Handbuch zur Nutzung des Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)

Fabian Barteld, Katharina Dreessen,
Sarah Ihden, Ingrid Schröder
(Institut für Germanistik – Universität Hamburg)

30. Januar 2021

Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200-1650)

gefördert durch die

DFG

Inhaltsverzeichnis

1 Zugänge zum ReN **3**

2 Korpusumfang **3**

3 CorA(-ReN)-XML **3**

4 Leseversionen der ReN-Texte **4**

5 TEI **4**

6 Das ReN mit ANNIS durchsuchen **4**

6.1 Allgemeine Hinweise 4

6.2 Oberfläche 5

6.3 Textansichten 6

6.4 Korpusinformationen 8

6.4.1 Node Annotations 10

6.4.2 Meta Annotations 10

6.5 Die Annotationen 10

6.6 Suchanfragen 13

6.6.1 Beispielanfragen in ANNIS 13

6.6.2 Vereinfachte Suche über die Projekthomepage 14

6.6.3 Suche über die ANNIS-Oberfläche 14

6.6.3.1 Suche nach konkretem Inhalt 14

6.6.3.2 Suche mithilfe regulärer Ausdrücke 14

6.6.3.3 Kombination mehrerer Suchausdrücke 15

6.6.3.4 Suchen innerhalb von Satzeinheiten 16

6.6.3.5 Einbinden von Metadaten in die Suchanfrage 16

6.6.4 Ausgewählte Suchanfragen zur Illustration 17

6.6.4.1 Groß- und Kleinschreibung von Substantiven 17

6.6.4.2 Komposita mit dem Grundwort *man* 18

6.6.4.3 Personalpronomen im Genitiv 18

6.6.4.4 Genitivattribute: voran- oder nachgestellt 18

6.6.4.5 Verbstellung im zweigliedrigen Verbalkomplex in *dat*-Sätzen 19

6.6.4.6 Das *wērden*-Passiv 20

6.6.4.7 Die *wēsen*-Periphrase 20

6.6.4.8 Der *tô*-Infinitiv 21

6.6.4.9 Doppelte Negation 22

6.6.4.10 Großschreibung am Satzanfang 22

6.7 Weiterverarbeiten der Suchergebnisse 22
6.7.1 Frequenzanalyse (frequency analysis) 22
6.7.1.1 Personalpronomen im Genitiv 22
6.7.1.2 Groß- und Kleinschreibung von Substantiven 23
6.7.1.3 Modalverben 25
6.7.1.4 Lemmaliste und Vollformlexikon 26
6.7.2 Export 27
6.7.2.1 CSVExporter 27
6.7.2.2 CSVMultiTokExporter 27
6.7.2.3 WekaExporter 28
6.7.2.4 TextColumnExporter 29
6.7.2.5 GridExporter 30
6.8 Weitere Informationen zu ANNIS 30

1 Zugänge zum ReN

Das Korpus besteht aus zwei Teilkorpora: einem annotierten Teil (***ren-anno***: in der Version **ReN 1.0** mit **146 Texten** und **1.415.362 Token**) und einem nur als Transkripte vorliegenden Teil (***ren-trans***: in der Version **ReN 1.0** mit **89 Texten** und **908.682 Token**). Beide Teilkorpora lassen sich in unterschiedlichen Formaten herunterladen:

- CorA-XML
- CorA-ReN-XML
- Leseversion
- TEI
- relAnnis
- relAnnis-split

2 Korpusumfang

Das ReN ist in mehreren Versionen (ab ReN 0.1 vom 23.08.2016) publiziert. Der Umfang der für die eigenen Suchabfragen verwendeten Korpusversion lässt sich einfach über eine entsprechende Abfrage in ANNIS ermitteln. Dabei ist zu beachten, dass es verschiedene Formen von Token gibt (z.B. *tok_anno* vs. *tok_dipl*, genauer hierzu s.u.). Soll z.B. die Anzahl der annotierten Token (hier sind z.B. Getrennt- und Zusammenschreibung normalisiert) ermittelt werden, muss im ANNIS-Suchfenster lediglich *tok_anno* eingegeben werden. Auf diese Weise kann auch der Umfang einzelner Subkorpora ermittelt werden. Für die beiden letzten Versionen ReN 1.0 und ReN 1.1 bspw. ergeben sich für die beiden oben erwähnten Teilkorpora die folgenden Zahlen für *tok_anno*:

ReN 1.0
- ren-anno: 1.415.362 Token
- ren-trans: 908.682 Token

ReN 1.1
- ren-anno: 1.485.963 Token
- ren-trans: 838.400 Token

Hinweis zur Qualitätssicherung der Daten:

Das ReN ist ein Korpus semi-automatisch annotierter Texte. Wie in der Korpus- und Computerlinguistik gut bekannt ist, enthalten auch manuell erstellte Ressourcen Annotationsfehler (vgl. u.a. Eskin 2000, Nakagawa/Matsumoto 2002, Dickinson 2005).

Um eine möglichst hohe Qualität der Annotationen sicherzustellen, wurden mehrere Verfahren angewendet:

- Jeder Text wurde zunächst automatisch annotiert und i.d.R. danach von zwei Annotator*innen nacheinander korrigiert, d.h., dass jede annotierte Wortform durch zwei Augenpaare geprüft wurde.
- Um eine möglichst konsistente Anwendung der Annotationsrichtlinien durch die unterschiedlichen Annotator*innen sicherzustellen, wurden Inter-Annotator Agreement-Experimente durchgeführt (vgl. Barteld el. 2018).
- Vor jeder Veröffentlichung wurden im sogenannten Postprocessing mithilfe manuell erstellter Regeln mögliche Fehler identifiziert. Hierbei wurde v.a. die interne Konsistenz zwischen unterschiedlichen Annotationsebenen (z.B. Lemma und dazu gehörige Wortart und Flexionsmorphologie) geprüft.
- Version 1.0 enthält im Vergleich zu Version 0.9 keine zusätzlichen annotierten Texte. Mithilfe von Suchanfragen nach unmöglichen bzw. unwahrscheinlichen Annotationsabfolgen wurden in Version 0.9 systematisch Annotationsfehler gesucht und diese für Version 1.0 korrigiert.

3 CorA(-ReN)-XML

CorA-XML ist das von dem zur Annotation des Korpus genutzen Annotationstool CorA generierte Dateiformat. Eine Beschreibung von CorA-XML findet sich in der Anleitung zu CorA.

Die CorA-XML-Version enthält die direkt aus CorA exportierten Daten, die lediglich um projektinterne Kommentare bereinigt wurden. Falls die Texte mit CorA bearbeitet werden sollen, empfiehlt es sich diese Version zu verwenden.

Ein Hinweis zur Darstellung der Token: CorA verwendet intern drei Darstellungen, die "trans", "utf" und "ascii" genannt werden (für eine Beschreibung s. die CorA-Anleitung). Bei ReN enthält "trans" die Tokendarstellung, wie sie in den Transkriptionsrichtlinien beschrieben ist. "utf" enthält eine um Metazeichen und gestrichene Teile bereinigte Version der Transkription. Bei "ascii" sind zusätzlich "=" und "-" entfernt – anders als es der Name nahelegt, handelt es sich also nicht um eine reine Ascii-Version des Textes.

Die CorA-ReN-XML-Version enthält einige Anpassungen, wodurch es sich nicht mehr um von CorA lesbare Dateien handelt. Unter anderem wurden die mod(ernized) Token zu anno(tierten) Token umbenannt, wie es auch bei den als CorA-XML vorliegenden Korpora Referenzkorpus Mittelhochdeutsch und St. Anselmi Fragen an Maria der Fall ist. Zusätzlich wurden weitere Annotationen automatisch aus den existierenden Annotationen abgeleitet und hinzugefügt (z.B. die unterschiedlichen Lemmavarianten). Falls mit den Daten weitergearbeitet werden soll, aber nicht CorA verwendet wird, empfiehlt es sich, diese Version des Korpus zu verwenden. Diese Version bildete auch die Grundlage für die Konvertierung der Daten in die anderen verfügbaren Formate.

(Hinweis: Die in vorherigen Korpus-Versionen enthaltene CorA-XML-Version der Texte entspricht CorA-ReN-XML).

4 Leseversionen der ReN-Texte

Volltextversionen der Texte sind als pdf und html verfügbar. Zeilenumbrüche und Leerzeichen richten sich hier nach der Orignialhandschrift bzw. dem Druck (d.h. es wurde die diplomatische Tokenisierung verwendet). Jede Zeile beginnt mit der Angabe der Blattnummer sowie, falls vorhanden, der Spalte und der Zeilenzählung.

Aufgelöste Kürzel sind kursiv dargestellt, Paratext ist durch Fettdruck ausgezeichnet, gestrichene Passagen sind als durchgestrichen formatiert und schlecht lesbare Textstücke stehen in eckigen Klammern.

5 TEI

Die TEI-Version ist eine Version des Korpus, die den Konventionen der Text Encoding Initiative P5 folgt. Sie ist keine vollständige Umsetzung der im CorA-XML-Format enthaltenen Informationen.

Der Textkörper ist als ein Block umgesetzt. Seiten-, Spalten- und Zeilenumbrüche sind durch die Tags pb, cb und lb ausgezeichnet.

Der Text ist als eine Folge von Wörtern gegeben, die den annotierten Token im CorA-XML entsprechen. Die Annotationen sind als Attribute umgesetzt. Falls zwei oder mehr Wörter in der Handschrift zusammengeschrieben sind, ist dies durch das Attribut join ausgezeichnet. Die Getrenntschreibung eines Wortes ist durch das Vorhandensein eines Zeilenumbruchs oder des Tags space gekennzeichnet. Die Satzgrenzenmarkierung ist mit s umgesetzt.

Unlesbare Passagen sind als gap dargestellt. Schlecht lesbare Passagen sind mit unclear, Auflösungen von Kürzeln mit expan, Streichungen mit del und Textfortsetzungen, Korrekturen sowie Ergänzungen sind mit add markiert.

Metatextuelle Paratexte sowie editorische Kommentare sind als note ausgezeichnet. Editorische Kommentare sind durch das Attribut type=editorial gekennzeichnet.

6 Das ReN mit ANNIS durchsuchen

6.1 Allgemeine Hinweise

Für die Nutzung des ReN in ANNIS gibt es zwei Möglichkeiten:

1. Die aktuellste Version des Korpus ist über das Hamburger Zentrum für Sprachkorpora (HZSK) unter https://corpora.uni-hamburg.de/annis/ren online verfügbar.

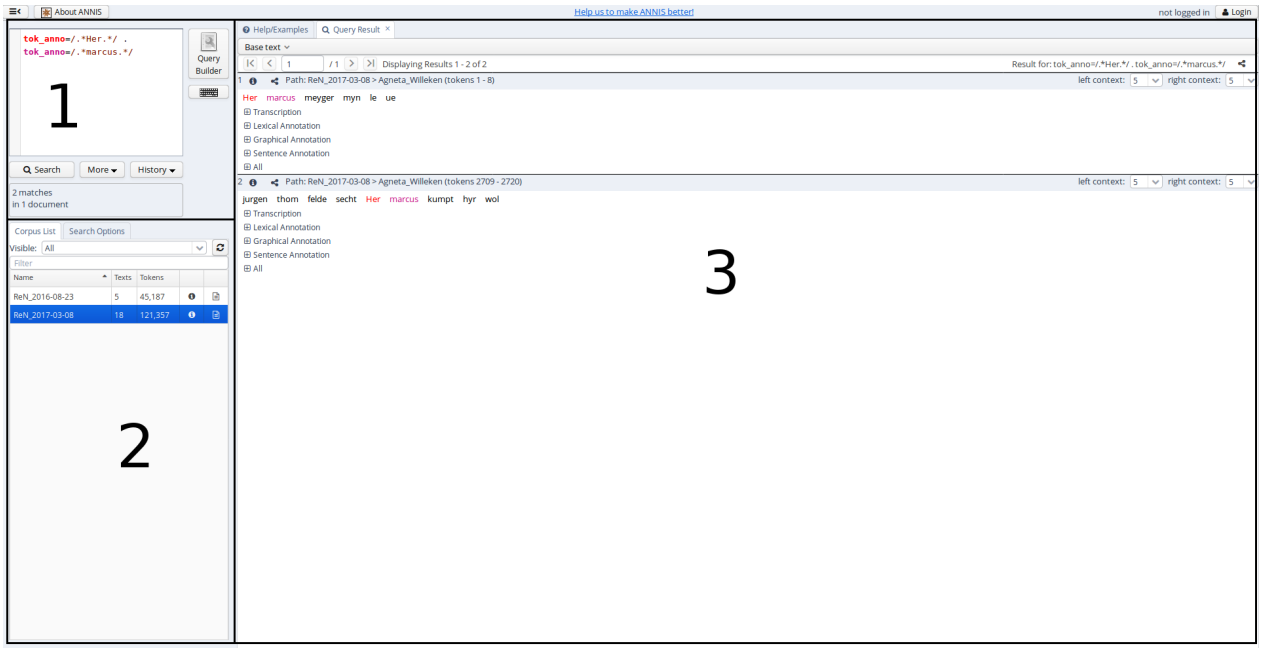
2. Alle Korpus-Versionen des ReN lassen sich im relAnnis-Format herunterladen und mithilfe des ANNIS Kickstarter auf dem eigenen Computer nutzen.

Ab der Version 0.9 (2019-05-28) ist das Korpus in zwei Varianten im relAnnis-Format verfügbar: Einmal komplett als ein Korpus (wie die vorherigen Versionen auch) und einmal als nach Sprach- und Zeitraum in einzelne Korpora aufgesplittete Version. Falls Abfragen mit dem vollständigen Korpus zu lange dauern, empfiehlt es sich, die aufgesplittete Version zu nutzen. Um im gesamten ReN zu suchen, müssen hierbei alle Teilkorpora in der Korpusliste ausgewählt sein (s. die Beschreibung im nächsten Abschnitt).

Ab Version 1.0 besteht das Korpus zusätzlich aus den Teilkorpora mit annotierten Texten ("anno") und reinen Transkriptionen ("trans").

6.2 Oberfläche

Die Oberfläche von ANNIS ist dreigeteilt:



1. Eingabefläche für die Suchanfragen
2. Korpusliste/ Suchoptionen
3. Inhaltsbereich

In der Korpusliste befinden sich die Korpusnamen. Im Falle von ReN ist das der Name *ReN* gefolgt von einem Unterstrich und dem Veröffentlichungsdatum der entsprechenende Korpusversion (z.B. *ReN_2016-08-23* für die erste publizierte Version). Hier gibt es zwei Optionen:

1. Korpusinformationen: zeigt Informationen zum Korpus an; hier lassen sich z.B. die Metadaten zu den einzelnen Texten finden.
2. Textliste: zeigt eine Liste der im Korpus enthaltenen Texte an; hier lassen sich Volltextansichten aufrufen.

Um in einem Korpus zu suchen, muss es in der Korpusliste ausgewählt sein. Dies ist daran erkennbar, dass es blau unterlegt ist.

6.3 Textansichten

Über einen Klick auf den Textliste-Button (2) öffnet sich im Inhaltsbereich die Textliste. Für jeden der aufgelisteten Texte lassen sich über den Info-Button rechts die Metadaten anzeigen.

Help/Examples | ReN_2017-09-05

Filter documents by name

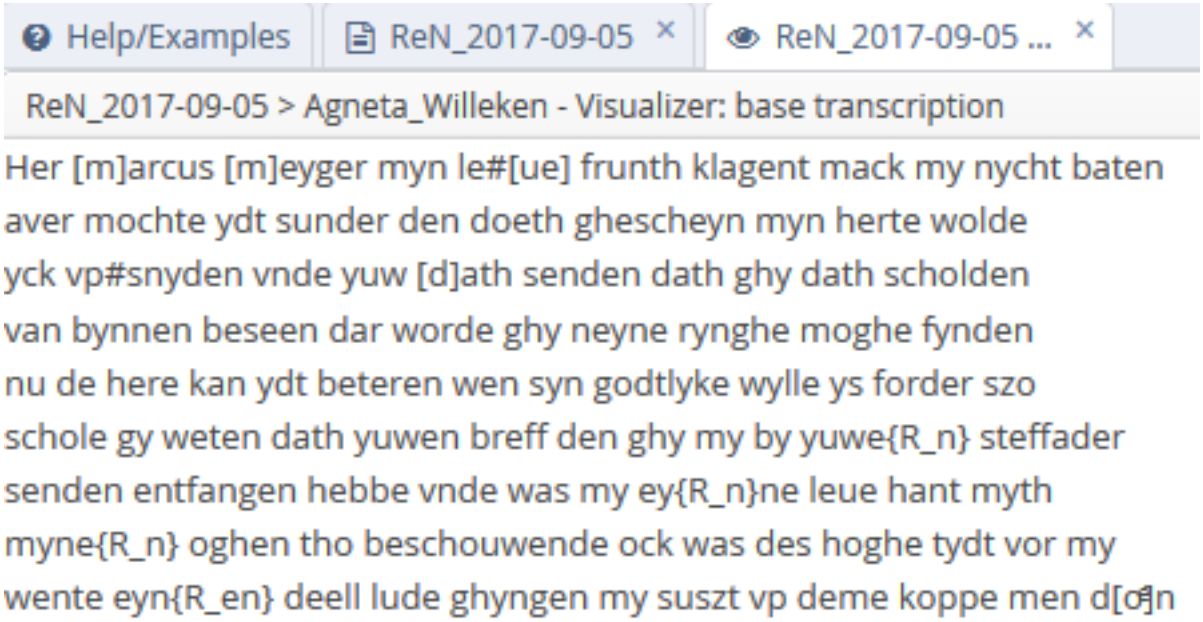
document name	corpus path	visualizer	info
Agneta_Willeken	ReN_2017-09-05 > Agneta_Willeken	base transcription diplomatic transcription modernized transcription	
Alexander_Helmst.			
Bord._Marien-Kl.			
Brem._Ssp.			
Brem._StR_1303,08			
Brem._Uk._1301-1350			
Brem._Uk._1351-1400			
Brem._Uk._1401-1450			
Brem._Uk._1451-1500			
Brs._Ius_Otton._1227			
Buxteh._Ev.			

metadata doc Agneta_Willeken

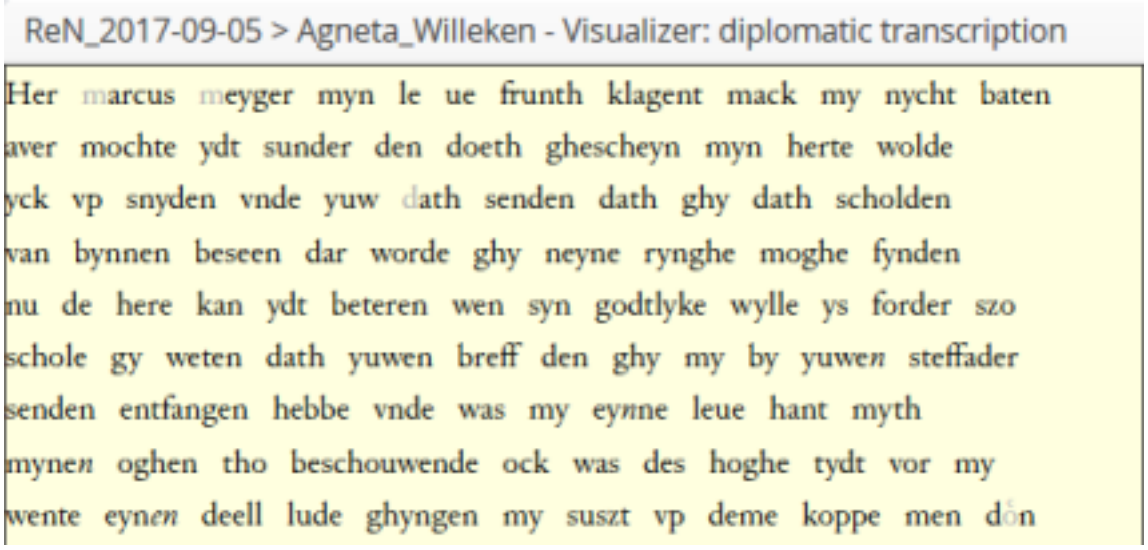
Name	Value
abbr_ddd	Agneta Willeken
annotation_by	SIH (Hamburg)
author	Agneta Willeken
base_for_transcription	Handschrift/Druck
collation_by	HMI (Hamburg)
columns	---
corpus	---
date_ReN	28.08.1535 (vgl. Reincke 1928)
digitization_by	ARE (Hamburg)
drawer	---
edition	a) Fischer, Christian (1988): Untersuchungen zur Stadtsprache Hamburgs im 16. Jahrhundert. Abschrift eines Liebesbriefes von Agneta Willeken an Marx Meyer. Wiss. Hausarbeit zur Erlangung des akad. Grades eines Magister Artium. Hamburg: Universität Hamburg.; b) Meier, Jürgen/ Möhn, Dieter (2008): Spuren der Vergangenheit für die Gegenwart. Hundert niederdeutsche Texte zwischen dem 9. und 17. Jahrhundert. Bremen: Verlag Schuster Leer (Schriften des INS).; c) Reincke, Heinrich (1928): Agneta Willeken. Ein Lebensbild aus Wullenwevers Tagen. Lübeck (Pfingstblätter des Hansischen Geschichtsvereins; Blatt 19)
extent	7
external_source	---
extract	gesamt
genre	P
hands	---
illustration	---
language_area	nordniedersaechsisch
library	Hamburg: Staatsarchiv Hamburg
library-shelfmark	Nr. 211-2; Reichskammergericht, W30 I
literature	Schampier, Andrea (2011): Agneta Willeken: Ein historischer Roman aus Hamburg und Lübeck. 2. Aufl. Fischerhude: Verl. Atelier im Bauernhaus - darin enthalten: Teilübersetzung.
medium	Handschrift
notes_annotation	
notes_language	
notes_manuscript	Es handelt sich um eine zeitgenössische Abschrift.
notes_transcription	Im Schriftbild schwierig zu erkennen: Groß- und Kleinschreibung, Unterschied zwischen <w>, <v> und <m> sowie zwischen <i> und <j>

Außerdem gibt es links vom Info-Button in der Spalte *visualizer* drei Möglichkeiten der Volltextansicht:

base transcription
Die Transkriptionsansicht enthält den Text mit dem Markup (z.B. Auflösung von Abbreviaturen in geschweiften Klammern, schlecht Lesbares in eckigen Klammern etc.), wie er transkribiert wurde (s. Transkriptionshandbuch). Die Ansicht orientiert sich stark an der projektinternen Transkription, bildet sie aber nicht eins zu eins ab (z.B. erscheinen in der *base transcription* keine Blattangaben und keine Markierungen der Satzgrenzen; Blattangaben finden sich in der *diplomatic transcription*, Satzgrenzenmarkierungen in der *modernized transcription*, s.u.)



diplomatic transcription
Die diplomatische Ansicht zeigt eine manuskriptnahe Version des Textes: Seiten, Spalten und Zeilen werden dem Original entsprechend abgebildet; auch die Getrennt- und Zusammenschreibung erfolgt wie im Original.



Folgendes Markup aus der Transkription wird durch Formatierungen in der diplomatischen Textansicht umgesetzt:

- schlecht lesbare Textabschnitte `[...]`: hellgraue Schriftfarbe
- Streichungen `‡...‡`: ~~gestrichen~~
- Auflösungen von Abbreviaturen `{*_...}`: *kursiv*
- Paratext `*[I/L/R/O/U/T][N/K/E]_...*`: ^{hochgestellt}

modernized transcription
Diese Ansicht orientiert sich an der linguistischen Annotation: Die Getrennt- und Zusammenschreibung basiert auf den annotierten Worteinheiten. Eine Zeile entspricht hier einer Satzeinheit (s. Transkriptionshandbuch). Im Fall getrennter Satzeinheiten (z.B. durch eingeschobene Nebensätze) wird am Ende der Zeile *(I)* für den initialen, *(F)* für den finalen und – falls vorhanden – *(M)* für den medialen Bestandteil der getrennten Satzeinheit angezeigt. PoS- und flexionsmorphologische Annotation sowie Lemma werden angezeigt, wenn man mit der Maus auf eine Wortform fährt.

6.4 Korpusinformationen

Über einen Klick auf den Korpusinformations-Button (1) werden die Korpusinformationen in einem kleinen Fenster geöffnet.

Corpus information for ReN_2017-09-05 (ID: 3283)

No metadata available

Select corpus/document: ReN_2017-09-05

ReN_2017-09-05
Agneta_Willeken
Alexander_Helmst.
Bord._Marien-Kl.
Brem._Ssp.
Brem._StR_1303,08
Brem._Uk._1301-1350
Brem._Uk._1351-1400
Brem._Uk._1401-1450
Brem._Uk._1451-1500

1-10/41

(no metadata)

Available annotations

Node Annotations

Name	Example (click to use query)	URL
bound_head	bound_head="Überschrift"	
bound_sent	bound_sent="Satz"	
column	column="<none>"	
comment	comment="unflektiert"	
deleted	deleted="to"	
editorial_commen	editorial_comment="null"	
expanded	expanded="n"	
lemma	lemma="<none>"	
lemma_simple	lemma_simple="<none>"	
lemma_var	lemma_var="<none>"	
lemma_wsd	lemma_wsd="<none>"	
line	line="01"	
morph	morph="<none>"	

Edge Annotations

Edge Types

Meta Annotations

Link to corpus: http://annis.corpora.uni-hamburg.de/gui/ren#_c=UmVOXzIwMTctMDktMDU

Die linke Hälfte des Fensters enthält die Metadaten. Über eine Dropdown-Liste lassen sich hier die einzelnen Texte auswählen, darunter werden dann die Metadaten für den entsprechenden Text angezeigt.

Corpus information for ReN_2017-09-05 (ID: 3283)

Metadata

Select corpus/document: Agneta_Willeken

Name	Value
abbr_ddd	Agneta Willeken
annotation_by	SIH (Hamburg)
author	Agneta Willeken
base_for_transcription	Handschrift/Druck
collation_by	HMI (Hamburg)
columns	---
corpus	---
date_ReN	28.08.1535 (vgl. Reincke 1928)
digitization_by	ARE (Hamburg)
drawer	---
edition	a) Fischer, Christian (1988): Untersuchungen zur Stadtsprache Hamburgs im 16. Jahrhundert. Abschrift eines Liebesbriefes von Agneta Willeken an Marx Meyer. Wiss. Hausarbeit zur Erlangung des akad. Grades eines Magister Artium. Hamburg: Universität Hamburg.; b) Meier, Jürgen/ Möhn, Dieter (2008): Spuren der Vergangenheit für die Gegenwart. Hundert niederdeutsche Texte zwischen dem 9. und 17. Jahrhundert. Bremen: Verlag Schuster Leer (Schriften des INS).; c) Reincke, Heinrich (1928): Agneta Willeken. Ein Lebensbild aus Wullenwevers Tagen. Lübeck (Pfingstblätter des Hansischen

Available annotations

Node Annotations

Name	Example (click to use query)	URL
bound_head	bound_head="Überschrift"	
bound_sent	bound_sent="Satz"	
column	column="<none>"	
comment	comment="unflektiert"	
deleted	deleted="to"	
editorial_commen	editorial_comment="null"	
expanded	expanded="n"	
lemma	lemma="<none>"	
lemma_simple	lemma_simple="<none>"	
lemma_var	lemma_var="<none>"	
lemma_wsd	lemma_wsd="<none>"	
line	line="01"	
morph	morph="<none>"	

Edge Annotations

Edge Types

Meta Annotations

Link to corpus: http://annis.corpora.uni-hamburg.de/gui/ren#_c=UmVOXzIwMTctMDktMDU

Die rechte Hälfte zeigt die vorhandenen Annotationsebenen, die abgefragt werden können. Relevant für ReN sind *Node*

Annotations und *Meta Annotations*. Mit einem Klick darauf wird jeweils die Liste mit den Namen der Annotationsebenen und jeweils einem von ANNIS automatisch generierten Beispiel ausgeklappt (s. Abb. zu Metadaten des Textes *Agneta Willeken*).

6.4.1 Node Annotations

Hierunter verbergen sich bei ReN alle Annotationen, die unterhalb der Textebene liegen. Dazu zählen sowohl der transkribierte Text in unterschiedlichen Versionen (s.u.), Informationen zum Layout der Handschrift wie Blatt und Zeilennummern sowie alle linguistischen Annotationen auf Tokenebene. Die einzelnen Annotationsebenen werden unten genauer vorgestellt.

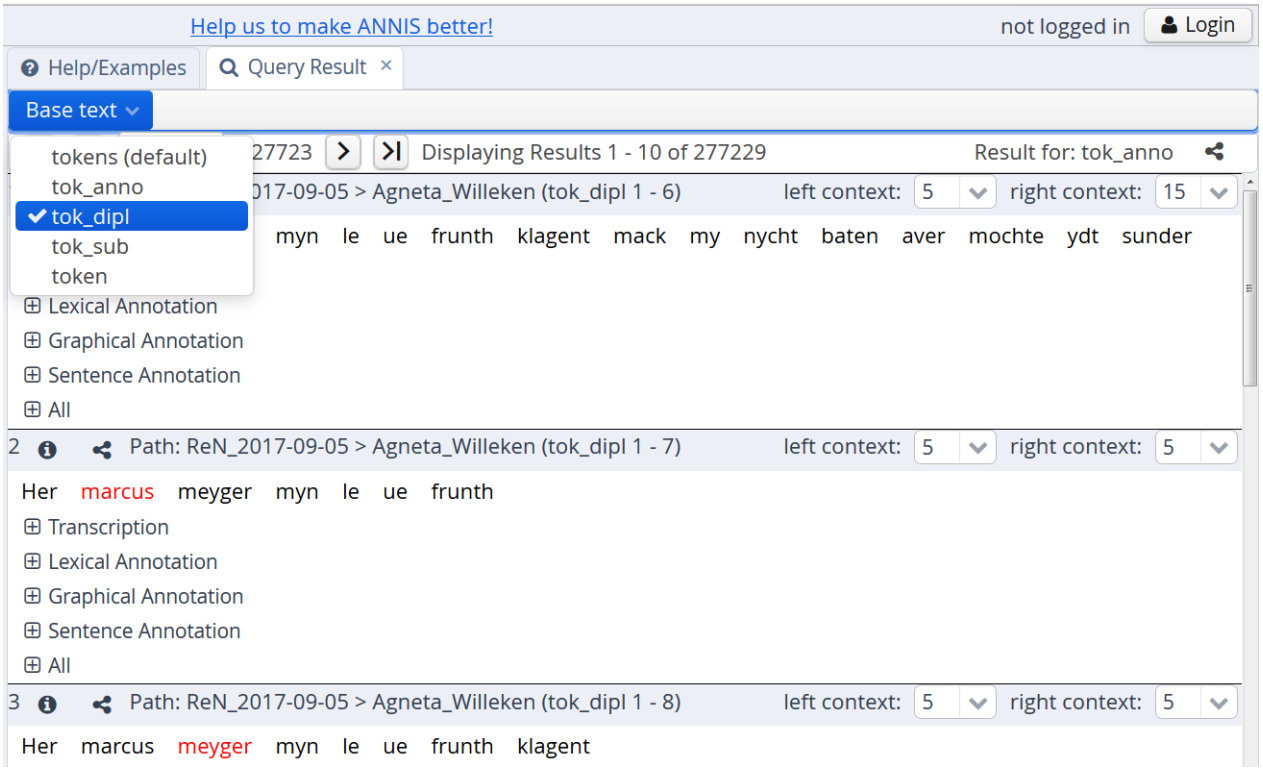
6.4.2 Meta Annotations

Hierbei handelt es sich um textbezogene Metadaten, z.B. Textkurztitel (doc), Zeitraum (time), Sprachraum (language_area) und Feld der Schriftlichkeit (topic_ReN). Diese Metadaten können in Suchanfragen eingebunden werden (s.u.).

Hinweis: Neben den üblichen Metadaten zum Text gibt es unter *notes*annotation_ Hinweise zu morphologischen und syntaktischen Auffälligkeiten im Text und zum Umgang mit textindividuellen Zweifelsfällen in der Annotation. Für grammatische Analysen mit dem Korpus ist es sinnvoll, diese Notizen ergänzend einzusehen.

6.5 Die Annotationen

Die Ergebnisse einer Suchanfrage (s.u.) öffnen sich im Inhaltsbereich der ANNIS-Oberfläche in einem neuen Reiter *Query Result*.



Oben links im Bereich *Base text* kann mithilfe einer Dropdownliste ausgewählt werden, mit welcher Tokenisierung die Suchergebnisse mit linkem und rechten Kontext dargestellt werden. Für ReN sind relevant:

tok_dipl

- Token in diplomatischer Transkription, d.h. die Tokengrenzen entsprechen dem Vorhandensein von Spatien in der Handschrift (vgl. Textansicht *diplomatic transcription*)
- v.a. sinnvoll für Suchen, bei denen die graphematischen Wörter im Fokus stehen (graphematische Analysen)

tok_anno

- Token entsprechen den annotierten Worteinheiten, d.h. die Tokengrenzen entsprechen einem syntaktischen/ morphologischen Verständnis von der Einheit „Wort“ (vgl. Textansicht *modernized transcription*)
- v.a. sinnvoll für Suchen, bei denen die syntaktischen Wörter im Fokus stehen (grammatische Analysen, inhaltliche Analysen etc.)

Beispiel: *leue* ist im Original mit Spatium zwischen /e/ und /u/ geschrieben:

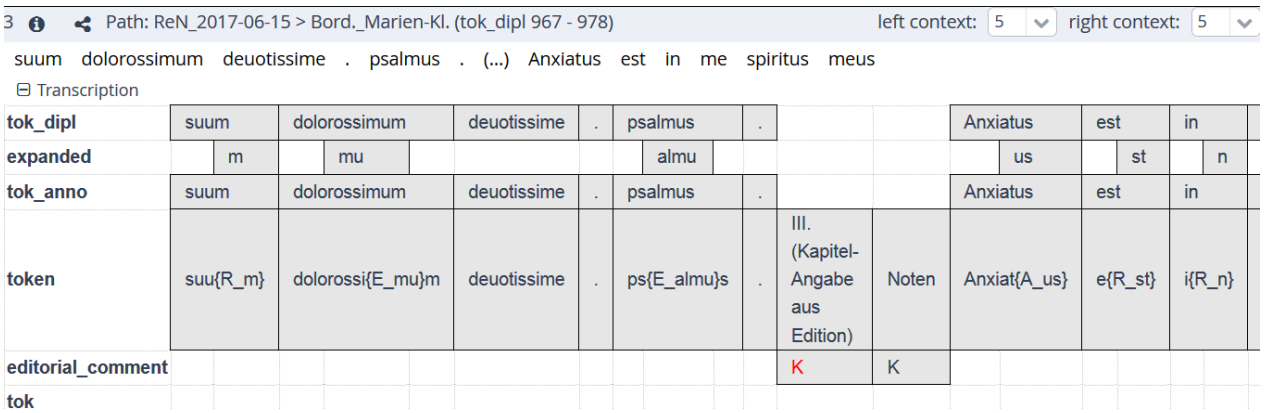
tok_dipl: *myn le ue frunth*
tok_anno: *myn leue frunth*

Die Suchergebnisse werden unter Angabe einer Treffernummer und mit dem Korpusnamen und dem Textkurztitel (doc) aufgeführt. Rechts daneben lässt sich der Umfang des linken und rechten Kontextes um den Treffer aus der Suchanfrage herum auswählen. Der Text darunter, bestehend aus Treffer und Kontext, erscheint in der unter *Base text* (s.o.) ausgewählten Tokenform. Unmittelbar unter dem Text werden die verschiedenen Gruppen von Annotationsebenen aufgeführt, die sich mit einem Klick auf das Pluszeichen links daneben aufklappen lassen:

Der Bereich **Transcription** umfasst die folgenden Annotationsebenen:

- `tok_dipl`: enthält den Text in diplomatischer Tokenisierung, d.h. Tokengrenzen dem Original entsprechend (s.o.)
- `unreadable`: enthält schwer oder nicht lesbare Elemente des Textes
- `deleted`: enthält Elemente des Textes, die im Original durchgestrichen erscheinen
- `expanded`: enthält Auflösungen von Kürzungszeichen (Abbreviaturen), z.B. das mithilfe eines Nasalstriches im Original angezeigte *n* am Ende des Wortes *lychtlyken*
- `para`: enthält paratextuelle Elemente, z.B. interlineare Korrekturen oder Notizen am Rand (s. Transkriptionshandbuch)
- `tok_anno`: enthält den Text tokenisiert entsprechend den annotierten Worteinheiten (s.o.)
- `token`: enthält Token in der projektinternen Transkription, d.h. mit Markup
- `editorial comment`: zeichnet Teile der *token*-Spur als editorischen Kommentar (d.h. nicht Teil des eigentlichen Textes) aus, so z.B. den Teil *Lateinische Textpassage*; die Auszeichnung erfolgt in der *editorial comment*-Spalte stets mit dem Buchstaben *K*
- `tok`: eine ANNIS-interne Spur, die oberflächlich leer ist und nicht für Suchanfragen verwendet wird

Allgemeiner Hinweis: Bei den Suchergebnissen werden nur diejenigen Annotationsspuren angezeigt, die auch Annotationen enthalten. Es kommt also vor, dass in einigen Belegen bestimmte Spuren nicht erscheinen, oft z.B. die Spuren *unreadable*, *deleted*, *editorial comment* oder auch *posMWU* (s.u.).



Der Bereich **Lexical Annotation** umfasst die folgenden Annotationsebenen:

- `tok_anno`: s.o.
- `lemma_wsd`: Lemma mit word sense disambiguation, d.h. Homonymendifferenzierung mithilfe hochgestellter Zahlen, und mit Diakritika z.B. für die Länge von Vokalen, Bsp.: *mîn*[2]
- `lemma`: Lemma mit Diakritika, aber ohne Homonymendifferenzierung, Bsp.: *mîn*
- `lemma_simple`: Lemma ohne Diakritika und Homonymendifferenzierung, Bsp.: *min*
- `lemma_var`: Lemma mit Variantenauflösung: Lemmata mit Klammerformen, wie z.B. *up(pe)*[1], werden hier in sämtliche Varianten (mit Komma getrennt) aufgelöst, Bsp.: *up,uppe*
- `pos`: PoS-Annotation nach dem HiNTS-Tagset (s. Annotationshandbuch Teil 1), Bsp.: *VVFIN* (Vollverb, finit)
- `posLemma`: Reduktion des PoS-Tags auf die Information zur Grundwortart ohne Angaben zu Distribution, Funktion, Flexion etc., Bsp.: *VV* (Vollverb); außerdem wird diese Spur für die Angabe der Grundwortart bei einem Wortartenwechsel verwendet, Bsp.: das Token *bedroweden* im Kontext *mennygen bedroweden dach* trägt auf der Ebene *pos* das Tag *ADJA* und auf der Ebene *posLemma* das Tag *VVPP*.
- `morph`: Flexionsmorphologische Annotation nach dem HiNTS-Tagset (s. Annotationshandbuch Teil 1), Bsp.: _Irr.3.Pl.Past.*_
- `comment`: enthält Kommentare der Annotator*innen z.B. zu ungewöhnlichen Formen oder zum Verständnis der Annotation, Bsp.: *Subjektellipse*
- `posMWU`: Spanne mit einem PoS-Tag über mehrere Token, die zusammen eine Mehrworteinheit (multi word unit) bilden, Bsp.: *KOUS*

2 Path: ReN_2017-06-15 > Agneta_Willeken (tok_dipl 247 - 259) left context: 5 right context: 5

froude schal to herten ghan wo wol dath godt de here do szo

Transcription

Lexical Annotation

tok_anno	froude	schal	to	herten	ghân	wowol	dath	godt	de	
lemma_wsd	vröude	schölen[1]	tô[1]	herte[1]	gân	wô	dat[2]b	got[1]	dê[1]/dê[1]/dat[2]a	
lemma	vröude	schölen	tô	herte	gân	wô	dat	got	dê/dê/dat	
lemma_simple	vroude	scholen	to	herte	gan	wo	dat	got	de/de/dat	
lemma_var	vroude	scholen	to	herte	gan	wo	dat	got	de/de/dat	
pos	NA	VMFIN	APPR	NA	VVINF	KOUS	KOUS	NA	DDARTA	
posLemma	NA	VM	AP	NA	VV	KO	KO	NA	DD	
morph	Fem.Nom.Sg	Prpr.3.Sg.Pres.Ind	Dat	Neut-Fem.Dat.Sg	<none>	<none>	<none>	Masc.Nom.Sg	Masc.Nom.Sg	
posMWU						KOUS				

Der Bereich **Graphical Annotation** umfasst die folgenden Annotationsebenen:

- `tok_dipl`: s.o.
- `unreadable`: s.o.
- `page`: Nummer des Blattes, auf dem der Text steht, Bsp.: *01*; besteht keine Differenzierung zwischen recto und verso, handelt es sich (gleichzeitig) um die Nummer der Seite
- `side`: Benennung der Seite (Vorder- oder Rückseite), auf der der Text steht: *r* für die Vorderseite (recto), *v* für die Rückseite (verso)
- `column`: Spalte, in der der Text steht, Bsp.: *a*, *b* oder auch *<none>* (wenn es keine Einteilung in Spalten gibt) ...
- `line`: Nummer der Zeile, in der der Text steht, Bsp.: *08*, *42* ...

1 Path: ReN_2017-06-15 > Brem._Ssp. (tok_dipl 128 - 309) left context: 10 right context: 5

. dar vmbe ist eme recht lef . dar l vmbe sen se sic vore al de den ghe richte van godes haluen be ualen is

⊞ Transcription
⊞ Lexical Annotation
⊟ Graphical Annotation

tok_dipl	.	dar	vmbe	ist	eme	recht	lef	.	dar	l	vmbe	sen	se	sic	vore	al	de	den	ghe	richte
expanded																		n		
para										l										
page	15																			
side	r									v										
column	b									<none>	a									
line	20		21							01	01								02	

Der Bereich **Sentence Annotation** umfasst die folgenden Annotationsebenen:

- `tok_anno`: s.o.
- `bound_head`: Angabe von Spannen über mehrere Token, die eine Überschrift (zur Definition s. Transkriptionshandbuch) bilden; der Inhalt ist stets *Überschrift*
- `bound_sent`: Angabe von Spannen über mehrere Token, die eine Satzeinheit (zur Definition und Auszeichnung s. Transkriptionshandbuch) bilden; folgende Inhalte sind möglich:
  - Satz: eine Satzeinheit
  - Satzteil_I: initialer Bestandteil einer Satzeinheit, die durch mindestens eine andere Satzeinheit getrennt ist
  - Satzteil_M: mediale/r Bestandteil/e einer Satzeinheit, die durch mindestens eine andere Satzeinheit getrennt ist/sind
  - Satzteil_F: finaler Bestandteil einer Satzeinheit, die durch mindestens eine andere Satzeinheit getrennt ist

1 Path: ReN_2017-06-15 > Alexander_Helmst. (tok_dipl 1 - 7) left context: 5 right context: 10

van alloxander Nectabius was en koningk to egipten vnde was en mes

⊞ Transcription
⊞ Lexical Annotation
⊞ Graphical Annotation
⊟ Sentence Annotation

tok_anno	van	alloxander	Nectabius	was	en	koningk	to	egipten	vnde	was	en	mester
bound_sent			Satz						Satz			
bound_head	Überschrift											

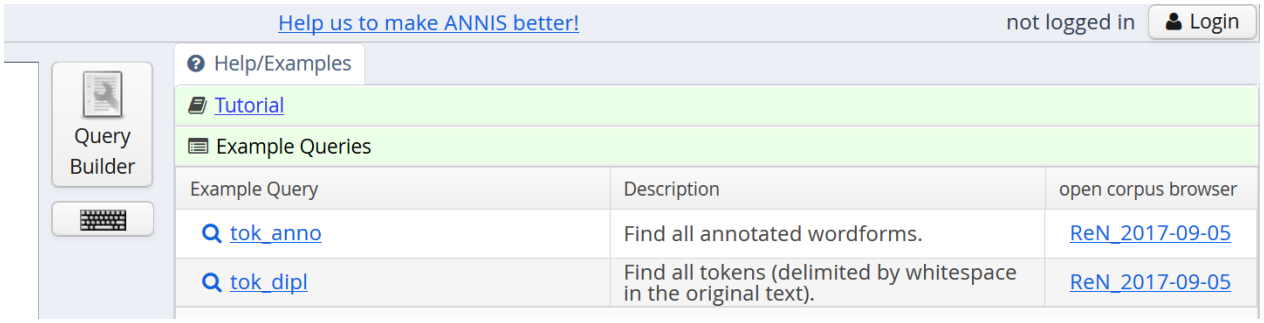
⊞ All

Der Bereich **All** führt alle der eben beschriebenen Annotationsebenen auf.

6.6 Suchanfragen

6.6.1 Beispielanfragen in ANNIS

Im Inhaltsbereich der ANNIS-Oberfläche kann unter dem Reiter *Help/Examples* im Bereich *Example Queries* zwischen verschiedenen Beispielsuchanfragen, z.B. nach sämtlichen annotierten Worteinheiten (tok_anno), ausgewählt werden. Mit einem Klick darauf werden die Suchanfragen direkt gestartet.



6.6.2 Vereinfachte Suche über die Projekthomepage

Auf der ReN-Homepage (www.referenzkorpus-mnd-nrh.de) wird unter *Datenzugang* neben dem direkten Zugang zum Korpus in ANNIS auch eine sogenannte *Einfache Suche* angeboten. Hier kann nach Wortformen (entspricht *tok_anno*, s.u.) und Lemmata (entspricht *lemma_simple*, s.u.) gesucht und die Suche mit Metadaten zu Entstehungszeit, Feld der Schriftlichkeit und Sprachlandschaft verknüpft werden.

6.6.3 Suche über die ANNIS-Oberfläche

6.6.3.1 Suche nach konkretem Inhalt

In ANNIS kann sowohl nach konkreten Inhalten als auch nach unterspezifizierten Inhalten mithilfe regulärer Ausdrücke gesucht werden. Für die Suche nach einem konkreten Inhalt einer Annotationsebene werden die doppelten Anführungsstriche am Beginnn und Ende des Suchausdruckes verwendet.
Bsp.:

```
tok_anno="wech"
lemma_wsd="wech¹"
pos="NA"
morph="Masc.Dat.Sg"
```

6.6.3.2 Suche mithilfe regulärer Ausdrücke

Zur Suche nach unterspezifizierten Inhalten dienen reguläre Ausdrücke. Hierzu zählen Operatoren, Platzhalter und Quantoren:

Ein gebräuchlicher **Operator** ist der *oder*-Operator (|) für Alternativen; die alternierenden Wörter/Buchstaben werden dafür in Klammern gesetzt, Bsp.: go(d|t) -> findet *god* und *got*
Ein häufig verwendeter **Platzhalter** ist der Punkt (.), der für ein beliebiges Zeichen steht, Bsp.: *go.* -> findet *god*, *got*, aber nicht *gode* etc.
Ein Beispiel für einen **Quantor** ist der Asterisk (*), der für keinmaliges, einmaliges oder beliebig häufiges Auftreten des vorangehenden Zeichens (auch Platzhalter) steht, Bsp.: *go.** findet *god*, *got* und auch *go* sowie *gode* etc.

Soll in einem Text nach Zeichen, die in regulären Ausdrücken als Zeichen für Operatoren, Platzhalter oder Quantoren fungieren, gesucht werden, müssen diese mithilfe des Backslashs (\) geschützt werden, Bsp.: \. findet den Punkt

Für die Suche mithilfe regulärer Ausdrücke in ANNIS wird der Schrägstrich am Beginn und Ende des Suchausdruckes verwendet.
Bsp.:

```
tok_anno=/go(d|t)/
-> findet „got“ und „god“
pos=/AV.*/
-> findet „AVD“, „AVKO“, „AVREL“, „AVNEG“, „AVW“
(sämtliche Tags, die mit „AV“ beginnen)
```

6.6.3.3 Kombination mehrerer Suchausdrücke

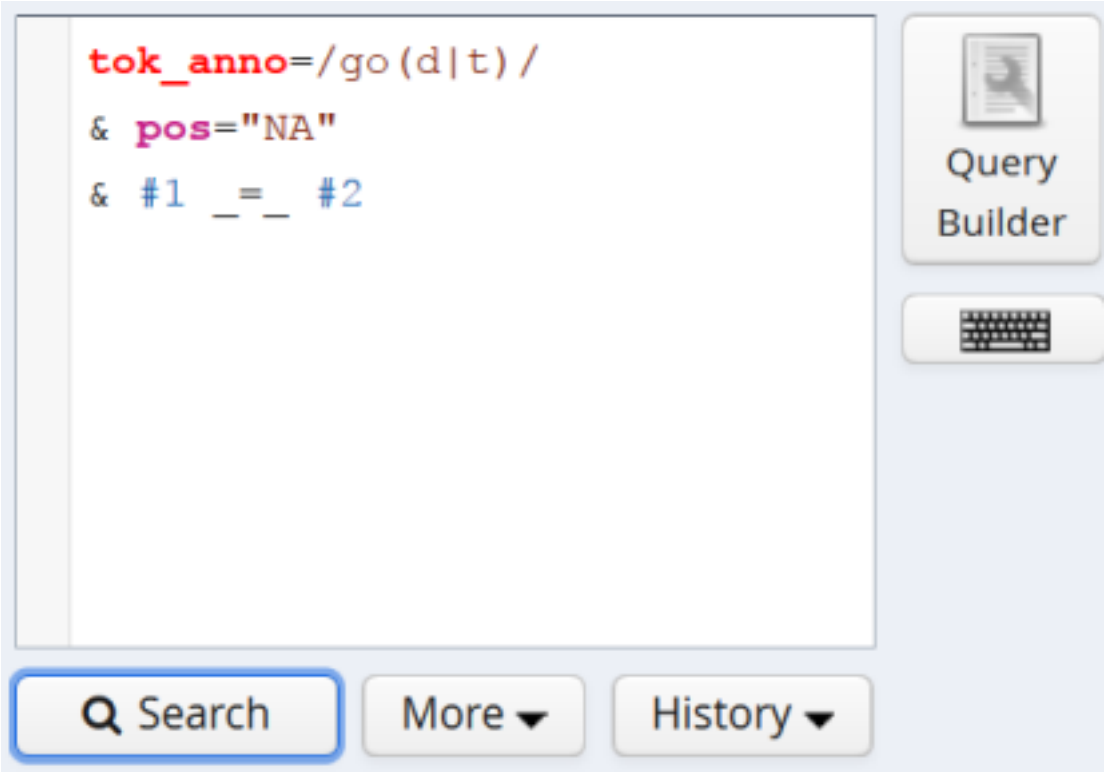
Für viele Suchanfragen ist eine Kombination mehrerer Suchkriterien (in ANNIS: *node*) nötig. Hierfür werden die einzelnen Ausdrücke mit dem &-Zeichen verbunden. Anschließend werden die Nummern der Kriterien (#1, #2 usw.) angefügt und ihr Verhältnis zueinander wird durch Operatoren angegeben, z.B. durch

- den Punkt (.) für direktes Aufeinanderfolgen
- _=_ für Übereinstimmung

Es ist auch möglich, die einzelnen Knoten direkt durch die Operatoren miteinander zu verbinden (s. Bsp. unten).

Eine Liste aller Operatoren findet sich bei ANNIS unter *Help/Examples* im *Tutorial*
Bsp.:

```
tok_anno=/go(d|t)/
& pos="NA"
& #1 _=_ #2

oder:

tok_anno=/go(d|t)/ _=_ pos="NA"

-> findet nur die Substantive, nicht die Adjektive

pos=/V.PP/
& pos=/V.FIN/
& #1 . #2

oder:

pos=/V.PP/ . pos=/V.FIN/

-> findet die Folge Partizip + Finitum
```



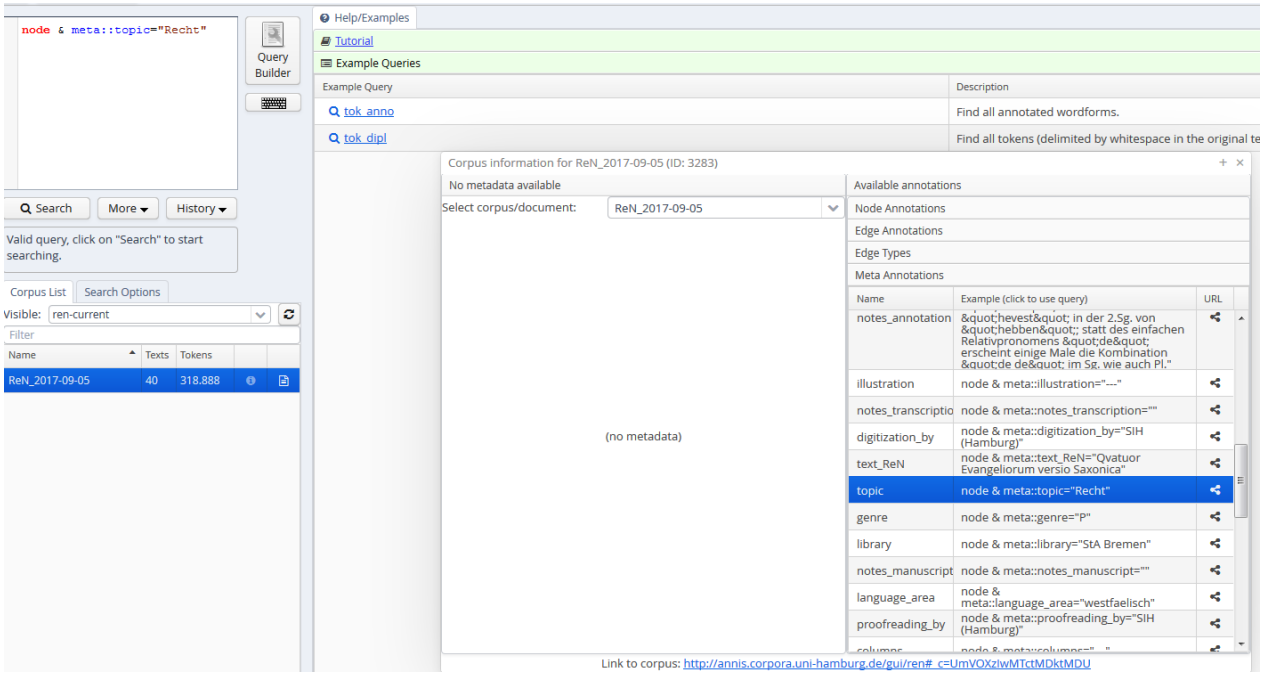
6.6.3.4 Suchen innerhalb von Satzeinheiten

Für viele v.a. syntaktische Fragestellungen sind Suchen innerhalb von Sätzen sinnvoll. Hierfür kann die Auszeichnung von Satzeinheiten auf der Annotationsspur `bound_sent` genutzt werden. Um auszudrücken, dass die Suchkriterien innerhalb einer Satzeinheit stehen sollen, muss jedes betreffende Kriterium mithilfe des Operators `_i_` (i = include) mit dem Suchkriterium `bound_sent` verbunden werden.
Bsp.:

```
pos="VMFIN"
& pos="VAINF"
& pos="VVPP"
& bound_sent
& #4 _i_ #1
& #4 _i_ #2
& #4 _i_ #3
-> findet dreigliedrige Verbalkomplexe
aus finitem Modalverb, infinitem Auxiliarverb
und Vollverb im Partizip Präteritum
innerhalb einer Satzeinheit
```

6.6.3.5 Einbinden von Metadaten in die Suchanfrage

Um eine Suchanfrage auf bestimmte Texte des Korpus einzugrenzen, können die Metadaten mit der Suchanfrage verknüpft werden. Hierfür werden am Ende des Suchausdruckes mit einem &-Zeichen der Name der gewünschten Metadatenspur und der Inhalt angefügt, z.B. `meta::doc="Griseldis"` für eine Suche innerhalb des Textes mit dem Textkurztitel *Griseldis*. Am einfachsten ist ein direktes Einfügen der Metadaten mithilfe der Auflistung der Metadatenspuren über den Info-Button (s.o.). Mit einem Klick auf die gewünschte Metadatenspur wird das dort aufgeführte Beispiel automatisch in das Suchfeld übertragen.



Achtung: Zuvor ins Suchfeld Eingetragenes wird überschrieben. Am besten ist es daher, zuerst die Metadaten einzufügen und anschließend den Bereich *node* vor dem &-Zeichen durch die gewünschten Suchausdrücke zu ersetzen, z.B. `lemma_wsd="erve²" & morph=/.*Nom.*/ & #1 _=_ #2 & meta::topic="Recht"`. Da die eingefügten Metadaten auf die gesamte Suchanfrage angewendet werden, muss ihr Verhältnis zu den einzelnen Kriterien der Suchanfrage nicht mithilfe von Operatoren bestimmt werden.

6.6.4 Ausgewählte Suchanfragen zur Illustration

Im Folgenden werden exemplarische Suchanfragen, die in der Korpusversion *ReN_2017-09-05* durchgeführt wurden, vorgestellt.

6.6.4.1 Groß- und Kleinschreibung von Substantiven

Suchanfrage (Beispielgraphem <f> und <F>):

```
tok_dipl=/(F|f).+/
& pos="NA"
& #1 _=_ #2
```

Erläuterung:
Um die Groß- und Kleinschreibung bestimmter Buchstaben bspw. am Anfang eines Substantives zu untersuchen, muss als erstes Kriterium in der Annotationsspur `tok_dipl` (Token in diplomatischer Transkription, s.o.) nach einem Token gesucht werden, das mit einem bestimmten Buchstaben beginnt, und zwar entweder mit dem Groß- oder dem Kleinbuchstaben. Für diese beiden Alternativen wird der *oder*-Oprerator verwendet: `(F|f)`. Da auf den ersten Buchstaben mindestens ein weiterer folgen soll, wird der Platzhalter `.` mit dem Quantor + (für ein oder mehrere beliebige Zeichen) gesetzt. Für die Einschränkung der Suche auf einfache Substantive wird das PoS-Tag `NA` genutzt. Da beide Bedingungen auf dasselbe Token zutreffen sollen, wird zur Angabe des Verhältnisse der Suchkriterien der Operator _=_ für Übereinstimmung verwendet.

Für eine Gegenüberstellung von Groß- und Kleinschreibung in den verschiedenen Wörtern der Belege bietet sich eine Frequenzanalyse (s.u.) an.

6.6.4.2 Komposita mit dem Grundwort *man*

Suchanfrage:

```
lemma=/.+man/
& pos="NA"
& #1_=_#2
```

Erläuterung:
Um ein Kompositum mit dem Grundwort *man* zu finden, wird vor dem Grundwort *man* der Platzhalter `.` mit dem Quantor + (für ein oder mehrere beliebige Zeichen) gesetzt. Um Wörter wie das Indefinitum *îderman* oder den Eigennamen *herman* auszuschließen, wird die Suche zusätzlich auf einfache Substantive (`NA`) in der Annotationsspur *pos* eingegrenzt. Da beide Bedingungen auf dasselbe Token zutreffen sollen, wird zur Angabe des Verhältnisse der Suchkriterien der Operator _=_ für Übereinstimmung verwendet.

6.6.4.3 Personalpronomen im Genitiv

Suchanfrage:

```
pos="PPER"
& morph=/.*\.Gen/
& #1 _=_ #2
```

Erläuterung:
Um Personalpronomen im Genitiv zu finden, müssen die Wortartbedingung, dass es sich um ein Personalpronomen (`PPER`) handeln soll, und die flexionsmorphologische Bedingung, dass der Kasus Genitiv (`Gen`) sein soll, auf dasselbe Token angewendet werden; daher wird der Operator _=_ für Übereinstimmung verwendet.

Hinweis zur Suche nach Kasus (hier: Genitiv):
Im ReN werden Ambiguitäten auf der Ebene der Flexionsmorphologie mithilfe sogenannter Portmanteau-Tags ausgezeichnet (s. Annotationshandbuch Teil 1), z.B. Masc-Neut.Dat-Akk.Sg für ein Token, dessen Genus Maskulinum oder Neutrum und dessen Kasus Dativ oder Akkusativ sein kann. Bei der Suche nach einem bestimmten flexionsmorphologischen Wert muss daher berücksichtigt werden, dass er von anderen Werten umgeben sein kann. Eine Suche nach `morph=/.*Gen.*/"` würde daher auch Belege mit z.B. der Annotation *Gen-Dat-Akk* oder *Nom-Gen* ausgeben. Daher muss der Punkt zur Abgrenzung der verschiedenen Flexionskategorien in die Suche einbezogen werden. Da im Falle der Personalpronomen die Kasusangabe an letzter Stelle steht, folgt nach ihr kein Punkt, dieser muss also nur davor gesetzt werden. Um den Punkt, der sonst als regulärer Ausdruck dient, zu schützen, muss davor ein Slash stehen. Um außerdem Annotationen wie *Gen-Dat-Akk* auszuschließen, genügt es, den Suchausdruck hinter `Gen` enden zu lassen, dort also keine weiteren Platzhalter zu setzen.

Um weiter zu ermitteln, welche Personalpronomen welcher Personen und Numera am häufigsten im Genitiv verwendet werden, bietet sich eine Frequenzanalyse (s.u.) an.

6.6.4.4 Genitivattribute: voran- oder nachgestellt

Suchanfrage (pränominale Genitive):

```
pos="NA"
& morph=/.+\.Gen\..+/
& pos="NA"
& morph!= /.+\.Gen\..+/
& #1_=_#2
& #3_=_#4
& #1.#3
```

Erläuterung:
Um Konstruktionen mit pränominalen Genitivattributen wie *gades willen* zu finden, muss gesucht werden nach:

- einem Substantiv (`pos="NA"`) im Genitiv (`morph=/.+\.Gen\..+/`) -> Da in dem flexionsmorphologischen Tag vor dem Kasus die Angabe zum Genus (z.B. *Fem*) und nach dem Kasus die Angabe zum Numerus (z.B. *Sg*) steht, werden die Platzhalter `.+` (ein oder beliebig viele Zeichen) links und rechts vom Ausdruck `Gen` gesetzt. Zur Abgrenzung der Angabe `Gen` von den Flexionskategorien links (hier: Genus) und rechts (hier: Numerus) wird davor und dahinter jeweils ein Punkt gesetzt. Um den Punkt, der sonst als regulärer Ausdruck dient, zu schützen, muss davor ein Slash stehen, d.h. `\.Gen\..`

- einem Substantiv (`pos="NA"`), das nicht im Genitiv steht (`morph!=/.+\.Gen\..+/`) -> das Ausrufezeichen ! steht für die Negation

- die Inhalte der Suchkriterien 1 und 2 einerseits und 3 und 4 andererseits sollen jeweils auf dasselbe Token zutreffen (#1_=_#2 & #3_=_#4)

- das im Suchkriterium 1 gesuchte Wort soll unmittelbar vor dem im Suchkriterium 3 gesuchten stehen (#1.#3)

Suchanfrage (postnominale Genitive):

```
pos="NA"
& morph=/.+\.Gen\..+/
& pos="NA"
& morph!= /.+\.Gen\..+/
& #1_=_#2
& #3_=_#4
& #3.#1
```

Erläuterung:
Um Konstruktionen mit postnominalen Genitivattributen wie *gebot godes* zu finden, muss die Folge der beiden gesuchten Substantive einfach vertauscht werden; hierzu wird die Angabe des Verhältnisses zu `#3.#1` verändert.

6.6.4.5 Verbstellung im zweigliedrigen Verbalkomplex in *dat*-Sätzen

Suchanfrage (Finitum + Infinitum):

```
lemma_wsd="dat²b"
& pos=/(VA|VM)FIN/
& pos=/VV(INF|PP)/
& bound_sent
& #2.#3
& #4 _i_ #1
& #4 _i_ #2
& #4 _i_ #3
```

Erläuterung:
Um in Nebensätzen die Stellung der verbalen Elemente eines zweigliedrigen Verbalkomplexes – also Finitum plus Infinitum wie in *mochte weszen* oder Infinitum plus Finitum wie in *hebben scholde* – zu ermitteln, muss innerhalb einer Satzeinheit (s.o.) gesucht werden. Da die Unterscheidung zwischen Haupt- und Nebensatz aufgrund der noch freieren Verbstellung im Mittelniederdeutschen in einigen Fällen schwierig ist, werden im ReN die PoS-Tags *KON* und *KOUS* rein stellungsbedingt vergeben (s. Annotationshandbuch Teil 1). Daher kann auch ein Satz mit dem Tag *KOUS* ein nebengeordneter Satz sein (z.B. *Vnd mit går tǒrnichliken dingen Synen helm he do beth vp bandt*) und ein Satz mit dem Tag *KON* ein untergeordneter (z.B. *dat vnse here were eyn droghenere*). Aus diesem Grund können Nebensätze nicht mithilfe des PoS-Tags *KOUS* ermittelt werden. Um in Nebensätzen zu suchen, sollte daher mithilfe entsprechender

Lemmata in prototypischen Nebensätzen, z.B. *dat*-Sätzen, gesucht werden.
Um sowohl ein finites Auxiliar- als auch Modalverb zu finden, wird ein regulärer Ausdruck mit dem *oder*-Operator verwendet: `(VA|VM)FIN` findet `VAFIN` und `VMFIN`. Ähnlich kann beim infiniten Bestandteil vorgegangen werden: `VV(INF|PP)` findet `VVINF` und `VVPP`.
Da die Junktion *dat* und die beiden Elemente des Verbalkomplexes innerhalb derselben Satzeinheit stehen sollen, müssen alle drei Suchkriterien mithilfe des Operators `_i_` mit dem Suchkriterium `bound_sent` verbunden werden.

Suchanfrage (Infinitum + Finitum):

```
lemma_wsd="dat²b"
& pos=/(VA|VM)FIN/
& pos=/VV(INF|PP)/
& bound_sent
& #3.#2
& #4 _i_ #1
& #4 _i_ #2
& #4 _i_ #3
```

Erläuterung:
Um Verbalkomplexe der Folge Infinitum plus Finitum zu finden, muss die Folge der beiden gesuchten Verben einfach vertauscht werden; hierzu wird die Angabe des Verhältnisses zu `#3.#2` verändert.

6.6.4.6 Das *wērden*-Passiv

Suchanfrage:

```
pos="VAFIN"
& lemma_wsd="wērden¹"
& pos="VVPP"
& bound_sent
& #1_=_#2
& #4 _i_ #1
& #4 _i_ #3
```

Erläuterung:
Um nach dem *werden*-Passiv zu suchen, eignet sich eine Verbindung des PoS-Tags `VAFIN` mit dem Lemma, wie es im Wörterbuch steht (mitsamt Diakritika und Homonymendifferenzierung), d.h. der Annotationsspur `lemma_wsd`. So kann die Verwendung von *wērden* als Vollverb in Sätzen wie *he wart koninc* ausgeschlossen werden. Nach dem zweiten Bestandteil der Passivperiphrase kann – da es sich um ein Verb im Partizip Präteritum handeln muss – mithilfe des PoS-Tags `VVPP` gesucht werden. Da das Auxiliarverb *wērden* und das Verb im Partizip Präteritum innerhalb derselben Satzeinheit stehen sollen, müssen beide Suchkriterien mithilfe des Operators `_i_` mit dem Suchkriterium `bound_sent` verbunden werden.

6.6.4.7 Die *wēsen*-Periphrase

Suchanfrage 1 (zweigliedrige Verbalkomplexe):

```
pos="VAFIN" _=_ lemma_wsd="wēsen²"
& pos="VVPP"
& bound_sent & #4_i_#1 & #4_i_#3
```

Erläuterung:
Um nach der *wēsen*-Periphrase zu suchen, d. h. nach einer Konstruktion aus einer finiten Form des Verbs *wēsen* und einem infiniten Verb in Form eines Vergangenheitspartizips, wird nach einer Satzeinheit gesucht (s.o.), die ein Token

mit dem PoS-Tag `VAFIN` und dem *lemma_wsd* (s.o.) `wēsen²` sowie ein Token mit dem PoS-Tag `VVPP` enthält. Um auch dreigliedrige Verbalkomplexe vom Typ Auxiliar oder Modalverb + Vergangenheitspartizip + *wēsen*-Infinitiv zu finden, wird die folgende Suchanfrage verwendet:

Suchanfrage 2 (dreigliedrige Verbalkomplexe):

```
pos=/V.FIN/
& pos="VVPP"
& pos=/V.INF/ _=_ lemma_wsd="wēsen²"
& bound_sent & #5_i_#1 & #5_i_#2 & #5_i_#3
```

Erläuterung:
Gesucht wird nach einer Satzeinheit, die drei verbale Elemente enthält: 1) Ein finites Verb: Hierbei kann es sich um ein Modal- oder Auxiliarverb handeln; der Platzhalter Punkt steht für ein beliebiges Zeichen, hier z.B. für `M` (Modalverb) und `A` (Auxiliaralverb); 2) ein Verb im Partizip Präterium; 3) ein Verb mit dem *lemma_wsd* `wēsen²`, das im Infinitiv steht: Auch hier wird der Platzhalter Punkt verwendet, um verschiedene Verbtypen einzubeziehen.

6.6.4.8 Der *tô*-Infinitiv

Um die Frequenz des flektierten Infinitivs wie in *tô dônde* (annotiert als Präposition mit substantiviertem Infinitiv) und des unflektierten Infinitivs wie in *tô dôn* (annotiert als Partikel mit Verb im Infinitiv) miteinander zu vergleichen, kann man beide Formen mithilfe einer Suchanfrage ermitteln:

Suchanfrage 1:

```
(zu#pos="PTKZU" . vb#pos=/V.INF/)|(zu#pos="APPR" _=_ lemma_wsd="tô¹" . vb#pos="NA" _=_ p
```

Erläuterung:
Mithilfe des oder-Operators | wird eine Abfrage für beide gesuchten Formen gestellt. Hierfür werden die entsprechenden Suchausdrücke in runde Klammern gesetzt und durch | getrennt. Für den unflektierten Infinitiv (links) wird nach einer Verbpartikel *tô* (pos-Tag `PTKZU`) gesucht, auf welche unmittelbar ein Verb im Infinitiv folgt. Für das unmittelbare Aufeinanderfolgen steht der Punkt als Operator zwischen beiden Suchausdrücken. Da es sich bei dem Verb nicht zwingend um ein Vollverb (`VVINF`) handelt, sondern auch ein Modalverb (`VMINF`) oder ein Auxiliarverb (`VAINF`) möglich sind, wird mithilfe des Punktes als Platzhalter für ein beliebiges Zeichen nach der unterspezifizierten Form `V.INF` gesucht. Im Gegensatz zum Suchausdruck für die Partikel, die eine konkrete Form (`PTKZU`) enthält und daher gemäß der ANNIS-Abfragesprache in Anführungsstrichen steht, muss die unterspezifizierte Form, die durch einen regulären Ausdruck realisiert ist, in Schrägstriche einklammert werden. Der flektierte Infinitiv (rechts) besteht aus der Präposition *tô* und einem substantivierten Infinitiv. Die Präposition wird anhand des pos-Tags `APPR` und des Lemmas *tô¹* ermittelt. Da beide Annotationen auf ein Token zutreffen sollen, wird der Operator für Übereinstimmung verwendet. Unmittelbar auf die Präposition folgt ein substantivierter Infinitiv, d. h., dass auf der Belegebene das Tag `NA` und auf der Lemmaebene das Tag `VVINF` vergeben ist. Die Suchausdrücke für beide Varianten enthalten unterschiedlich viele Knoten – beim unflektierten Infinitiv (links) sind es zwei, beim flektierten (rechts) dagegen vier. Damit diese korrekt zugeordnet werden können, müssen diejenigen Knoten, die in den beiden Varianten einander entsprechen, mit derselben individuell zu vergebenden Benennung versehen werden. Hierfür wird vor den Knoten eine Raute und davor der jeweilige Kurzname gesetzt, z. B. *zu* für das Element *tô* und *vb* für den verbalen Bestandteil. (Hinweis: Diese Suchanfrage ist erst ab der ANNIS-Version 3.5.0 möglich.)

Um auch Partikelverben mit inkorporiertem *tô* wie in *aftônēmen*, die im ReN nicht getrennt, sondern im Ganzen als Infinitiv annotiert werden, zu finden, wird die folgende Suchanfragte verwendet:

Suchanfrage 2 (Partikelverben):

```
pos=/V.INF/ _=_ tok_anno=/.+to.+/
```

Erläuterung:
Gesucht werden Token, die als infinites Verb annotiert sind (pos=/V.INF/) und die ein *to* im Wortinneren enthalten. Die Kombination aus den regulären Ausdrücken `.+` wird stellvertretend für ein oder mehrere Zeichen vor und hinter dem *to* gesetzt. Da eine genauere Suche nicht möglich ist, enthält das Ergebnis auch Fehlbelege, die manuell aussortiert werden müssen.

6.6.4.9 Doppelte Negation

Suchanfrage:

```
lemma_wsd=/(ne²|en⁴)/
& lemma_wsd="nicht¹"
& bound_sent
& #3_i_#1
& #3_i_#2
```

Erläuterung:
Um nach Satzeinheiten zu suchen, die sowohl die Negationspartikel *en/ne* als auch die Negationspartikel *nicht* enthalten, wird mithilfe der Lemmata, wie sie im Wörterbuch stehen (mitsamt Diakritika und Homonymendifferenzierung), gesucht, d.h. in der Annotationsspur `lemma_wsd`. Da es für die klitische Negationspartikel zwei verschiedene Lemmata gibt, wird der *oder*-Operator verwendet. Da die Negationspartikel *en/ne* und die Negationspartikel *nicht* innerhalb derselben Satzeinheit stehen sollen, müssen beide Suchkriterien mithilfe des Operators `_i_` mit dem Suchkriterium `bound_sent` verbunden werden.

6.6.4.10 Großschreibung am Satzanfang

Suchanfage:

```
bound_sent
& tok_anno=/[A-Z].*/
& #1 _l_ #2
```

Erläuterung:
Um nach großgeschriebenen Wortformen zu suchen, wird mithilfe eines regulären Ausdrucks nach allen `tok_anno` gesucht, die mit einem Buchstaben zwischen A und Z beginnen. Zusätzlich wird mit dem Operator `_l_` (left aligned) festgelegt, dass das entsprechende `tok_anno` an der gleichen Stelle wie ein `bound_sent` beginnen soll, also am Anfang eines Satzes steht.

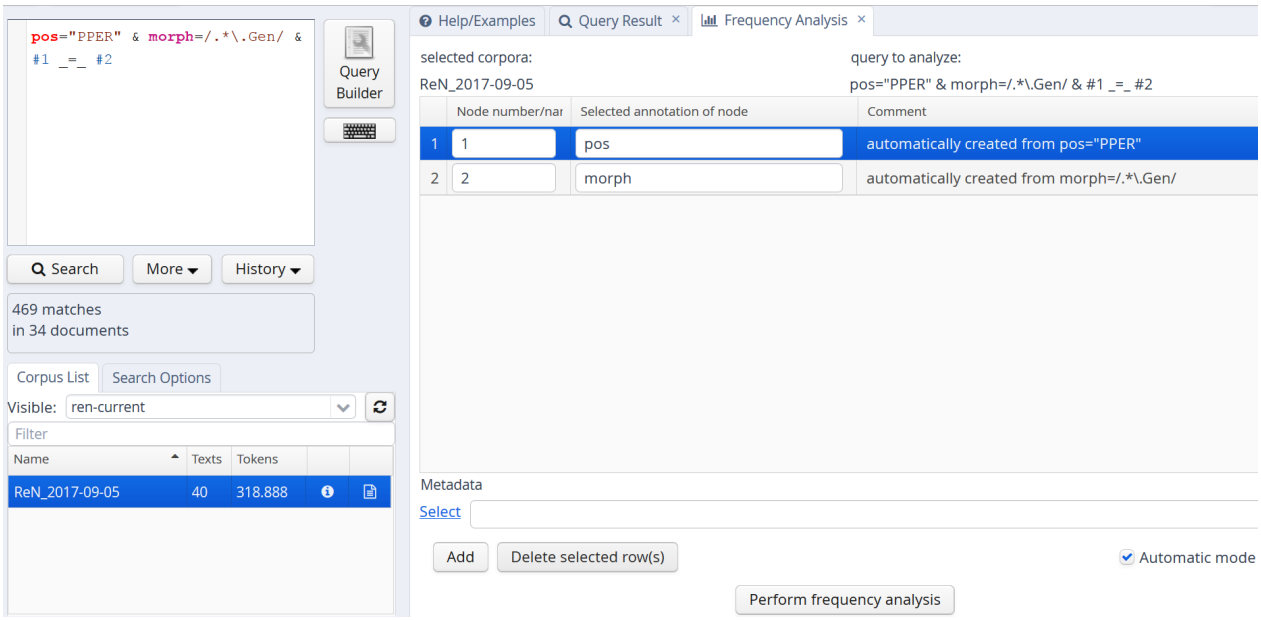
6.7 Weiterverarbeiten der Suchergebnisse

6.7.1 Frequenzanalyse (frequency analysis)

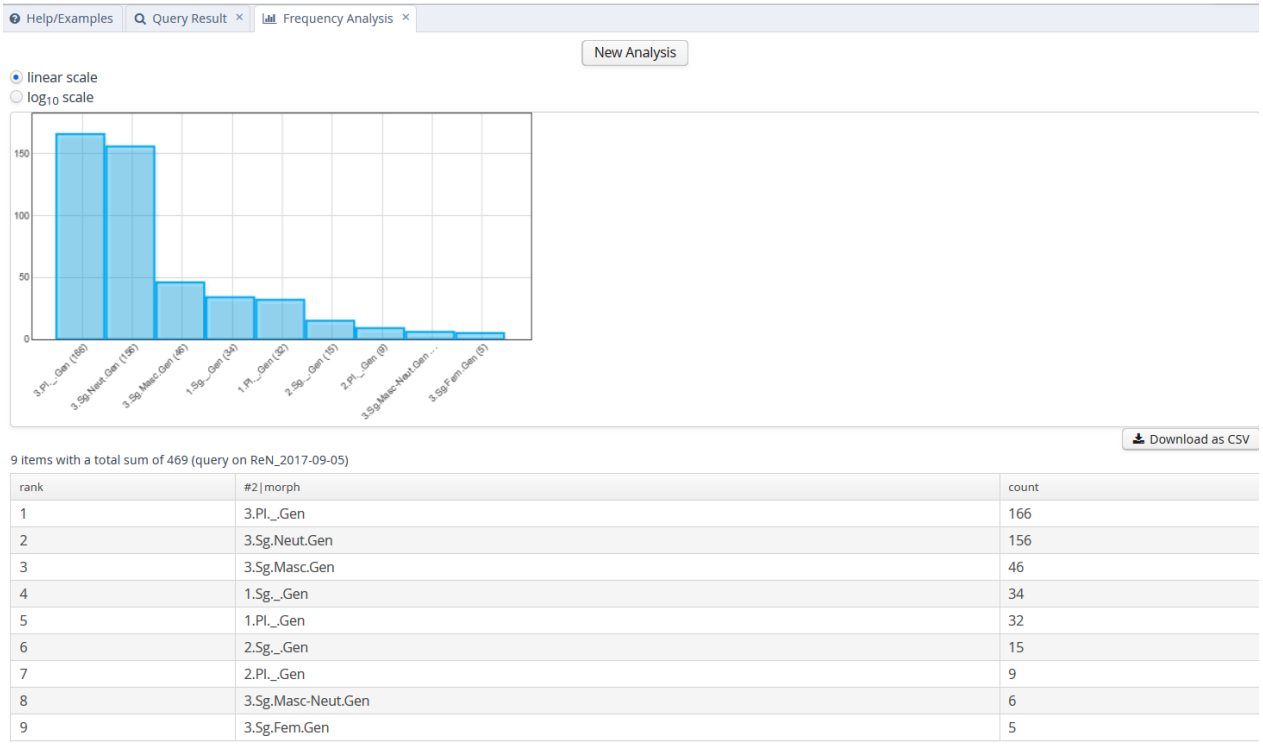
Die Frequenzanalyse dient dazu, Häufigkeiten von Annotationen in den verschiedenen Annotationsspuren zu ermitteln. Unterhalb der Eingabefläche für Suchanfragen kann über den Button *More* die Option *Frequency Analysis* ausgewählt werden. Anschließend öffnet sich ein neues Fenster, in dem die Nummer der Suchkriterien (nodes) und die Namen der Annotationsspuren, für welche die Frequenzanalyse durchgeführt werden soll, angegeben werden. Im Folgenden soll dies anhand konkreter Beispiele zu z.T. oben bereits erläuterten Suchanfragen veranschaulicht werden:

6.7.1.1 Personalpronomen im Genitiv

Die Frequenzanalyse kann, basierend auf den Ergebnissen der oben vorgestellten Suchanfrage, u.a. genutzt werden, um zu ermitteln, welche Personalpronomen welcher Personen und Numera am häufigsten im Genitiv verwendet werden. Nach dem Anklicken der Option *Frequency Analysis* erscheinen in dem neu geöffneten Fenster die zwei auf Basis der Suchanfrage automatisch generierten Zeilen mit Kriterium 1 und der Annotationsspur *pos* sowie Kriterium 2 mit der Spur *morph* (s.u.). Soll nun die Häufigkeit von Person und Numerus der Personalpronomen im Genitiv ermittelt werden, genügt eine Analyse in der Spur *morph*, die Spur *pos* muss daher gelöscht werden. Hierfür genügt ein Klick auf die entsprechende Zeile, die dann blau unterlegt erscheint (s.u.), und anschließend ein Klick auf den Button *Delete selected row(s)*.



Mit einem Klick auf den Button *Perform frequency analysis* wird die Frequenzanalyse durchgeführt und das Ergebnis im selben Fenster geöffnet. Es liefert ein Balkendiagramm und darunter eine Tabelle mit sämtlichen Annotationen in der gesuchten Spur und deren absoluten Häufigkeiten. In diesem Beispiel zeigt sich, dass die Personalpronomen in der 3. Person Plural sowie der 3. Person Singular Neutrum am häufigsten im Genitiv verwendet werden:



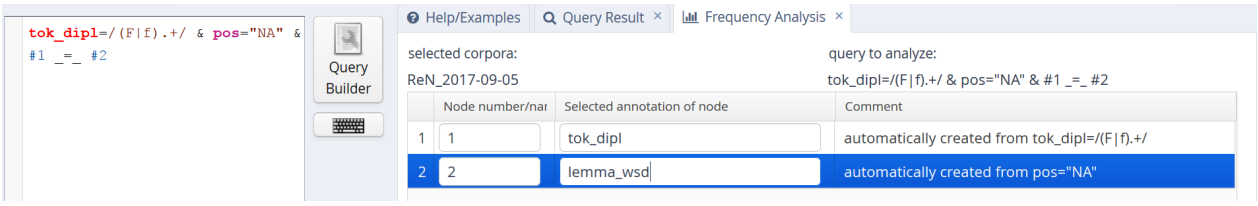
rank	#2|morph	count
1	3.Pl._.Gen	166
2	3.Sg.Neut.Gen	156
3	3.Sg.Masc.Gen	46
4	1.Sg._.Gen	34
5	1.Pl._.Gen	32
6	2.Sg._.Gen	15
7	2.Pl._.Gen	9
8	3.Sg.Masc-Neut.Gen	6
9	3.Sg.Fem.Gen	5

Die Tabelle kann mit einem Klick auf den Button *Download as CSV* als csv-Datei heruntergeladen und weiterverarbeitet werden.

6.7.1.2 Groß- und Kleinschreibung von Substantiven

Um die Groß- und Kleinschreibung in den verschiedenen Wörtern der Belege aus der oben beschriebenen Suchanfrage genauer betrachten zu können, bietet sich eine Übersicht sämtlicher Belege ohne den Kontext an. Auch hierfür kann

die Frequenzanalyse genutzt werden. Auf Basis der obigen Suchanfrage werden automatisch zwei Zeilen mit Kriterium 1 und Annotationsspur *tok*dipl_ sowie Kriterium 2 und Spur *pos* generiert. Sinnvoll wäre eine Auflistung aller Belege in der diplomatischen Transkription mit den dazugehörigen Lemmata. Daher wird in der zweiten automatisch generierten Zeile die Benennung der Spur von *pos* manuell zu *lemma_wsd* geändert:



Das Ergebnis liefert sämtliche Types für Substantive, die mit <f> oder <F> beginnen, mit den zugehörigen Lemmata und kann als csv-Datei gespeichert werden. Die folgende Abbildung zeigt einen Ausschnitt dieser csv-Datei:

	A	B	C
1	#1|tok_dipl	#2|lemma_wsd	count
2	frouwen	vrouwe²	26
3	frowden	vröude	17
4	frowde	vröude	14
5	frunde	vrünt	11
6	firmament	firmament	9
7	frucht	vrucht²	8
8	frist	vrist	7
9	frede	vrēde	7
10	frouwe	vrouwe²	7
11	forsten	vörste²	7
12	frowde	vröude	6
13	frouw	vrouwe²	6
14	frow	vrouwe²	6
15	frunden	vrünt	5
16	Fürsten	vörste²	5
17	figuren	figure	4
18	forme	forme	4
19	flyd	vlît	4
20	fruwen	vrouwe²	4
21	frucht	vruchte	4
22	Forste	vörste²	4

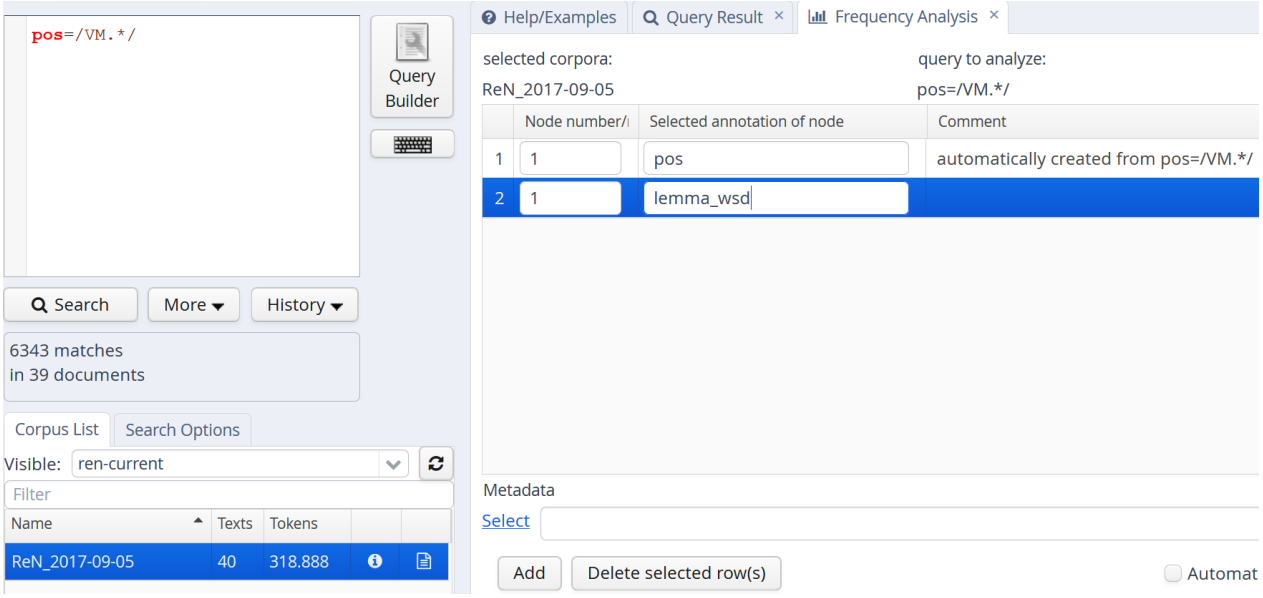
In einer solchen Ergebnistabelle könnten die Belege z.B. nach den Lemmata gefiltert und so ermittelt werden, ob bestimmte Wörter häufiger großgeschrieben werden als andere, ob. z.B. Belege des Lemmas *vrouwe²* (nhd. 'Frau') eher kleingeschrieben werden (hier insgesamt 49 Belege für Kleinschreibung, kein Belege für Großschreibung), Belege des Lemmas *vörste²* (nhd. 'Fürst') hingegen eher großgeschrieben werden (hier insgesamt 9 Belege für Großschreibung, 7 Belege für Kleinschreibung) (vgl. Tabelle).

6.7.1.3 Modalverben

Um bspw. zu ermitteln, welche Wörter am häufigsten als Modalverben verwendet werden und in welcher Form – d.h. finit, infinit, imperativisch, im Partizip Präteritum oder Partizip Präsens –, muss zunächst mithilfe der folgenden Suchanfrage nach sämtlichen Modalverben im Korpus gesucht werden:

```
pos=/VM.*/
```

Da beim Öffnen der Eingabe für die Frequenzanalyse nur eine Zeile automatisch generiert wird (da die Suchanfrage nur ein Kriterium enthält), jedoch sowohl nach den Lemmata als auch den PoS-Tags gesucht werden soll, muss über Klick auf den Button *Add* eine zusätzliche Zeile eingefügt und der Name der Annotationsspur zu *lemma*wsd_ bearbeitet werden:



Auch hier kann das tabellarisch dargestellte Ergebnis als csv-Datei heruntergeladen werden:

	A	B	C
1	#1|pos	#1|lemma_wsd	count
2	VMFIN	schōlen¹	3071
3	VMFIN	willen¹	1274
4	VMFIN	mōgen	1124
5	VMFIN	mōten²	506
6	VMFIN	künnen	224
7	VMFIN	dörven	80
8	VMFIN	dōren²	26
9	VMIMP	willen¹	14
10	VMINF	mōgen	12
11	VMINF	willen¹	4
12	VMINF	künnen	3
13	VMPP	künnen	2
14	VMINF	mōten²	1
15	VMFIN	bedörven	1
16	VMINF	dōren²	1

Die Tabelle zeigt, dass bei den finiten Modalverben (VMFIN) das Lemma *schōlen*[1] am häufigsten vertreten ist und bei den infiniten (VMINF) das Lemma *mōgen*. Im Imperativ (VMIMP) gibt es ausschließlich Belege mit dem Lemma *willen*[1] und im Partizip Präteritum (VMPP) nur Belege mit dem Lemma *künnen*.

6.7.1.4 Lemmaliste und Vollformlexikon

Mithilfe der Frequenzanalyse ist es einfach möglich, eine Liste aller im Korpus vorkommenden Lemmata zu erstellen. Hierzu sucht man einfach nach `lemma_wsd`. Führt man nun eine Frequenzanalyse aus, ohne die vorgeschlagenen Einstellungen anzupassen, wird eine Liste aller Lemmata mit der Häufigkeit des Vorkommens im Korpus erstellt.

Ein Vollformlexikon, das Wortformen mit entsprechenden POS-, Morphologie-Tags und Lemmata enthält und z.B. zum Lemmatisieren mithilfe des dem RFTagger beiliegenden Lemmatisierungsskript verwendet werden kann, lässt sich ebenfalls einfach mit der Frequenzanalyse erstellen. Hierzu fügt man bei der gleichen Suchanfrage zur Spur `lemma_wsd` noch drei weitere Zeilen zur Frequenzanalyse hinzu und ändert die Namen dieser Annotationsspuren zu *tok_anno*, *pos* und *morph*. Um das resultierende Vollformlexikon mit dem RFTagger zu nutzen, müssen in der exportierten CSV-Datei die Spalten *rank* und *count* gelöscht und die Titelzeile entfernt werden. Die Spalten *pos* und *morph* müssen zu einer Spalte zusammengefasst werden, wobei das POS- und der Morphologie-Tag mit einem Punkt (.) voneinander getrennt werden. Das Lemmatisierungsskript des RFTaggers erwartet mit Tabulator getrennte Spalten.

Nicht im Wörterbuch enthaltene Lemmata

Die Lemmatisierung im ReN basiert im Wesentlichen auf dem Mittelniederdeutschen Handwörterbuch von Lasch et al. (1956ff.) sowie dem Mittelniederdeutschen Handwörterbuch von Lübben/ Walther (1995) (s. hierzu das Annotationshandbuch Teil 2: Lemmatisierung). Nicht in diesen zugrunde gelegten Wörterbüchern vorkommende Lemmata wurden im ReN neu aufgenommen. Diese Lemmata sind in der Annotation daran zu erkennen, dass sie in der Spur `lemma_wsd` mit einem Asterisk enden, z.B. `bergisch*`. Um eine Liste aller neu aufgenommenen Lemmata zu erhalten, kann die folgende Suchanfrage genutzt werden:

```
lemma_wsd=/.+\*/
```

Gesucht wird nach einem Lemma, das auf `*` endet. Da der Asterisk in regulären Ausdrücken eine besondere Bedeutung hat, muss bei einer Suche nach diesem Zeichen vor das Zeichen zum Schutz ein Backslash gesetzt werden (Erläuterung s.o. unter `Suche mithilfe regulärer Ausdrücke`)

Für nicht im Wörterbuch enthaltene Eigennamen wurden keine neuen Lemmata angesetzt. Stattdessen wurde als Lemma die Wortform des Belegs übernommen und am Ende um das Zeichen ° ergänzt, z.B. *Johan°*. Auf diese Weise können diese Eigennamen in einer Gesamtliste aller Lemmata einfach identifiziert werden.

6.7.2 Export

Eine Form des Datenexports wurde im vorangehenden Kapitel bereits angesprochen: das Herunterladen der tabellarischen Ergebnisse aus der Frequenzanalyse als csv-Datei. Daneben liefert ANNIS auch die Möglichkeit, die Ergebnisse einer Suchanfrage (ohne Frequenzanalyse) zu exportieren. Zur Auswahl stehen verschiedene Exporter, die im Folgenden anhand von Beispielen in der Korpusversion *ReN_2017-09-05* beschrieben werden.

Generell ist zu beachten, dass das ReN mehrere Tokenisierungen enthält und damit nicht alle Exporter gleichermaßen umgehen können. Erscheint bei dem gewünschten Exporter die Meldung **This exporter does not work well with dialog data (corpora that have more than one primary text). Use the GridExporter instead.**, sollte dieser Exporter nicht mit dem Referenzkorpus verwendet werden. Hierbei handelt es sich um den **SimpleTextExporter** und **TokenExporter**, die daher nicht beschrieben werden.

Eine weitere Besonderheit ist durch das Vorkommen von Kommata (,) in Textnamen (z.B. *Brs._Dud._1279a,b*) zu beachten: in ANNIS-Versionen vor 3.5.0 kommen der **GridExporter** und der **CSVMultiTokExporter** damit nicht zurecht und liefern leere oder unvollständige Exporte, wenn Treffer aus einem solchen Text kommen.

6.7.2.1 CSVExporter

Dieser Exporter exportiert nur die Annotationen der gesuchten Knoten und keinerlei Kontext. Er ist daher v.a. für kontextunabhängige Analysen – bspw. zur Schreibung eines Wortes – geeignet. Beim Export erhält man eine txt-Datei, die mit einem Tabellenprogramm geöffnet werden kann. Für die ReN-Daten besser geeignet als der CSV-Exporter ist der CSV-Multi-Tok-Exporter (s.u.).

6.7.2.2 CSVMultiTokExporter

Dieser Exporter exportiert ebenfalls nur die Annotationen der gesuchten Knoten und keinerlei Kontext. Der Vorteil dieses Exporters liegt darin, dass er für Korpora mit mehreren Tokenisierungen wie das ReN auch Text, der unter Spannenannotationen liegt, exportieren kann. Da im ReN an verschiedenen Stellen Spannen verwendet werden (z.B. zur Angabe der Satzeinheiten oder bei Mehrworteinheiten), sollte dieser Exporter dem einfachen CSVExporter grundsätzlich vorgezogen werden. Beim Export erhält man eine csv-Datei, die mit einem Tabellenprogramm geöffnet werden kann. Als Beispiel dient das Ergebnis der folgenden Suchanfrage nach dem Lemma *vrünt* im Nominativ Singular:

```
lemma_wsd="vrünt"
& morph="Masc.Nom.Sg"
& #1 _=_ #2
```

Die folgende Abbildung zeigt einen Ausschnitt aus der Export-Datei:

	A	B	C	D	E
1	1_id	1_span	1_anno_annotation::lemma_wsd	1_anno_annotation::morph	1_anno_annotation::pos
2	1850509	frunth	vrünt	Masc.Nom.Sg	NA
3	1853227	frunt	vrünt	Masc.Nom.Sg	NA
4	1861375	frunt	vrünt	Masc.Nom.Sg	NA
5	1863659	frunt	vrünt	Masc.Nom.Sg	NA
6	1905026	vrund	vrünt	Masc.Nom.Sg	NA
7	1911245	Vrunt	vrünt	Masc.Nom.Sg	NA
8	2032561	vrent	vrünt	Masc.Nom.Sg	NA
9	2282920	vrunt	vrünt	Masc.Nom.Sg	NA
10	2334253	vrunt	vrünt	Masc.Nom.Sg	NA
11	2576889	Frundt	vrünt	Masc.Nom.Sg	NA
12	2591391	frundt	vrünt	Masc.Nom.Sg	NA
13	2885565	vrunt	vrünt	Masc.Nom.Sg	NA
14	3043269	fründ	vrünt	Masc.Nom.Sg	NA
15	3078569	vrunt	vrünt	Masc.Nom.Sg	NA
16	3191448	urint	vrünt	Masc.Nom.Sg	NA
17	3246619	urunt	vrünt	Masc.Nom.Sg	NA
18	3292201	vrunt	vrünt	Masc.Nom.Sg	NA
19	3293347	vrund	vrünt	Masc.Nom.Sg	NA
20	3296432	vrunt	vrünt	Masc.Nom.Sg	NA
21	3303669	Vrunt	vrünt	Masc.Nom.Sg	NA
22	3303899	vrunt	vrünt	Masc.Nom.Sg	NA
23	3390220	frundt	vrünt	Masc.Nom.Sg	NA

Um auch Kontext mit zu exportieren, kann man zur Suchanfrage einfach beliebig oft *bound_sent* hinzufügen, z.B.:

```
lemma_wsd="vrünt" _=_ morph="Masc.Nom.Sg"
& bound_sent . bound_sent . bound_sent
& #4_i_#1
```

Hier würden nun der Satz, in dem sich das Lemma *vrünt* befindet, sowie jeweils ein Satz davor und dahinter mit ausgegeben.

Möchte man die Annotationen nicht nur der eigentlich gesuchten Token, sondern auch der folgenden Token bspw. exportieren, kann man zur Suchanfrage einfach beliebig oft *tok_anno*, das auf das gesuchte Element folgt, hinzufügen, z.B.:

```
lemma_wsd="vrünt" _=_ morph="Masc.Nom.Sg" . tok_anno . tok_anno
& bound_sent . bound_sent . bound_sent
& #6_i_#1
```

Hier würden nun nicht nur die Annotationen der Token mit dem Lemma *vrünt*, sondern auch der zwei darauf folgenden Token exportiert. Dies kann z.B. für Abfragen innerhalb von Phrasen, z.B. Präpositionalphrasen, sinnvoll sein.

6.7.2.3 WekaExporter

Dieser Exporter erstellt Dateien, die für die Verwendung mit der data-mining Software "Weka" gedacht sind. Exportiert werden die gleichen Daten wie beim CSVExporter (Komma-separiert, Texttrenner: '). Zusätzlich zu den in einer csv-Datei vorhandenen Informationen enthalten die erzeugten arff-Dateien im oberen Teil der Exportdatei eine Auflistung der Spaltennamen (Annotationsspuren). In der folgenden Abbildung ist das Ergebnis der oben genannten Suchanfrage nach dem Lemma *vrünt* im Nominativ Singular zu sehen:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	@relation name																			
2																				
3	@attribute #1_id string																			
4	@attribute #1_span string																			
5	@attribute #1_anno_annotation:lemma string																			
6	@attribute #1_anno_annotation:lemma_simple string																			
7	@attribute #1_anno_annotation:lemma_var string																			
8	@attribute #1_anno_annotation:lemma_wsd string																			
9	@attribute #1_anno_annotation:morph string																			
10	@attribute #1_anno_annotation:pos string																			
11	@attribute #1_anno_annotation:posLemma string																			
12	@attribute #1_anno_default_ns:tok_anno string																			
13	@attribute #2_id string																			
14	@attribute #2_span string																			
15	@attribute #2_anno_annotation:lemma string																			
16	@attribute #2_anno_annotation:lemma_simple string																			
17	@attribute #2_anno_annotation:lemma_var string																			
18	@attribute #2_anno_annotation:lemma_wsd string																			
19	@attribute #2_anno_annotation:morph string																			
20	@attribute #2_anno_annotation:pos string																			
21	@attribute #2_anno_annotation:posLemma string																			
22	@attribute #2_anno_default_ns:tok_anno string																			
23																				
24	@data																			
25																				
26	1850509		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunth	1850509		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunth
27	1853227		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt	1853227		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt
28	1861375		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt	1861375		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt
29	1863659		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt	1863659		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frunt
30	1905026		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrund	1905026		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrund
31	1911245		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Vrunt	1911245		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Vrunt
32	2032561		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrent	2032561		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrent
33	2282920		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	2282920		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
34	2334253		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	2334253		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
35	2576889		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Frundt	2576889		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Frundt
36	2591391		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frundt	2591391		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frundt
37	2885565		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	2885565		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
38	3043269		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	fründ	3043269		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	fründ
39	3078569		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	3078569		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
40	3191448		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	urint	3191448		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	urint
41	3246619		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	urunt	3246619		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	urunt
42	3292201		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	3292201		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
43	3293347		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrund	3293347		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrund
44	3296432		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	3296432		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
45	3303669		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Vrunt	3303669		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	Vrunt
46	3303899		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt	3303899		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	vrunt
47	3390220		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frundt	3390220		vrünt	vrunt	vrunt	vrünt	Masc.Nom.Sg	NA	NA	frundt
48																				

6.7.2.4 TextColumnExporter

Dieser Exporter exportiert die zu den gesuchten Knoten gehörenden Token und eine auswählbare Anzahl an Token als linker und rechter Kontext. Es wird jeweils nur die Tokenspur (ohne Annotationen) exportiert, die zu den gesuchten Knoten passt. Bei Annotation wie *pos* oder *lemma* ist dies die Tokenspur *tok_anno*. Der Name dieser Spur ist in der Spalte *speaker* aufgeführt. Beim Export erhält man eine txt-Datei, die mit einem Tabellenprogramm geöffnet werden kann. Als Beispiel dient das Ergebnis der oben genannten Suchanfrage nach dem Lemma *vrünt* im Nominativ Singular. Die folgende Abbildung zeigt den gesamten Output der Export-Datei:

	A	B	C	D	E
1	match_number	speaker	left_context	match_column	right_context
2	1	tok_anno	meyger myn leue	frunth	klagent mack my nycht baten
3	2	tok_anno	hebben myn alder leueste	frunt	des dancke yck yuw dennoch
4	3	tok_anno	marcus meyger myn leue	frunt	szo scholde jck yuw wol
5	4	tok_anno	vore myn alder leueste	frunt	hir myt geue yck yuw
6	5	tok_anno	respondit Johannes myn vil leue	vrund	Twyde my an desser stund
7	6	tok_anno	Maria cantat	Vrunt	iohannes neue gute
8	7	tok_anno	he mit ter beteringhe got	vrent	wille wesen thes ghenen the
9	8	tok_anno	brudegamen is De	vrunt	des brudegames stat vnde
10	9	tok_anno	to ene lazarus vnse	vrunt	slapet sunder ik wil
11	10	tok_anno	synt gy hyr herkamen	Frundt	dat hedd ick gern vormamen
12	11	tok_anno	genesen Jck wil iuw sunder	frundt	dar umm wesen Vnd synt
13	12	tok_anno	Jhesus sprack to eme .	vrunt	war to bist u kamen
14	13	tok_anno	nicht eenmal eer Oehm und	fründ	to beden . Sold he
15	14	tok_anno	werden . de gude leve	vrunt	mengit . vnde nemelike
16	15	tok_anno	vnde dhe pars nageste	urint	. dhe is nagest uoremunt
17	16	tok_anno	mach dhe andere ofte nen	urunt	dhe dhar tohort allene
18	17	tok_anno	. Du byst myn gude	vrunt	. dar umme
19	18	tok_anno	sede . Johan myn gude	vrund	. Jck bidde dy /
20	19	tok_anno	vnde sede . O leue	vrunt	en dode my nicht /
21	20	tok_anno	sere vnde sprack .	Vrunt	Wor kame gy her mit
22	21	tok_anno	Her Frederick seede . Gude	vrunt	/ segge vns de auenture
23	22	tok_anno	GODT will syn eyn ewich	frundt	Nicht reppen mehr de sunde

6.7.2.5 GridExporter

Dieser Exporter exportiert nicht nur die Annotationen der gesuchten Knoten, sondern auch den Kontext. Der Umfang des gewünschten Kontextes kann unter *Left Context* und *Right Context* über Auswahl aus einer Dropdownliste angegeben werden. Ohne weitere Einschränkungen werden stets sämtliche Annotationsspuren und deren Annotationen exportiert. Möchte man die Annotationen auf bestimmte Spuren beschränken – z.B. nur *pos* und *lemma_wsd* –, kann man die Namen dieser Spuren unter *Annotation Keys* mit Komma getrennt eingeben. Es werden dann nur diese Annotationsspuren exportiert. Der Grid-Exporter ist v.a. für Analysen geeignet, für die eine Betrachtung des Kontextes notwendig ist, z.B. syntaktische oder inhaltliche Fragestellungen.

6.8 Weitere Informationen zu ANNIS

- ANNIS-Handbuch: http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.4.3.pdf
- Dipper, S. (2015): Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch. Zeitschrift für Germanistische Linguistik, 43(3), 516–563.