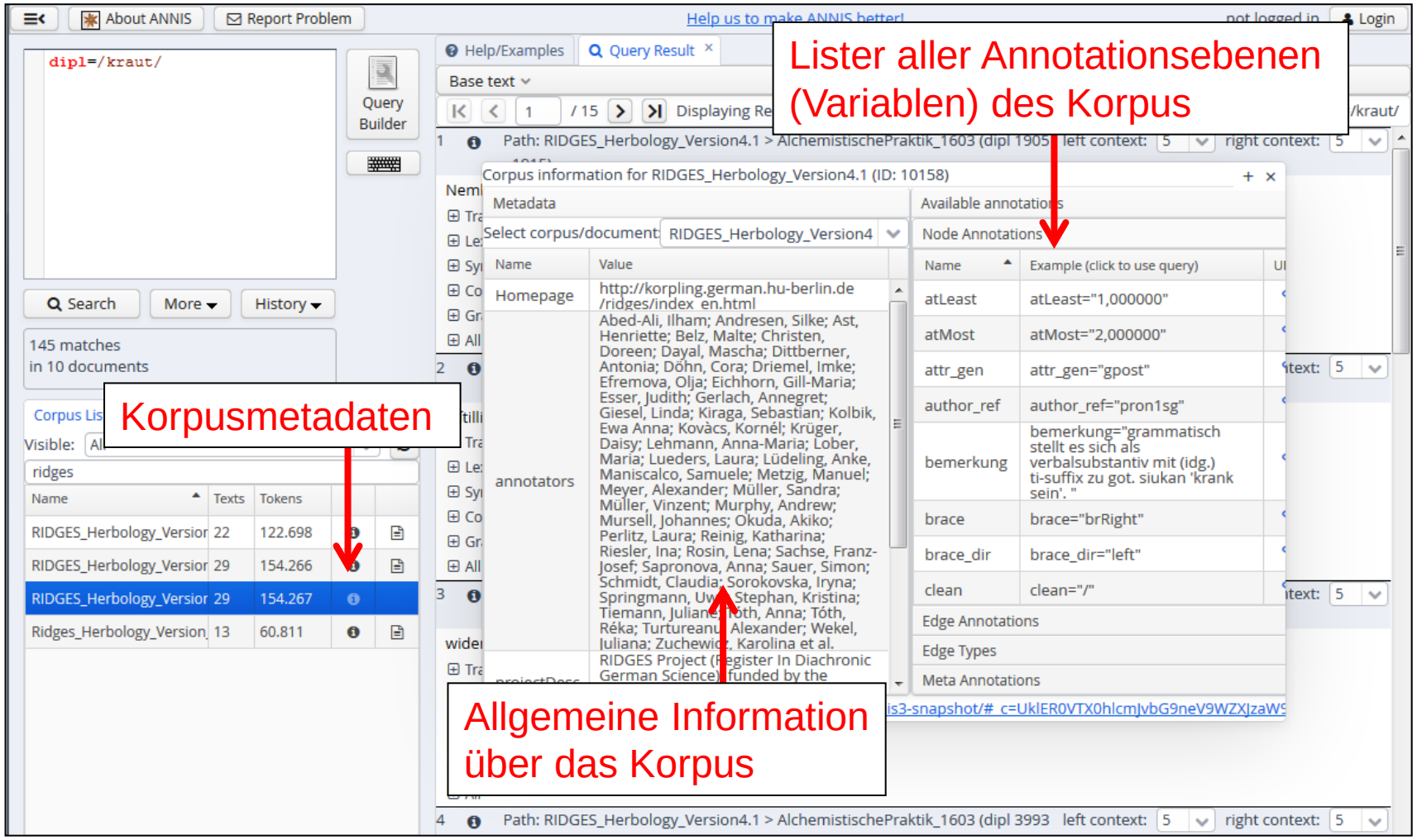
dipl	ein	ſchönes	Saltz	von	dem	Kraut	/	das	du	genommen	haſt
clean	ein	schönes	Saltz	von	dem	Kraut	/	das	du	genommen	hast
norm	ein	schönes	Salz	von	dem	Kraut	/	das	du	genommen	hast

norm	ein	schönes	Salz	von	dem	Kraut	/	das	du	genommen	hast
pos	ART	ADJA	NN	APPR	ART	NN	$(	PRELS	PPER	VVPP	VAFIN
pos_klein	ART	ADJ	N	APPR	ART	N	ZEICHEN	REL	N	VINF	VFIN
lemma	ein	schön	Salz	von	d	Kraut	/	d	du	nehmen	haben

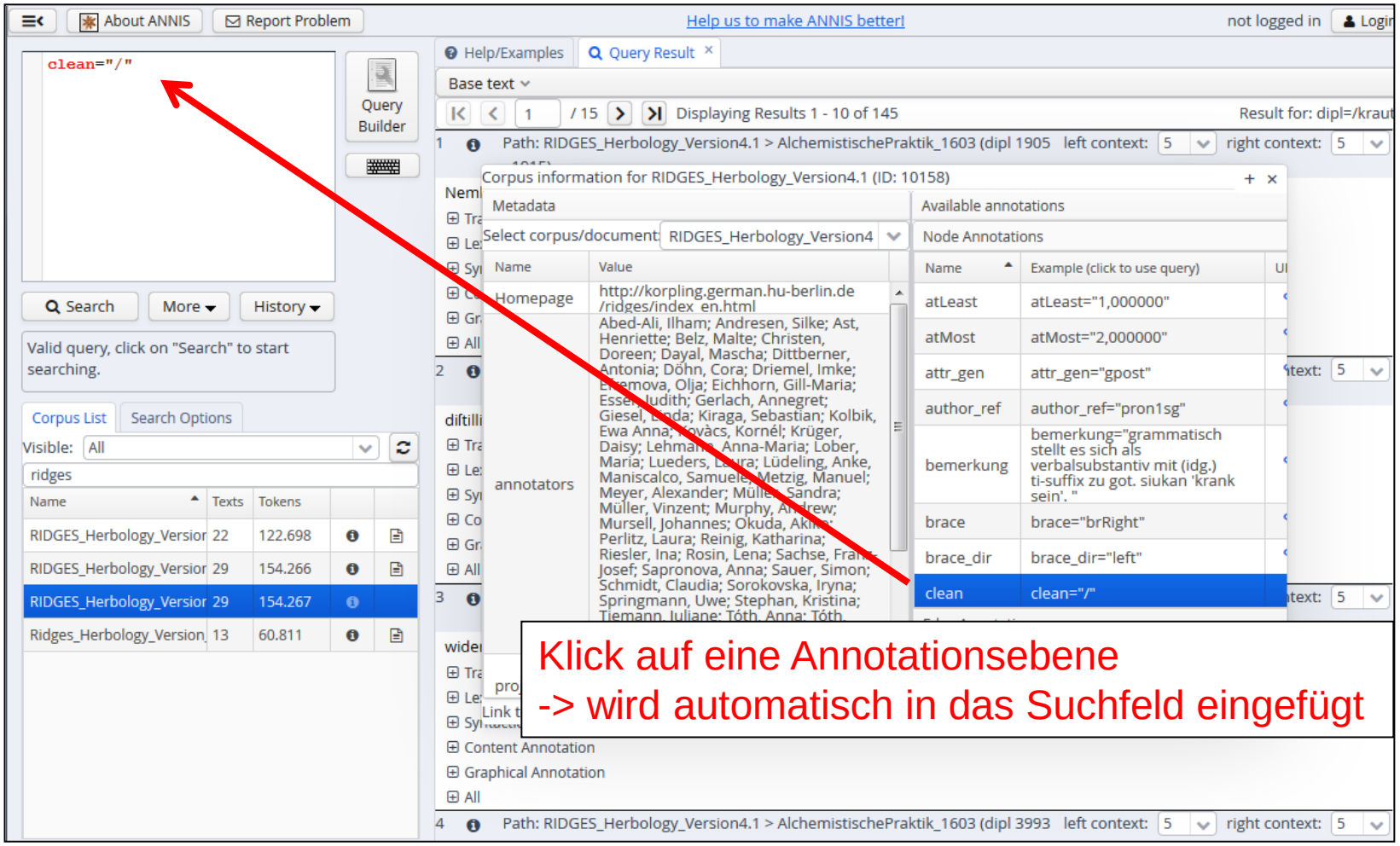
Interface Metadaten



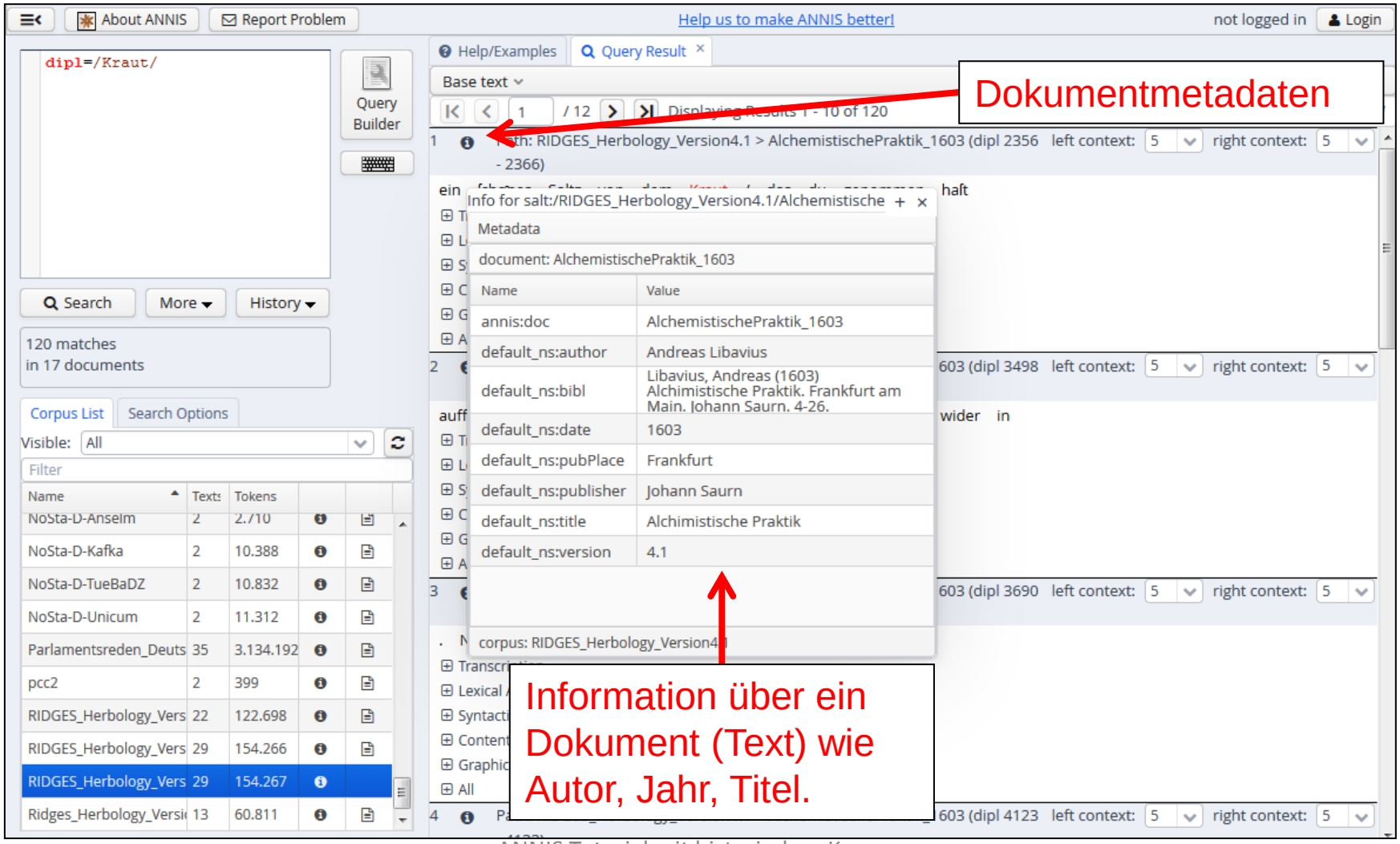
Interface Korpusmetadaten



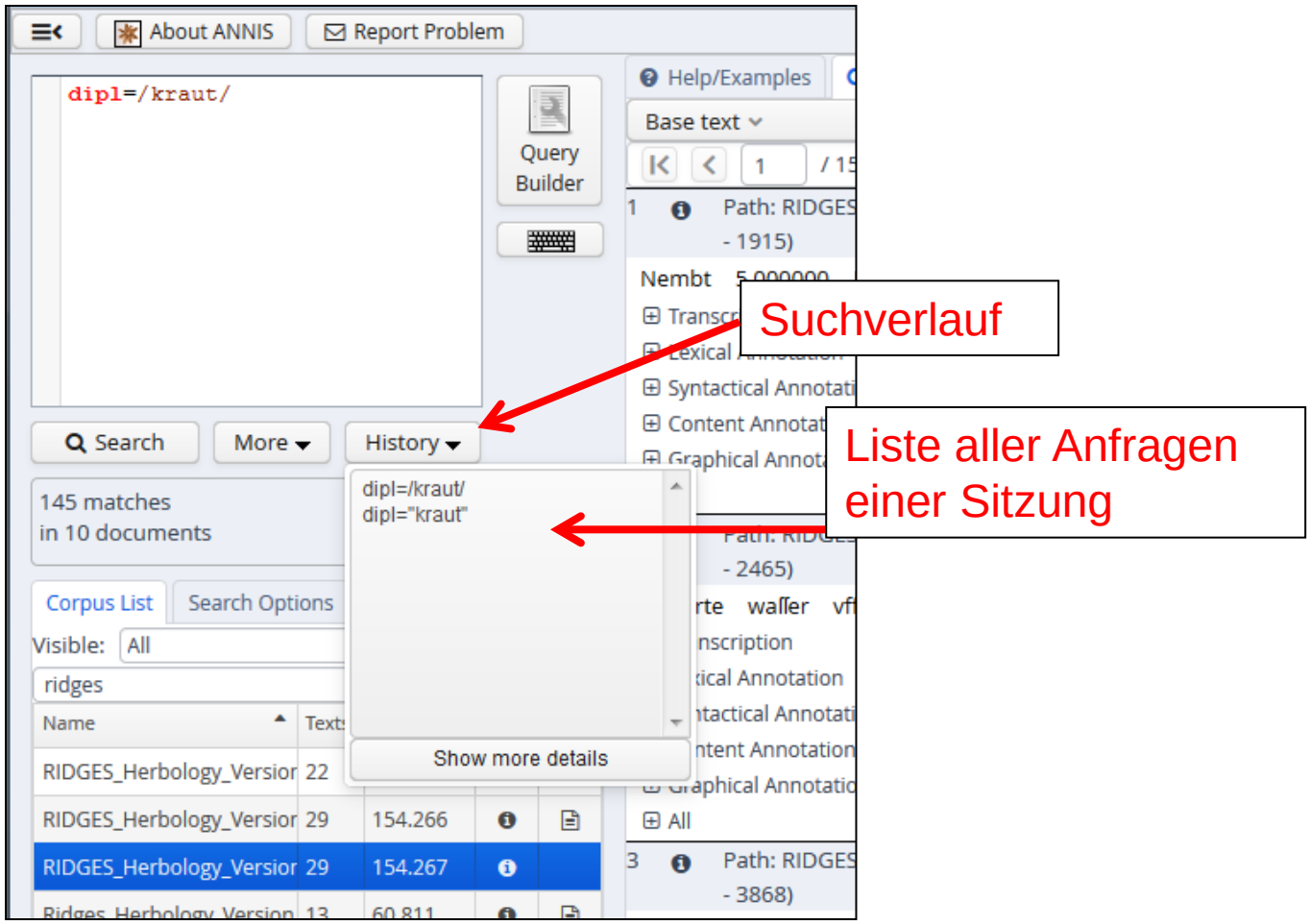
Interface Korpusmetadaten II



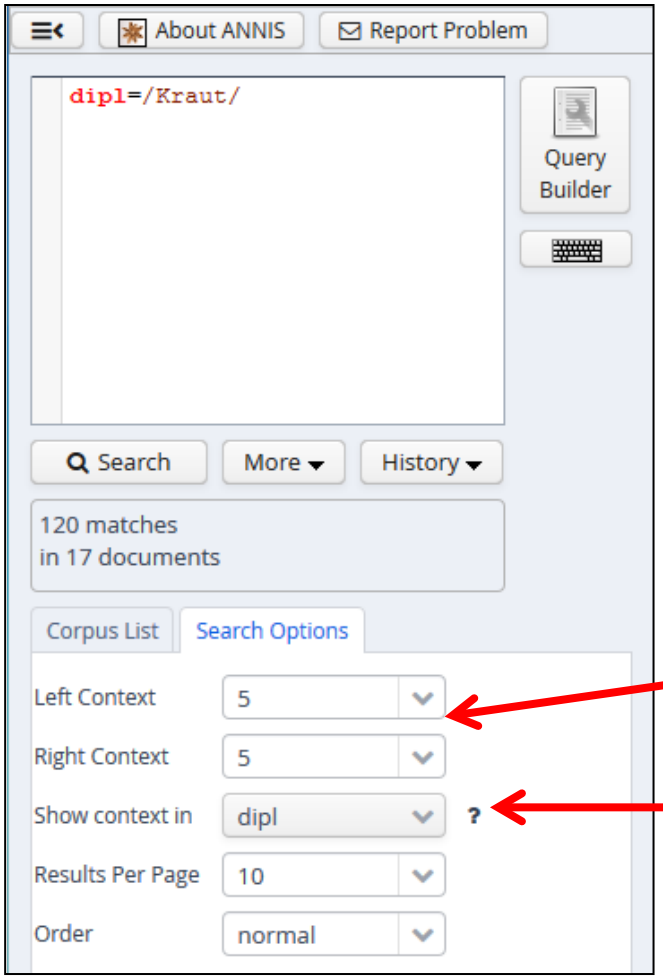
Interface Dokumentmetadaten



Interface
Suchverlauf der Anfragen



Interface
Anfrage Optionen



Größe des Kontexts festlegen

Welche Transkriptionsebene soll in der Trefferliste angezeigt werden?

Prinzip I: Variable-Wert-Paar

- AQL
  - Namen der Annotationsebene →Variable
  - annotierte Kategorien →Wert

dipl = /kraut/

- Wissen, welche Annotationen vorhanden sind
  - Korpusmetadaten
  - Annotationsrichtlinien des Korpus (für RIDGES vorhanden http://korpling.german.hu-berlin.de/ridges/download/pubs/annotationGuidelines_v4.1.pdf)

Prinzip I: Variable-Wert-Paar

1) Voraussetzung ist das Vorhandensein einer Ebene namens „dipl“.
2) Erwartetes Ergebnis ist exakt alle Vorkommen dieser Zeichenkette in „dipl“ im gesamten Korpus zu finden.

dipl= /kraut/

Variable (Layer, Tier, Ebene ...)

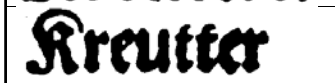
Wert (Wort, Lemma, Satz, Wortart ...)

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

Historische Texte

Ridges
Register in Diachronic German Science

- Varianz der Orthographie bzw. Setzung
- vieles nicht wirklich „vorhersehbar“
- graphische/(text-) strukturelle Informationen
  - Zeilenumbruch, Seitenumbruch, Setzung

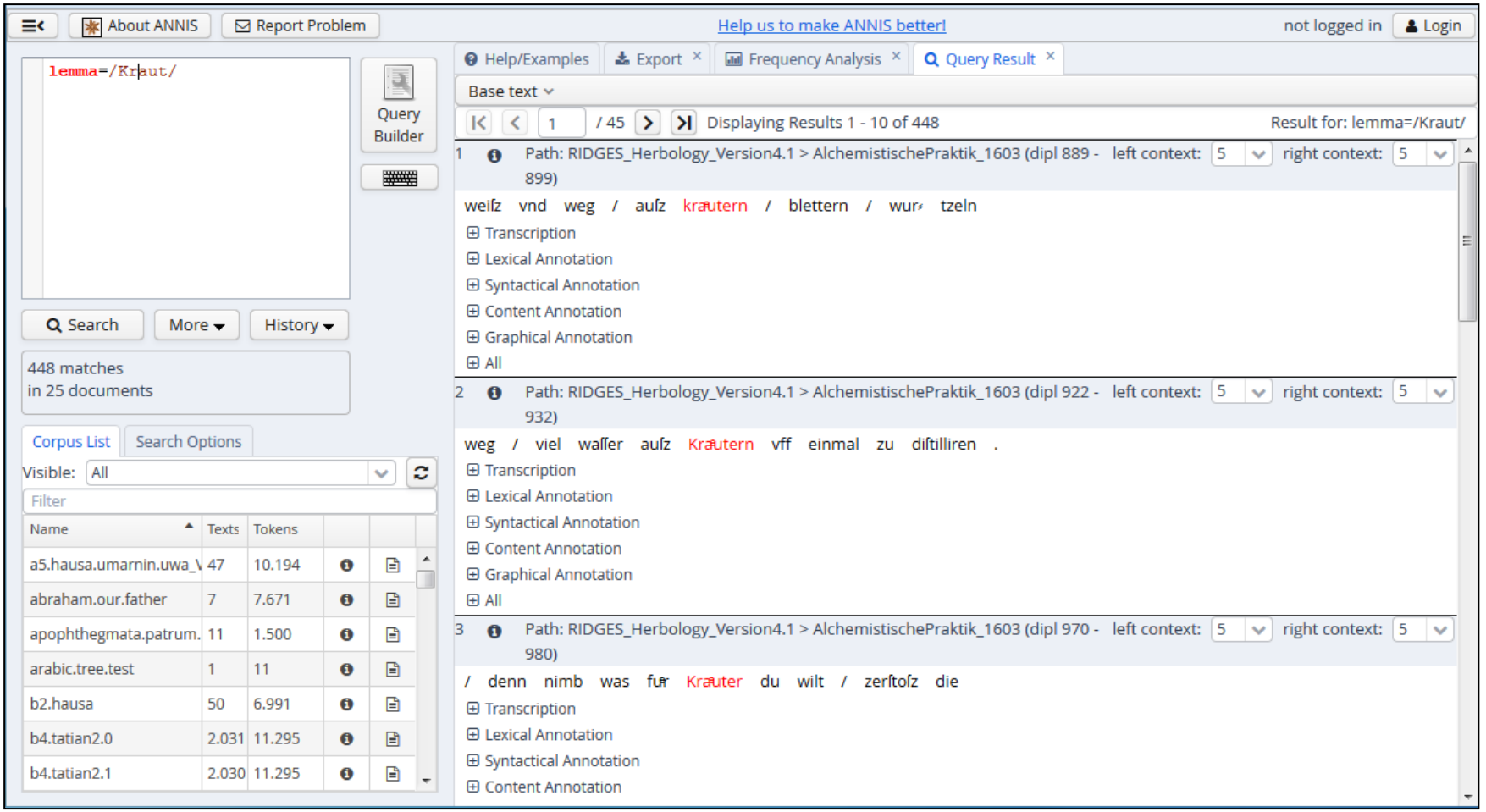
	Kräutern	Alchimistische Praktik 1603
	Kraut	Alchimistische Praktik 1603
	kraut	Alchimistische Praktik 1603
	Kreutern	Alchimistische Praktik 1603
	Kreutter	Alchimistische Praktik
	kreüter	Fuchs New Kreüterbuch 1543
	Kräuteren	Pflantz.Gart. 1639
	Kreuter	Alchmistische Praktik 1603
	Kräuter	Deutsche Pflanzennamen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

Aufgabe Schreibvarianten

- Finden Sie alle Schreib- und Flexionsvarianten von *Kraut*!
  - für die Suche nach allen Varianten wählen Sie die passende Annotationsebene (Variable) aus
    - passende Variable wäre hier **lemma**
  - und setzen den Wert **/Kraut/** ein

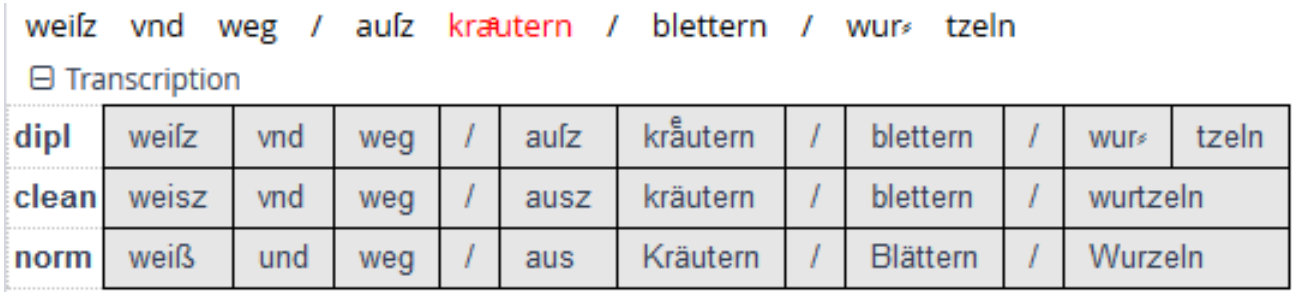
➢lemma=/Kraut/

Aufgabe
Schreibvarianten



Normalisierung

- Schauen Sie sich die Treffer zu lemma=/Kraut/ genau an!


weiſz vnd weg / auſz kræutern / blettern / wurʒ tzeln

⊟ Transcription

dipl	weiſz	vnd	weg	/	auſz	kr̊äutern	/	blettern	/	wurʒ tzeln
clean	weisz	vnd	weg	/	ausz	kräutern	/	blettern	/	wurtzeln
norm	weiß	und	weg	/	aus	Kräutern	/	Blättern	/	Wurzeln

  - Es werden alle historischen Wortformen von Kraut gefunden!

Aufgabe Normalisierung

- Suchen Sie das Lemma von *zusammensetzen*!
  - Finden von ganz unterschiedlichen historischen Schreibvarianten!
  - ➢ lemma=/zusammensetzen/

  - Treffer 1

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

  - Treffer 2

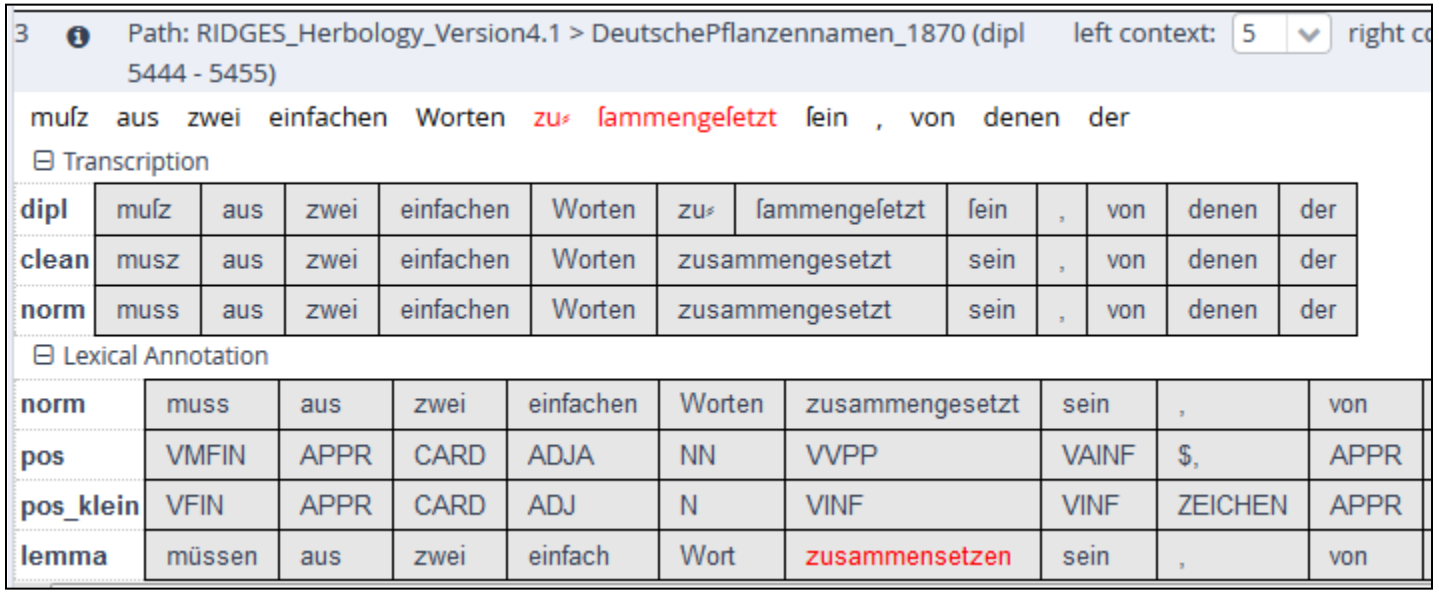
    , die aus drei Worten zuſammen- geſetzt ſind , wie die oben

  - Treffer 3

    muſz aus zwei einfachen Worten zu- ſammengeſetzt ſein , von denen der

Aufgabe Normalisierung

- Schrittweise Normalisierung (dipl>clean>norm) erlaubt
  - Zusammenführung von historischen Schreibweisen (dipl)
  - Zuweisung von Wortarten und Lemmatisierungen auf der normierten Ebene (norm)



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

muſz aus zwei einfachen Worten zuſammengeſetzt ſein , von denen der

⊟ Transcription

dipl	muſz	aus	zwei	einfachen	Worten	zuſammengeſetzt	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

Operatoren für die Mustersuche

.	Ein beliebiges Zeichen
?	0 oder 1 Zeichen (des vorherigen Elementes)
*	0 bis unendlich viele Zeichen (d. vorh. E.)
+	1 bis unendlich viele Zeichen (d. vorh. E.)
\\	wörtlich (folgendes Zeichen)
!	nicht
(a|b)	a oder b (auch: [ab])

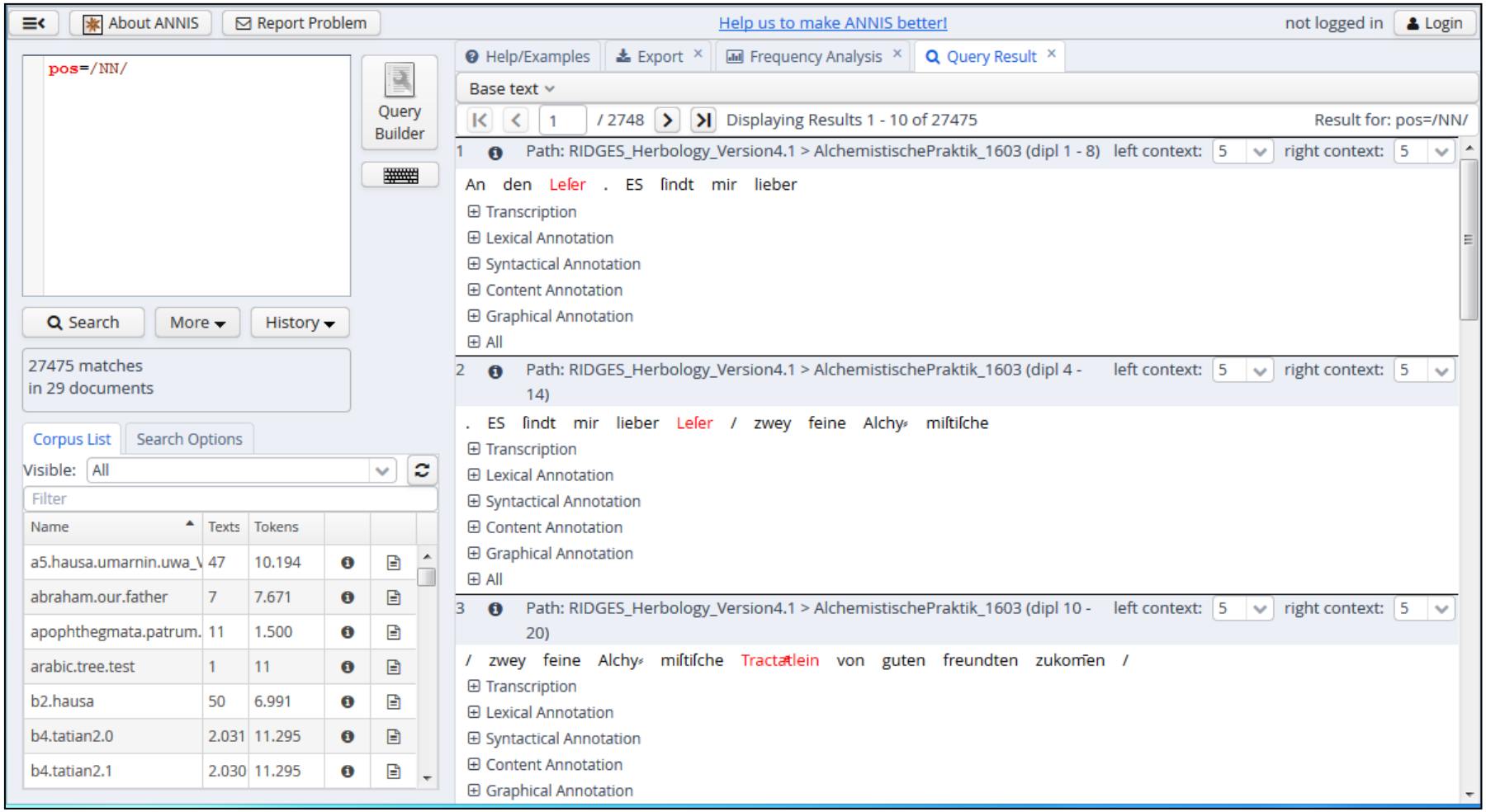
Aufgabe
Operatoren

- Welche Ergebnisse erwarten Sie für folgende Anfragen, sogenannte Mustersuchen?
  - norm=/g.b./
    - ➢ *gebe, gibt* (für RIDGES, theoretisch noch andere möglich)
  - dipl=/r(a|o)t/
    - ➢ *rot, rat* (für RIDGES)
  - dipl=/meint?/
    - ➢ *mein, meint* (für RIDGES)

Aufgabe
Wortart

- Wie finden Sie alle Appellativa in RIDGES?
  - passende Variable wäre hier**pos**
  - ➢pos=/NN/

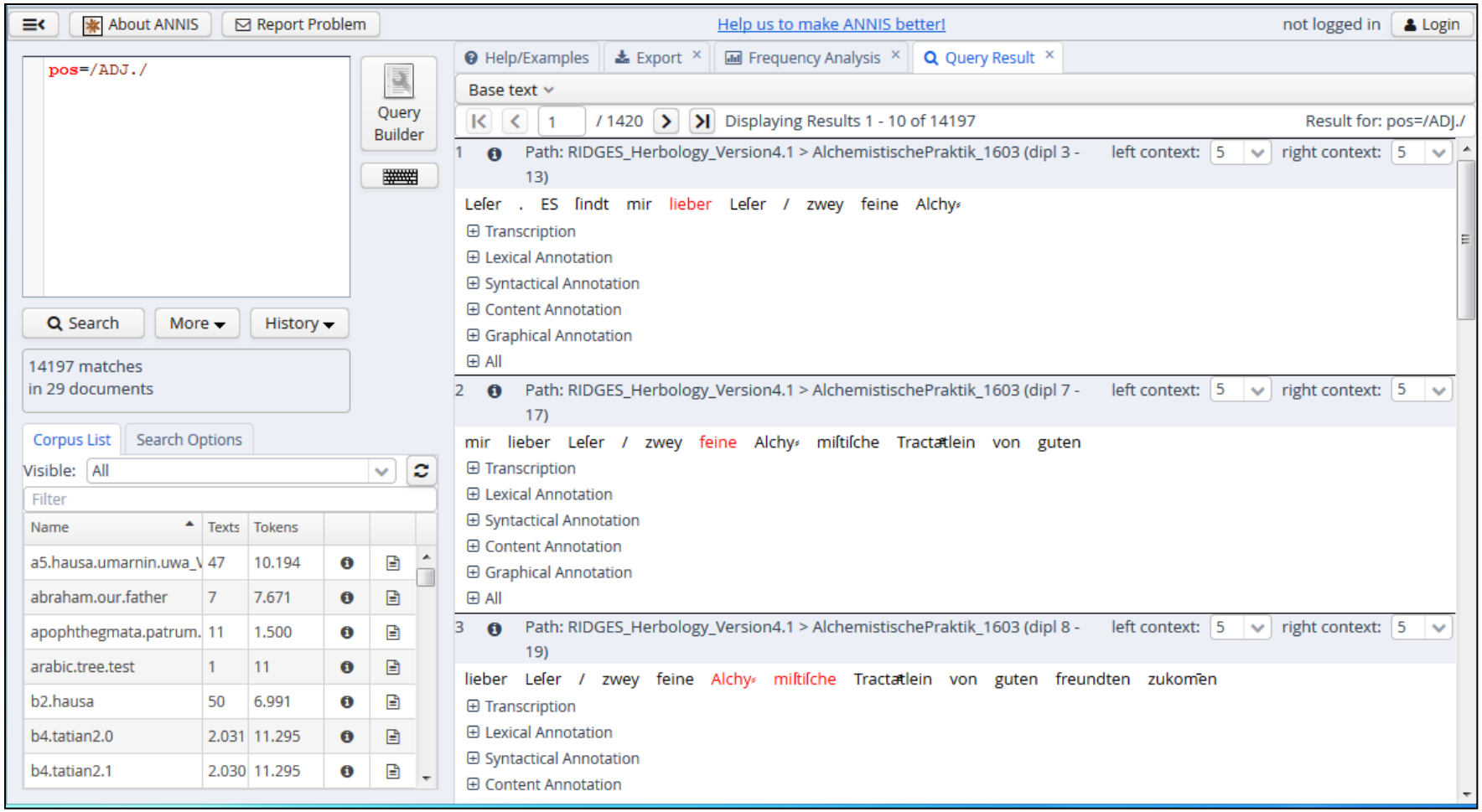
Aufgabe
Wortart



Aufgabe
Wortarten

- Wie finden Sie alle Adjektive unabhängig von ihren Bezugswörtern in RIDGES?
  - Was sagen die Richtlinien (STTS)?
  - passende Variable wäre hier **pos**
  - STTS: Unterschied ADJA und ADJD?
  - ➢ pos=/ADJ./

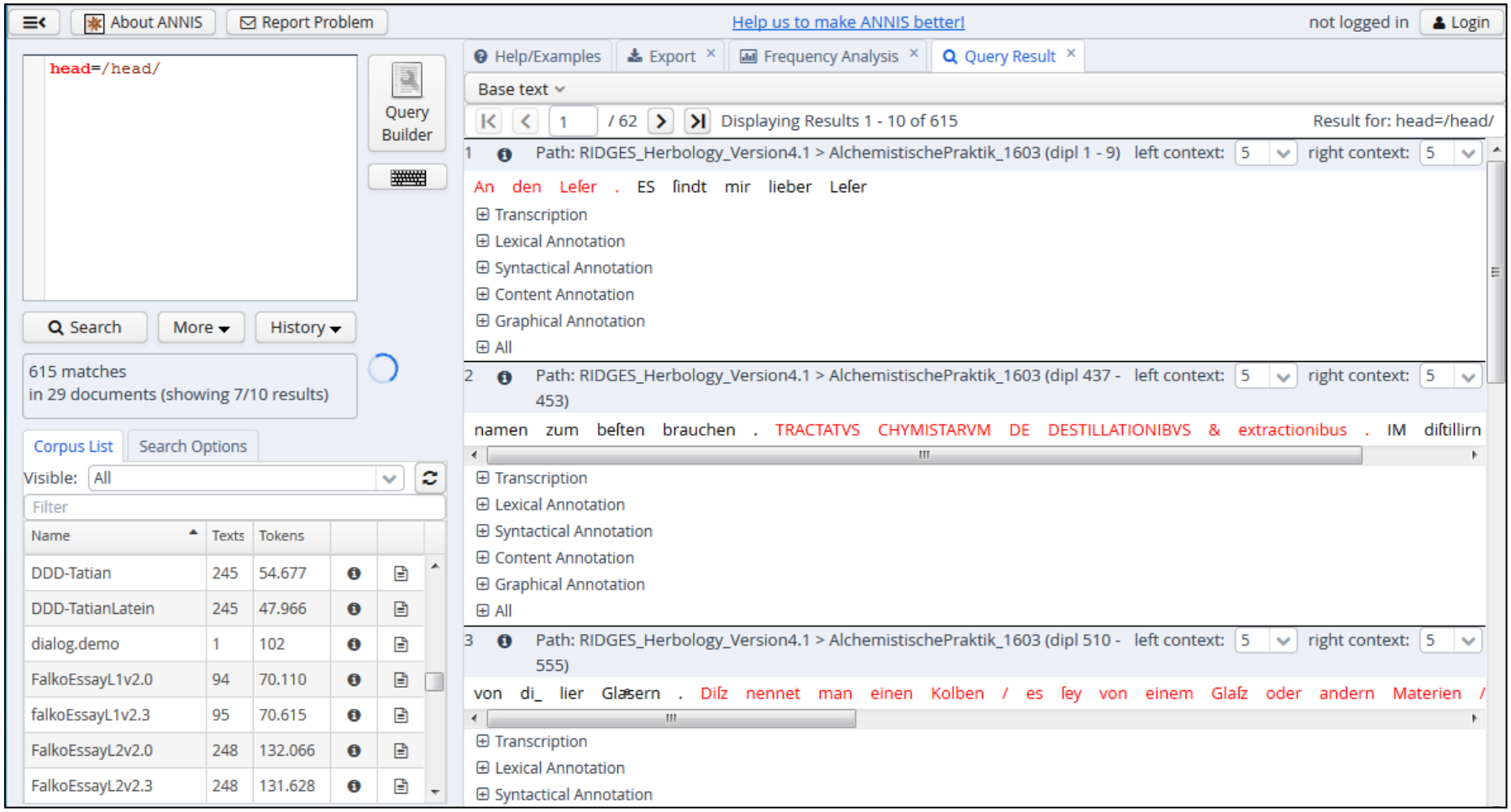
Aufgabe
Wortarten



Aufgabe
Suchen von Annotationsebenen

- Finden Sie heraus, ob es in RIDGES Herbology Annotationen für Überschriften gibt?
  - Wie sieht eine Anfrage aus, um Überschriften zu finden?
  - passende Variable wäre hier **head**
  - ➢ head=/head/

Aufgabe
Suchen von Annotationsebenen

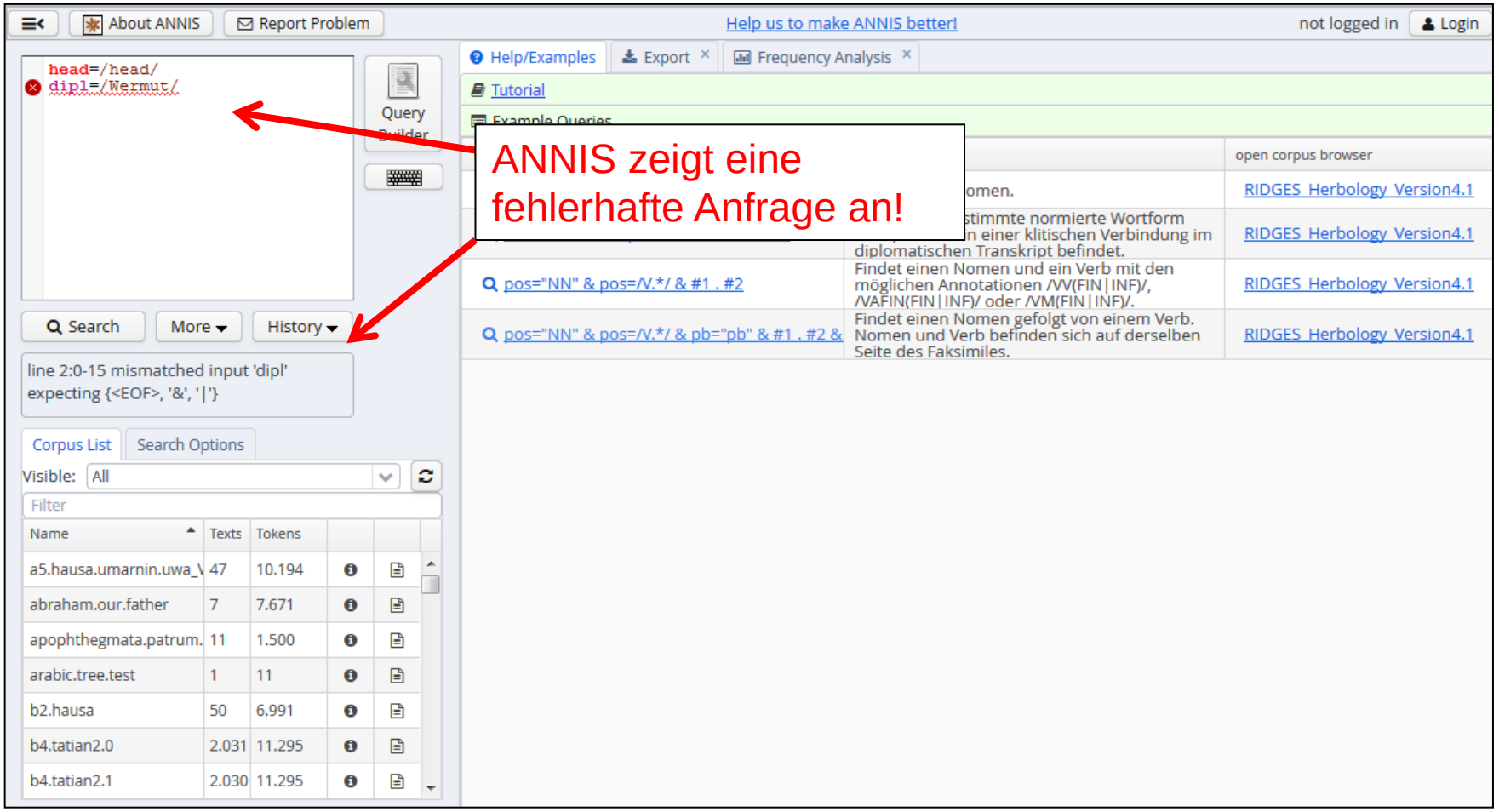


Aufgabe
Kombinierte Suche

- Finden Sie die diplomatische Wortform *Wermut*, die in einer Überschrift vorkommt!
  - passende Variablen wären hier **head** und **dipl**
  - head=/head/
  - dipl=/Wermut/

  Was passiert?

Aufgabe
Kombinierte Suche



Prinzip II: Relationen

head=/head/ **VW-Paar1**

_o_ **Relation**

dipl=/Wermut/ **VW-Paar2**

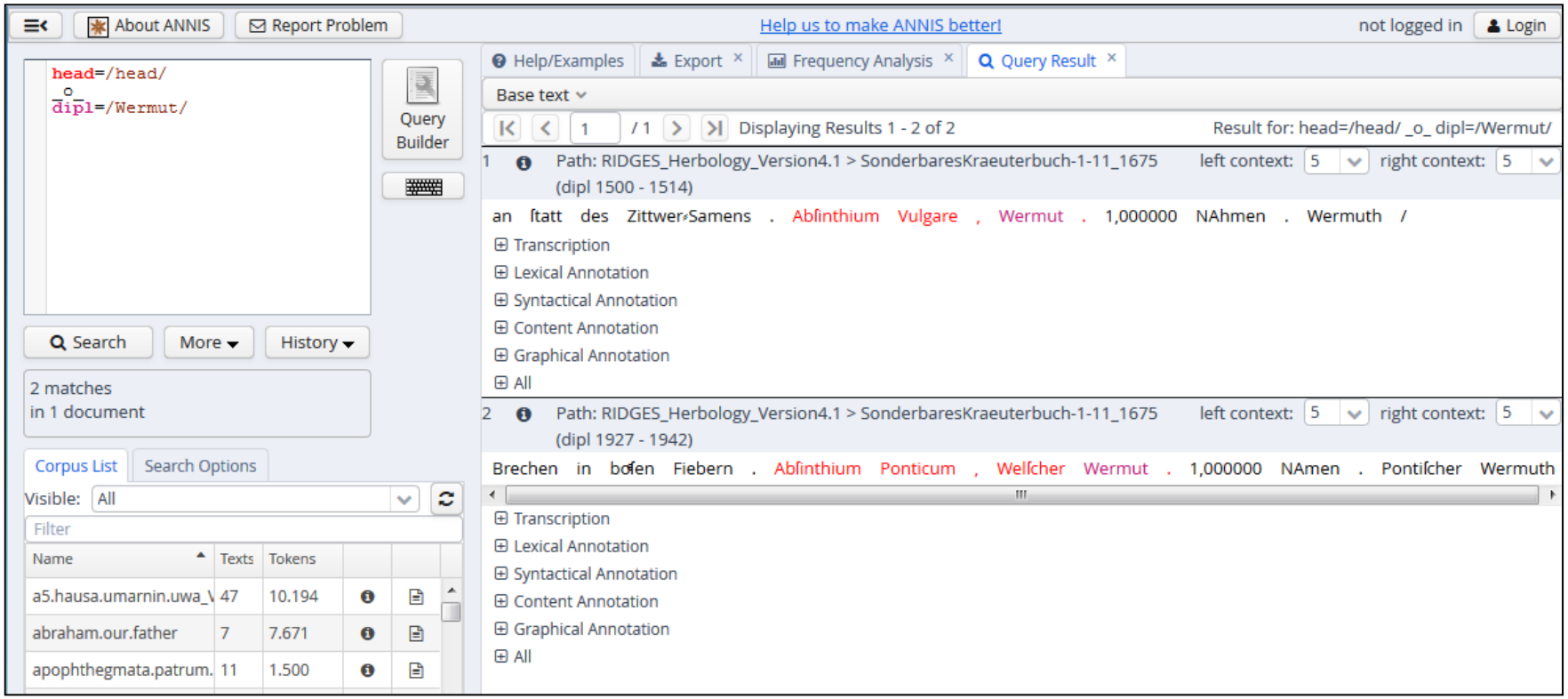
1) Es gibt mehrere Arten von Annotationen!
2) Wissen, wie diese Annotationen in Beziehung zu einander stehen können!
3) Spannen (head) überlappen Tokenannotationen (dipl)!

XVIII. Saturniniſche Kræuter . Von ħ Kræutern ihrer Natur

⊟ Transcription

dipl	XVIII.	Saturniniſche	Kr̊auter	.	Von	ħ	Kr̊autern	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

- pro VW-Paar eine Farbe – hier unser Beispiel:
  - Farbe Rot für alle Werte der Variable **head**
  - Farbe Lila für alle Werte der Variable **dipl**

```
head=/head/
  o
dipl=/Wermut/
```

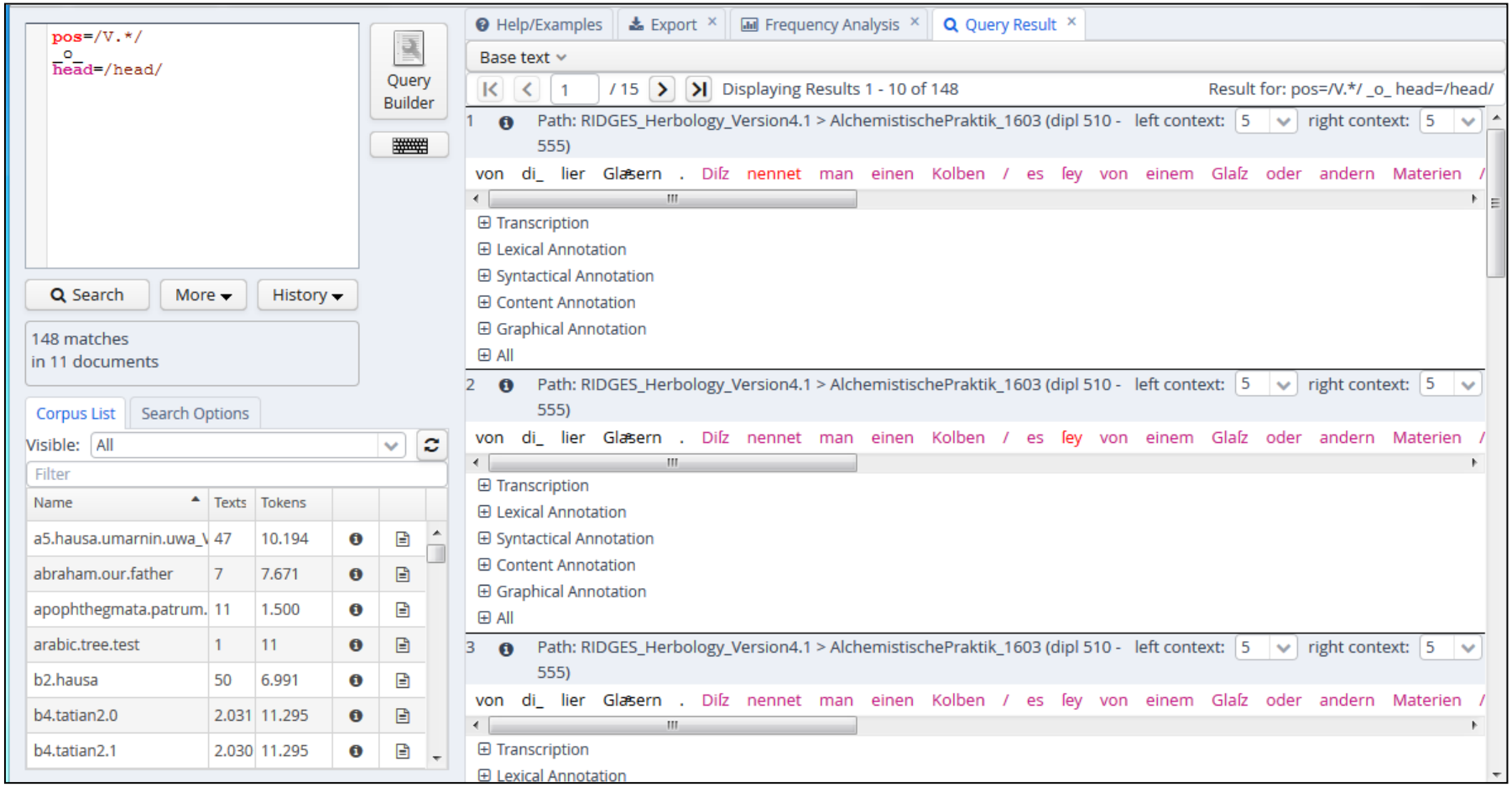
- Treffer erhalten genau diese Farben:
  - mehrere Token rot, da sie zusammen in einer Überschrift stehen
  - ein Token lila, da dieses der gesuchte dipl-Wert

Abſinthium Vulgare , Wermut .

Aufgabe Überlappung

- Suchen Sie ein Verb, das in einer Überschrift vorkommt!
  - passende Variablen wären **pos** und **head**
  - Operator **_o_**
  - ➢ pos=/V.*/ _o_ head=/head/

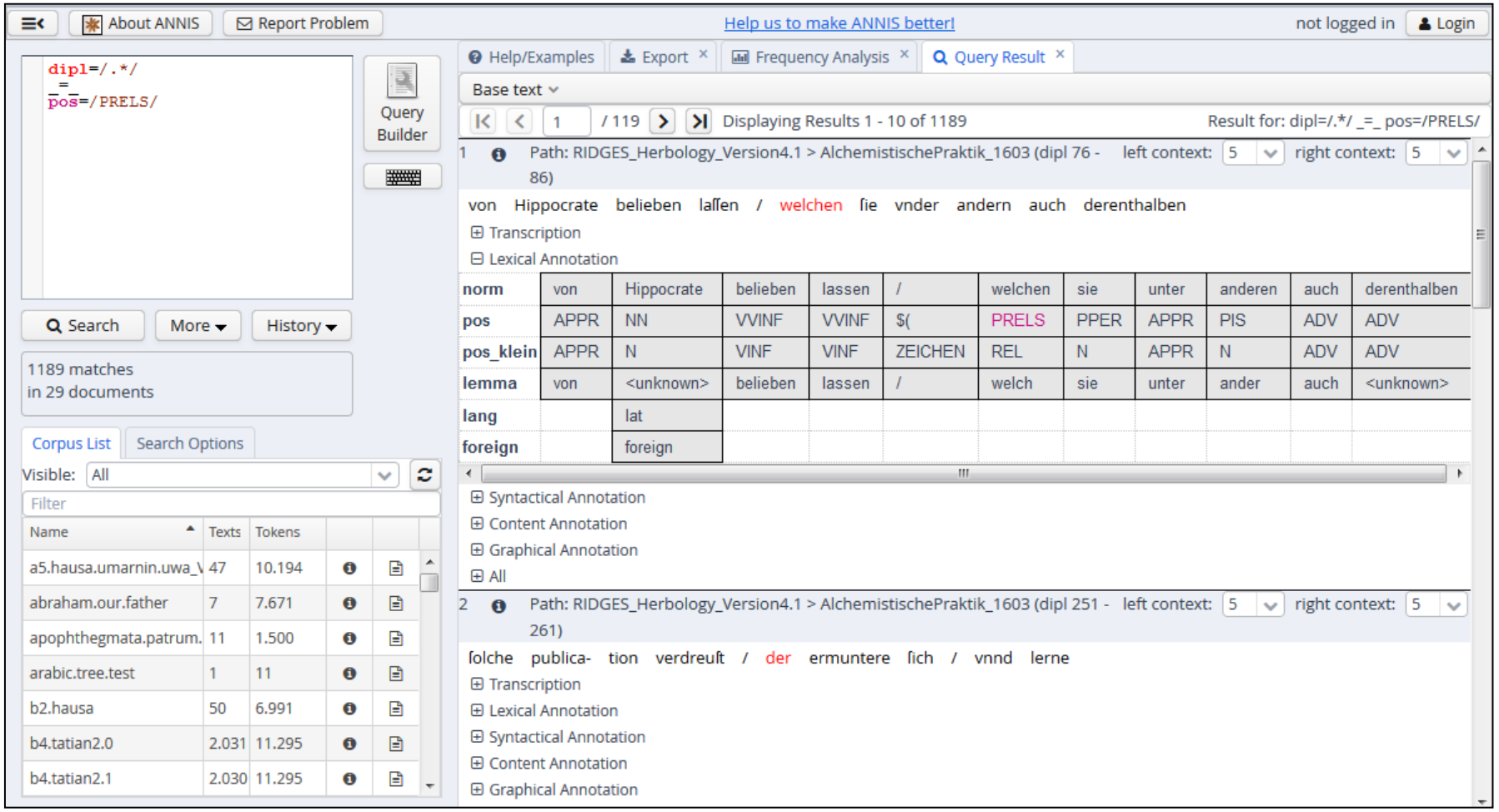
Aufgabe
Überlappung



Aufgabe Identität

- Suchen Sie alle diplomatischen Wortformen, die als substituierende Relativpronomen annotiert worden sind!
  - passende Variablen wären **dipl** und **pos**
  - Operator _=_
  - ➢ dipl=/.*/ _=_ pos=/PRELS/

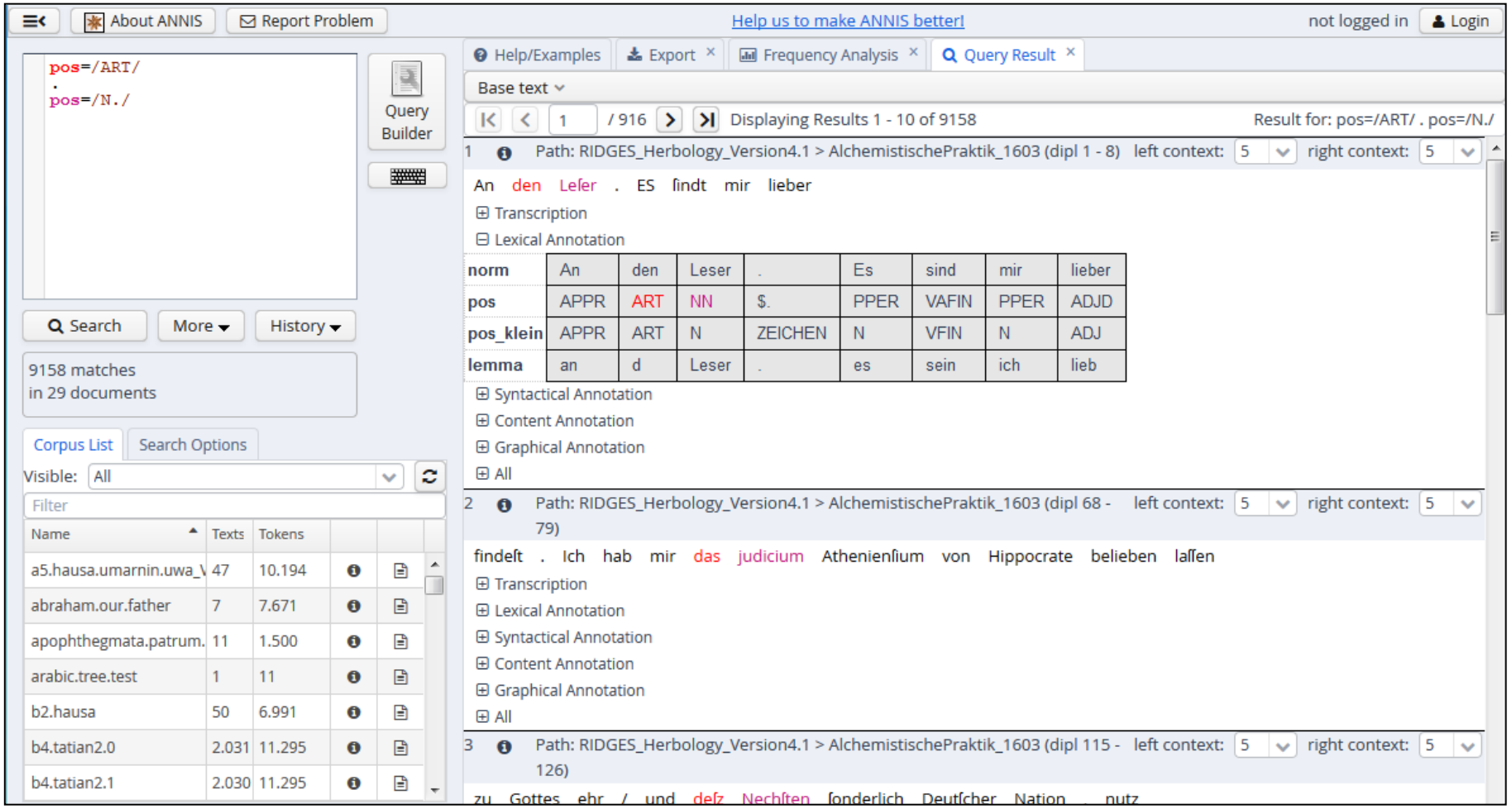
Aufgabe Identität



Aufgabe
Direkte Präzedenz

- Suchen Sie einen Artikel, der ein Nomen direkt präzediert!
  - passende Variable wäre **pos**
  - Operator **.**
  - ➢ pos=/ART/ . pos=/N./

Aufgabe
Direkte Präzedenz

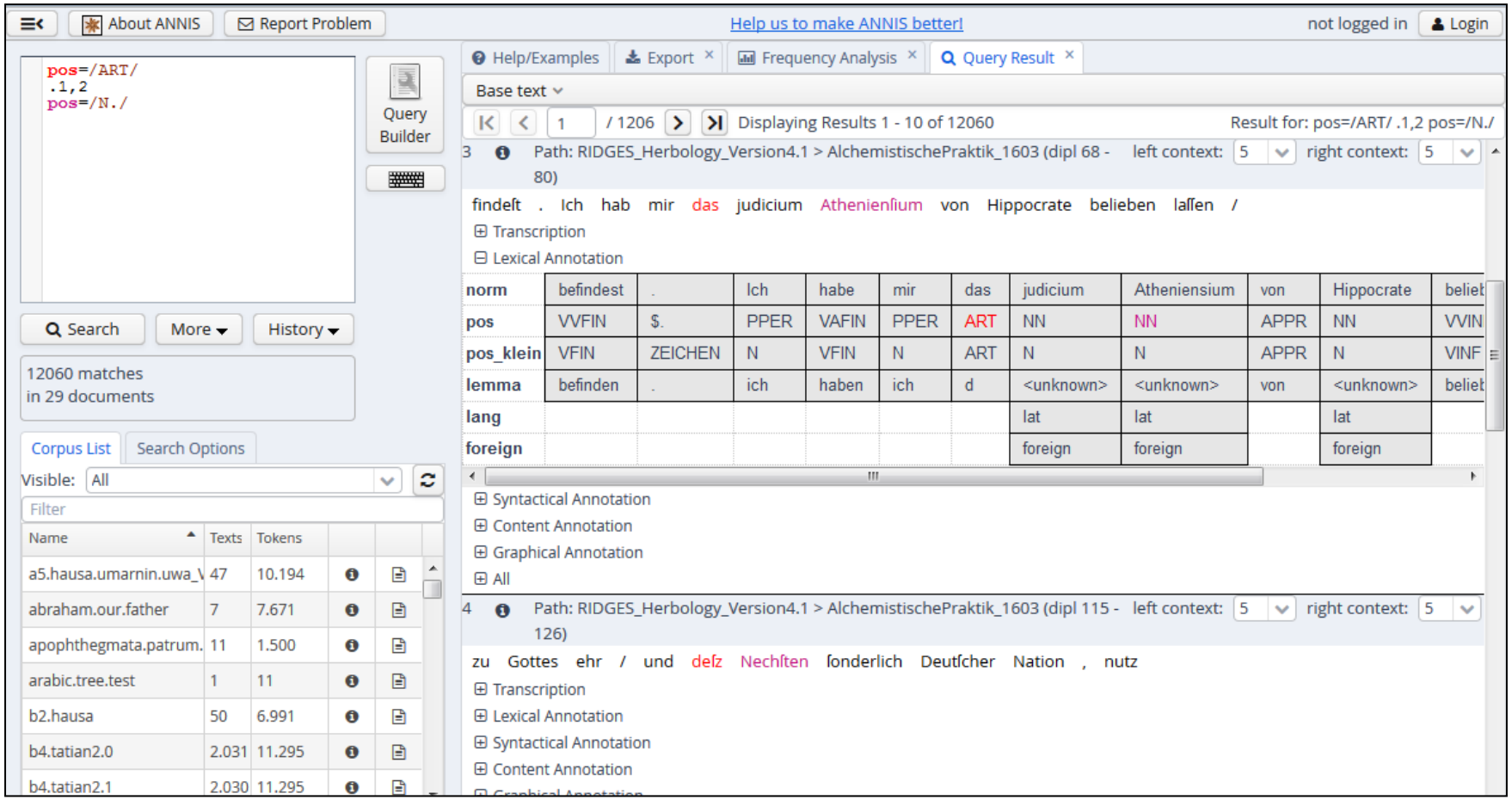


Aufgabe
Indirekte Präzedenz

- Suchen Sie einen Artikel **(A)**, der ein Nomen **(B)** indirekt präzediert. Sie wollen ebenfalls einen möglichen pränominalen Modifikator **(C)** in ihrer Trefferliste abfangen!
  - passende Variablen wäre **pos**
  - Operator **.** und Abstand 1,2 (zu lesen: Ich suche **A** und **B** direkte aufeinanderfolgend oder es kann ein Token **C** dazwischen stehen)

➢ pos=/ART/ .1,2 pos=/N./

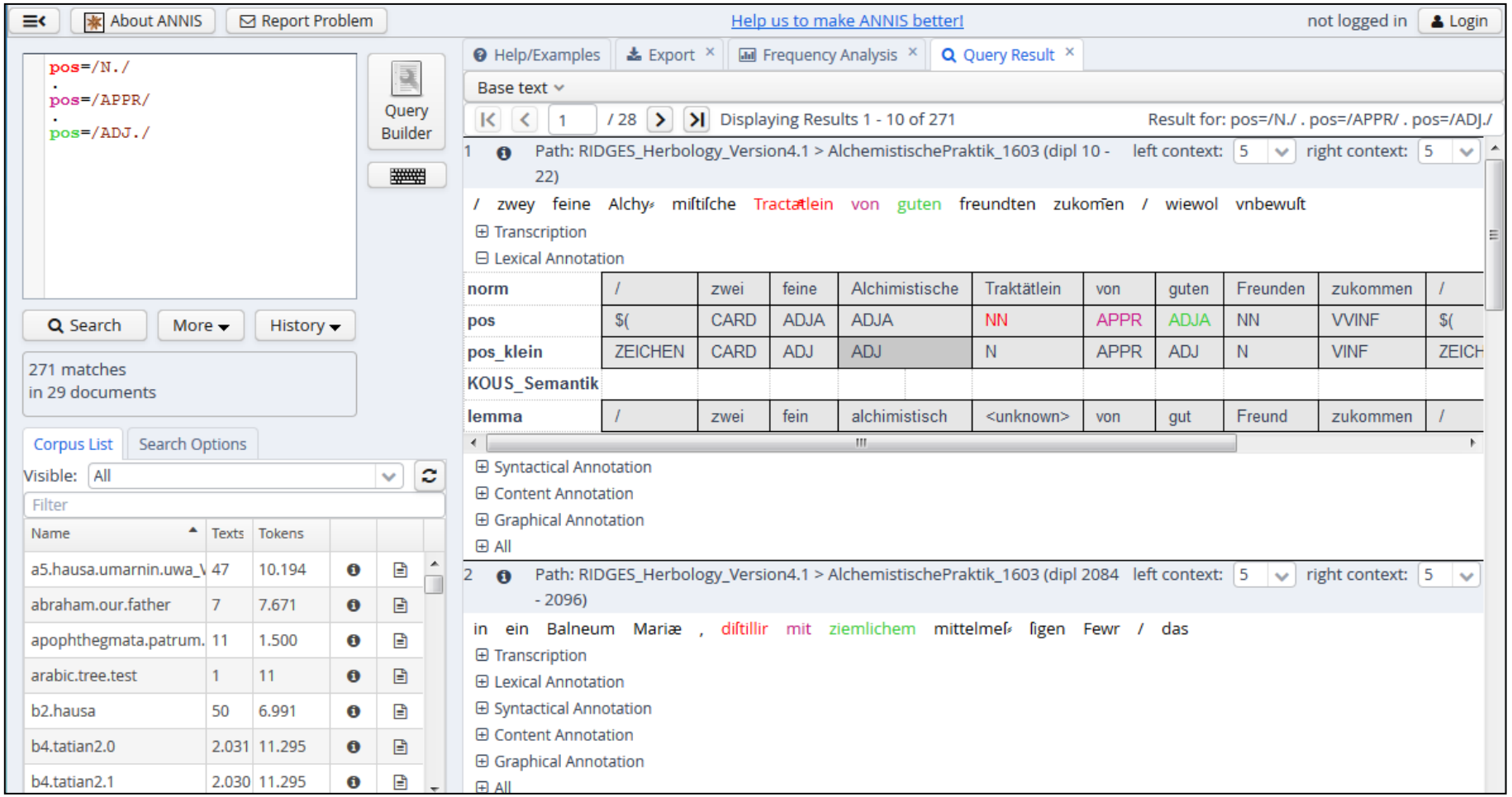
Aufgabe
Indirekte Präzedenz



Aufgabe
Abfolgen

- Suchen Sie ein Nomen direkt gefolgt von einer Präposition direkt gefolgt von einem Adjektiv!
  - passende Variable wäre **pos**
  - Operator **.**
  - ➢ pos=/N./ . pos=/APPR/ . pos=/ADJ./

Aufgabe Abfolgen



Aufgabe
Abfolgen

- Suchen Sie ein Nomen direkt gefolgt von einer Präposition direkt gefolgt von einem Adjektiv!
  - passende Variable wäre **pos**
  - Operator **.**

➢pos=/N./ . pos=/APPR/ . pos=/ADJ./

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

⊞ Transcription

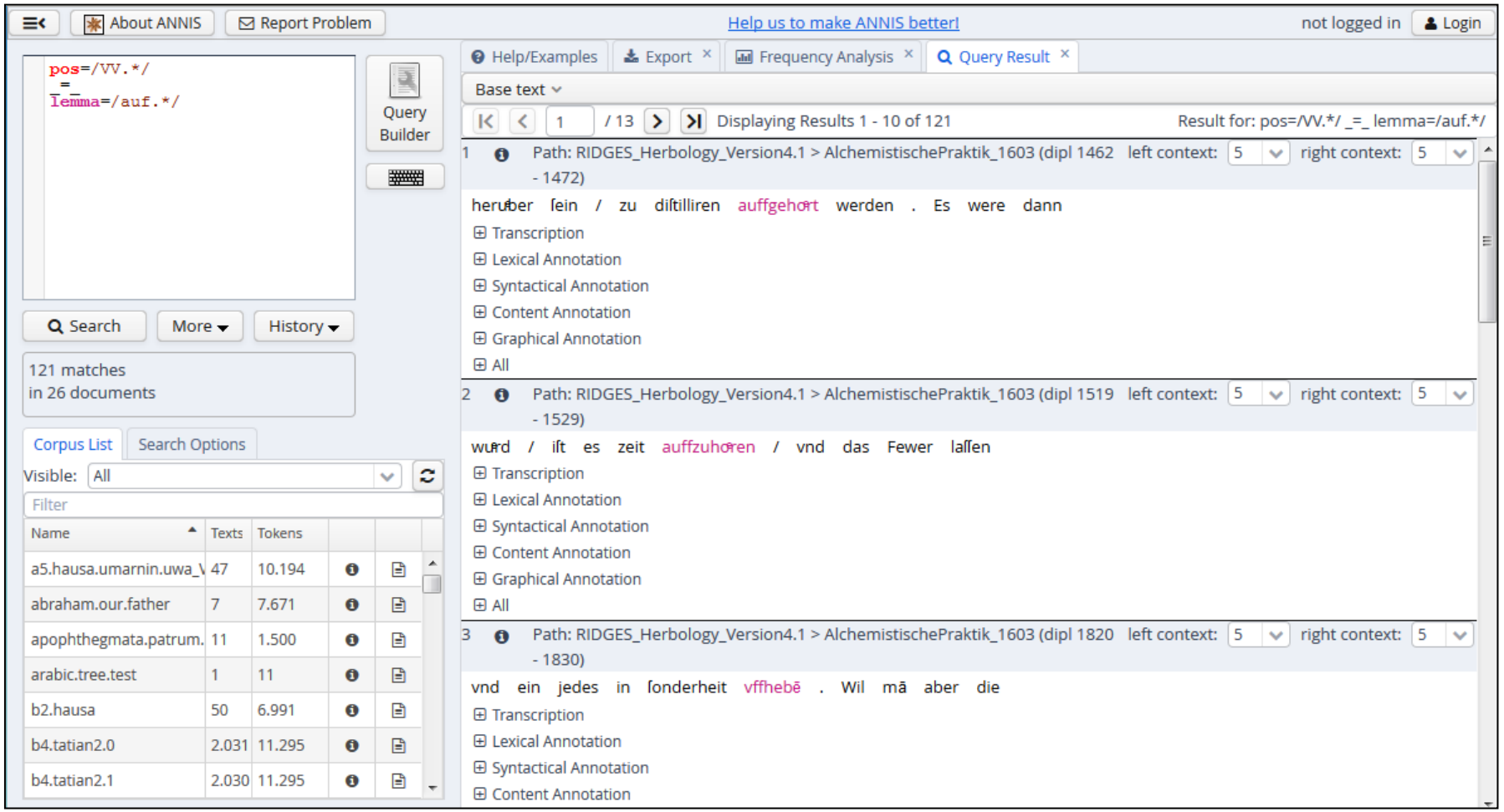
⊟ Lexical Annotation

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

Aufgabe Identität

- Suchen Sie Partikelverben, die die Partikel *auf* beinhalten!
  - passende Variablen wären **lemma** und **pos**
  - Operator **_=_**
  - ➢ pos=/VV.*/ _=_ lemma=/auf.*/

Aufgabe Identität



Aufgabe Identität

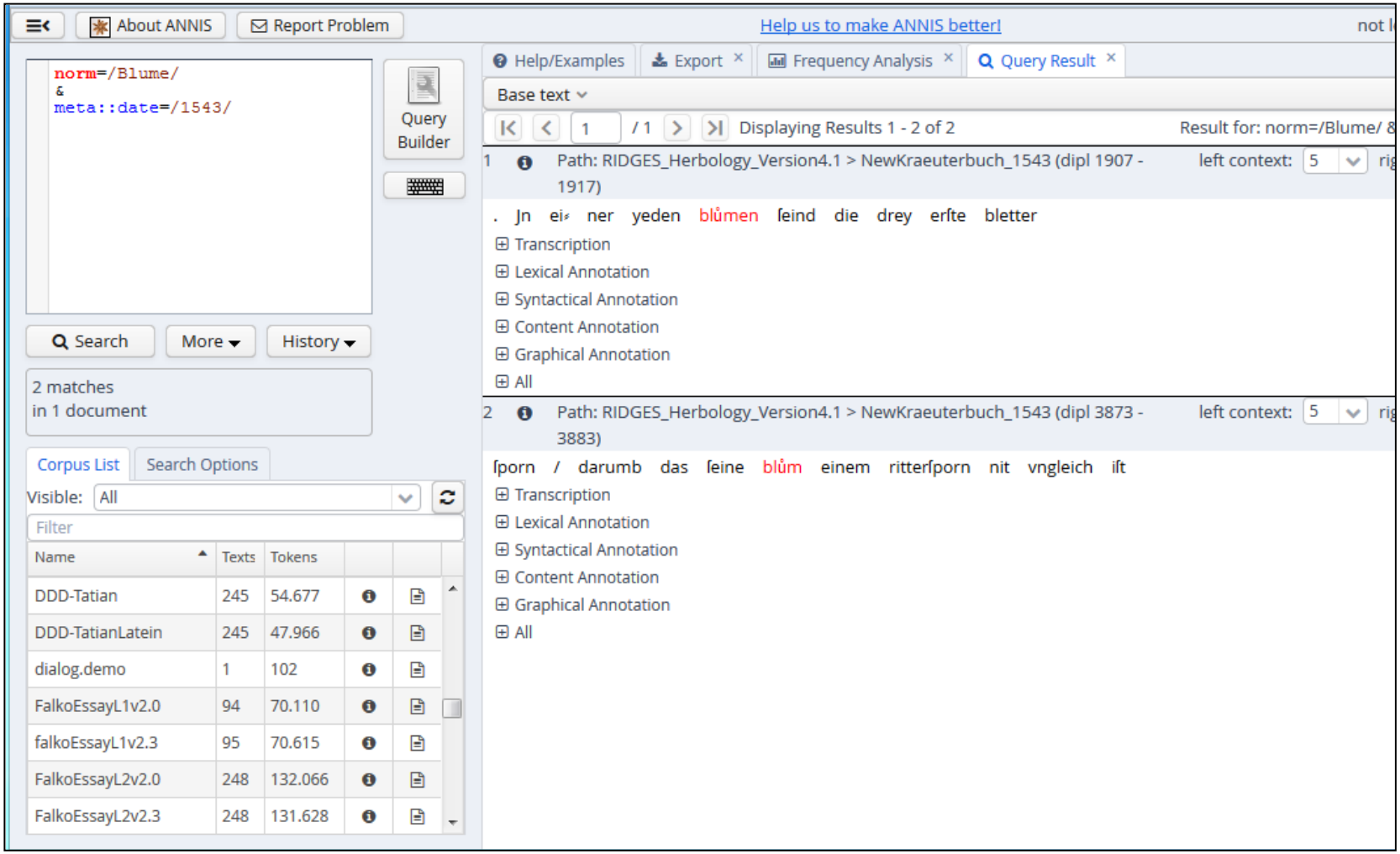
- Suchen Sie Partikelverben, die die Partikel *auf* beinhalten!
  - passend wären **lemma** und **pos**
  - Operator **_=_**
  - ➢ pos=/V.*/ _=_ lemma=/auf.*/

  Was findet man damit nicht?

Aufgabe Metadaten

- Finden Sie heraus, ob die normierte Wortform *Blume* in einem Text aus dem Jahr 1543 zu finden ist!
- Wenn ja, wie oft?
  - passende Variablen wären **norm** und **meta::date**
  - Verknüpfungszeichen **&**
  - ➢ norm=/Blume/ & meta::date=/1543/
  - ➢ 2 Treffer

Aufgabe Metadaten



Tipps

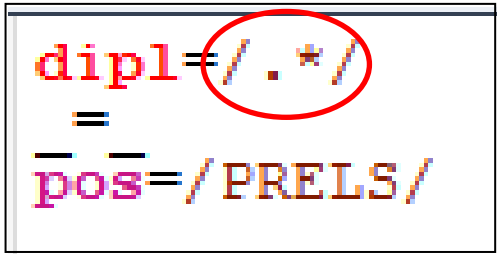
- Die Trefferliste zeigt im Kontext nur die Annotationsebenen (Variablen) an, die auch genau in diesem Trefferkontext annotiert worden sind!
- Schauen Sie in die Korpusmetadaten und Annotationsrichtlinien, um zu wissen, welche Annotationsebenen im Korpus vorhanden sind!
- Nicht alle Dokumente in einem Korpus müssen die gleichen (Anzahl und Typ) Annotationsebenen besitzen!

Zusammenfassung

- Suche in ANNIS basiert auf
  - Variablen (Annotationsebenen) und Werten (Kategorien in den Annotationsebenen), z.B.:
    - Suche von **exakten Werten**, z.B. pos=/ADJA/
    - Suche von **Mustern**, z.B. pos=/ADJ./
  - Variable-Wert-Paare könne in Relation zu einander gesucht werden, z.B.:
    - Annotationen **überlappen** sich
    - Annotationen decken einen **identischen Bereich** ab
  - Variable-Wert-Paare können in Verbindung mit Metadaten gesucht werden, z.B.:
    - pos=/ADJA/ & meta::date=/1870/

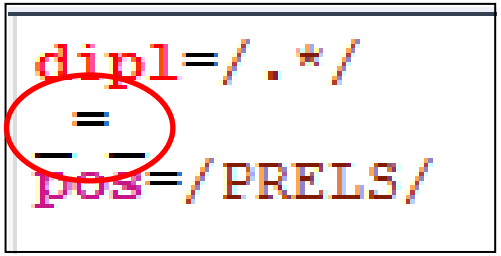
wichtige **Operatoren** für Mustersuche (VW-Paar)

.	Ein beliebiges Zeichen
?	0 oder 1 Zeichen (des vorherigen Elementes)
*	0 bis unendlich viele Zeichen (d. vorh. E.)
+	1 bis unendlich viele Zeichen (d. vorh. E.)
\\	wörtlich (folgendes Zeichen)
!	nicht
[abc]	Menge (oder [^abc]=alles außer abc)
(a|b)	a oder b (auch: [ab])
a{2,3}	a 2 bis 3mal



wichtige **Relationen** zwischen VW-Paaren

.	Direkte Präzedenz
.*	Indirekte Präzedenz
_=_	Identische Abdeckung
_i_	Inklusion
_o_	Überlappung
_l_	linksseitige Überlappung
_r_	rechtsseitige Überlappung



Vielen Dank!

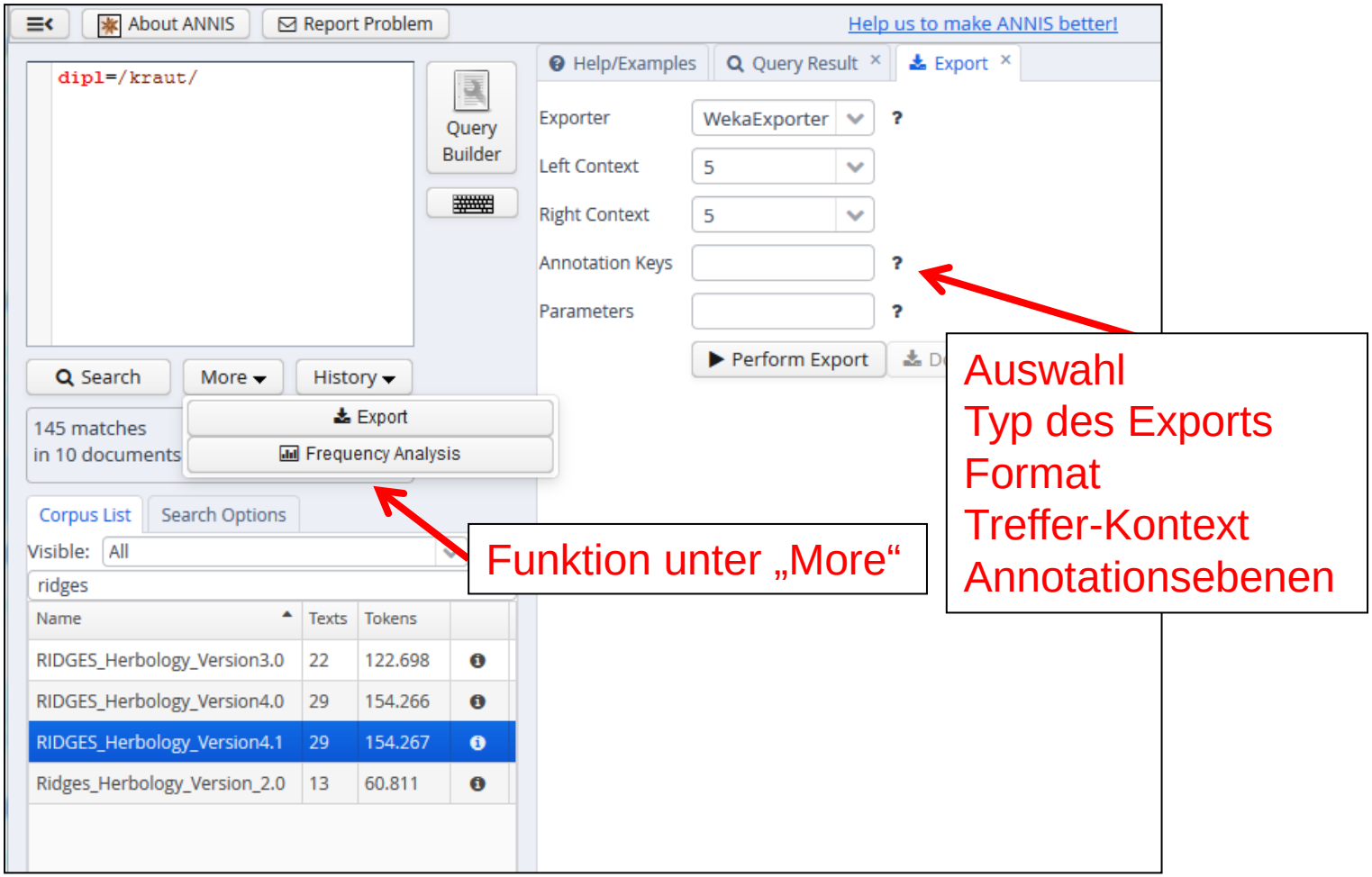
Anhang

- Interface Export
- Interface Frequency Analysis

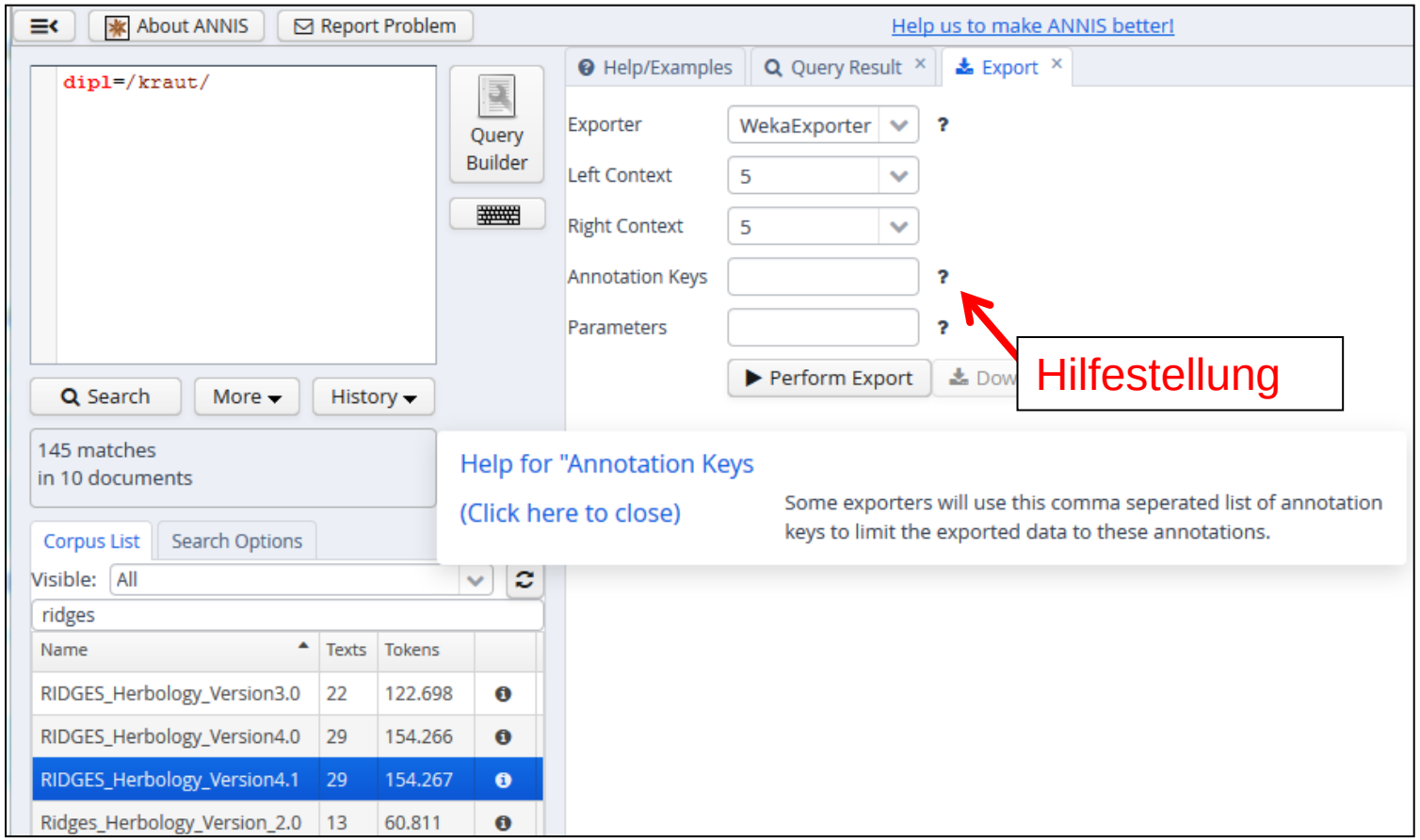
Interface

Export, Frequenzanalyse

Interface
Export von Treffern



Interface
Export von Treffern



Interface
Export von Treffern

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
Export von Treffern

Wekaexporter

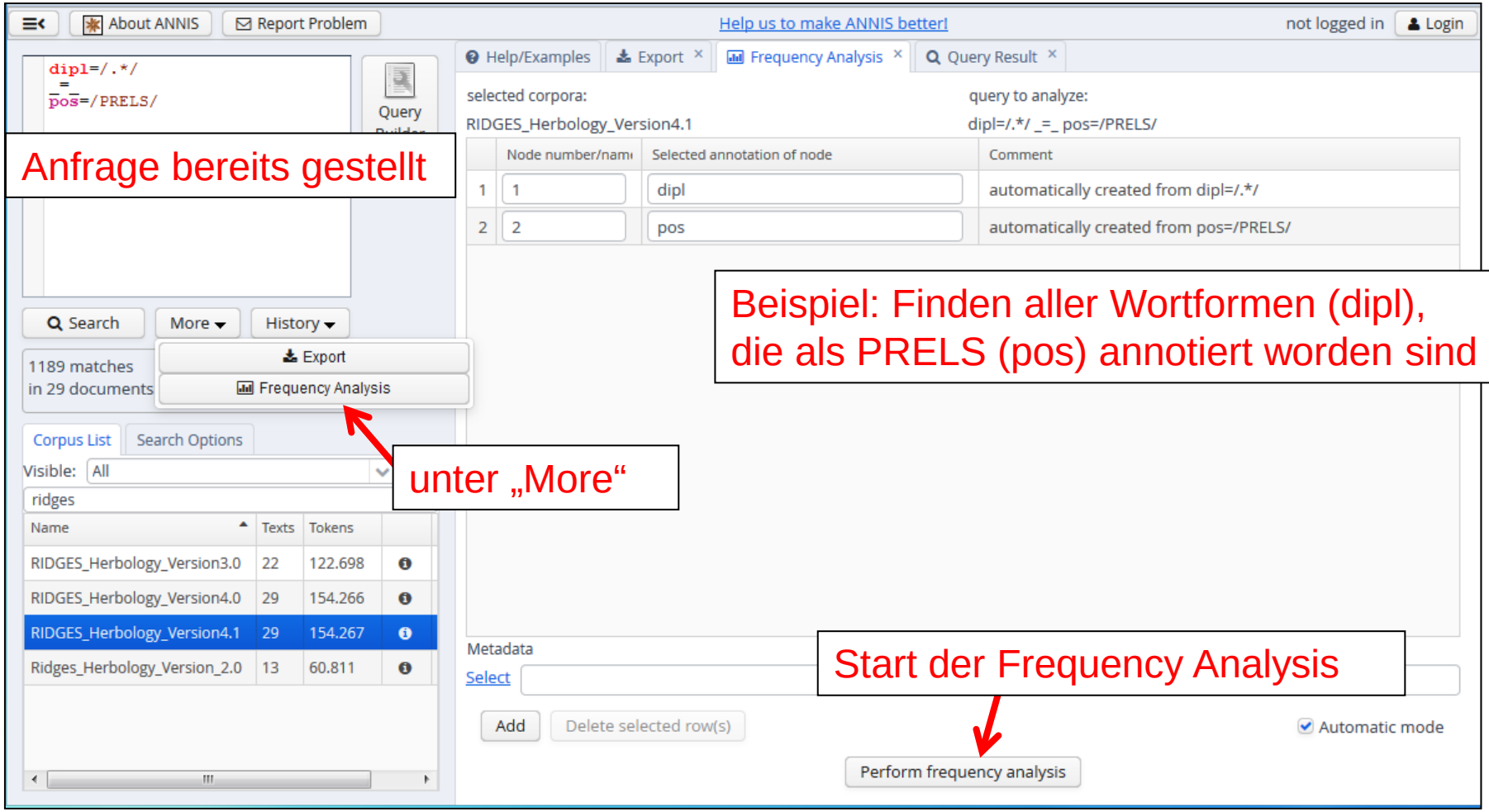
```
@relation name

@attribute #1_id string
@attribute #1_span string
@attribute #1_anno_const:cat string
@attribute #2_id string
@attribute #2_span string
@attribute #2_anno_GUM:claws5 string
@attribute #2_anno_GUM:lemma string
@attribute #2_anno_GUM:pos string

@data

'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



Interface
Frequency Analysis