Stefanie Dipper*

Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch

DOI 10.1515/zgl-2015-0020

Abstract: This article addresses the use of annotated corpora for research in historical German syntax, introducing the *Reference Corpus of Middle High German* (REM) as a showcase. The article shows how to search the corpus with the corpus tool ANNIS. It takes up a range of popular research questions from historical syntax, in particular inflection of predicative adjectives, pre- and postnominal genitive constructions, relative order of pronouns and of verbs and auxiliaries. It is shown in detail how REM can be exploited to retrieve corpus evidence of these phenomena. The article also discusses challenges and opportunities of using annotated corpora. It argues that use of annotated corpora opens up new ways for research in historical syntax.

***Kontaktperson: Prof. Dr. Stefanie Dipper:** Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, D-44780 Bochum, E-Mail: dipper@linguistics.ruhr-uni-bochum.de

1 Einleitung

In diesem Artikel möchte ich einen Überblick über verfügbare Korpora historischer Sprachstufen des Deutschen geben und insbesondere auf deren Einsatz für die historische Syntaxforschung eingehen.[1] In verschiedenen Anwendungsszenarien werde ich am Beispiel des *Referenzkorpus Mittelhochdeutsch* (REM) illustrieren, wie solche Korpora für linguistische Untersuchungen, insbesondere von syntaktischen Phänomenen, genutzt werden können.

Der Artikel beginnt mit einer kurzen Einführung zu annotierten historischen Korpora und zu Annotation im Allgemeinen (Abschnitt 2). Danach gebe ich einen Überblick über verfügbare Korpora historischer Sprachstufen des Deutschen (Abschnitt 3). Es folgen Anwendungsbeispiele, die die Nutzung von REM für verschiedene linguistische Untersuchungen zeigen (Abschnitte 4 und 5). Dabei werde ich auch auf das Annotationsschema *HiTS* („Historisches Tagset“) eingehen, das für die Wortart-Annotation von REM und weiteren deutschsprachigen historischen Korpora entwickelt wurde. Der Artikel schließt mit einigen Überlegungen zu Problemen und Chancen annotierter Korpora (Abschnitte 6 und 7). Im Appendix I sind die Siglen der Korpusbelege aufgeschlüsselt. Appendizes II und III listen die in REM genutzten Annotationskategorien für Wortart bzw. Morphologie.

Ziel dieses Artikels ist es nicht, tatsächliche Ergebnisse für die historische Linguistik zu liefern. Das ist schon deswegen nicht möglich, weil die verwendete Version von REM nur einen Ausschnitt aus einer Vorversion darstellt, bei der u. a. letzte Korrekturschritte noch fehlen (s. Fußnote 21). Stattdessen soll der

1 Für hilfreiche Diskussionen und Kommentare danke ich Marcel Bollmann, Julia Krasselt, Sarah Kwekkeboom, Florian Petran und Sandra Waldenberger sowie den Gutachtern und den Herausgebern der Zeitschrift und insbesondere Jürg Fleischer, dem Herausgeber dieses Themenhefts. Die Arbeit an diesem Artikel wurde unterstützt durch die DFG (Geschäftszeichen DI-1558/1). Alle URLs in diesem Artikel wurden am 16.6.2014 abgerufen.

Artikel einen Beitrag dazu leisten, die Möglichkeiten (und Grenzen) annotierter historischer Korpora zu beleuchten und anhand von REM zu illustrieren.

2 Annotierte historische Korpora

Die historische Linguistik ist natürlicherweise schon immer eine korpusbasierte Wissenschaft. Auch ist eines der ältesten digitalisierten Korpora ein historisches: der von Pater Roberto Busa in den Jahren 1949–1974 erstellte lateinisch-sprachige *Index Thomisticus*, eine Konkordanz aller Wörter der Werke Thomas von Aquins, die zunächst auf Lochkarten und später auf Magnetbändern gespeichert wurde (Busa 1974–1980, 1980). Das Korpus enthält mehr als 10 Millionen Wörter.

Viele Jahre nach dem Index Thomisticus entstanden ab den 1980er Jahren die ersten digitalen *diachronen* Korpora für das Englische, zum einen das *Helsinki Corpus of English Texts*, zum andern *A Representative Corpus of Historical English Registers* (ARCHER).[2]

Die digitalen Korpora der ersten Generation bestanden aus ausgewogen zusammengestellten historischen Texten, waren aber noch nicht mit zusätzlicher linguistischer Information wie Morphologie oder Wortart angereichert, d. h. sie waren noch nicht linguistisch *annotiert*. Aufbauend auf dem Helsinki Corpus wurden ab den 1990er Jahren die ersten morphologisch und syntaktisch annotierten Korpora erstellt: die *Penn-Helsinki Parsed Corpora* sowie das *York-Toronto-Helsinki Parsed Corpus of Old English Prose*.[3]

In diesem Artikel geht es ausschließlich um *annotierte* Korpora. Wie unten ausführlich dargelegt und illustriert, bieten Annotationen einen gezielten Zugriff auf relevante Daten und ermöglichen damit auch die Untersuchung komplexer Phänomene.

In historischen Korpora finden sich typischerweise Annotationen von Lemmata, Morphologie und Wortart und gelegentlich Syntax, meist in Form von Dependenzanalysen. Üblicherweise werden die Annotationen semi-automatisch erstellt, d. h. ein Programm erzeugt vollautomatisch Vorschläge, die manuell nachkorrigiert werden.

Annotation und Disambiguierung

Ein wichtiger Aspekt von Annotation ist, dass dabei ambige Wortformen disambiguiert werden. In (1) fungiert beispielsweise die Wortform *diz* einmal

2 http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/generalintro.html.

3 www.ling.upenn.edu/hist-corpora; http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

als selbstständiges Pronomen (1a) und einmal als Demonstrativ-Artikel (1b). Durch die Zuweisung unterschiedlicher Wortart-Tags kann dieser Unterschied kodiert werden.

(1) a. do iheſus *diz* horte do wndirte her ſich vnde ſprach czu den dy ym volgeten. (M402-10r,22f)[4]
,Da Jesus *das* hörte, wunderte er sich und sprach zu denen, die ihm folgten'
b. In der czit. ſaite vnſe herre ſynen iungeren *diz* glichniſſe. (M402-12r,17f)
,In der Zeit sagte unser Herr seinen Jüngern *dieses* Gleichnis'

Im Index Thomisticus wurden alle Wortformen mit Hilfe eines elektronischen Lexikons semi-automatisch lemmatisiert und morphologisch analysiert. Allerdings wurden die Annotationen an isolierten Wortformen, d. h. außerhalb ihres Kontextes, vorgenommen. Daraus ergibt sich, dass bei ambigen Wortformen keine eindeutige Analyse erfolgen kann, sondern stattdessen alle theoretisch möglichen Lemmata und morphologischen Tags aufgezählt werden müssen. Ein Beispiel hierfür ist die ambige Form *quod*, die als Pronomen (‚welches') oder als Konjunktion (‚weil; dass') analysiert werden kann. *quod* erhält im Index Thomisticus also systematisch zwei Analysen: einmal als Pronomen, dem das Lemma *qui, quae, quod* sowie die morphologischen Analysen *Nom.Sg.Neut* bzw. *Akk.Sg.Neut* zugeordnet werden; und außerdem als Konjunktion mit dem Lemma *quod* und der Markierung *invariant* als morphologische Analyse.[5]

Annotationen stellen Abstraktionen dar, die die konkreten Wortformen auf allgemeinere, linguistisch motivierte Klassen abbilden. Die Kategorie, die die wenigsten Abstraktionsschritte involviert, ist das Lemma, das von den verschiedenen Flexionsformen eines Wortes abstrahiert. Morphologische Annotation kann in gewisser Weise als dazu komplementär angesehen werden: Hier wird von verschiedenen Wortstämmen abstrahiert und es werden Klassen gleich flektierter Wortformen zusammengefasst. Auf einer höheren Abstraktionsstufe steht die Wortart-Annotation, die Klassen morphologisch und distributionell äquivalenter Wörter bildet.

Vorteile von Annotationen

Annotationen stellen linguistisch relevante Informationen bereit, auf die in Suchanfragen direkt zugegriffen werden kann. Dank solcher Annotationen genügt *eine*

4 Alle Beispiele in diesem Artikel stammen aus einer Vorversion des *Referenzkorpus Mittelhochdeutsch* (REM), s. Abschnitt 3. Die Siglen der Quellenangabe (z. B. M402) sind im Appendix I aufgelöst. Die Stellenangabe (10r,22f) bezieht sich, sofern nicht anders angegeben, immer auf die Handschrift.

5 Siehe http://www.corpusthomisticum.org/it/.

Abfrage für alle Formen einer Kategorie, wie z. B. alle flektierten Formen des Verbs *sein* oder alle Wortformen im Dativ Singular oder alle Konjunktionen im Korpus.

Möchte man komplexere Phänomene untersuchen, die z. B. Beziehungen zwischen mehreren Wörtern involvieren (z. B.: welche Verben erlauben Genitivkomplemente?), so lassen sich dafür Korpora ohne Annotationen nicht sinnvoll nutzen. Ohne Annotationen bleibt nur die Option, Verben (d. h. ihre flektierten Formen) reihenweise einzeln abzufragen und die jeweiligen Fundkontexte manuell zu sichten. Insbesondere müssen alle abzufragenden Verb-Lemmata vom Nutzer vorher festgelegt werden. Es kann also nicht ergebnisoffen im Korpus gesucht werden.

Disambiguierte Annotationen erlauben es zudem, unmittelbar nach *relevanten* Belegen zu suchen. Im (nicht disambiguierten) Index Thomisticus kann man zwar automatisch beispielsweise nach allen Belegen von *quod* suchen und bekommt sie im KWIC-Format (*keyword in context*) angezeigt. Interessiert man sich aber nur für *quod* in der Verwendung als Konjunktion, muss man sämtliche Belege manuell sichten und jeweils entscheiden, ob es sich um die gewünschte Konjunktion oder das aktuell uninteressante Pronomen handelt. Untersucht man komplexere Phänomene (z. B.: welche Tempusformen tauchen nach der Konjunktion *quod* auf?), so potenziert sich das Problem.

Potenzielle Probleme von Annotationen

Als problematischer Punkt ist zu nennen, dass Annotation und Disambiguierung immer auch *Interpretation* erfordern. Bei modernen Korpora werden oft linguistische Tests formuliert, z. B. in Form von Paraphrasentests, die zur Bestimmung der konkret vorliegenden Form genutzt werden können. So kann beispielsweise ein Nomen, das nicht overt kasusmarkiert ist, durch ein overt flektiertes Pronomen ersetzt werden. In historischen Korpora sind solche Tests mangels muttersprachlicher Intuition natürlich nicht möglich. Hier muss man sich also an den overten Markierungen orientieren. Zusätzlich können unterstützend die Muster, die sich aus analogen Konstruktionen ergeben, als Interpretationshilfe hinzugezogen werden.

Ein weiterer Aspekt ist, dass kategoriale Grenzen nicht immer klar zu ziehen sind bzw. sich auch im Laufe der Sprachentwicklung z. B. aufgrund von Grammatikalisierungsprozessen verändern können. Oft sind solche „Grenzfälle“ die linguistisch interessanten. Zur Markierung solcher Fälle können *unterspezifizierte* Annotationen zum Einsatz kommen, die mögliche Ambiguitäten nicht oder nicht vollständig auflösen. Das in diesem Artikel vorgestellte Annotationsschema *HiTS* unterscheidet zudem zwischen der Wortart eines Lemmas an sich und der Wortart eines konkreten Belegs dieses Lemmas. Diese doppelte Annotation dient ebenfalls zur Markierung kategorialer Grenzverschiebungen.

Trotz aller Umsicht und Sorgfalt besteht die Gefahr, dass bei der Annotation falsche Entscheidungen getroffen werden. Es bieten sich hier verschiedene Lösungswege an, auf die ich gegen Ende dieses Artikels eingehen werde.

Editionen

Viele historische Korpora basieren auf mehr oder weniger stark normalisierten Editionen. Für lexikalisch-semantische Untersuchungen stellt dies weniger ein Problem dar. Für morphologisch-syntaktische Untersuchungen hingegen, bei denen beispielsweise die genaue Form einer Flexionsendung relevant ist, sind solche Korpora oft nur bedingt brauchbar (vgl. Wegera 2015). In jüngeren Korpusprojekten wird daher vermehrt auf die Digitalisierung einzelner Handschriften und Überlieferungsträger gesetzt, die *diplomatisch*, d. h. so handschriftennah wie möglich, transkribiert werden.

3 Historische Korpora des Deutschen

Im Folgenden stelle ich eine Reihe von linguistisch annotierten Korpora historischer deutscher Sprachstufen vor, beginnend mit den ältesten Sprachstufen. Der Fokus liegt dabei auf den Annotationen der Korpora sowie ihrer Verfügbarkeit (s. auch den Überblick in Kroymann et al. 2004).

Einige Korpora nutzen als Tagset[6] für die Wortart- und Morphologie-Annotation das STTS („Stuttgart–Tübingen Tagset“), das für moderne deutschsprachige Korpora das Standardtagset darstellt (Schiller et al. 1999). Ein Teil der Korpora nutzt das speziell für historische deutschsprachige Korpora entwickelte HiTS („Historisches Tagset“, Dipper et al. 2013) bzw. ein daran angelehntes Tagset. Beim Großteil der aufgeführten Korpora sind die Annotationen manuell überprüft.

Einige der Korpora können über das Korpussuchtool ANNIS (Zeldes et al. 2009) abgefragt werden. Das Tool sowie das Tagset HiTS werden in den Abschnitten 4 und 5 anhand von Beispiel-Anwendungen näher vorgestellt.

6 Als Tagset bezeichnet man das Inventar von Annotationskategorien. Die zugewiesenen Kategorien werden Tags genannt.

Deutsche Diachrone Baumbank (DDB) (Hirschmann und Linde 2010)

Dieses Korpus setzt sich aus religiösen Texten aus dem bairischen Sprachraum zusammen. Es enthält Texte aus dem Althochdeutschen (Ahd.), dem Mittelhochdeutschen (Mhd.) und dem Frühneuhochdeutschen (Fnhd.), mit einem Gesamtumfang von etwas über 8000 Wörtern. Die Texte aus dem Ahd. und Mhd. beruhen auf Editionen, die in unterschiedlichem Maß normalisiert sind. Annotiert ist das Korpus mit Lemma, Morphologie und Wortart nach dem STTS sowie mit syntaktischen Strukturen nach dem TIGER-Schema (Albert et al. 2003).[7] Das Korpus kann über ANNIS abgefragt und über das LAUDATIO-Repository heruntergeladen werden.[8]

Kali-Korpus

Das Kali-Korpus entstand unter der Leitung von Gabriele Diewald (Hannover). Es besteht aus Texten des Ahd. und des Mhd. und enthält über 200.000 Wörter. Alle Verbvorkommen sind lemmatisiert und morphologisch annotiert, zwei Texte sind zudem vollständig glossiert gemäß den Leipzig Glossing Rules (Comrie et al. 2008). Ein Teil des Korpus kann über ein Suchinterface abgefragt werden.[9]

Referenzkorpora

Die sog. Referenzkorpora wurden bzw. werden im Rahmen mehrerer DFG-finanzierter Projekte erstellt. Die Korpusstruktur sowie die Annotationsebenen und -prinzipien wurden in enger Absprache festgelegt, so dass die Korpora letztlich für diachrone Abfragen nutzbar sein sollen. Als gemeinsames Suchinterface wird das Korpussuchtool ANNIS (Zeldes et al. 2009) dienen. Annotiert sind die Referenzkorpora mit Lemma, Morphologie und Wortart nach bzw. angelehnt an HiTS. Es ist geplant, die Lemmata über gemeinsame IDs diachron zu verlinken.

Das **Referenzkorpus Altdeutsch (750—1050)** entstand unter der Projektleitung von Karin Donhauser (HU Berlin), Jost Gippert (Frankfurt) und Rosemarie

7 Zum TIGER-Schema, s. Abschnitt 5.5.

8 URL zur Dokumentation: http://korpling.german.hu-berlin.de/ddb-doku/; URL zur Korpusabfrage: http://korpling.german.hu-berlin.de/ddd/search.html; URL zum Korpus: http://hdl.handle.net/11022/0000-0000-2107-3.

9 URL der Projektseite: http://www.kali.uni-hannover.de.

Lühr (Jena). Es enthält sämtliche überlieferten Sprachdenkmäler des Althochdeutschen und des Altsächsischen mit einem Gesamtumfang von 650.000 Wörtern. Die Texte beruhen auf Editionen.[10]

Das **Referenzkorpus Mittelhochdeutsch (1050—1350)** entstand unter der Leitung von Thomas Klein (Bonn), Klaus-Peter Wegera (Bochum), Stefanie Dipper (Bochum) und Claudia Wich-Reif (Bonn). Es enthält eine zeitlich und räumlich ausgewogene Auswahl von Texten des Mittelhochdeutschen, die zudem die Überlieferungsformen Vers, Prosa und Urkunde angemessen repräsentiert. Das Korpus hat einen Gesamtumfang von 2 Millionen annotierten Wörtern (plus weitere nur transkribierte, aber nicht annotierte Wörter). Die Texte beruhen auf Handschriften. Dieses Korpus wird in den nachfolgenden Abschnitten genauer vorgestellt.[11]

Das **Referenzkorpus Frühneuhochdeutsch (1350—1650)** wird aktuell unter der Leitung von Hans-Joachim Solms (Halle), Klaus-Peter Wegera (Bochum), Ulrike Demske (Potsdam) und Stefanie Dipper (Bochum) erstellt. Die Auswahl der Texte beruht auf denselben Kriterien wie beim Referenzkorpus Mhd. Geplant ist ein Gesamtumfang von 4,4 Millionen Wörtern.[12]

Das **Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650)** entsteht aktuell unter der Leitung von Ingrid Schröder (Hamburg) und Robert Peters (Münster). Auch hier ist die Auswahl der Texte durch zeitliche und räumliche Parameter bestimmt sowie durch die Textsorte (wie z. B. Urkunden, Literatur, Inschriften).[13]

Mittelhochdeutsche Begriffsdatenbank (MHDBDB) (Springeth 2009)

Die ersten Planungen für dieses Korpus reichen zurück bis in die 1970er Jahre. In einer Kooperation zwischen Klaus M. Schmidt (Ohio) und Horst P. Pütz (Kiel) entstand dann ab den 1990er Jahren ein Korpus mit heute 1 Million Wortformen, die lemmatisiert und mit Wortart-Annotation (nach einem eigenen Tagset) versehen sind. Die Lemmata sind außerdem Bedeutungsfeldern zugeordnet.

10 URL zur Dokumentation: http://www.deutschdiachrondigital.de; URL zur Korpusabfrage: https://korpling.german.hu-berlin.de/annis3/ddd. Aktuell (Juli 2015) ist das Notker-Subkorpus noch nicht in ANNIS abfragbar.

11 URLs zu den Projektseiten: http://referenzkorpus-mhd.uni-bonn.de, http://www.rub.de/wegera/rem.

12 URL zur Projektseite: http://www.rub.de/wegera/ref.

13 URL zur Projektseite: https://vs1.corpora.uni-hamburg.de/ren.

Die Texte basieren auf Editionen. Sie können über ein Online-Tool abgefragt werden.[14]

Bonner Frühneuhochdeutsch-Korpus (Schmitz und Wegera 2013)

Die Anfänge dieses Korpus gehen ebenfalls bis in die 1970er Jahre zurück. Somit gehört es zu den ältesten elektronisch erfassten Korpora. Es entstand in Bonn unter der Leitung von Werner Besch, Winfried Lenders, Hugo Moser und Hugo Stopp und enthält Texte im Umfang von 500.000 Wortformen aus dem Zeitraum 1350–1700 in einer strukturierten Auswahl nach Zeit und Raum (nicht allerdings nach Textsorte). Das Korpus ist partiell (d. h. für die Hauptwortarten Nomen, Verben, Adjektiv) annotiert mit Lemma, Morphologie und den drei genannten Hauptwortarten. Es ist über ein Such-Interface online abfragbar und auch frei als Download verfügbar. Das Korpus geht in überarbeiteter Form in das Referenzkorpus Frühneuhochdeutsch ein.[15]

Fuerstinnenkorrespondenz (Prutscher und Seidel 2012)

Dieses Korpus entstand unter der Leitung von Rosemarie Lühr (Jena/Berlin). Sie besteht aus Briefen von Fürstinnen aus dem 16.–18. Jahrhundert im Umfang von 250.000 Wortformen. Es ist reichhaltig annotiert, u. a. mit nhd. Wortformen, Lemma, Morphologie und Wortart nach STTS und grammatischen Funktionen. Das Korpus kann über ANNIS abgefragt werden und ist über das LAUDATIO-Repository frei verfügbar.[16]

Mercurius-Baumbank (Demske 2007)

Diese Baumbank entstand unter der Leitung von Ulrike Demske (Potsdam). Sie enthält Texte des Frühneuhochdeutschen aus zwei Zeitungen aus den Jahren 1597 und 1667 mit einem Gesamtumfang von 170.000 Wortformen. Annotiert ist das Korpus mit Morphologie und Wortart nach dem STTS sowie mit syntaktischen

14 URL zur Dokumentation: http://www.mhdbdb.sbg.ac.at:8000/; URL zur Korpusabfrage: http://www.mhdbdb.sbg.ac.at:8000/mhdbdb/App?action=TextQueryModule.

15 URL zur Dokumentation und zum Download: http://www.korpora.org/Fnhd/; URL zur Korpusabfrage: http://korpora.org/fnhd/Suche.

16 URL zur Projektseite: http://dwee.eu/Rosemarie_Luehr/?Projekte_DFG-Projekte_Fruehneuzeitliche_Fuerstinnenkorrespondenz_im_ mitteldeutschen_Raum; URL zur Korpusabfrage: https://korpling.german.hu-berlin.de/annis3/#_c=RnVlcnN0aW5uZW5rb3JyZXNwb25kZW56MS4x; URL zum Korpus: http://hdl.handle.net/11022/0000-0000-82A0-7.

Strukturen nach dem TIGER-Schema (Albert et al. 2003). Das Korpus kann über ANNIS abgefragt sowie über das LAUDATIO-Repository heruntergeladen werden. In Abschnitt 5.5 wird es näher vorgestellt.[17]

GermanC (Scheible et al. 2011)

Das Korpus entstand unter der Leitung von Martin Durrell (Manchester) und enthält Texte des Frühen Neuhochdeutsch (1650–1800) im Umfang von 800.000 Wortformen. Auch hier ist die Auswahl der Texte bestimmt durch die Parameter Zeit, Raum und Textsorte (von Drama, Zeitung und Briefen bis hin zu wissenschaftlichen und Rechtstexten). Die Texte sind annotiert mit Lemma, Morphologie und Wortart gemäß dem STTS. Das Korpus ist frei herunterladbar.[18]

Kasseler Junktionskorpus (KAJUK) (Hennig 2013)

Das Korpus enthält Texte aus dem 17. und 19. Jahrhundert im Umfang von 96.000 Wortformen und entstand unter der Leitung von Vilmos Ágel und Mathilde Hennig (Gießen). Das Korpus ist partiell annotiert mit Information zu Junktoren, Prädikaten, Subjekten und ausgewählten weiteren Konstituenten. Es ist abfragbar über ANNIS und über das LAUDATIO-Repository verfügbar.[19]

Deutsches Textarchiv (DTA) (Haaf und Schulz 2014)

Dieses sehr umfangreiche Korpus mit Texten von 1600–1900 wird an der Berlin-Brandenburgischen Akademie der Wissenschaften unter der Leitung von Wolfgang Klein (Nijmegen) und Alexander Geyken (Berlin) erstellt. Es enthält aktuell über 1300 Texte mit rund 100 Millionen Wortformen. Die Digitalisate beruhen auf den jeweiligen Erstausgaben. Der Schwerpunkt liegt auf überregionalen Texten, mit einer ausgewogenen Auswahl von drei Textsorten (Belletristik,

17 URL zur Projektseite: http://www.uni-potsdam.de/guvdds/projekte/abgproj/mercurius.html; URL zur Korpusabfrage: https://korpling.german.hu-berlin.de/annis3/#_c=TWVyY3VyaXVz; URL zum Korpus: http://hdl.handle.net/11022/0000-0000-467D-6. Aktuell (Juli 2015) enthalten die verfügbaren Versionen keine morphologische Information.

18 URL zur Projektseite: http://www.llc.manchester.ac.uk/research/projects/germanc/files/; URL zum Korpus: http://ota.ox.ac.uk/desc/2544.

19 URL zur Projektseite: http://www.uni-giessen.de/kajuk/; URL zur Korpusabfrage: https://korpling.german.hu-berlin.de/annis3/#_c=S0FKVUs; URL zum Korpus: http://hdl.handle.net/11022/0000-0000-2102-8.

Gebrauchsliteratur, Wissenschaft). Die Texte werden automatisch annotiert mit Lemma und Wortart nach dem STTS, d. h. die Annotationen sind nicht manuell geprüft. Das Korpus kann über ein Suchtool online abgefragt werden.[20]

4 Anwendungsbeispiele

In den beiden folgenden Abschnitten geht es um die Nutzung annotierter Korpora. Anhand beispielhafter Anfragen möchte ich illustrieren, welche Recherche-Möglichkeiten gängige Suchtools bieten. Die Suchanfragen erfolgen in ANNIS (Zeldes et al. 2009), als Datengrundlage dient das Referenzkorpus Mittelhochdeutsch (REM) in einer Vorversion.[21] Eine Liste aller in REM enthaltenen Texte findet sich unter http://www.ruhr-uni-bochum.de/wegera/rem/Uebersicht.htm.

Die Darstellung beginnt mit einführenden Suchanfragen auf den Ebenen der Wortform, Lemma und Morphologie. Anfragen nach syntaktischen Phänomenen, die sich typischerweise auf die Ebenen der Wortart und gelegentlich der Morphologie beziehen, werden im nachfolgenden Abschnitt 5 behandelt. Eine Reihe von Suchbeispielen im Abschnitt zur Syntax sind angelehnt an Beispielphänomene aus Fleischer (2011).

4.1 Suche nach Wortformen

Die einfachste Suche ist die nach (diplomatischen) Wortformen. In ANNIS muss dazu der Name des betreffenden linguistischen Merkmals („tok_dipl“ in der

20 URL zur Dokumentation und Korpusabfrage: http://www.deutschestextarchiv.de/.

21 Das REM-Korpus ist zum aktuellen Zeitpunkt noch nicht veröffentlicht. Es fehlen noch letzte manuelle Korrekturdurchgänge auf den Annotationen, außerdem ist die Metainformation zu den einzelnen Texten noch nicht in ANNIS abfragbar. Damit kann z. B. bei dieser Vorversion keine Unterscheidung zwischen Prosa- und Verstexten getroffen werden, was aber natürlich insbesondere für syntaktische Untersuchungen ein sehr relevanter Faktor ist. Schließlich kann in der Vorversion in den Fällen, wo historische und moderne Wortgrenzen divergieren, bei der Suchanfrage nur auf die moderne Version Bezug genommen werden, und die manuell gesetzten modernen Satzgrenzen können ebenfalls nicht abgefragt werden. Alle diese Beschränkungen werden im offiziellen Release aufgehoben.

Die hier verwendete Version von REM umfasst nur einen Teil des Korpus im Umfang von rund 670.000 Wortformen.

Für eine umfassende Einführung in die Benutzung von ANNIS, s. http://annis-tools.org/documentation.html.

REM-Vorversion) und die gesuchte Oberflächenform eingegeben werden, wie in (2) gezeigt.

(2) `tok_dipl = "durch"`

Als Antwort auf diese Suche liefert ANNIS alle Textstellen mit der entsprechenden Wortform, zusammen mit allen Annotationen. Die Größe des angezeigten Kontexts kann vom Nutzer konfiguriert werden. Abbildung 1 zeigt einen Beispieltreffer für die Suche nach *durch*.[22] Im oberen Teil steht Information zur Belegstelle, darunter zur Wortform selbst, und im unteren Teil stehen die linguistischen Annotationen.

Annotations

page	p5								
no	48								
side	r								
column	c5								
line	10			11					
tok_dipl	alſo	iu	gót	durch	ſinen	ſun	uergeben	habe	iuuer
tok_mod	also	iu	got	durch	sinen	sun	uergeben	habe	iuuer
pos	AVD	PPER	NA	APPR	DPOSA	NA	VVINF	VAFIN	DPOSA
posLemma	AVD	PPER	NA	AP	DPOS	NA	VV	VA	DPOS
inflection	–	Dat_Pl_2	Nom_Sg		Masc_Akk_Sg_st	Akk_Sg		Subj_Pres_Sg_3	Neut_Akk_Sg_0
inflectionClass	–	–	st_Masc	–	–	stu_Masc	st5	wk	–
inflectionClassLemma	–	–	st_Masc	–	–	stu_Masc	st5	wk	–
lemma	all-sô	ir	got	durh	sîn	sun	ver-gëben	haben	iuwer

Abbildung 1: Beispieltreffer in ANNIS für die Anfrage `tok_dipl = "durch"`.

Für die Suche in historischen Korpora ergibt sich das Problem, dass aufgrund der stark variierenden Schreibungen an sich äquivalente Schreibungen nicht unmittelbar abfragbar sind. Hierfür gibt es unterschiedliche Lösungswege, die in den folgenden Abschnitten gezeigt werden.

4.2 Suche nach Lemmata

In vielen historischen Korpora ist zu jeder Wortform ein *standardisiertes Lemma* angegeben. In REM wird als Referenz für die Lemmatisierung das Wörterbuch von Lexer (1872) genutzt.[23] Die Lexer-Lemmata sind konstruierte standardisierte Lemmata eines normalisierten Mittelhochdeutsch.

Für die Anfrage nach dem Lemma *durh* ‚durch' (3a) liefert die REM-Vorversion beispielsweise insgesamt 1612 Belege, die sich auf 48 verschiedene (diplomatische)

22 Im Artikel werden die morphologischen Tags durch Punkte getrennt dargestellt, also z. B. „Nom.Sg" statt „Nom_Sg" wie in der Abbildung.

23 Das Wörterbuch ist online abfragbar: http://woerterbuchnetz.de/Lexer/. Die Form der Lemmata in REM unterscheidet sich z. T. in Details von denen im Lexer. Beispielsweise werden in Lexer nur zwei Arten von „e" unterschieden: *ë* für original germanisches „e" und *e* für den Rest (z. B. *her-bërge*

Schreibungen verteilen. Die 10 häufigsten Schreibungen mit ihren absoluten Frequenzen sind in (3b) angegeben. Das Beispiel zeigt auch, dass bei den Wortformen zwischen Groß- und Kleinschreibung unterschieden wird.

(3) a. `lemma = "durh"`
b. durch (915), durh (335), Durch (55), dur (47), thurh (39), dvrch (32), Thurh (31), durc (30), dvr (22), dvrh (12)

Die Lemmasuche liefert bei flektierenden Wortarten sämtliche Formvarianten. Um die Suche auf die gewünschte Form einzuschränken, können morphologische Bedingungen hinzugefügt werden (s. unten).

4.3 Suche nach vereinfachten Wortformen

Manche Korpora enthalten eine *vereinfachte (simplifizierte) Wortform*, bei der beispielsweise Diakritika gelöscht oder hochgestellte Zeichen normal gestellt werden. Die REM-Suche nach der vereinfachten Form *bluote* ‚Blut' (4a) liefert z. B. die in (4b) gezeigten diplomatischen Wortformen als Treffer.[24]

(4) a. `tok_mod = "bluote"`
b. blůte (34), bluote (2)

Einige Suchtools unterstützen *fuzzy matching*-Suchen. D. h. der Nutzer kann vereinfachte Wortformen eingeben, die dann expandiert werden zu Formen mit plausiblen Diakritika und Superskripten, nach denen letztlich faktisch gesucht wird. In diesem Szenario enthält nicht das Korpus das Wissen über die Beziehung zwischen diplomatischen und vereinfachten Formen, sondern der Suchprozess, der bei jeder Anfrage mit vereinfachten Formen die entsprechenden hypothetischen diplomatischen Formen automatisch neu erzeugt.

4.4 Suche mit Hilfe regulärer Ausdrücke

Schließlich können gezielt *reguläre Ausdrücke* für verallgemeinerte Suchanfragen eingesetzt werden. Reguläre Ausdrücke enthalten spezielle Symbole, sogenannte

‚Herberge'). REM nutzt zusätzlich *è*, um das Resultat eines *i*-Umlauts zu markieren. *e* in REM kennzeichnet „e" als Schwa (*hèr-bërge*), und *É* markiert unklare Fälle (z. B. *tÉmpel* ‚Tempel').

24 `tok_mod` steht für „modernisiertes Token". Zum einen verwenden diese Formen keine historischen Sonderzeichen. Zum andern entsprechen die Wortgrenzen den modernen Konventionen. `tok_dipl` hingegen gibt die Wortgrenzen des Originals wieder. Sämtliche wortbasierten Annotationen beziehen sich auf `tok_mod`-Formen, vgl. Fußnote 27.

Wildcards, und Operatoren, die für unterspezifizierte Suchanfragen genutzt werden können.

Z. B. steht der Punkt in der Anfrage in (5a) für genau ein beliebiges Zeichen, die Kombination Punkt–Fragezeichen (5b) für ein optionales beliebiges Zeichen (das Fragezeichen macht das vorhergehende Zeichen optional). Der Ausdruck in (5a) sucht also nach einer vereinfachten Wortform, die mit *bl-* beginnt und auf *-te* endet und insgesamt 5 oder 6 Buchstaben hat.

Die häufigsten Treffer sind in diplomatischer Form in (5b) gezeigt. (6a) zeigt eine noch allgemeinere Anfrage, die mit Hilfe von Disjunktionen „(...|...)" alternativ *b-* oder *p-* am Wortanfang und *-d* oder *-t* am Ende zulässt; Beispieltreffer werden in (6b) gelistet.

(5) a. `tok_mod = /bl..?te/`
b. blůte (34), blute (10), blŏte (6), blihte (3), blvte (3)

(6) a. `tok_mod = /(b|p)l..?(t|d)e/`
b. blůte (34), blute (10), plute (10), plůte (10), blŏte (6), blude (5)

Generell ist es besser, nach vereinfachten statt nach diplomatischen Wortformen zu suchen, da man auf diese Art z. B. auch (diplomatische) Vokale mit Diakritika finden kann. Interessiert man sich jedoch für die genaue Schreibung, z. B. die Verwendung von Diakritika oder des Schaft-*s* vs. rundem *s* (*ſ*, *s*), so ist die diplomatische Ebene relevant.

4.5 Suche mit mehreren Bedingungen

Um die Suche auf Formen des Lemmas *bluot* einzuschränken, kann die Anfrage wie in (7a) gezeigt erweitert werden. Die Kombination Punkt-Stern steht für eine beliebig lange Folge beliebiger Zeichen (der Stern erlaubt beliebig lange Folgen des vorhergehenden Zeichens). Einzelne Bedingungen können durch „&" miteinander kombiniert werden (zur besseren Lesbarkeit wird hier und im Folgenden jede Bedingung in eine eigene Zeile geschrieben). Der Teilausdruck „#1 _=_ #2" besagt, dass die an erster und zweiter Stelle genannten Bedingungen auf genau dieselbe Wortform zutreffen sollen. (7a) sucht also nach Belegen des Lemmas *bluot*, die auf *-e* enden. (6c) zeigt Beispieltreffer der Anfrage in (6a).

(7) a.
```
tok_mod = /.*e/
& lemma = "bluot"
& #1 _=_ #2
```
b. blůte (34), plute (10), plůte (10), blute (9), blŏte (6), pluote (4)

Es fällt auf, dass hier nur noch 9 Treffer der Form *blute* gelistet werden, im Gegensatz zu 10 Treffern in (5b) und (6b). Der Grund dafür ist, dass diese Wortform nicht immer dem Lemma *bluot* ‚Blut' zuzuordnen ist, sondern auch als Präteritum des Verbs *blüejen* lemmatisiert werden kann. (8a) zeigt einen Beleg mit dem Lemma *bluot*, (8b) mit dem Lemma *blüejen*.[25]

(8) a. du frowe gute. biwollin in deme *blute*. (M055-1c,1f)
‚du gute Frau, beschmutzt in diesem *Blut*'
b. wurzel allir gute uon dir der ast *blute* (M107S-295f; Editionszählung)
‚Wurzel aller Güte, von dir der Ast *blühte*'

4.6 Suche nach Morphologie

In Korpora, die mit morphologischer Information annotiert sind, kann statt nach Endungen auch direkt nach bestimmten morphologischen Belegen gesucht werden. (9a) sucht beispielsweise nach Belegen des Lemmas *bluot* im Dativ Singular in REM.[26] Entsprechende diplomatische Formen sind in (9b) gelistet. Unter den selteneren Formen finden sich auch zwei, die nicht auf *-e* enden (9c). Nach solchen Ausnahmen kann auch direkt mit Hilfe des Negationsoperators „!" gesucht werden (10) („!=" bedeutet: ist ungleich).

(9) a.
```
lemma = "bluot"
& inflection = "Dat.Sg"
& #1 _=_ #2
```
b. blůte (34), plute (10), plůte (9), blute (9), blŏte (6)
c. blůt (2), plv̊ti (1)

(10)
```
lemma = "bluot"
& inflection = "Dat.Sg"
& tok_mod != /.*e/
& #1 _=_ #2 & #1 _=_ #3
```

25 Hinweise zu den Beispielbelegen: Die in den Beispielen wiedergegebenen Wortgrenzen und Satzzeichen entsprechen der diplomatischen Transkription, bis auf eine Ausnahme: Trennungen von Wortformen aufgrund des Zeilenendes werden hier nicht wiedergegeben. Alle Satzzeichen werden außerdem einheitlich als Punkt wiedergegeben. Sequenzen in eckigen Klammern ([...]) stellen Ergänzungen dar.

Historische Wortsequenzen, die mehreren modernen Wortformen entsprechen, werden in den Übersetzungen durch Unterstrich getrennt übersetzt, z. B. *fanter* ‚sandte_er'. Die Übersetzungen sind so gewählt, dass sie möglichst nah am Original bleiben.

26 Die morphologische Annotation wird in REM unter dem Merkmal „inflection" gelistet. Alle morphologischen Tag-Bestandteile sind im Appendix III aufgeführt und erläutert.

Da Nomen prinzipiell nicht overt genusmarkiert sind und zudem in den älteren Sprachstufen das Genus vieler Nomen weniger stark festgelegt ist als im heutigen Deutsch, sind in REM die Nomen nicht für Genus markiert. Man kann stattdessen nach genusmarkierten Artikeln und/oder Adjektiven suchen, die das Genus nachfolgender Nomen anzeigen. Dafür benötigen wir den Punkt-Operator „#1 . #2“, der für unmittelbare Präzedenz steht. D. h. die an erster und zweiter Stelle genannten Bedingungen müssen von zwei aufeinanderfolgenden Wortformen erfüllt werden.

Der Suchausdruck in (11a) erfasst beispielsweise morphologisch als feminin markierte Wortformen, die dem Lemma *slange* ‚Schlange' direkt vorangehen. Zufälligerweise enthält der Treffer in (11b) gleich zwei Vorkommen dieses Lemmas, und zwar eines in maskuliner und eines in femininer Verwendung. (11c,d) zeigen die morphologische Annotation des unmittelbaren Kontexts beider Verwendungen.

(11) a.
```
inflection = /.*Fem. */
& lemma = "slange"
& #1 . #2
```

b. Do hiezhe *eínín kupírín flangín* gízín. Vndí gibot dat man *dí flange*. hinge vor dat volc an eíní ſtangen. (M239-127r,9ff)
‚da hieß_er *eine kupferne Schlange* gießen und gebot, dass man *die Schlange* hinge vor das Volk an eine Stange'

c.

eínín	kupírín	flangin
Masc.Akk.Sg.st	Pos.Masc.Akk.Sg.0	Akk.Sg

d.

dí	flange
Fem.Akk.Sg	Akk.Sg

Vergleicht man die Anzahl der als maskulin bzw. feminin markierten vorausgehenden Wortformen, so ergibt sich ein klares Bild: die maskuline Verwendung mit 23 Belegen überwiegt ganz klar gegenüber der femininen Verwendung mit gerade einmal einem Beleg. Allerdings erfasst unser Suchausdruck natürlich nur einen Teil der Verwendungen des Lemmas *slange*. Gleichzeitig können natürlich auch einige falsche Treffer durch den Suchausdruck erfasst sein, in denen z. B. das vorausgehende Wort gar keine Konstituente mit dem Lemma *slange* bildet. Die Zahl der unerwünschten Treffer könnten wir einschränken, indem beispielsweise nur bestimmte Wortarten zugelassen werden wie Artikelwörter oder Adjektive. Dies birgt aber immer die Gefahr, dass damit auch echte Treffer ausgeschlossen werden.

Zurück zu den nicht erfassten Verwendungen: Die maskuline Verwendung in (12a) ist beispielsweise nicht abgedeckt, da das Adjektiv, das dem Nomen unmittelbar vorausgeht, nicht flektiert ist.[27] Ebenso unerfasst sind

27 In der diplomatischen Transkription sind das Adjektiv und das Nomen in einem Wort geschrieben. Für die Annotation mit linguistischer Information werden aber Einheiten mit

Pluralverwendungen wie in (12b), bei denen das Genus prinzipiell nicht morphologisch bestimmbar ist.

(12) a. den uerflant *der altflange*! (M024-120rb,46)
‚den verschlang *die alte_Schlange*'
b. Alſ wir id wale wizzen. dad ſi *flangen* bizzen. (M045-1r,5f)
‚also wir es wohl wissen, dass sie *Schlangen* bissen'

Liegen bei Adjektiven und Artikelwörtern Wortformen vor, die selbst nicht eindeutig genusmarkiert sind, so sind sie in REM unterspezifiziert annotiert, z. B. durch Aufzählung der möglichen Werte („Masc-Neut" in (13a)) oder durch die Wildcard „*", die hier „unbestimmbar" bedeutet (13b).[28] Das scheinbare Possessivpronomen *íro* ist tatsächlich eine genitivische, also nicht-kongruierende Form, ähnlich zu deren im heutigen Deutsch. Die „0" am Ende des morphologischen Tags („*.Gen.Pl.0") zeigt an, dass eine Nullendung vorliegt. Die Endung „echter" Possessivpronomen wie *sîn* ‚sein' wie in (13c) wird als „stark" („st") markiert.

(13) a. an dem ende (M004-2v,9)
– Masc-Neut.Dat.Sg Dat.Sg
‚an dem Ende'
b. ín ír zûnon. ún ín íro hêrzon. (M155-32r,11f)
– *.Gen.Pl.0 Dat.Pl – – *.Gen.Pl.0 Dat.Pl
‚in ihren Zungen und in ihren Herzen'
c. mit finem finne. (M004-2v,3)
– Masc.Dat.Sg.st Dat.Sg
‚mit seinem Sinn'

5 Syntaktische Phänomene

Viele Korpora sind mit Informationen zur Wortart annotiert. In modernen Korpora werden dabei nicht nur die „grobkörnigen" Wortarten wie Nomen, Verben, Adjektive annotiert, sondern es wird zusätzliche syntaktisch relevante Information kodiert, z. B. ob es sich um eine finite oder infinite Verbform handelt oder um ein flektiertes oder unflektiertes Adjektiv. In REM wird bei flektierten Adjektiven außerdem festgehalten, ob es seinem Bezugsnomen voran- oder nachgestellt ist.

---

modernen Wortgrenzen zu Grunde gelegt, vgl. Fußnote 24. D. h. bei Abfragen nach Lemmata, morphologischer Information etc. werden die diplomatischen Wortgrenzen ignoriert.

28 *ín ír* in (13b) ist in der Handschrift in einem Wort geschrieben, hier aber wegen der morphologischen Annotationen in zwei Wörtern wiedergegeben.

Wie im Folgenden gezeigt wird, kann solche Information in bestimmtem Umfang als Ersatz für eine „volle“ syntaktische Annotation (z. B. in Form eines Konstituentenbaumes oder einer Dependenzanalyse) dienen. Im letzten Teil dieses Abschnittes gehe ich noch kurz auf syntaktisch tief annotierte Korpora ein.

5.1 HiTS: Historisches Tagset

In den Referenzkorpus-Projekten Altdeutsch und Mittelhochdeutsch wurde ein gemeinsames Tagset, d. h. ein Inventar von Annotationskategorien, für historische Sprachdaten des Deutschen entwickelt: „HiTS“ (Historisches Tagset) (Dipper et al. 2013).

HiTS orientiert sich am STTS, dem Standard-Tagset für moderne deutschsprachige Korpora, und übernimmt das hierarchische Design der Tagnamen: der erste Buchstabe bzw. die Buchstabengruppe des Tagnamens zeigt die Hauptwortart an, die zweite Gruppe eine Unterklasse davon etc. Mit Hilfe regulärer Ausdrücke kann so nach generelleren oder nach spezifischeren Wortarten gesucht werden. Die Suchanfrage in (14a) sucht beispielsweise nach der Wortart (Englisch „part of speech“ = „POS“) *ADJA* („Adjektiv, attributiv“), während (14b) sämtliche Adjektive des Korpus erfasst. (Eine Auflistung sämtlicher HiTS-Tags findet sich im Appendix II.)

(14) a. `pos = "ADJA"`
b. `pos = /ADJ.*/`

HiTS unterscheidet sich vom STTS in verschiedenen Punkten. U. a. sind die definiten und indefiniten Artikel nicht wie beim STTS einer eigenen Kategorie „ART“ zugeordnet, sondern Unterklassen der Kategorie „D“ (Determinativa): „DDA“ („Determinativ, definit, attributivisch/vorangestellt“) kennzeichnet den definiten Artikel, „DIA“ („Determinativ, indefinit, attributiv/vorangestellt“) den indefiniten. Die Sonderstellung des Artikels, die im STTS durch ein eigenes Tag „ART“ betont wird, ist eine neuere Entwicklung.

Des Weiteren fehlt bei den historischen Daten die Möglichkeit, über die muttersprachliche Intuition ambige Strukturen zu disambiguieren. Es können beispielsweise keine linguistischen Paraphrasentests vorgenommen werden. In solchen Fällen soll „konservativ“ annotiert werden, d. h. es soll prinzipiell die historisch ältere Analyse bevorzugt werden. Beispielsweise soll bei Konstruktionen, die (noch) als pränominaler Genitiv oder (schon) als Kompositumserstglied analysiert werden könnten, die Genitiv-Lesart annotiert werden (dies ist in REM aus historischen Gründen allerdings abweichend umgesetzt, vgl. Abschnitt 5.3.2, „Prä- vs. postnominale Genitive“).

Schließlich wird in HiTS unterschieden zwischen der „generellen“ Wortart des Lemmas und der spezifischen Wortart des konkreten Belegs. Eine Lemma-Wortart wäre z. B. „ADJ“ (allgemeines Adjektiv), eine Beleg-Wortart „ADJA“ (attributiv verwendetes Adjektiv). Die beiden Wortarten können auch verschiedenen Klassen entstammen. Beispielsweise kann ein Adjektiv (Lemma-Wortart) auch in der Funktion eines Adverbs verwendet werden (Beleg-Wortart). Diese Doppelannotation ermöglicht einen direkten Zugriff auf Sprachwandelprozesse, die mit einem Wortartwechsel einhergehen: Die Lemma-Wortart kodiert dabei die „originale“, „konservative“ Wortart, die Beleg-Wortart die „aktuelle“, „progressive“. (15) zeigt zwei entsprechende Beispiele. In (15a) ist das Wort *man* ursprünglich ein Nomen („NA“, „Nomen appelativum“), wird aber hier als Indefinitpronomen verwendet („PI“, „Pronomen indefinit“). In (15b) fungiert das Partizip Präteritum („VVPP“) als attributives, nachgestelltes Adjektiv („ADJN“). In der mittleren Zeile ist jeweils die Wortart des Belegs, in der unteren die Wortart des Lemmas angegeben.

(15) a. do man daz kint beſniden ſcolte. (M004-1r,33)
KOUS PI DDART NA VVINF VMFIN
KO NA DD NA VV VM
‚als *man* das Kind beschneiden sollte‘

b. diu chindelin niu gebornen. (M014-131rb,21f)
DDART NA ADJD ADJN
DD NA ADJ VVPP
‚die neu *geborenen* Kinder‘

Belege, in denen die beiden Wortarten wie in (15) differieren, können mit Abfragen wie in (16) gefunden werden.

(16)
```
pos = "PI"
& posLemma != /P.*/
& #1 _=_ #2
```

In den folgenden Abschnitten spielt nur die Wortart des Belegs eine Rolle.

5.2 Suche nach Wortart

Die einfachste Anfrage ist die Suche nach Belegen einer bestimmten Wortart. In (17)–(19) sind Anfragen nach Tags gelistet, die in HiTS speziell für die historischen Daten definiert wurden, zu denen es also keine Entsprechung im STTS gibt. In den (b)-Beispielen wird jeweils ein beispielhafter Treffer angegeben.

„ADJN“ markiert nachgestellte attributive Adjektive. „DPOSGEN“ ist die Kategorie für die bereits im Abschnitt 4 erwähnten genitivischen Possessivpronomen, aus denen sich später echte Possessivpronomen entwickelten (Kategorie „DPOSA“).

(17) a. `pos = "ADJN"` („Adjektiv, attributiv, nachgestellt“):
b. do fanter fil drate. ze einer magede vil *here*. den engel gabrielem. (M004-1r,15f)
‚da sandte_er sehr schnell zu einer sehr *herrlichen* Magd den Engel Gabriel‘

(18) a. `pos = "DPOSGEN"` („Determinativ, personal-possessiv, Genitiv“)
b. ín*ír* zûnon. ún ín *íro* hêrzon. (M155-32r,11f)
‚in_ *ihren* Zungen und in *ihren* Herzen‘

(19) a. `pos = "DPOSA"` („Determinativ, possessiv, attributiv/vorangestellt“)
b. *iren* rehten ſcheiffere. (M045-1r,2f)
‚*ihren* rechten Schöpfer‘

5.3 Wortarten und reguläre Ausdrücke

Man kann nun Suchen nach Wortarten und gegebenenfalls auch nach morphologischer Information mit Hilfe regulärer Ausdrücke so kombinieren, dass nach interessanten syntaktischen Phänomenen gesucht werden kann.

5.3.1 Flexion prädikativer Adjektive

Zunächst die Anfrage nach einem Phänomen, das u. a. in Fleischer (2007, S. 307) thematisiert wird: die Flexionseigenschaften prädikativ verwendeter Adjektive (vgl. Fleischer 2011, S. 70–71). Die Anfrage in (20a) sucht nach prädikativen Adjektiven (gekennzeichnet durch das Wortart-Tag „ADJD“), die morphologisch als Nominativ markiert sind.[29] Insgesamt überwiegen in REM die unflektierten Verwendungen überaus deutlich (mit 7377 Belegen = 95%) gegenüber den flektierten (363 Belege = 5%). Schaut man nur die nominativischen flektierten Formen an, sind maskuline und Plural-Formen am häufigsten, wie im (spätahd./frühmhd.) Beispiel (20b). (20c) listet die morphologischen Typen, die unter den

29 Mit „ADJD“ markierte Adjektive können auch in anderen Kasus als Nominativ verwendet werden, wie in (i). Durch die Beschränkung auf den Nominativ sollen nur die Fälle erfasst werden, in denen das Adjektiv prädikativ mit den Verben *sein* und *werden* steht.

(i) er fůrtin *lemptigen* inſin lant. (M009-111rb20f)
‚er führt_ihn *lebendig* in_sein Land‘

nominativisch flektierten die häufigsten sind (mit einer Mindestfrequenz von 10). (Ich werde in Abschnitt 5.4 darauf eingehen, wie Frequenztabellen wie in (20c) erstellt werden können.)

(20) a.
```
pos = "ADJD"
& inflection = /.*Nom.*/
& #1 _=_ #2
```

b. Alſámo unſer fatér adám. unzér *nakedêr* uuaſ in paradyſo (M155-33r,21f)
‚wie unser Vater Adam, als_er *nackt* war im Paradies'

c.

Morphologischer Typ	Frequenz
Pos.Masc.Nom.Sg.st	54
Pos.*.Nom.Pl.st	30
Pos.Masc.Nom.Pl.st	22
Pos.Masc.Nom.Sg.wk	19
Pos.Fem.Nom.Sg.wk	12
Pos.Fem.Nom.Sg.st	10

Die Gegenprobe – wie oft diese morphologischen Typen (z. B. „Pos.Masc.Nom.Sg") unflektiert vorkommen – ist allerdings nicht durchführbar, da die entsprechenden Adjektive unterspezifiziert annotiert sind (z. B. unflektierte Adjektive im Positiv mit „Pos.*.*.*.0"). Daher lassen sich die Ergebnisse auch nicht mit denen von Fleischer (2007) vergleichen (unter den nominativischen Adjektiven in den althochdeutschen Texten, die Fleischer (2007) untersucht, kommen im Durchschnitt noch 21% flektierte Formen vor).

5.3.2 Prä- vs. postnominale Genitive

Ein weiteres Phänomen ist die Verteilung von Genitivattributen, die prä- oder postnominal stehen können (vgl. Fleischer 2011, S. 56–58, 78–80). Nach einer Untersuchung von Barufke (1995, zitiert nach Prell 2000, S. 31) anhand von mhd. Verstexten stehen durchschnittlich nur 6.3% aller Genitivattribute nachgestellt, d. h. hinter ihrem Bezugsnomen. Prell (2000, S. 32) findet in mhd. Prosatexten deutlich abweichende Ergebnisse, mit einem durchschnittlichen Anteil an nachgestellten Genitivattributen von 58.8%. Die Anteile innerhalb der verschiedenen Prosatexte schwanken allerdings stark.

REM enthält keine Annotationen zu Konstituentengrenzen oder Dependenzrelationen. D. h. man kann nicht gezielt nach Genitivattributen suchen (dafür benötigt man eine *Baumbank*, s. Abschnitt 5.5). Man kann aber die Wortart-Annotation in REM dazu nutzen, mit Hilfe heuristischer Anfragen zum einen

grobe Trends festzustellen, zum andern die Menge der manuell zu überprüfenden Fälle drastisch zu reduzieren. In nicht annotierten Korpora müssten theoretisch sämtliche Sätze einzeln durchgegangen und auf Vorkommen von Genitiv-Attributen hin untersucht werden. In annotierten Korpora hingegen kann man gezielt nach Belegen von Nomen im Genitiv suchen und braucht dann „nur noch" die Treffer (oder eine Stichprobe daraus) einzeln zu überprüfen.

Ich beschränke mich hier auf die Darstellung der groben Trends. Die Anfrage in (21) sucht nach einer Wortform im Genitiv, die keines der in (18) illustrierten genitivischen Possessivpronomen ist (= erste und zweite Bedingung, die auf dieselbe Wortform zutreffen müssen), gefolgt von einem Nomen oder Eigennamen (Tag „NA" bzw. „NE"), die selbst nicht im Genitiv stehen (= dritte und vierte Bedingung). Die Anfrage für postnominale Genitive vertauscht die Abfolge der beiden Bedingungsblöcke.[30]

(21)
```
inflection = /.*Gen.*/
& pos != "DPOSGEN"
& inflection != /.*Gen.*/
& pos = /N.*/
& #1 _=_ #2
& #1 . #3
& #3 _=_ #4
```

Insgesamt gibt es 4785 Treffer (= 62%) für die heuristische pränominale Anfrage und 2953 (= 38%) für die postnominale. (22a,b) zeigen Beispielsätze für beide Verwendungen. Um die Resultate mit den oben zitierten vergleichbar zu machen, müssten zum einen die Anteile pro Text erhoben werden, zum andern nach Prosa- vs. Verstexten unterschieden werden, was beides in der hier genutzten REM-Version nicht möglich ist (s. Fußnote 21).

(22) a. ſtant hir inne. důrc *deſ heil*[*igen*] *criſtef* willeN. (M003-69E,2f; Editionszählung)
‚stand hierin durch *des Heiligen Christus* Willen'

b. er iſt Leibeigener und Knecht *des heiligen Mannes Sankt Johannes*' (M004-2v,4f)

30 D. h. die Bedingung `#1 . #3` wird ersetzt durch `#3 . #1`.
Es soll nochmals betont werden, dass diese Anfragen nicht erfassen können, ob die genitivische Wortform tatsächlich als Genitivattribut zu dem nicht-genitivischen Nomen oder Eigennamen fungiert.

Es finden sich unter den pränominalen Belegen allerdings auch Beispiele, in denen der Genitiv eher als Kompositumserstglied zu interpretieren ist. Dazu gehören Belege wie *den gotef fun* ‚den Gottes-Sohn' im ersten Teilsatz von (23a), in denen dem Genitiv ein Artikel und/oder Adjektiv vorausgeht, das mit dem nachfolgenden Kopfnomen kongruiert. (23a) enthält außerdem noch ein ambiges Beispiel: *ane mannef minne* ‚ohne Mannes-Minne', das sowohl als pränominaler Genitiv als auch als Kompositum analysiert werden könnte. Nach HiTS soll „konservativ" annotiert werden, d. h. es sollte dann der Genitiv als die historisch vorausgehende Konstruktion annotiert werden.

In REM wurde die Entscheidung zwischen pränominalen Genitiven und Komposita allerdings nicht anhand von Kongruenzrelationen (wie eigentlich in HiTS vorgesehen), sondern mit Bezug auf Standard-Wörterbücher wie Lexer (1872) getroffen: Was dort als Kompositum eingetragen ist, wird auch in REM so analysiert. *gotef fun* ‚Gottes-Sohn' (23a) ist im Lexer nicht eingetragen. Hingegen gibt es zu *gotfhuf* ‚Gotteshaus' (23b) einen entsprechenden Eintrag im Lexer unter dem Stichwort *got-hûs*.[31] Folglich werden beide Fälle in (23a) als Genitive analysiert, der Fall in (23b) aber als Kompositum. (Zu den Wortgrenzen in REM, s. Fußnote 21.)

(23) a. ſiv ſcolte den *gotef fun* gewinnen. ane *mannef minne*. (M004-1r,17)
‚sie sollte den *Gottes-Sohn* gewinnen ohne *Mannes-Minne*'

b. dů wrden ſi zeu[u]oret. unde ir *gotfhuf* zeſtoret. (M119-103va,40f)
‚da wurden sie zerstreut und ihr *Gotteshaus* zerstört'

Schließlich noch einige Frequenzvergleiche: Die weit überwiegende Mehrheit der Genitive sind Singular-Formen (83% der pränominalen, 73% der postnominalen). Bei den pränominalen Genitiven ist das mit weitem Abstand meistgenutzte Lemma got ‚Gott' (23% aller Fälle); bei den postnominalen Genitiven ist das Lemma *hêrre* ‚Herr' am häufigsten, das in 6% aller Fälle vorkommt.[32] Laut Ebert (1999, S. 93–96), der vorwiegend Daten des Fnhd. untersucht, wird die Stellung des Genitivs v. a. durch den Faktor Personenbezeichnung (präferiert pränominal) vs. Nichtpersonenbezeichnung (präferiert postnominal) bestimmt. Allerdings treten die Lemmata *Gott, Herr* häufiger auch nachgestellt auf. Die Ergebnisse aus dem REM-Korpus bestätigen bei

31 Im Trierer Online-Wörterbuchnetz unter der URL http://woerterbuchnetz.de/Lexer?lemma=gothus.

32 Das genitivische Kopfnomen wird von der postnominalen Version der Anfrage in (21) erstmal nicht erfasst. Um das häufigste Nomen-Lemma unter den postnominalen Genitiven zu ermitteln, ist eine komplexere Anfrage notwendig, die z. B. nach genitivischen postnominalen Nomen im Abstand von bis zu zwei Wörtern sucht (vgl. die Anfrage in (30)).

hêrre die besondere Nachstellung, zeigen allerdings bei *got* ein entgegengesetztes Bild mit weit überwiegender Voranstellung.[33]

5.3.3 Abfolge von Pronomen

Ein weiteres Beispielphänomen ist die relative Abfolge von Personalpronomen im Dativ und Akkusativ (vgl. Fleischer 2011, S. 73). In seinen Texten beobachtet Fleischer (2010, S. 527–531) als insgesamt dominierende Abfolge Akkusativ > Dativ (64% der Fälle), allerdings findet sich v. a. bei Pronomen der 1. und 2. Person Singular auch die Abfolge Dativ > Akkusativ.

Die Abfrage in (24a) sucht nach zwei adjazenten Personalpronomen mit beliebigem Kasus in REM. (24b,c) zeigen Beispiel-Belege mit der Abfolge Akkusativ > Dativ bzw. Dativ > Akkusativ, jeweils mit Pronomen der 3. Person. Beide Belege stammen aus Verstexten.

(24) a.
```
pos = "PPER"
& pos = "PPER"
& #1 . #2
```
b. gegeben habete *fi ime* got. (M028-95va,6)
‚gegeben hatte *sie*[Sg] *ihm* Gott'
c. chod wolte fin mendente. daz *ime fi* got hête gigeben in ellente. (M088-90v,9f)
‚(er) sprach, (er) wollte froh sein, dass *ihm sie*[Pl] Gott hatte gegeben im Elend'

(25) zeigt die am häufigsten vorkommenden Abfolgen der Kasus (mit mehr als 10 Vorkommen), mit ihrer absoluten und relativen Frequenz. Ganz deutlich dominieren die Abfolgen, in denen vorne ein Nominativ-Pronomen steht, mit insgesamt 90% aller Fälle. Die Abfolge Akkusativ > Dativ nimmt im Vergleich zur Abfolge Dativ > Akkusativ 60% der Fälle ein, also ein ähnliches Ergebnis wie das in Fleischer (2010) berichtete.

(25)

Abfolge der Kasus	Frequenz abs.	Frequenz rel.
Nom > Akk	2966	46%
Nom > Dat	2504	39%
Nom > Gen	343	5%

33 Im REM gibt es von *hêrre* 99 pränominale und 121 postnominale Belege; von got gibt es 1110 pränominale und 43 postnominale Belege.

Akk > Dat	**246**	**4%**
Dat > Akk	**166**	**3%**
Dat > Gen	58	< 1%
Akk > Gen	56	< 1%
Gen > Dat	33	< 1%
Akk > Akk	28	< 1%
Gen > Akk	20	< 1%
Dat > Nom	18	< 1%
Nom > Nom	17	< 1%
Akk > Nom	13	< 1%

Fleischer (2010) folgend schauen wir uns außerdem die Verteilungen der beiden Abfolgen Akkusativ > Dativ und Dativ > Akkusativ genauer an. (26) zeigt die Person-Merkmale der Pronomen in diesen Abfolgen (mit mehr als 10 Vorkommen). Die Daten bestätigen auch hier die Ergebnisse von Fleischer: In der Abfolge Akkusativ > Dativ stehen in 89% der Fälle Pronomen der 3. Person vorne, in der Abfolge Dativ > Akkusativ nur in 32% der Fälle.

(26)

Akkusativ > Dativ			Dativ > Akkusativ		
Person	Frequenz		Person	Frequenz	
3 > 3	118	48%	1 > 3	62	37%
3 > 2	67	28%	3 > 3	53	32%
3 > 1	33	13%	2 > 3	51	31%

Als nächstes betrachten wir noch die relative Abfolge von drei direkt aufeinander folgenden Personalpronomen, s. die Abfrage in (27a). Der mit Abstand häufigste Typ ist der in (27b) gezeigte. Schaut man sich wieder die Kasus der Pronomen an, so ergibt sich die Rangfolge in (27c) (aufgeführt sind nur die Abfolgen mit mehr als einem Beleg). In 97% der Fälle stehen Pronomen im Nominativ an erster Stelle. Die häufigste Abfolge, Nominativ > Akkusativ > Dativ, entspricht der dominanten Abfolge im modernen Deutsch.

(27) a.
```
pos = "PPER"
& pos = "PPER"
& pos = "PPER"
& #1 . #2
& #2 . #3
```

b. zeware fagen *ich ez ev*. (M028–95va,25)
‚zur_Wahrheit/wahrheitsgemäß sage *ich es euch*'

c.

Abfolge	Frequenz	
Nom > Akk > Dat	102	49%
Nom > Dat > Akk	52	25%

Nom > Dat > Gen	18	9%
Nom > Akk > Gen	13	6%
Nom > Gen > Dat	8	4%
Dat > Nom > Akk	4	2%
Nom > Gen > Akk	3	1%
Nom > Akk > Akk	3	1%

Der erste Abfolgetyp, bei dem kein Nominativ-Pronomen vorne steht, ist Dativ > Nominativ > Akkusativ mit vier Belegen. Schaut man sich die entsprechenden Beispiele an, so sieht man, dass drei der vier Beispiele tatsächlich keine relevanten Treffer darstellen: hier steht das Dativ-Pronomen nämlich innerhalb einer Präpositionalphrase, wird also von einer Präposition regiert, wie im Fall *zu ime* ‚zu ihm' in (28a).

Der einzige echte Treffer (28b) stammt aus M116, wieder ein Text in Versen.

(28) a. zu *ime er in* ſazte. daz er in ioſebef irgazte. (M116–81rb,11f)
‚zu *ihm er ihn* setzte, dass er ihn des Josephs ergötzte'
b. *inerz* ane legete. zů deme gewalte er in ſta[be]te. (M116–80vb,25f)
‚*ihnen_er_es anlegte*, zu der Gewalt/Macht er ihn einwies'

Um unerwünschte Treffer wie in (28a) auszuschließen, könnte man als zusätzliche Bedingung einführen, dass keine Präposition vorausgehen darf (29).

(29)
```
pos != "APPR"
& pos = "PPER"
& pos = "PPER"
& pos = "PPER"
& #1 . #2
& #2 . #3
& #3 . #4
```

Das löst zwar das aktuelle Problem insofern, als bei dieser Anfrage tatsächlich nur noch der erwünschte Treffer (28b) übrig bleibt. Klar ist aber, dass auch die Anfrage in (29) theoretisch noch Platz für falsche Treffer lässt. Beispielsweise könnte das erste Pronomen nach wie vor von einer Präposition regiert werden, die allerdings nicht unmittelbar vorangeht, wie in diesem Fall: *Präposition* > Personalpronomen > Konjunktion > *Personalpronomen*.

Um wirklich sicher zu sein, dass man drei Schwesterkonstituenten abfragt, benötigt man einen komplexeren Typ von Annotation: syntaktische Strukturen (s. Abschnitt 5.5).

5.3.4 Abfolge von Nominalphrasen

Wollte man die Untersuchungen des vorhergehenden Abschnitts ausweiten auf volle Nominalphrasen, so würden die Anfragen deutlich komplexer. REM und ähnlich annotierte Korpora bieten sich v. a. für die Suche nach (Sequenzen von) Einzelwörtern mit bestimmten Eigenschaften an. Möchte man nach Konstituenten suchen, muss man die Anfrage in eine entsprechende Einzelwort-Sequenz „übersetzen“ und die Struktur der Konstituenten mit Hilfe regulärer Ausdrücke beschreiben. In unserem konkreten Fall bedeutet das, dass man eine kleine Grammatik für Nominalphrasen schreibt. (30a) zeigt eine Beispielanfrage und (30b) einen dazugehörigen Treffer (die Wortart-Tags sind mit angegeben).[34]

(30) a.
```
pos = /(D.*ART|D.*A)/
& pos = "AVD"
& pos = "ADJA"
& pos = /N./
& pos = "AVD"
& pos = "ADJN"
& ( #1 . #2 & #2 . #3 & #3 . #4
  | #1 . #3 & #3 . #4
  | #1 . #4
  | #3 . #4
  | #4 . #5 & #5. #6
  | #4 . #6 )
```

b. dife Uil gute wŏrze. (M121V-29vb,19f)
DDA AVD ADJA NA
‚diese sehr guten Wurzeln‘

Ein Problem dieses Ansatzes ist, dass zum einen die Suchausdrücke sehr komplex und dadurch fehleranfällig werden können. Zum anderen ist insbesondere

34 Die Anfrage spezifiziert Artikelwörter (1. Bedingung), gefolgt von einfachen APs (bestehend aus einem Adverb „AVD“ und einem attributiven Adjektiv „ADJA“; 2. und 3. Bedingung), gefolgt vom Kopfnomen (Nomen appelativum oder Eigenname: „N.“; 4. Bedingung). Nach dem Kopfnomen kann wiederum eine einfache AP stehen. Die nachfolgenden alternativen Präzedenzbedingungen bewirken, dass einzelne Bedingungen faktisch optional sind. Beispielsweise kann das Adverb in der ersten Adjektivphrase fehlen (Bedingung „#1 . #3 & #3 . #4“) oder die gesamte Adjektivphrase fehlt („#1 . #4“).

Tatsächlich ist die eigentliche Anfrage noch komplexer, da in ANNIS keine ungebundenen Variablen vorkommen dürfen, in der ersten Disjunktion aber z. B. die 5. und 6. Bedingung nicht aufgenommen werden. Die Formulierung in (30) dient v. a. dazu, zu illustrieren, wie reguläre Ausdrücke als Approximation für Konstituenten genutzt werden können.

problematisch, dass man beim Formulieren des regulären Ausdrucks schon vorhersehen muss, was für Typen von Phrasen vorkommen können. Oft gehört dieser Punkt aber zur Forschungsfrage dazu. Eine sinnvolle Strategie ist daher, im ersten Schritt großzügige, unterspezifizierte Anfragen zu formulieren, die möglichst alle Belege im Korpus auffinden können, auf Kosten vieler falscher Treffer. Nach und nach verfeinert man den Suchausdruck, um die Anzahl der falschen Treffer zu reduzieren, aber gleichzeitig alle gewünschten Treffer beizubehalten. Am Ende dieses Prozesses steht dann die manuelle Sichtung der verbliebenen Treffer.

5.3.5 Abfolge von verbalen Elementen

Als letztes Beispiel ein Phänomen aus dem verbalen Bereich: die relative Abfolge von Vollverb (V) und Auxiliar (Aux) (vgl. Fleischer 2011, S. 77–78, 163–170). Nach einer Studie von Ebert (1998) an Daten aus dem Fnhd. hängt es u. a. vom Typ der Konstruktion ab, wie oft die moderne Abfolge V > Aux auftritt. Die Konstruktionen (mit ihren V–Aux–Vorkommenshäufigkeiten) sind: *werden*-Passiv (95%), *sein*-Passiv (88%), *haben*-Perfekt (78%), Infinitivkonstruktionen (59%), *sein*-Perfekt (47%) (Ebert 1998, S. 65).

(31a,b) zeigen entsprechende (noch zu verfeinernde) REM-Anfragen: Gesucht wird hier ein finites Auxiliar („VAFIN“) oder Modalverb („VMFIN“) und ein Vollverb (d. h. der Tag-Name beginnt mit „VV“) bzw. Modalverb („VM“), das entweder im Infinitiv (der Tag-Name endet auf „INF“) oder im Partizip Präteritum („PP“) steht.

(31) a.
```
pos = /(VA|VM)FIN/
& pos = /(VV|VM)(INF|PP)/
& #1 . #2
```
(Abfolge Aux > V)

b.
```
pos = /(VA|VM)FIN/
& pos = /(VV|VM)(INF|PP)/
& #2 . #1
```
(Abfolge V > Aux)

Insgesamt gibt es 4352 Treffer (= 58%) für die Anfrage nach „Aux > V“ und 3178 (= 42%) für „V > Aux“, also ein relativ ausgeglichenes Bild. (32a,b) zeigen Beispielsätze für beide Verwendungen. Bei genaueren Untersuchungen müsste natürlich berücksichtigt werden, dass Abfolgen auch reimbedingt sein könnten, vgl. (32a).

(32) a. daz daz kint werde geborn. daz got darzv *hat erchorn*. (M004-1r,2f)
‚dass das Kind werde geboren, das Gott dazu *hat erkoren*‘

b. nune iſt nie man in diner geburte. der ſo *genant ſi*. (M004-1v,2f)
‚Nun_[Neg.Part.] ist niemand in deiner Verwandtschaft, der so *genannt sei*'

Was die Anfragen in (31) allerdings nicht erfassen, ist der Nebensatzstatus. Beispielsweise erhalten wir mit diesen Anfragen auch unerwünschte Treffer wie in (33), in denen die finiten Verbformen in Hauptsätzen in Verbzweit-Position stehen.

(33) a. min roſ *iſt erręhet*. (M001-10r,15)
‚mein Ross *ist steif*'
b. *geſegenet ſiſ* tu herre aller ifrahel[e]. (M004-1v,15f)
‚*gesegnet seist*_du, Herr aller Israeliten'

Wie kann man die Suchanfragen auf Nebensätze einschränken? Eine Möglichkeit ist, im vorausgehenden Kontext eine subordinierende Konjunktion oder ein Relativpronomen zu fordern. Die entsprechende Anfrage wird in (34) gezeigt. Der Operator „.1,10" (in der letzten Bedingung) besagt, dass das subordinierende Element (1. Bedingung) 1–10 Wörter vom finiten Auxiliar oder Modalverb (2. Bedingung) entfernt sein kann. Diese Bedingung soll zur Folge haben, dass das subordinierende Element zum gleichen Teilsatz wie die verbalen Elemente gehört.

(34)
```
pos = /(KOUS|DREL.*)/
& pos = /(VA|VM)FIN/
& pos = /(VV|VM)(INF|PP)/
& #2 . #3
& #1 .1,10 #2
```

Auch hier ist wieder klar, dass es sich um eine heuristische Abfrage handelt, die nicht alle gewünschten Fälle korrekt erfasst. Z. B. kann das subordinierende Element zu einem anderen Teilsatz gehören als die verbalen Elemente.

Die Trefferanzahl wird durch die genannte Einschränkung ungefähr halbiert. Es entfallen noch 1882 (= 55%) auf Aux > V und 1560 (= 45%) auf V > Aux. Schaut man sich, Ebert (1998) folgend, die Typen von Konstruktionen an, so zeigt sich folgendes Bild (35):[35] Im Gegensatz zu Ebert, bei dem das *werden*-Passiv fast ausschließlich die moderne Stellung V > Aux zeigt (zu 95%), steht diese Konstruktion in den REM-Daten nur zu 56% in dieser Abfolge (in Nebensätzen). Ähnliches kann auch für die anderen (vergleichbaren, s. Fußnote 35) Konstruktionen gesagt werden: Die moderne Stellung ist stets noch deutlich seltener als in Eberts Daten aus dem Fnhd.

35 Das *sein*-Perfekt und das *sein*-Passiv lassen sich nicht automatisch unterscheiden und müssen deshalb hier vereinfachend als eine Klasse behandelt werden.

(35)

Konstruktion	Frequenz	
	Aux > V	V > Aux
sein-Perfekt/-Passiv (*sîn/wësen* + VVPP)	239 (41%)	343 (59%)
werden-Passiv (*wërden* + VVPP)	195 (44%)	253 (56%)
haben-Perfekt (*haben* + VVPP)	429 (52%)	392 (48%)
Infinitivkonstruktion (Aux/Modal + VVINF)	1019 (64%)	572 (36%)

5.4 Frequenztabellen

Im Verlauf dieses Abschnittes zu Syntaxabfragen habe ich Ergebnisse verschiedentlich in Form von Frequenz-Tabellen zusammengefasst. Wie erstellt man solche Tabellen?

Eine naheliegende, aber aufwändige Lösung ist, für jeden zu unterscheidenden Fall eine eigene Anfrage zu stellen und sich die jeweilige Zahl der Treffer zu notieren. Beispielsweise würde die Anfrage für zwei adjazente Pronomen in der Abfolge Dativ > Akkusativ wie in (36) lauten.

(36)
```
pos = "PPER"
& inflection = /.*Dat.*/
& pos = "PPER"
& inflection = /.*Akk.*/
& #1 _=_ #2
& #1 . #3
& #3 _=_ #4
```

Interessiert man sich zusätzlich für den Faktor der Person (s. Abschnitt 5.3.3), so potenziert sich allerdings die Anzahl der zu stellenden Anfragen. ANNIS bietet daher für einzelne linguistische Merkmale einfache Frequenzvergleiche an. Unter dem Menüpunkt „Frequency Analysis“ lassen sich die Merkmale auswählen, deren Frequenzen man vergleichen möchte. Für die Anfrage nach zwei beliebigen adjazenten Pronomen (24a) könnte man als Vergleichsmerkmale die jeweiligen morphologischen Analysen („inflection“) auswählen. Abbildung 2 zeigt das resultierende Säulendiagramm. Die häufigste Kombination ist demnach „Masc.Nom.Sg.3 > Masc.Akk.Sg.3“ mit 166 Treffern.

Möchte man nur den Kasus oder Kasus und Person der beiden Pronomen vergleichen, wie in (25) bzw. in (26) gezeigt, so ist eine weitere Verarbeitung der Daten außerhalb von ANNIS notwendig. Dazu kann man die Treffer exportieren und mit einem externen Programm bearbeiten.

Beispielsweise bietet ANNIS verschiedene Exportformate, die eine tabellarische Auflistung aller Treffer und ihrer Eigenschaften produzieren. Daten in

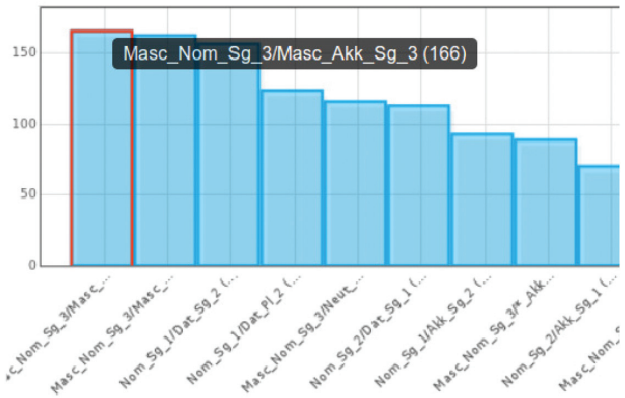

Abbildung 2: Frequenzvergleiche der morphologischen Merkmale zweier adjazenter Pronomen in ANNIS (Ausschnitt).

solchen Formaten können z. B. in ein Programm zur Tabellenkalkulation eingelesen und ausgewertet werden.

5.5 Baumbanken

Wie mehrfach erwähnt, lassen sich einige der in diesem Abschnitt angesprochenen Phänomene bequemer untersuchen, wenn eine reichere syntaktische Annotation vorliegt. Syntaktisch annotierte Korpora heißen *Baumbanken*, d. h. Datenbanken mit Syntaxbäumen. Die syntaktischen Annotationen können aus Konstituentenstrukturen, aus Dependenzrelationen oder auch aus hybriden Strukturen bestehen, die im Folgenden kurz vorgestellt werden.

Ein Beispiel für eine Baumbank mit reinen Konstituentenstrukturen ist das *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al. 2003). Abbildung 3 zeigt einen Beispielbaum aus diesem Korpus.[36] Neben Annotationen für Wortart und Flexionsmorphologie auf der untersten Ebene (z. B. „PRO^N“ für Pronomen im Nominativ) kodiert der Baum Konstituenten mit Kategorie und Funktion bzw. Kasus (z. B. „NP-NOM“ für Nominativ, d. h. Subjekt-NP). Formal unterspezifizierte Wortformen bleiben morphologisch unmarkiert (z. B. „N“ für Nomen allgemein).

36 Entnommen aus http://www-users.york.ac.uk/~lang22/YCOE/doc/annotation/YcoeLite.htm.

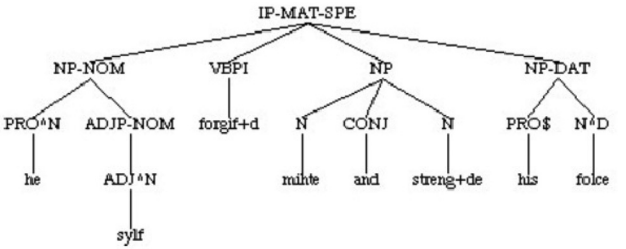

Abbildung 3: Beispielsatz aus dem *York-Toronto-Helsinki Parsed Corpus of Old English Prose*.

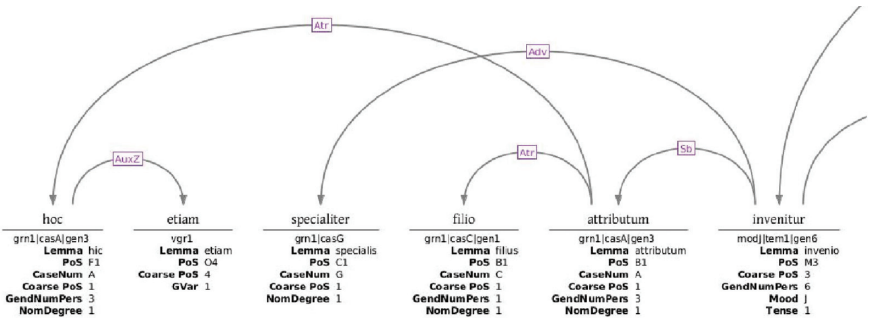

Abbildung 4: Beispielsatz aus der *Index Thomisticus Treebank*.

Abbildung 4 zeigt einen Satz mit Dependenzrelationen. Es handelt sich um den Anfang des ersten Satzes der *Index Thomisticus Treebank* (McGillivray et al. 2009), die syntaktische Analysen für rund 60.000 Wortformen des Index Thomisticus enthält (s. Abschnitt 2).[37] Überkreuzende Abhängigkeiten werden in Dependenzanalysen durch entsprechende überkreuzende Kanten dargestellt. In einer Konstituentenanalyse würden stattdessen z. B. Spuren angenommen.

Für Baumbanken des (modernen) Deutsch wird meist eine hybride Analyse, das *TIGER-Schema* (Albert et al. 2003), verwendet, das einerseits Konstituentenknoten vorsieht, andererseits nicht-lokale Abhängigkeiten durch kreuzende Kanten darstellt. Die beiden verfügbaren historischen Baumbanken des Deutschen, die *Deutsche Diachrone Baumbank* (Hirschmann und Linde 2010) sowie die *Mercurius-Baumbank* (Demske 2007) (s. Abschnitt 3), nutzen ebenfalls dieses Schema.

37 Die Abbildung ist mit Hilfe des Online-Tools *TüNDRA* erzeugt; https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tundra.

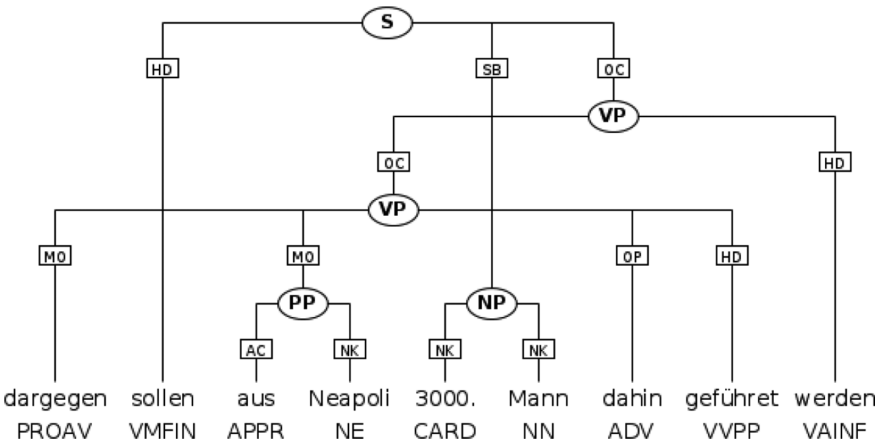

Abbildung 5: Beispielsatz aus der *Mercurius-Baumbank* (‚dagegen sollen aus Neapel 3000 Mann dahin geführt werden').

Abbildung 5[38] zeigt einen Beispielsatz aus der *Mercurius Baumbank*, in dem die Konstituenten *dargegen*, *aus Neapoli* und *dahin* vom eingebetteten Verb *geführet* abhängen und nicht von den übergeordneten Auxiliar- bzw. Modalverben *werden* und *sollen*. Dadurch ergeben sich kreuzende Kanten.

Die Anfragen im Abschnitt 5.3 nutzen reguläre Ausdrücke über Wortart-Annotationen. Wie wir gesehen haben, eignen sich diese „flachen" Annotationen oft nur für heuristische Suchen, bei denen sowohl bestimmte Belege nicht erfasst als auch bestimmte ungewollte Treffer nicht ausgeschlossen werden können. Baumbanken hingegen enthalten „tiefe" Annotationen, so dass Anfragen präziser formuliert werden können, wie im folgenden für die Problemfälle von oben skizziert wird. Als Korpus für die Abfragen wird die Mercurius-Baumbank in ANNIS genutzt. Die im Folgenden gezeigten Abfragen nutzen zusätzliche Operatoren, die speziell für Baumbank-Abfragen vorgesehen sind.

Allerdings zeigt sich auch, dass gerade Adjazenzbedingungen über dem TIGER-Schema oft nicht einfach zu formulieren sind, da bei kreuzenden Kanten eine Konstituente durch „intervenierende" andere Konstituenten unterbrochen wird. Beispielsweise werden beide VP-Konstituenten in Abbildung 5 durch das Modalverb unterbrochen. Das bedeutet, dass beispielsweise die Konstituenten *dargegen* und *aus Neapoli* zwar benachbarte Schwesterkonstituenten sind

38 Die Abbildung ist mit Hilfe des Tools *TIGER-Search* erzeugt; http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.html.

(beides sind direkte, adjazente Töchter des unteren VP-Knotens), gleichzeitig aber auf der Terminalebene durch das Modalverb getrennt sind. Das erschwert die Abfrage nach adjazenten Nichtterminal-Knoten, wie beispielsweise benachbarte Nominalphrasen.

5.5.1 Prä- vs. postnominale Genitive

(37a,b) zeigt die Genitiv-Abfragen für das Mercurius-Korpus in der ANNIS-Demoversion.[39] „node2“ steht für eine beliebige Konstituente oder Wortform und „cat="NP"“ für Nominalphrasen. „>[label="GL"]“ bedeutet, dass eine Dominanzrelation besteht und die untere Konstituente als „GL“ („Genitiv-Attribut links“), d. h. als pränominaler Genitiv fungiert. „GR“ markiert entsprechend postnominale Genitive. In der Mercurius-Baumbank gibt es 899 Treffer für prä- und 1185 für postnominale Genitive; hier überwiegt also die jüngere Konstruktion. Beispiele dafür stehen in (37c,d).

(37) a. `node >[ label="GL"] cat= "NP"`
b. `node >[ label="GR"] cat= "NP"`
c. zu [NP [NP$_{gen}$ *des Hofes*] grosser Freude]
‚zu *des Hofes* großer Freude‘
d. [NP mächtige Summen [NP$_{gen}$ *baares Geldes*]]
‚mächtige Summen *baren Geldes*‘

5.5.2 Abfolge von Pronomen und Nominalphrasen

Die Anfrage in (38a) sucht nach zwei benachbarten Personalpronomen. Davon gibt es in der Baumbank insgesamt 70. In der Mehrzahl der Fälle (57) steht dabei ein Pronomen in der Funktion eines Subjekts an erster Stelle. In den restlichen Fällen steht das Pronomen *es* an erster Stelle, oft in expletiver Funktion wie in (38b).[40]

(38) a. `pos = "PPER" . pos = "PPER"`
b. wie *es jhnen* aber vor Pöldten ergangen
‚wie *es ihnen* aber vor Pölten ergangen (ist)‘

39 https://korpling.german.hu-berlin.de/annis3/#_c=TWVyY3VyaXVz

40 Morphologische Information ist nicht abfragbar (s. Fussnote 17), so dass hier kein direkter Vergleich mit den Daten aus Abschnitt 5.3.3 möglich ist.

Die Anfrage nach benachbarten Nominalphrasen nutzt den Schwester–Operator „$“. Allerdings müsste man hier noch zusätzlich ausschließen, dass weitere Schwesterkonstituenten intervenieren zwischen den beiden Nominalphrasen, d. h. selbst auf einer Baumbank ist diese Anfrage nicht richtig elegant formulierbar (jedenfalls nicht mit den gegebenen Such-Operatoren und dem TIGER-Schema).[41]

(39) `cat = "NP" $ cat = "NP"`

5.5.3 Abfolge von verbalen Elementen

Auch die Anfrage nach der relativen Abfolge von Verb und Auxiliar ist nicht ganz einfach zu formulieren, s. (40a) für die Anfrage nach Auxiliar > Verb (für die umgekehrte Abfolge muss die letzte Zeile entsprechend modifiziert werden). Hierfür ist u. a. das TIGER-Schema verantwortlich, das Verb und Auxiliar nicht als eine Konstituente analysiert, sondern hierarchisch deutlich verschieden behandelt. Für die Abfolge Auxiliar > Verb gibt es 266 Treffer, s. (40b), die meisten davon allerdings innerhalb von 3-Verb-Clustern wie in (40c). Für die Abfolge Verb > Auxiliar gibt es dagegen 1285 Treffer, s. (40d). Im Vergleich zu den REM-Daten steht das Mercurius-Korpus dem modernen Deutsch also bereits sehr nahe.

(40) a.
```
cat = "S"
& pos = /(VA|VM)FIN/
& pos = /KOUS|PREL.*/
& pos = /(VV|VM)(INF|PP)/
& #1 >[label="HD"] #2
& #1 > #3
& #1 >* #4
& #2 . #4
```

b. daß beyde sich forthin *sollen enthalten* von aller Plünderung
‚dass beide sich künftig *enthalten sollen* von aller Plünderung‘

c. daß Ordre *möchte gestellet* werden / seine Verwundete abzuholen
‚dass der Befehl *gegeben* werden *sollte*, seine Verwundeten abzuholen‘

d. daß er sich nach Deutschlandt *begeben werde*
‚dass er sich nach Deutschland *begeben werde*‘

41 Auch auf einer reinen Dependenz-Baumbank sind solche Anfragen typischerweise nicht einfacher zu formulieren. Der Grund dafür ist, dass in Dependenz-Strukturen Adjazenzbedingungen auch immer nur mit Bezug auf die Wortebene ausgewertet werden können – anders als in „reinen“ Baumbanken, die keine kreuzenden Kanten erlauben.

6 Umgang mit problematischen Analysen

Im Anschluss an die Anwendungsszenarien folgen nun einige Überlegungen dazu, wie man als Nutzer eines Korpus damit umgehen sollte, dass die Annotationen gelegentlich unintuitive Analysen wiedergeben können (wie im Beispiel der Komposita, die nicht gemäß HiTS annotiert sind) oder sogar fehlerhaft sein können (z. B. wenn ein Annotator sich bei der Interpretation einer Textstelle geirrt hat).

Beginnen wir mit den unintuitiven Analysen. Ist klar dokumentiert, wie die entsprechende Konstruktion annotiert ist, so ist sie häufig trotz ihrer Analyse gut auffindbar. Beispielsweise kann man alle HiTS-Komposita auffinden, indem man nach Genitivkonstruktionen mit den entsprechenden Kongruenzmerkmalen sucht. In solchen Fällen stellen unintuive Analysen also kein echtes Hindernis für die Nutzung des Korpus dar. Allerdings muss der Nutzer sich dafür einigermaßen intensiv mit der Dokumentation auseinandersetzen.

Zum Fall der fehlerhaften Annotationen ist zunächst festzuhalten, dass es *das* fehlerfreie Korpus nie geben wird. Es werden immer Transkriptions- und/oder Annotationsfehler im Korpus enthalten sein. Dasselbe kann aber auch von der traditionellen manuellen Auswertung gesagt werden: Auch hier ist zu erwarten, dass bei der manuellen Auswertung größerer Datenmengen einzelne Belege quasi zwangsläufig übersehen werden.

Ist das untersuchte Phänomen einigermaßen frequent und interessiert man sich hauptsächlich für die größeren Trends, so stellen einzelne Fehlannotationen auch kein allzu großes Problem dar, sondern gehen im Rauschen unter. Kommt es allerdings auf jeden einzelnen Beleg an, z. B. weil es sich um ein seltenes Phänomen handelt, so bietet es sich an, auf verschiedene Arten im Korpus danach zu suchen und die Ergebnisse der Suchen miteinander zu vergleichen. Beispielsweise kann man nach Genitiv-Komplementen suchen, indem man entweder Abfragen nach Nominalphrasen mit morphologischer Markierung als Genitiv stellt oder nach Nominalphrasen mit typischen Genitiv-Endungen oder nach Verben, von denen man weiß, dass sie Genitiv-Komplemente regieren können.

Allgemein bietet es sich dabei an, mit eher generellen Anfragen zu beginnen, die schrittweise verfeinert und eingeschränkt werden.

7 Zusammenfassung und Ausblick

Ziel dieses Artikel war es, die Nutzung annotierter Korpora für die historische Syntaxforschung zu illustrieren. Das geschah vorwiegend am Beispiel des *Referenzkorpus Mittelhochdeutsch*, REM. REM ist annotiert mit Lemmata,

Morphologie und Wortart-Information gemäß dem Tagset HiTS. Es wurde anhand ausgewählter syntaktischer Phänomene illustriert, wie diese Annotationen für die gezielte Suche im Korpus genutzt werden können und wie die entsprechenden Anfragen an REM im Korpussuchtool ANNIS zu formulieren sind.

Bei komplexeren Phänomenen kann mit Hilfe der Anfragen oft nur eine Vorauswahl getroffen werden, die anschließend manuell gesichtet werden muss. Ohne ein annotiertes Korpus wären solche Phänomene allerdings überhaupt nur dann auffindbar (und damit systematisch untersuchbar), wenn die Texte komplett manuell gesichtet würden. Die Vorauswahl stellt also einen wichtigen Schritt dar.

Syntaktisch annotierte Korpora, sogenannte Baumbanken, erlauben oft einen einfacheren und gezielteren Zugriff auf syntaktische Phänomene. Beispielsweise ließen sich die Anfragen für prä- vs. postnominale Genitive oder adjazente Pronomen sehr einfach formulieren. Abhängig vom gewählten Annotationsschema und dem verwendeten Suchtool können die Abfragen aber auch hier umfangreicher werden und gegebenenfalls auch nur eine heuristische Annäherung sein, so bei adjazenten Nominalphrasen oder der Abfolge verbaler Elemente.

Wie deutlich wurde, sind die notwendigen Anfragen sowohl auf REM als auch auf der Mercurius-Baumbank häufig recht komplex. Geeignete Anfragen zu formulieren, erfordert daher einige Übung. D. h. möchte man den Reichtum annotierter Korpora voll ausnutzen, setzt das eine recht gründliche Auseinandersetzung mit den verwendeten Tagsets und mit der Abfragesprache voraus. Der Aufwand der Einarbeitung zahlt sich aber nach kurzer Zeit aus.

Vereinfachtes Such-Interface

Für Erst- und Gelegenheitsnutzer stellen diese Anforderungen aber natürlich ein beachtliches Hindernis dar. Für die Referenzkorpora wurde daher ein besonderes Interface geschaffen, das einfach zu bedienen ist und einen schnellen, intuitiven Zugang zu den Korpora bietet.[42] Konkret kann man mit dem Such-Interface in REM nach Wortformen („tok_mod“) oder nach Lemmata suchen und die Suche auf einen bestimmten Zeitraum oder bestimmte Sprachregionen einschränken. Abbildung 6 zeigt eine Beispielabfrage nach der Wortform *durh*. Die Suche ist eingeschränkt auf westoberdeutsche Texte (d. h. alemannisch, schwäbisch, elsässisch) des 12. Jahrhunderts aus dem Themenbereich Religion.

42 Der erste Prototyp des Interfaces wurde von Tom Ruette, HU Berlin, entwickelt. Für ANNIS existiert außerdem ein Interface zur grafischen Eingabe der Suchanfrage, der sogenannte „Query Builder“, s. http://annis-tools.org/documentation.html.

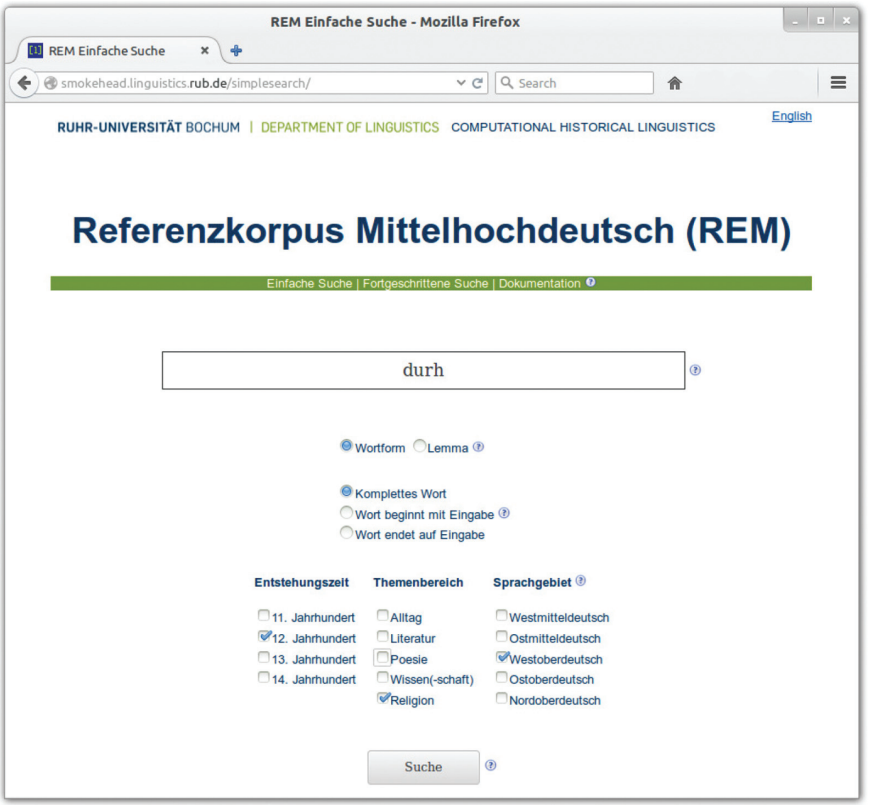

Abbildung 6: Beispielanfrage auf REM im vereinfachten Such-Interface.

Es ist offensichtlich, dass für Suchanfragen nach syntaktischen Phänomenen im Allgemeinen solche vereinfachten Interfaces nicht genügend Ausdrucksmittel zur Verfügung stellen können. Ziel dieses Artikels war es daher v. a. auch, anhand von beispielhaften Anfragen konkret zu zeigen, wie nach interessanten historischen Syntax-Phänomenen in REM gesucht werden kann und welche Fülle an Informationen REM und andere historische Korpora bieten, die für die Untersuchung solcher Phänomene genutzt werden kann.

Vorteile annotierter Korpora

Abschließend möchte ich die Vorteile und Chancen zusammenfassen, die annotierte Korpora für die historische Syntaxforschung bieten.

Die Spirale empirisch-wissenschaftlichen Vorgehens: (*Forschungsliteratur- und) Datensichtung → Aufstellung von Hypothesen → Überprüfung der Hypothesen an weiteren Daten → gegebenfalls Modifikation der Hypothesen → erneute Überprüfung an weiteren Daten etc.* – wird durch den Einsatz annotierter, durchsuchbarer Korpora stark unterstützt und deutlich vereinfacht, und zwar gleich in mehreren Punkten.

(i) Datensichtung

Annotierte Korpora dienen dazu, im optimalen Fall die relevanten Belege im Korpus gezielt aufspüren zu können oder, im schlechteren Fall, die zu sichtende Datenmenge immerhin auf eine handhabbare Größe zu reduzieren.

Liegen die historischen Texte beispielsweise nur auf Papier oder als (digitale) Abbildungen vor, muss manuell durch alle Texte gegangen und nach relevanten Belegen gesucht werden. Die Fundstellen müssen entsprechend markiert und gezählt werden.

Liegen die Texte in digitalen (aber unannotierten) Transkriptionen vor, so kann nach Vollformen und mittels regulärer Ausdrücke auch nach Wortfragmenten gesucht werden, um die zu sichtende Datenmenge zu reduzieren. Ein solches Vorgehen kann z. B. angewendet werden, wenn es um einzelne, bekannte Lexeme geht (z. B. alle Vertreter einer geschlossenen Wortklasse) oder um morphologisch klar markierte Formen, so dass mit regulären Ausdrücken nach den entsprechenden Affixen gesucht werden kann.[43]

Handelt es sich jedoch um ein Phänomen, das Vertreter offener Wortklassen (z. B. Nomen, Verben, Adjektive) und keine morphologisch klar markierten Wortformen involviert, so helfen reine Transkriptionen nicht weiter. Hier benötigt man grammatische Information, die in Form von Annotationen abgefragt werden kann, um aus den Daten die relevanten Belege zu extrahieren.[44]

(ii) Kombination von Merkmalen

Viele linguistisch interessante Phänomene zeichnen sich dadurch aus, dass Eigenschaften mehrerer Ebenen betroffen sind und interagieren, beispielsweise: eine morphologische Form und ihre Realisierung (wie z. B. Vorkommen von Dativ Singular, die nicht auf *-e* enden, (9c)); eine Wortart in einer bestimmten

43 (2) und (5) in Abschnitt 4 zeigen entsprechende Beispielanfragen.

44 Der Großteil der Anfragen in Abschnitt 5 sind von diesem Typ, beispielsweise die Anfragen in (20), (21) oder (27).

grammatischen Funktion und morphologischen Form (z. B. prädikative Adjektive im Nominativ, (20)); Vorkommen mehrerer Wortarten, die in einer bestimmten Abfolge stehen (z. B. Genitivattribute (21), Sequenzen von Pronomen (27) oder Verb-Auxiliar-Abfolgen (34)); und nicht zuletzt Kombinationen grammatischer Merkmale mit soziolinguistischen Merkmalen wie dem Textgenre oder zeit- und sprachräumlichen Merkmalen. Solche komplexen Bedingungen manuell in Texten zu überprüfen, ist zum einen aufwändig, zum andern fehleranfällig.

(iii) Modifizierte Hypothesen

Gemäß dem Spiralmodell müssen in Abhängigkeit von den gefundenen Belegen vorläufige Hypothesen so lange modifiziert werden, bis eine zufrieden stellende Form gefunden ist. Konkret kann das bedeuten, dass die Faktoren (Merkmale), auf die sich die Hypothesen beziehen, modifiziert werden oder neue Faktoren hinzukommen oder auch alte Faktoren wegfallen. Beispielsweise untersucht Sapp (2011) Verbcluster aus drei verbalen Elementen und berücksichtigt dabei u. a. folgende Faktoren (Sapp 2011, S. 76–90): die Kategorie des vorhergehenden Wortes; NP- und PP-Extraposition; Typ der Konstruktion (vgl. (35)); Verbpräfix; Zeit- und Sprachraum; sozialer Status des Schreibers; Textgenre.

Da solche Faktoren sich oft erst sukzessive im Laufe einer detaillierten Untersuchung als interessant und potenziell relevant herausstellen, müssten bei einer rein manuellen Datensichtung sämtliche Daten jeweils wieder komplett von Neuem durchgegangen und auf die jeweiligen Faktoren hin überprüft werden. Liegt hingegen ein Korpus vor, das mit geeigneten Merkmalen annotiert ist, brauchen die neuen Faktoren nur entsprechend in die Suchanfrage integriert zu werden und die veränderte Anfrage kann nochmals über das Korpus laufen gelassen werden. Danach kann verglichen werden, welche Belege bei der ursprünglichen und bei der aktuellen Anfrage gefunden wurden.

(iv) Gegenprobe

Untersucht man die Eigenschaften einer Konstruktion, so muss man darauf achten, die Gesamtheit der relevanten Fälle zu berücksichtigen. Untersucht man beispielsweise die Abfolge von Verb und Auxiliar, so ist das prozentuale Verhältnis der beiden Abfolgen zueinander relevant (s. Abschnitt 5.3.5). Untersucht man flektierte prädikative Adjektive im Nominativ, so muss man sich auch das „Gegenstück“ dazu, nämlich unflektierte prädikative Adjektive im Nominativ anschauen (s. Abschnitt 5.3.1).

Je nach Fragestellung und Art der Annotation kann die Gegenprobe durch eine leichte Abwandlung der Suchanfrage durchgeführt werden, z. B. durch Negation einer der Bedingungen oder, im Fall einer Abfolge, durch Vertauschen der Argumente der Präzedenzbedingung (`#1 . #2 wird zu #2 . #1`). Allerdings gibt es Fälle, in denen sich die Gegenprobe nicht so einfach durchführen lässt: falls die notwendige Information nicht geeignet annotiert ist (wie im Fall der prädikativen Adjektive).

(v) Datenbasis

Wie beschrieben, schaffen die Annotationen die Voraussetzungen dafür, gezielt und effizient nach relevanten Daten zu suchen. Mit den neu geschaffenen Ressourcen der Referenzkorpora eröffnen sich damit ganz neue Möglichkeiten für die historische Syntaxforschung: Zum einen können nun Untersuchungen auf einer breiten, ausgewogenen Datenbasis durchgeführt werden. Zum andern sind exhaustive Überprüfungen der Treffer mit erheblich weniger Aufwand verbunden. Insgesamt ergibt sich dadurch eine solidere Datengrundlage für die Forschung als bei einer rein manuellen Auswertung der Daten.

(vi) Reproduzierbarkeit

Nicht zuletzt sind Ergebnisse, die aus (frei verfügbaren) Korpora gewonnen werden, reproduzierbar. Quantitative Resultate lassen sich somit auch „belegen“ und nachvollziehen, insbesondere, wenn die verwendeten Suchanfragen mit angegeben werden (wie im vorliegenden Artikel). Die Wege und potenziellen Irrwege der Forschung werden transparenter, und nachfolgende Untersuchungen können auf Ergebnissen früherer Forschung aufbauen und sie z. B. korrigieren oder verfeinern, indem zusätzliche Faktoren berücksichtigt werden.

Als ein Fazit kann festgehalten werden, dass annotierte Korpora die systematische Untersuchung vieler Phänomene wesentlich vereinfachen, indem effizient nach Belegen gesucht werden kann. Entsprechende manuelle Untersuchungen wären mit großem Aufwand verbunden. Bestimmte Phänomene sind überhaupt erst durch den Einsatz annotierter Korpora systematisch untersuchbar. Mit Ressourcen wie den Referenzkorpora stehen der historischen Syntaxforschung daher neue Wege und Phänomenbereiche offen.

Appendix I: Quellen des Referenzkorpus Mittelhochdeutsch

Die nachfolgende Liste wurde der Übersicht unter http://www.ruhr-uni-bochum.de/wegera/rem/Uebersicht.htm entnommen.

Sigle	Titel	Aufbewahrungsort	Signatur
M001	Ad equum errehet	Paris, Bibl. Nationale	Ms. nouv. acq. lat. 229
M003	Ad restringendum sanguinem [=Blutsegen, Erfurter]	Erfurt, Universitätsbibl.	Cod. Ampl. 8° 62b
M004	Adelbrecht: Johannes Baptista	St. Paul im Lavanttal, Stiftsbibl.	Fragm. 25/8
M009	Lambrechts Alexander (Vorauer Alexander)	Vorau, Stiftsbibl.	Cod. 276
M013O	Annolied (O: Opitz)	Breslau / Wroclaw (?)	[(?) nicht mehr nachweisbar]
M024	Ava: Leben Jesu	Vorau, Stiftsbibl.	Cod. 276
M028	Balaam, Vorauer (Bücher Mosis 5)	Vorau, Stiftsbibl.	Cod. 276
M045	Christi Geburt, Von	Innsbruck, Universitäts- und Landesbibl.	Fragm. 69
M055	Crescentia	Colmar, Archives Départementales du Haut-Rhin	Fragments de Ms. no. 559
M088	Wiener Genesis	Wien, Österr. Nationalbibl.	Cod. 2721
M107S	Heinrich: ‚Litanei' (S)	Straßburg, Seminarbibl.	Cod. C. V. 16.6. 4° [verbrannt]
M116	Vorauer Joseph (Bücher Mosis 2)	Vorau, Stiftsbibl.	Cod. 276
M119	Judith, Die Jüngere	Vorau, Stiftsbibl.	Cod. 276
M121V	Kaiserchronik A (V) [Ausschnitt]	Vorau, Stiftsbibl.	Cod. 276
M155	Physiologus (Älterer/Ahd. Physiologus)	Wien, Österr. Nationalbibl.	Cod. 223
M239	Wernher v. Niederrhein: Die vier schiven	Hannover, Landesbibl.	Ms. I 81 (Teil II)
M402	Berliner Evangelistar	Berlin, Staatsbibl.	mgq 533

Appendix II: HiTS: Historisches Tagset

Die nachfolgende Liste wurde Dipper et al. (2013) entnommen.

Tag	Beschreibung
ADJA	*Adjektiv, attributiv, vorangestellt*
ADJD	*Adjektiv, prädikativ*
ADJN	*Adjektiv, attributiv, nachgestellt*
ADJS	*Adjektiv, substituierend*
APPO	*Postposition*
APPR	*Präposition*
AVD	*Adverb*
AVG	*Relativadverb, generalisierend*
AVNEG	*Adverb, negativ*
AVW	*Adverb, interrogativ*
CARDA	*Kardinalzahl, attributiv, vorangestellt*
CARDD	*Kardinalzahl, prädikativ*
CARDN	*Kardinalzahl, attributiv, nachgestellt*
CARDS	*Kardinalzahl, substituierend*
DDA	*Determinativ, definit, attributiv, vorangestellt* (ad.)
DDART	*Determinativ, definit, artikelartig* (mhd.)
DDD	*Determinativ, definit/demonstrativ, prädikativ*
DDN	*Determinativ, definit/demonstrativ, attributiv, nachgestellt*
DDS	*Determinativ, definit/demonstrativ, substituierend*
DGA	*Determinativ, generalisierend, attributiv, vorangestellt*
DGD	*Determinativ, generalisierend, prädikativ*
DGN	*Determinativ, generalisierend, attributiv, nachgestellt*
DGS	*Determinativ, generalisierend, substituierend*
DIA	*Determinativ, indefinit, attributiv, vorangestellt* (ad.)
DIART	*Determinativ, indefinit, artikelartig* (mhd.)
DID	*Determinativ, indefinit, prädikativ*
DIN	*Determinativ, indefinit, attributiv, nachgestellt*
DIS	*Determinativ, indefinit, substituierend*
DNEGA	*Determinativ, negativ, attributiv, vorangestellt*
DNEGD	*Determinativ, negativ, prädikativ*
DNEGN	*Determinativ, negativ, attributiv, nachgestellt*
DNEGS	*Determinativ, negativ, substituierend*
DPOSA	*Determinativ, possessiv, attributiv, vorangestellt*
DPOSD	*Determinativ, possessiv, prädikativ*
DPOSGEN	*Determinativ, personal-possessiv, Genitiv*
DPOSN	*Determinativ, possessiv, attributiv, nachgestellt*

(*continued*)

Tag	Beschreibung
DPOSS	*Determinativ, possessiv, substituierend*
DRELS	*Determinativ, relativisch, substituierend*
DWA	*Determinativ, interrogativ, attributiv, vorangestellt*
DWD	*Determinativ, interrogativ, prädikativ*
DWN	*Determinativ, interrogativ, attributiv, nachgestellt*
DWS	*Determinativ, interrogativ, substituierend*
FM	*Fremdsprachliches Material*
ITJ	*Interjektion*
KO*	*Konjunktion, neben- oder unterordnend*
KOKOM	*Vergleichspartikel*
KON	*Konjunktion, nebenordnend*
KOUS	*Konjunktion, unterordnend*
NA	*Nomen appelativum*
NE	*Eigenname*
PAVAP	*Pronominaladverb, präpositionaler Teil*
PAVD	*Pronominaladverb, pronominaler Teil*
PAVG	*Pronominaladverb, pronominaler Teil, generalisierend*
PAVREL	*Pronominaladverb, pronominaler Teil, relativisch*
PAVW	*Pronominaladverb, pronominaler Teil, interrogativ*
PG	*Pronomen, generalisierend*
PI	*Pronomen, indefinit*
PNEG	*Pronomen, indefinit, negativ*
PPER	*Pronomen, personal, irreflexiv*
PRF	*Pronomen, personal, reflexiv*
PTKA	*Partikel bei Adjektiv oder Adverb*
PTKANT	*Antwortpartikel*
PTKINT	*Fragepartikel* (ad.)
PTKNEG	*Negationspartikel*
PTKREL	*Relativpartikel*
PTKVZ	*Verbzusatz*
PW	*Pronomen, interrogativ*
VAFIN	*Auxiliar, finit*
VAIMP	*Auxiliar, Imperativ*
VAINF	*Auxiliar, Infinitiv*
VAPP	*Auxiliar, Partizip Präteritum, im Verbalkomplex*
VAPS	*Auxiliar, Partizip Präsens, im Verbalkomplex*
VMFIN	*Modalverb, finit*
VMIMP	*Modalverb, Imperativ*
VMINF	*Modalverb, Infinitiv*
VMPP	*Modalverb, Partizip Präteritum, im Verbalkomplex*
VMPS	*Modalverb, Partizip Präsens, im Verbalkomplex*

(*continued*)

Tag	Beschreibung
VVFIN	*Vollverb, finit*
VVIMP	*Vollverb, Imperativ*
VVINF	*Vollverb, Infinitiv*
VVPP	*Vollverb, Partizip Präteritum, im Verbalkomplex*
VVPS	*Vollverb, Partizip Präsens, im Verbalkomplex*
$_	*originale Interpunktion*
$(	*sonstige Satzzeichen*

Appendix III: Morphologische Tags

Bereich	Tag	Beschreibung
Kasus	Nom	*Nominativ*
	Gen	*Genitiv*
	Dat	*Dativ*
	Akk	*Akkusativ*
Numerus	Sg	*Singular*
	Pl	*Plural*
Genus	Masc	*Maskulin*
	Fem	*Feminin*
	Neut	*Neutrum*
Person	1	*erste Person*
	2	*zweite Person*
	3	*dritte Person*
bei Adjektiven	Pos	*Positiv*
	Comp	*Komparativ*
	Sup	*Superlativ*
bei Verben	Ind	*Indikativ*
	Subj	*Konjunktiv*
	Pres	*Präsens*
	Past	*Präteritum*
bei Adjektiven, Artikelwörtern und Pronomen		
	St	*stark*
	wk	*schwach*
Sonstiges	0	*nicht zutreffend bzw. Nullendung*
	*	*unbestimmbar/unterspezifiziert*

Literatur

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pußel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith und Hans Uszkoreit. TIGER Annotationsschema. Technical report, Universität des Saarlandes, Universität Stuttgart, Universität Potsdam, http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/TIGERCorpus/annotation/tiger_scheme-syntax.pdf, 2003.

Birgit Barufke. *Attributsstrukturen des Mittelhochdeutschen im diachronen Vergleich*. Nummer 12 in Beiträge zur germanistischen Sprachwissenschaft. Buske, Hamburg, 1995.

Roberto Busa. *Index Thomisticus*. Frommann-Holzboog, Stuttgart–Bad Cannstatt, 1974–1980.

Roberto Busa. The annals of humanities computing: The Index Thomisticus. *Computers and the Humanities*, 14(2):83–90, 1980.

Bernard Comrie, Martin Haspelmath, und Balthasar Bickel. Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses. Technical report, Max Planck Institute for Evolutionary Anthropology, 2008. Revised version; https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Ulrike Demske. Das Mercurius-Projekt: eine Baumbank für das Frühneuhochdeutsche. In Gisela Zifonun und Werner Kallmeyer, Hgg., *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*, Jahrbuch des Instituts für deutsche Sprache 2006, S. 91–104. De Gruyter, Berlin, 2007.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller und Klaus-Peter Wegera. HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28(1):85–137, 2013. Special Issue.

Robert Peter Ebert. *Verbstellungswandel bei Jugendlichen, Frauen und Männern im 16. Jahrhundert*. Nummer 190 in Germanistische Linguistik. Niemeyer, Tübingen, 1998.

Robert Peter Ebert. *Historische Syntax des Deutschen II: 1300–1750*. Nummer 6 in Germanistische Lehrbuchsammlung. Weidler, Berlin, 2. Auflage, 1999.

Jürg Fleischer. Das prädikative Adjektiv und Partizip im Althochdeutschen und Altniederdeutschen. *Sprachwissenschaft*, 32(3):279–348, 2007.

Jürg Fleischer. Zur Abfolge akkusativischer und dativischer Personalpronomen im Prosalancelot (Lancelot I). In Arne Ziegler [unter Mitarbeit von Christian Braun], Hg., *Historische Textgrammatik und historische Syntax des Deutschen: Traditionen, Innovationen, Perspektiven*, S. 511–536. De Gruyter, Berlin/New York, 2010. 2 Bände.

Jürg Fleischer. *Historische Syntax des Deutschen. Eine Einführung*. Narr, Tübingen, 2011. Unter Mitarbeit von Oliver Schallert.

Susanne Haaf und Matthias Schulz. Historical newspapers & journals for the DTA. In *Proceedings of the LREC Workshop on Language Resources and Technologies for Processing and Linking Historical Documents and Archives—Deploying Linked Open Data in Cultural Heritage (LRT4HDA)*, S. 50–54, 2014.

Mathilde Hennig. The Kassel corpus of clause linking. In Paul Bennett, Martin Durrell und Silke Scheible, Richard J. Whitt, Hgg., *New Methods in Historical Corpus Linguistics*, Band 3 in

Corpus Linguistics an Interdisciplinary Perspectives on Language – CLIP. Narr, Tübingen, 2013.

Hagen Hirschmann und Sonja Linde. Annotationsguidelines zur Deutschen Diachronen Baumbank. Technical Report; Humboldt-Universität zu Berlin, 2010. http://korpling.german.hu-berlin.de/ddb-doku.

Emil Kroymann, Sebastian Thiebes, Anke Lüdeling und Ulf Leser. Eine vergleichende Analyse von historischen und diachronen digitalen Korpora. Technical Report 174; Institut für Informatik, Humboldt-Universität, Berlin, 2004. http://www.deutschdiachrondigital.de/publikationen/TRHistorischeKorpora.pdf.

Matthias Lexer. *Mittelhochdeutsches Handwörterbuch*. Leipzig, 1872. 3 Volumes 1872–1878. Reprint: Hirzel, Stuttgart 1992.

Barbara McGillivray, Marco Passarotti und Paolo Ruffolo. The Index Thomisticus Treebank Project: Annotation, parsing and valency lexicon. *Traitement Automatique des Langues*, 50(2):103–127, 2009.

Heinz-Peter Prell. Die Stellung des attributiven Genitivs im Mittelhochdeutschen. Zur Notwendigkeit einer Syntax mittelhochdeutscher Prosa. *Beiträge zur Geschichte der Deutschen Sprache und* Literatur, 122(1):23–29, 2000.

Daniela Prutscher und Henry Seidel. Mehrebenenannotation frühneuzeitlicher Fürstinnenkorrespondenzen – ein Arbeitsbericht. In Gisela Brandt, Hg., *Bausteine zu einer Geschichte weiblichen Sprachgebrauchs X: Texte – Zeugnisse des produktiven Sprachhandelns von Frauen in privaten, halböffentlichen und öffentlichen Diskursen vom Mittelalter bis in die Gegenwart*, Band 457 in *Stuttgarter Arbeiten zur Germanistik*, S. 109–124. Heinz, Stuttgart, 2012.

Christopher D. Sapp. *The verbal complex in subordinate clauses from medieval to modern German*, Band 173 in *Linguistik Aktuell/Linguistics Today*. Benjamins, Amsterdam/Philadelphia, 2011.

Silke Scheible, Richard Jason Whitt, Martin Durrell und Paul Bennett. A gold standard corpus of Early Modern German. In *Proceedings of the ACL-HLT 2011 Linguistic Annotation Workshop (LAW V)*, S. 124–128, Portland, Oregon, 2011.

Anne Schiller, Simone Teufel, Christine Stöckert und Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset), 1999. Technical report, Universitäten Stuttgart und Tübingen, http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf.

Hans-Christian Schmitz und Bernhard Schröder Klaus-Peter Wegera. Das Bonner Frühneuhochdeutsch-Korpus und das Referenzkorpus ‚Frühneuhochdeutsch'. In Ingelore Hafemann, Hg., *Perspektiven einer corpus-basierten historischen Linguistik und Philologie. Internationale Tagung des Akademienvorhabens „Altägyptisches Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften, 12.–13. Dezember 2011*, S. 205–219, Berlin, 2013.

Margarete Springeth. Der analytische Weg ist das Ziel: Die mittelhochdeutsche Begriffsdatenbank als Online-Textarchiv. In Wernfried Hofmeister und Andrea Hofmeister-Winter, Hgg., *Wege zum Text. Überlegungen zur Verfügbarkeit mediävistischer Editionen im 21. Jahrhundert. Grazer Kolloquium 17.–19. September 2008*, Nummer 30 in Beihefte zu editio, S. 185–202. Narr, Tübingen, 2009.

Ann Taylor, Anthony Warner, Susan Pintzuk und Frank Beths. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). Department of Language and Linguistic Science, University of York, 2003. http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.

Klaus-Peter Wegera. vnd machet sie mit gesehenden augen blind. Zum Problem von Editionen als Datenquelle für sprachhistorische Untersuchungen. *Zeitschrift für deutsche Philologie*, 134(1):77–95, 2015.

Amir Zeldes, Julia Ritz, Anke Lüdeling und Christian Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*, Liverpool, UK, 2009.