Manuel Burghardt & Sandra Reimann

Möglichkeiten der elektronischen Aufbereitung und Nutzung eines historisch syntaktischen Verbwörterbuchs des Deutschen

1. Einleitung

Mit den folgenden Ausführungen soll ein Überblick zu den Möglichkeiten der Nutzung eines elektronisch aufbereiteten, historisch syntaktischen Verbwörterbuchs des Deutschen gegeben und dabei deutlich gemacht werden, welche technischen Anforderungen sich für die Erstellung der Inhalte hinsichtlich des Dokument-Markups ergeben.

Ziel des internationalen Projekts „Historisch syntaktisches Verbwörterbuch des Deutschen“ (HSVW, https://histvw.wordpress.com/), für das derzeit Vorarbeiten laufen, ist es, die Valenzgeschichte von Verben zu ermitteln. Als Hintergrund ist festzuhalten: Verben zeichnen sich – abgesehen von ihren morphologischen und semantischen Merkmalen – besonders auch durch ihre Eigenschaft aus, bei der Satzbildung quantitativ und qualitativ definierte „Leerstellen“ für Satzglieder zu „eröffnen“ (Valenz), die durch entsprechende Formulierungen besetzt werden können/müssen. Wie diverse Valenzwörterbücher zeigen, gehört die Angabe des Valenzpotentials eines Verbs auch zu seiner vollständigen Beschreibung im Wörterbuch (vgl. z. B. Greule 1999; Helbig/Schenkel 1983; Schumacher u. a. 2004; Sommerfeldt/Schreiber 1996). Wir befinden uns somit an der Schnittstelle zwischen Grammatik (Syntax) und Lexikon. Die Ausweitung der Valenztheorie auf ältere Sprachstufen des Deutschen und die Ausarbeitung entsprechender Valenzwörterbücher zeigen bereits, welcher Gewinn nicht nur für die historische Grammatik, sondern auch für die Interpretation der Texte aus den herangezogenen Quellen selbst durch die Ermittlung der Verbvalenz auf den historischen Sprachstufen erzielt werden kann. Der Erkenntnisgewinn wird durch den diachronen Aspekt noch erweitert, das heißt, durch die Aufarbeitung der Valenzgeschichte der Verben vom Beginn ihres Auftretens bis zu ihrer syntaktischen, semantischen und stilistischen Verwendung in der heutigen Kommunikation. So können über die Valenzgeschichten von Verben nicht nur Aufschlüsse über den semantischen und syntaktischen Wandel eines Verbs herausgearbeitet, sondern daraus z. B. auch Erklärungen für das Verschwinden eines Verbs aus der aktuellen Kommunikation abgeleitet werden.

Die Projektbeteiligten (derzeit mit den Universitäten Graz, Heidelberg, Helsinki, Leipzig, Regensburg, Turku, Zürich) machen es sich zur Aufgabe, aus historischen Quellen die Valenzgeschichte der Verben des Deutschen zu erarbeiten und in Lexikonartikeln zusammenzufassen. Bezugsgröße des Vergleichs wird also das Signifikat/Semem des Verbs sein (vgl. Greule 2014, 57f.). Mehrere Sememe zu einem Ausdruck/Signifikant sind in eigenen Artikeln zu behandeln. Zugrunde gelegt werden außerdem Wortfelder, wie z. B. Verben der Emotion. Als Quellen dienen historische Wörterbücher und andere (digitale) Repositorien des Deutschen. Der Vorteil einer quellen-basierten Herangehensweise jenseits der Wörterbücher liegt gewissermaßen in der Bereitstellung des größeren Kontextes, hingegen liefern die Wörterbü-

cher bereits Zuordnungen zu Sememen, die allerdings nicht ungeprüft übernommen werden sollten (vgl. Rapp/Reimann 2016).

Für die Bearbeitung der auf den drei historischen Sprachstufen (alt-, mittel-, frühneuhochdeutsch) pro Wortfeld vorhandenen Verben ist je ein Team zuständig. Schließlich wird eine zusammenfassende Valenzgeschichte für jedes Verb verfasst – zuständig ist ein viertes Team –, und zwar durch Vergleich der Analyseergebnisse auf den Sprachstufen mit Bezug auf das Neuhochdeutsche. Ziel ist somit eine diachrone Sprachgeschichte von Verben ausgewählter Wortfelder.

Geplant ist zudem die elektronische Repräsentation des Lexikons mit einer Abfrageschnittstelle für die bearbeiteten Verben des HSVW. Die Struktur der Artikel wird von Beginn an auf eine mögliche quantitative Suche ausgerichtet. Das bedeutet auch, dass bereits bei der anfänglichen Konzeption des Projekts alle Entscheidungen im Hinblick auf die elektronische Aufbereitung getroffen werden sollten, da ansonsten später ein erheblicher Mehraufwand entstehen würde. Die Semantik wird – das sei vorab beispielhaft angesprochen – eine große Rolle spielen, wie auch schon an der Auswahl der Verben – nach Wortfeldern – ersichtlich ist. Auch der diachrone Aspekt wird, wie angesprochen, berücksichtigt und soll abgefragt werden können (z. B. Valenzänderungen). Klar ist, dass dieses Ziel exakte Absprachen nicht nur beispielsweise der Terminologie, sondern auch der Verschriftung erfordert (z. B. Langform oder Kurzwort bei Ergänzungen).

Die folgenden Ausführungen sollen somit als Vorschlag für die auf eine elektronische Version ausgerichtete Struktur der Valenz-Artikel verstanden werden. In Kapitel 2 des Beitrags werden zunächst das Potenzial und die Möglichkeiten der computergestützten Textanalyse und Annotation vorgestellt. Kapitel 3 gibt einen Überblick zu den technischen Grundlagen des Dokumentmarkups. Im vierten Kapitel werden verwandte Arbeiten im Bereich digitaler Lexikographie zusammengefasst und das Desiderat eines elektronisch verfügbaren, historisch syntaktischen Verbwörterbuchs des Deutschen herausgearbeitet. Im fünften Kapitel wird der methodische Prozess bei der Erstellung historischer Valenzartikel vorgestellt. Darüber hinaus werden systematisch die wesentlichen Informationselemente eines solchen Artikels an einem Beispiel erläutert. Kapitel 6 schließlich beschreibt die Konzeption einer Dokumentgrammatik mit XML Schema für die genannten Informationselemente. Zudem wird für ein beispielhaftes Verb ein XML-Dokument erstellt, welches valide hinsichtlich der im XML Schema definierten Markup-Regeln ist. Im siebten Kapitel werden abschließend Vorteile und Nutzen eines derart elektronisch aufbereiteten Wörterbuchartikels dargestellt, insbesondere die Möglichkeit der computergestützten Suche nach auffälligen sprachlichen Mustern.

2. Computergestützte Textanalyse und Annotation

Vor dem Hintergrund einer zunehmenden Verfügbarkeit maschinenlesbarer Textdaten erfreuen sich computerbasierte Verfahren der Textanalyse in unterschiedlichen wissenschaftlichen Disziplinen großer Beliebtheit[1]. Dabei können Computerprogramme unmittelbar auf große

[1] Vgl. etwa die Verwendung elektronischer Korpora in der Sprachwissenschaft oder distant reading-Ansätze (Moretti 2007) in der Literaturwissenschaft.

Mengen Text angewandt werden, um etwa bestimmte Schlüsselwörter zu identifizieren (Volltextsuche) oder aber um Konkordanzlisten und Kollokationsgraphen zu erstellen[2]. Auf der Zeichenebene ist es mit computergestützten Verfahren ohne weitere Vorverarbeitung eines Texts sofort möglich, bestimmte Wörter oder Zeichenfolgen zu identifizieren. So liefert eine Suchanfrage nach dem Wort *Hund* auf einer beliebigen Websuchmaschine unzählige Treffer von Seiten, in denen die Zeichenfolge „h-u-n-d" enthalten ist. Dabei werden meist grundlegende Varianten mitberücksichtigt, etwa Pluralformen (*die Hunde*) oder Deklinationen (*den Hunden*) sowie auch Komposita (*Schäferhund*). Allerdings würden Seiten, auf denen die Wörter *Dackel* oder *Dobermann* vorkommen und die damit auch Inhalte zum Wortfeld „Hund" beschreiben, von der obigen Suchanfrage nicht erfasst. Dieses einfache Beispiel macht deutlich, dass Computer ohne die Annotation expliziter Information, etwa auf semantischer Ebene (Dackel gehört zum Wortfeld „Hund"), nicht ohne Weiteres in der Lage sind, Informationen aus einem Text zu extrahieren bzw. Suchanfragen zu beantworten; oder anders formuliert:

> Suchmaschinen können nur eingesetzt werden, wenn der gesuchte Gegenstand, sei es eine Lösung, ein numerischer Wert, eine Figur, eine Information oder ein Angebot, mit einer Adresse verbunden werden kann, also auf systematische Weise etikettiert ist. Das Gesuchte muß beschriftet sein, um erreichbar und verfügbar zu werden. (Gugerli 2009, 15)

Die hier formulierten Anforderungen an explizite Annotation all derjenigen Parameter, die später computergestützt identifizierbar und durchsuchbar gemacht werden sollen, gelten selbstverständlich auch für den Bereich linguistischer Korpora. Soll also beispielsweise eine Suchanfrage wie „Finde alle Sätze, in denen eine beliebige Form des Verbs *sein*, unmittelbar gefolgt von einem beliebigen *Adjektiv,* vorkommt" an ein Textkorpus gestellt werden, so macht dies eine vorhergehende Annotation der Lemmata sowie auch der Wortarten erforderlich. Um neue Erkenntnisse aus einem Korpus ziehen zu können, müssen also zunächst grundlegende Informationen hinzugefügt, also annotiert, werden (Leech 1997, 4).

3. Kurze Einführung in die technischen Grundlagen des Dokumentmarkups

Nachdem im vorhergehenden Kapitel erklärt wurde, warum explizite Annotation in linguistischen Korpora wichtig und notwendig ist, soll an dieser Stelle kurz erläutert werden, wie diese technisch umgesetzt werden. Zentral ist hierbei die Idee der Auszeichnungssprachen (engl. *markup languages*) und Dokumentgrammatiken. Auszeichnungen – auch Annotation oder Markup genannt – sind im Wesentlichen Markierungen im Text, welche über sogenannte Tags (Etiketten) umgesetzt werden. Solche Tags stehen üblicherweise in spitzen Klammern und können so leicht von Softwaretools vom Primärtext unterschieden werden.

---

[2] Ein weit verbreitetes Tool für die elektronische Textanalyse, welches all diese Funktionen bietet und kostenlos als Webservice genutzt werden kann, ist Voyant-Tools (verfügbar unter http://voyant-tools.org/). Ein umfassender Überblick zu weiteren Textanalysewerkzeugen findet sich auf dem Portal TAPoR 2.0 (verfügbar unter http://www.tapor.ca/).

Hinweis: Alle URLs, die in diesem Artikel Erwähnung finden, wurden zuletzt im Januar 2016 auf Erreichbarkeit überprüft.

Beispiel:

```
<h1> Faust: Der Tragödie Erster Teil </h1>
...
<rede sprecher= "geist"> Wer ruft mir? </rede>
<rede sprecher="faust"> Schreckliches Gesicht </rede>
```

Tags treten dabei immer paarweise auf und markieren so den Anfang und das Ende eines auszuzeichnenden Bereichs im Text. Gleichzeitig werden für den markierten Textbereich über den Namen des Tags (*h1*, *rede*) sowie ggf. über optionale Attribute (*sprecher*) zusätzliche Informationen annotiert, die später auch von Tools ausgelesen und verarbeitet werden können. So könnten etwa Suchanfragen formuliert werden, welche nur alle Reden im Text zurückgeben (nicht aber die Überschriften erster Ordnung = *h1*) oder auch nur die Reden, die vom Sprecher Faust stammen. Die Gesamtheit aller Tags für einen abgeschlossenen Bereich wird als Markupsprache bezeichnet. So ist etwa die *Hypertext Markup Language* (HTML) eine Sammlung von Tags, mit deren Hilfe Webseiten beschrieben werden können.

HTML ist eine konkrete Anwendung der *Standard Generalized Markup Language* (SGML), einer Meta-Markupsprache, die 1986 als ISO-Standard (SGML ISO 8879) veröffentlicht wurde und die es erlaubt, konkrete Markupsprachen für spezifische Zwecke zu definieren. Wegen ihrer Mächtigkeit und der damit einhergehenden Komplexität der SGML fand sie außerhalb von Universitäten und Großkonzernen kaum Beachtung, weshalb 1997 eine vereinfachte Version in Form der *eXtensible Markup Language* (XML) veröffentlicht wurde, welche SGML mittlerweile weitestgehend abgelöst hat. Dementsprechend findet sich heute eine Vielzahl von XML-basierten Markupsprachen für die unterschiedlichsten Einsatzgebiete. Exemplarisch sind etwa die Auszeichnungsvokabularien der *Text Encoding Initiative* (TEI), welche als Defacto-Standard in den textbasierten Geisteswissenschaften gelten, sowie der speziell an die Anforderungen der Korpuslinguistik angepasste *Corpus Encoding Standard* (XCES), der als Teilmenge und Erweiterung der TEI zu verstehen ist, zu nennen.

Die Spezifikation solcher Markupsprachen erfolgt über sogenannte Dokumentgrammatiken, in welchen nicht nur das erlaubte Vokabular, also die Tags, definiert werden, sondern auch grundlegende Regeln über deren Auftretenshäufigkeit, Kombinationsmöglichkeiten und Reihenfolge. Solche Regeln können formal gut umgesetzt werden, da mit Markup versehene Textteile automatisch auch immer in einen hierarchischen Zusammenhang und somit über die bloße, sequenzielle Abfolge eines Texts zusätzlich in eine Teil-Ganzes-Beziehung gebracht werden (Lobin 2004, 52).

Beispiel:

```
<buch>
   <kapitel>
      <titel> ... </titel>
      <absatz> ... </absatz>
       <absatz> ... </absatz>
   </kapitel>
   <kapitel>  ... </kapitel>
</buch>
```

Durch die Verschachtelung der paarweise auftretenden Tags entsteht so eine Baumstruktur des Dokuments, die es erlaubt, Eigenschaften von Elementen von übergeordneten an untergeordnete Elemente zu vererben und durch Angabe spezifischer Knoten präzise durch die Dokumentstruktur zu navigieren und damit beliebige Teilmengen zu selektieren. Der XML-Standard beinhaltet neben grundlegenden Regeln, welche vorgeben, wie die Baumstruktur eines Dokuments formal korrekt[3] zu notieren ist, auch eine eigene Sprache zur Formulierung von Dokumentgrammatiken. *Document Type Definitions* (DTDs) sind dabei ein etabliertes Konzept zur Umsetzung solcher Grammatiken. Die nachfolgende Beispiel-DTD zeigt eine einfache Grammatik für das vorhergehende Buch-Beispiel:

```
<!ELEMENT buch (kapitel+)>
<!ELEMENT kapitel (titel, absatz+)>
<!ELEMENT titel (#PCDATA)>
<!ELEMENT absatz (#PCDATA)>
```

In der ersten Zeile wird spezifiziert, dass ein Dokument des Typs Buch aus beliebig vielen Kapiteln bestehen kann, mindestens aber aus einem (der Plusoperator hinter Kapitel bedeutet: Element kommt mindestens einmal oder öfter vor). In Zeile 2 wird festgelegt, dass unterhalb eines Kapitel-Elements jeweils genau ein Kapiteltitel (kein Operator bedeutet: Element kommt genau einmal vor) sowie beliebig viele Absätze, mindestens aber einer, auftreten dürfen. Die Reihenfolge der Elemente in den runden Klammern ist auch im Dokument einzuhalten, d. h. also, der Titel muss vor den Absätzen kommen. In den Zeilen 3 und 4 wird schließlich spezifiziert, dass innerhalb eines Titels sowie auch der einzelnen Absätze keine weiteren Elemente geschachtelt werden können, sondern eine beliebige Menge an Text (PCDATA = *parsed character data*) stehen muss.

Wird ein XML-Dokument mit einer DTD verknüpft, kann mit einem entsprechenden Validierungstool automatisch überprüft werden, ob alle in der DTD spezifizierten Regeln eingehalten werden und das Dokument damit valide ist[4]. Valide Dokumente sind wichtig, um später von Computerprogrammen – sogenannten Parsern – korrekt verarbeitet werden zu können. Eine weitere Möglichkeit zur Definition von Dokumentgrammatiken bietet *XML Schema*. Im Gegensatz zu DTDs, welche in einer proprietären Notation vorliegen, ist XML Schema selbst in XML notiert. Darüber hinaus liefert XML Schema reichhaltigere Datentypen und die Möglichkeit mächtigerer Strukturdefinitionen (vgl. Lobin 2004, 61).

---

[3] Man spricht in diesem Zusammenhang auch vom Kriterium der „Wohlgeformtheit“. Vgl. auch die Hinweise zu wohlgeformten XML-Dokumenten im entsprechenden W3C-Standard (verfügbar unter http://www.w3.org/TR/xml/#dt-wellformed). Zu diesen Wohlgeformtheitskriterien gehören etwa: Es gibt genau ein Haupt- oder Wurzelelement, alle Elemente mit Inhalt besitzen eine Beginn- und eine End-Kennung, verschachtelte Elemente dürfen nicht überlappen, etc.

[4] Ein bekanntes, frei im Netz verfügbares Validierungstool findet sich unter http://www.validome.org/.

4. Verwandte Arbeiten im Bereich der digitalen Lexikographie

XML-basierte Markupsprachen sind im Bereich der linguistischen Informationsmodellierung (vgl. Witt 2004, 39) weit verbreitet und eignen sich somit auch hervorragend für die elektronische Aufbereitung des eingangs beschriebenen HSVW-Projekts. Bevor im nächsten Kapitel konkrete Überlegungen zur Modellierung und späteren Nutzung elektronisch aufbereiteter historisch syntaktischer Lexikonartikel präsentiert werden, sollen zur weiteren Kontextualisierung zunächst kurz einschlägige, verwandte Arbeiten aus dem Gebiet der digitalen Lexikographie, insbesondere in den Bereichen historische Lexikographie und Valenzlexika, vorgestellt werden.

Ein Referenzprojekt im Bereich der digitalen Modellierung eines Valenzwörterbuchs von Verben des Gegenwartsdeutschen stellt hier sicherlich E-VALBU[5] dar. Es handelt sich dabei um eine elektronische Neubearbeitung des 2004 abgeschlossenen VALBU-Projekts (Schumacher u. a. 2004), welches sich der Erarbeitung eines didaktisch orientierten Valenzwörterbuchs deutscher Verben widmete. Über eine Eingabemaske kann E-VALBU nach bestimmten Komplementen, Satzbauplänen oder Schlüsselwörtern durchsucht werden. Daneben gibt es eine Vielzahl elektronisch verfügbarer Lexika, die zwar keine explizite Valenzinformation enthalten, die dafür aber miteinander vernetzt und somit als Wörterbuchkonglomerat durchsuchbar sind. Beispielhaft sind hier etwa das OWID-Projekt[6] des IDS Mannheim und das Projekt „Wörterbuchnetz“[7], des Trier Center for Digital Humanities zu nennen. Während OWID auf Lexika des Gegenwartsdeutschen fokussiert, finden sich im Wörterbuchnetz auch viele historische Wörterbücher. Hier ist vor allem das Mittelhochdeutsche Wörterbuch[8] zu nennen, welches eine wesentliche Datengrundlage für das im Rahmen des HSVW-Projekts zu erstellende historische Valenzwörterbuch darstellt. Ein weiteres wichtiges Projekt im Bereich historischer Korpora stellt das DDD (*Deutsch Diachron Digital*)[9] dar: Hier wurde bereits ein umfangreiches Referenzkorpus des Althochdeutschen erstellt, welches über das Informationssystem ANNIS[10] digital abgefragt und exploriert werden kann. Informationen zur Verbvalenz wurden bei ANNIS bislang allerdings nicht berücksichtigt. Ergänzend ist außerdem noch das am IDS ansässige, DFG-geförderte Netzwerk „Internetlexikografie“[11] als verwandte Arbeit zu

---

[5] Das E-VALBU-Portal ist online verfügbar unter http://hypermedia.ids-mannheim.de/evalbu/index.html.

[6] Online verfügbar unter http://www.owid.de/. Mehr Informationen zum Projekt finden sich unter http://www.owid.de/wb/owid/start.html.

[7] Das Wörterbuchnetz ist online verfügbar unter http://woerterbuchnetz.de/. Mehr Informationen zum Projekt finden sich unter http://kompetenzzentrum.uni-trier.de/de/projekte/projekte/woerterbuchnetz/.

[8] Online verfügbar unter http://www.mhdwb-online.de/.

[9] Online verfügbar unter http://www.deutschdiachrondigital.de.

[10] Mehr Informationen zu ANNIS finden sich online unter http://corpus-tools.org/annis/. Das DDD-Korpus ist ebenfalls online über ANNIS durchsuchbar: https://korpling.german.hu-berlin.de/annis3/ddd.

[11] Informationen zum Netzwerk sowie ein Überblick zu bisherigen Arbeitstreffen sind online verfügbar unter http://multimedia.ids-mannheim.de/mediawiki/web/index.php/Hauptseite.

nennen. Das Netzwerk veranstaltet seit 2011 regelmäßig Arbeitstreffen zu einschlägigen Themen rund um die Erstellung von elektronischen Wörterbüchern. Besonders interessant für das HSVW-Projekt sind die Erkenntnisse der 2015 abgehaltenen Tagung zum Thema „Valenz und Kollokation im (digitalen) Wörterbuch".

Der exemplarische Überblick zu verwandten Arbeiten im Bereich der digitalen Internetlexikographie macht deutlich, dass es für verschiedene Teilaspekte des HSVW-Projekts bereits erfolgreich umgesetzte Projektbeispiele gibt, die elektronische Aufbereitung eines historischen Valenzwörterbuchs jedoch weiterhin ein Desiderat bleibt.

5. Wesentliche Informationen in einem historischen Valenzlexikonartikel

Beispielhaft soll nun ein Lexikonartikel der Regensburger Arbeitsgruppe „Mittelhochdeutsch" vorgestellt werden, um zu verdeutlichen, was die wesentlichen linguistischen Parameter und inhaltlichen Elemente eines historischen Valenzwörterbuchartikels sind und wie diese im nächsten Schritt in ein adäquates XML-Dokumentmarkup überführt werden können. Ausführliche Informationen zur Korpusauswahl sowie zum Vorgehen bei der Erstellung von Valenzwörterbuch-Einträgen anhand lemmatisierter Konkordanzen finden sich im Beitrag von Albrecht Greule (in diesem Band) und sollen deshalb hier nicht im Detail wiedergegeben werden. Nachfolgend werden alle Informationen, die während des Entstehungsprozesses eines historischen Lexikonartikels gesammelt werden, systematisch aufgeführt.

Im ersten Schritt werden für jeden Lexikonartikel grundlegende Metadaten wie etwa Autor, Erstellungsdatum sowie das zugrundeliegende Korpus dokumentiert. Wenn es sich beim Korpus um eine Online-Ressource handelt, wird auch die URL angegeben. Als Nächstes werden für jedes Verb, ausgehend von konkreten Belegsätzen, systematische Einzelanalysen erstellt. Als Quelle für die Belegsätze dienen bestehende Wörterbücher, etwa das online verfügbare „Mittelhochdeutsche Wörterbuch". Soweit verfügbar, werden die Originalquellen der Belegsätze unter Angabe eines Kurztitels sowie auch der vollständigen bibliographischen Angaben, und – falls vorhanden – einer URL zum Volltext, mit dokumentiert. In jedem Belegsatz werden die inhaltlich relevanten Satzglieder identifiziert und anhand der Parameter „Form", „Funktion" und „Semantik" analysiert. Unter „Form" werden Wortarten bzw. die Klassifikation als Wortgruppe und Flexionsformen aufgeführt. Die „Funktion" enthält die „syntaktische Funktion" im Satz (Ergänzung oder Angabe, Attribute spielen für unsere Belange keine Rolle). Die Semantik umfasst den Tiefenkasus/die thematische Rolle der Satzglieder (vgl. Tarvainen 1987). Ergänzend zur Semantik der einzelnen Satzglieder können zudem als optionale Aspekte die beispielsatzbezogene „semantische Spezifikation", die frei klassifiziert werden kann, sowie die „semantische Restriktion", die dann angegeben wird, wenn ein Satzglied beispielübergreifend mit der gleichen übergeordneten semantischen Information belegt wird (z. B. immer Hum / Mensch oder immer Abstraktum) angeführt werden.

Eine solche Analyse auf Belegsatzebene soll nachfolgend in einer tabellarischen Darstellung beispielhaft für das Verb *antlâzen*[12] gezeigt werden (vgl. Tabellen 1–4).

Belegsatz: *helt, nu antlaze du mir, / daz min sele icht prinne!* (Rol, 6481)					
Satzglied	nu	antlaze	du	mir	daz … prinne
Form	Adverb	Finitum	Pronomen	Pronomen	Nebensatz
Funktion	Angabe	Prädikat	Ergänzung im Nominativ	Ergänzung im Dativ	Angabe
Semantik	temporal	Handlung	Agens	Adressat	final
Spezifikation	-	-	Vergebender	dem die Vergebung zugesprochen wird	-
Restriktion	-	-	Hum	Hum	-

Tab. 1: Beispielhafte Analyse des ersten Belegsatzes zum Verb *antlâzen*.

Belegsatz: *antlazzit alle ain andir, ôb ir wârn antlaz uon gote hiute welt gewinnin* (Spec, 53,16)				
Satzglied	antlazzit	alle	ain andir	ôb … gewinnin
Form	Finitum	Nomen	Pronomen	Nebensatz
Funktion	Prädikat	Ergänzung im Nominativ	Ergänzung im Dativ	Angabe
Semantik	Handlung	Agens	Adressat	konditional
Spezifikation	-	Vergebender	dem die Vergebung zugesprochen wird	-
Restriktion	-	Hum	Hum	-

Tab. 2: Beispielhafte Analyse des zweiten Belegsatzes zum Verb *antlâzen*.

Belegsatz: *mit dienste si sich flizzen, / daz si in [ihnen] der unmâze / geruochte antlâzen* (Wernh, 3690)					
Satzglied	si	in	der unmâze	geruochte	antlâzen
Form	Pronomen	Pronomen	Nominal-gruppe	Finitum	Infinitiv

[12] Anmerkung: Die Belegsätze samt Quellenangaben wurden aus der Online-Version des Mittelhochdeutschen Wörterbuchs übernommen (vgl. http://mhdwb-online.de/wb.php?buchstabe=A&portion=1920). Die Analyse erfolgte durch Albrecht Greule, Mitglied der Regensburger Arbeitsgruppe des HSVW-Projekts.

Funktion	Ergänzung im Nominativ	Ergänzung im Dativ	Ergänzung im Genitiv	Prädikatsteil 1	Prädikatsteil 2
Semantik	Agens	Adressat	Themativ	Handlung	Handlung
Spezifikation	Vergebender	dem die Vergebung zugesprochen wird	Verfehlung	-	-
Restriktion	Hum	Hum	Abstr	-	-

Tab. 3: Beispielhafte Analyse des dritten Belegsatzes zum Verb *antlâʒen*.

Belegsatz: *ir schulde si veriahen / unde baten in got antlazen* (Serv, 2321)				
Satzglied	ir schulde	in	got	antlazen[13]
Form	Nominalgruppe	Pronomen	Nomen	Finitum
Funktion	Ergänzung im Genitiv	Ergänzung im Dativ	Ergänzung im Nominativ	Prädikat
Semantik	Themativ	Adressat	Agens	Handlung
Spezifikation	Verfehlung	dem die Vergebung zugesprochen wird	Vergebender	-
Restriktion	Abstr	Hum	Gott	-

Tab. 4: Beispielhafte Analyse des vierten Belegsatzes zum Verb *antlâʒen*.

Im nächsten Schritt wird dann aus den Einzelanalysen der Belegsätze die generische Valenzinformation für das untersuchte Verb induktiv abgeleitet. Diese Szene besteht aus einer generischen Paraphrase der Belegsätze sowie aus Informationen zu den Funktionen und zur semantischen Interpretation der Ergänzungen. Im Falle von *antlâʒen* ergibt sich die folgende Szene (vgl. Tabelle 5).

[13] Hinweis: In diesem Belegsatz ist eine Rekonstruktion des Finitums notwendig: [*ir schulde in got antlaze*].

	Funktion	**Semantik**	**Spezifikation**	**Restriktion**
a (obligatorisch)	Nominativ	Agens	Vergebender	– [14]
b (obligatorisch)	Dativ	Adressat	dem die Vergebung zugesprochen wird	Hum
c (fakultativ)	Genitiv	Themativ	Verfehlung	Abstr
Generische Paraphrase: a spricht b die Vergebung (der Verfehlungen c) zu				

Tab. 5: Generische Valenzinformation für das Verb *antlâʒen*, abgeleitet aus den einzelnen Belegsätzen.

Zuletzt werden noch für jedes Verb grundlegende linguistische Informationen wie Lemma[15], Wertigkeit, Verbklasse und Sprachstufe dokumentiert (vgl. Tabelle 6). Um ein Verbindungselement für spätere diachrone Analysen zu schaffen, soll zudem für jedes Verb auf jeder Sprachstufe die neuhochdeutsche Übersetzung angegeben werden: Wir sprechen hier von einem „Hyperlemma" mit der Funktion, die Wiederauffindbarkeit eines Verbs über die Sprachstufen hinweg zu erleichtern. Verben werden außerdem anhand ihres jeweiligen Wortfelds klassifiziert. Dabei legen wir das Wortfeldinventar von Sommerfeldt/Schreiber (1996) zugrunde, welches insgesamt 13 grundlegende Wortfelder unterscheidet. Das Kriterium „Varietät" ist optional und ermöglicht die Erwähnung von Charakteristika des Sprachgebrauchs bzw. Ko(n)textes, im folgenden Beispiel (Varietät: Theolekt) handelt es sich um eine spezifische Kommunikationssituation.

Gibt es für ein Verb mehrere Sememe so wird jeweils ein eigener Wörterbucheintrag mit Belegsätzen und Valenzszene für jedes einzelne Semem angelegt (siehe Kap. 1).

Lemma	antlâʒen
Hyperlemma	Ablass erteilen
Wertigkeit	3-wertig
Verbklasse	Schwaches Verb
Sprachstufe	Mittelhochdeutsch
Wortfeld	Mitteilung (Sprachproduktion)[16]
Varietät	Theolekt

Tab. 6: Allgemeine linguistische Informationen für das Verb *antlâʒen*.

[14] Erläuterung: Da in den Belegsätzen sowohl Menschen (Hum) als auch einmal Gott auftreten, kann hier keine allgemeine semantische Restriktion für die Ergänzung im Nominativ abgeleitet werden.

[15] Falls es für ein Lemma orthografische Varianten gibt, so werden diese ebenfalls dokumentiert; Beispiel *angesten*/*angestigen*.

[16] Vgl. Sommerfeldt/Schreiber (1996, 144ff.).

6. Konzeption einer Dokumentgrammatik für historische Valenzartikel

Im vorhergehenden Kapitel wurden beispielhaft die wesentlichen Informationen eines Valenzwörterbuchartikels vorgestellt. Nachfolgend soll die Modellierung dieser Informationen als generische Dokumentgrammatik aufgezeigt werden. Die Grammatik versteht sich dabei als erster Entwurf, welcher in einer interdisziplinären Übung „Elektronische Aufbereitung eines historischen Verbwörterbuchs“ zwischen Germanistischer Sprachwissenschaft und Medieninformatik im Sommersemester 2016 an der Universität Regensburg getestet und iterativ verfeinert werden soll. In weiterführender Perspektive ist zudem denkbar, das finale Dokumentmodell – soweit möglich – auf bestehende Tagsets wie etwa *TEI Dictionaries*[17] abzubilden, um größere Interoperabilität mit bestehenden Analysetools zu gewährleisten.

Modellierung der Dokumentstruktur

Wie zuvor beschrieben, ist für XML-Dokumente durch die ineinander verschachtelbaren Tags eine baumartige Dokumentstruktur charakteristisch. Als konzeptionelle Vorarbeit für die Erstellung einer Dokumentgrammatik für historische Valenzartikel soll deshalb zunächst eine solche hierarchische Struktur modelliert werden. Abbildung 1 zeigt nochmals alle wesentlichen Informationselemente sowie deren hierarchische Beziehung. Zusätzlich sind über verschiedene Umrahmungen der Elemente Regeln für deren jeweils erlaubte Auftretenshäufigkeit kodiert. Grundsätzlich ist zu unterscheiden, ob ein Element zwingend auftreten muss, weil es einen essentiellen Bestandteil eines Lexikoneintrags darstellt (z. B. Lemmainformation), oder ob es ggf. weggelassen werden kann (z. B. Volltext-URL zum verwendeten Korpus; muss nicht zwangsweise immer vorhanden sein). Bei obligatorischen Elementen kann darüber hinaus unterschieden werden, ob diese genau einmal (z. B. ein Erstellungsdatum) oder mindestens einmal (z. B. ein Autor oder mehrere Autoren) auftreten können. Fakultative Elemente sollen in diesem Szenario entweder gar nicht oder genau einmal auftreten können.

Natürlichsprachlich formuliert, liest sich die Dokumentstruktur in Abbildung 1 wie folgt: Ein Lexikon besteht aus mindestens einem Artikel. Ein Artikel kann Einträge zu mehreren Sememen enthalten. Für jedes Semem werden grundlegende Metadaten, linguistische Informationen, die generische Valenzszene sowie die Belegsätze und deren Analyse dokumentiert. Als Metadaten müssen mindestens ein Autor und ein Korpus sowie genau ein Erstellungsdatum des Artikels angegeben werden. Für jedes Korpus muss genau ein Titel sowie optional eine URL zum Volltext des Korpus spezifiziert werden, usw.

[17] Online verfügbar unter http://www.tei-c.org/release/doc/tei-p5-doc/de/html/DI.html.

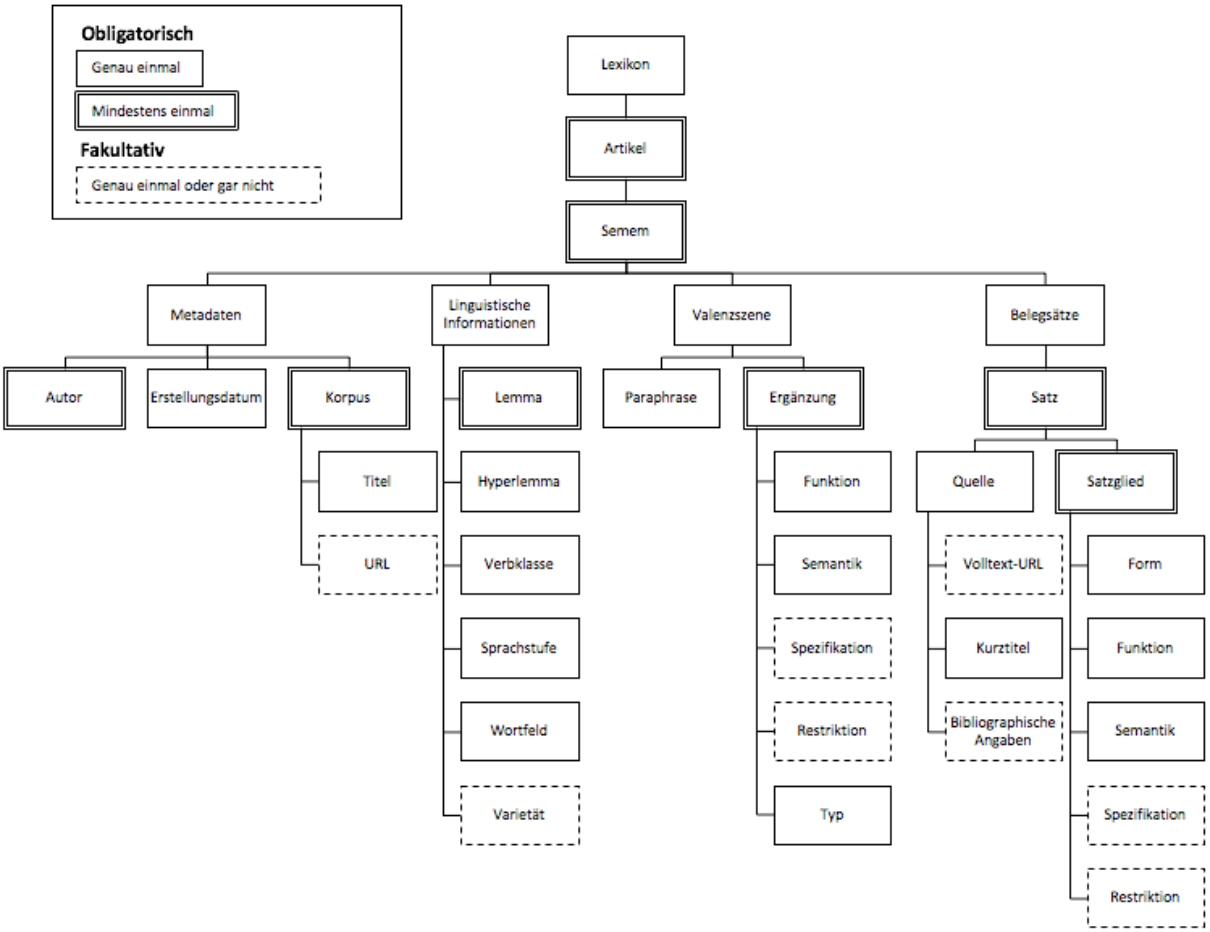

Abb. 1: Hierarchische Darstellung der wesentlichen Elemente eines historischen Lexikonartikels.

Ziel bei der Modellierung einer XML-Struktur ist die Vermeidung von Redundanz. Aus diesem Grund wird etwa das Element „Wertigkeit“ nicht explizit modelliert, weil die Information zur Wertigkeit eines Verbs auch automatisch aus der Anzahl der Ergänzungen abgeleitet und ausgegeben werden kann.

Definition der Wertebereiche

Neben der Definition grundlegender Dokumentelemente sowie deren hierarchischer Beziehung kann in einer Dokumentgrammatik auch festgelegt werden, welche Werte für die jeweiligen Elemente erlaubt sein sollen. Lexikonartikel können dann anhand einer Grammatik validiert werden, um automatisch zu prüfen, ob die angegebenen Werte zu einem Verb im Rahmen der definierten Wertebereiche liegen.

Für die Analyse der Belegsätze auf den Ebenen „Form“, „Funktion“ und „Semantik“ ergeben sich etwa die folgenden Wertebereiche[18] (vgl. Tabelle 7):

[18] Wie bereits erwähnt, handelt es sich hierbei um eine erste Sammlung plausibler Wertebereiche, die anhand erster beispielhafter Analysen erstellt wurden. Im Laufe des Projekts wird diese Liste sukzessive an die Analysepraxis angepasst und um weitere Werte ergänzt werden.

Parameter	Werteliste
Form	Finitum, Infinitiv, Partizip I, Partizip II, Verbgruppe, Nominalgruppe, Nomen, Pronominalgruppe, Pronomen, Adjektiv, Adjektivgruppe, Adverb, Adverbgruppe, Präpositionalgruppe, Konjunktionalgruppe, Nebensatz, Infinitivkonstruktion, Partizipialkonstruktion
Funktion	Prädikat/Prädikatsteil, Angabe, Ergänzung im Nominativ, Ergänzung im Genitiv, Ergänzung im Dativ, Ergänzung im Akkusativ, Ergänzung mit fester Präposition, Ergänzung mit unfester Präposition, Ergänzung mit Prädikat, Korrelat, Platzhalter-*es*
Semantik	Für Prädikate: Handlung, Zustand, Vorgang; Für Angaben: Lokalangabe, Temporalangabe, Kausalangabe, Konditionalangabe, Konzessivangabe, Finalangabe, Konsekutivangabe, Instrumentalangabe, Komitativangabe, Restriktivangabe, Adversativangabe, Kommentarangabe, Modalangabe, Prädikativangabe, Negationsangabe, Angabe mit freiem Dativ, Angabe mit freiem Akkusativ; Für Ergänzungen: Agens, Patiens, affiziertes Objekt, Resultat, Possessiv, Themativ, Perzeptiv, Finativ, Adressat, Kausativ, Komitativ, Instrument, Modativ, Lokativ, Temporativ, Denominativ, Direktiv, Resultativ, Handlung;

Tab. 7: Beispielhafte Wertebereiche für die Analyseparameter „Form“, „Funktion“ und „Semantik“.

Gleichzeitig gibt es einige Parameter, für die man vorab keine vollständige Werteliste definieren kann, da die möglichen Werte stark vom jeweiligen Belegsatz abhängig sind und damit sehr heterogen sein können. Zu diesen Parametern gehört etwa die semantische Spezifikation eines Satzglieds. In der Dokumentgrammatik wird an dieser Stelle deshalb kein konkreter Wertebereich spezifiziert, stattdessen können hier bei der Artikelerstellung beliebige Werte angegeben werden.

Umsetzung einer Dokumentgrammatik mit XML Schema

Aufbauend auf den Vorüberlegungen zu den wesentlichen Elementen, deren hierarchischer Beziehung zueinander, deren Auftretenshäufigkeit im Dokument sowie den zu erwartenden Wertebereichen kann nun eine erste Dokumentgrammatik für Valenzwörterbuchartikel mo-

delliert werden. Abbildungen 2 und 3 zeigen jeweils Ausschnitte aus einer Dokumentgrammatik, die mithilfe von XML Schema umgesetzt wurde.

In Abbildung 2 wird zunächst ein Element „Satzglied“ als komplexes Element (*complexType*) definiert. Ein Satzgliedelement kann laut der Dokumentgrammatik gemischte Inhalte (*mixed=“true“*), nämlich sowohl das betreffende Satzglied in textueller Form als auch diverse Annotationen in Form weiterer Elemente beinhalten. Die erlaubten Annotationselemente („Rekonstruktion“, „Form“, „Funktion“, „Semantik“, „Spezifikation“, „Restriktion“) werden dann als Sequenz (*xs:sequence*) definiert, d. h., sie müssen im Dokument auch genau in dieser Reihenfolge vorkommen, damit das Dokument valide hinsichtlich der Grammatik ist. Außerdem wird für jedes Element definiert, wie häufig es minimal (*minOccurs*) und maximal (*maxOccurs*) vorkommen darf.

```
<xs:element name="satzglied">
  <xs:complexType mixed="true">
    <xs:sequence>
      <xs:element ref="rekonstruktion" minOccurs="0" maxOccurs="1" />
      <xs:element ref="form" minOccurs="1" maxOccurs="1" />
      <xs:element ref="funktion" minOccurs="1" maxOccurs="1" />
      <xs:element ref="semantik" minOccurs="1" maxOccurs="1" />
      <xs:element ref="spezifikation" minOccurs="0" maxOccurs="1" />
      <xs:element ref="restriktion" minOccurs="0" maxOccurs="1" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

Abb. 2: Ausschnitt aus der XML Schema-Dokumentgrammatik zum Element „Satzglied“ sowie zu erlaubten Unterelementen.

In Abbildung 3 wird das Element „Form“ weiter spezifiziert, indem eine Liste erlaubter Werte (*value*), die innerhalb des Form-Tags stehen können, definiert wird.

```
<xs:element name="form">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="Fin" />
      <xs:enumeration value="Inf" />
      <xs:enumeration value="Part1" />
      <xs:enumeration value="Part2" />
      <xs:enumeration value="VG" />
      <xs:enumeration value="NG" />
      <xs:enumeration value="N" />

      <!-- weitere Werte -->

    </xs:restriction>
  </xs:simpleType>
</xs:element>
```

Abb. 3: Ausschnitt aus der XML Schema-Dokumentgrammatik zum Element „Satzglied“ sowie zu erlaubten Werten für dieses Element.

Das vollständige XML Schema ist online verfügbar unter https://histvw.wordpress.com/beispielanalysen/. Auf derselben Seite ebenfalls verfügbar ist ein beispielhaftes XML-Dokument für den Lexikonartikel zum Verb *antlâzen* (vgl. Abbildung 4).

```xml
<?xml version="1.0" encoding="UTF-8"?>
<lexikon>
    <artikel>
        <semem>
            <metadaten>
                <autor>Albrecht Greule</autor>
                <erstellungsdatum>01.01.2016</erstellungsdatum>
                <korpus>
                    <titel>Mittelhochdeutsches Wörterbuch Online</titel>
                    <url>http://mhdwb-online.de</url>
                </korpus>
            </metadaten>
            <linguistische-information>
                <lemma>antlâzen</lemma>
                <hyperlemma>Ablass erteilen</hyperlemma>
                <verbklasse>swV</verbklasse>
                <sprachstufe>Mhd</sprachstufe>
                <wortfeld>Mitteilung (Sprachproduktion)</wortfeld>
                <varietät>Theolexem</varietät>
            </linguistische-information>
            <valenzszene>
                <paraphrase>a spricht b die Vergebung (der Verfehlungen c) zu</paraphrase>
                <ergaenzung id="a">
                    <erg-funktion>Nominativ</erg-funktion>
                    <erg-semantik>Agens</erg-semantik>
                    <erg-spezifikation>Vergebender</erg-spezifikation>
                    <erg-typ>obligatorisch</erg-typ>
                </ergaenzung>
                <ergaenzung id="b">
                    <erg-funktion>Dativ</erg-funktion>
                    <erg-semantik>Adressat</erg-semantik>
                    <erg-spezifikation>dem die Vergebung zugesprochen wird</erg-spezifikation>
                    <erg-typ>obligatorisch</erg-typ>
                </ergaenzung>
                <ergaenzung id="c">
                    <erg-funktion>Genitiv</erg-funktion>
                    <erg-semantik>Themativ</erg-semantik>
                    <erg-spezifikation>Verfehlung</erg-spezifikation>
                    <erg-typ>fakultativ</erg-typ>
                </ergaenzung>
            </valenzszene>
            <!-- Weitere Informationen zu den Belegsatzanalysen ... -->
        </semem>
    </artikel>
</lexikon>
```

Abb. 4: Ausschnitt aus einem validen XML-Dokument für den Lexikonartikel zum Verbum *antlâzen*.

An dieser Stelle soll nochmals verdeutlicht werden, dass die Erstellung des hier vorgestellten XML-Codes später nicht zwangsweise in einem Texteditor geschehen muss, was in der Praxis nicht nur aufwendig und fehleranfällig ist, sondern auch nicht gängigen Anforderungen an Benutzerfreundlichkeit und Gebrauchstauglichkeit computergestützter Tools entspricht (vgl. auch die Besonderheiten der „humanist-computer interaction“, Burghardt/Wolff 2015). Vielmehr gibt es eine Vielzahl von Annotationswerkzeugen mit grafischen Benutzeroberflächen (Burghardt 2012; 2014), die zur Erstellung solcher XML-Dokumente verwendet werden kön-

nen und die auch im Rahmen des HSVW-Projekts zum Einsatz kommen sollen. In erster Linie ist die XML-Repräsentation der Lexikonartikel dazu gedacht, gut von Computerprogrammen verarbeitet werden zu können.

7. Vorteile und Nutzen elektronisch aufbereiteter Wörterbuchartikel

Vorrangiges Ziel der digitalen Modellierung der geplanten HSVW-Wörterbuchressource ist es, diese später komfortabel anhand verschiedener linguistischer Parameter durchsuchbar zu machen. Anders als beim vergleichbaren E-VALBU- Projekt soll die digitale Variante nicht nachträglich zur gedruckten Ausgabe erstellt werden – was ggf. zusätzlichen Aufwand in Form texttechnologischer und manueller Nachbearbeitung erzeugt (vgl. Schneider 2008, 3) –, sondern von Anfang an parallel zur sukzessiven Erarbeitung der Wörterbuchartikel umgesetzt werden. Durch diese frühzeitige digitale Modellierung ergeben sich auch schon bei der Bearbeitung und Erstellung der Lexikonartikel einige wesentliche Vorteile.

Konsistenz der Daten

Durch eine vorab spezifizierte Dokumentgrammatik, mit Regeln über Form und Struktur der historischen Wörterbuchartikel, insbesondere der Valenzinformationen, kann von Anfang an sichergestellt werden, dass die Artikel konsistent aufgebaut sind. Da sich das HSVW-Projekt wegen des Umfangs auf verschiedene Arbeitsstellen verteilt, ist eine konsistente bzw. kompatible Beschreibung der Verben von besonderer Wichtigkeit (vgl. Greule in diesem Band). Die Einhaltung der Dokumentgrammatik, und damit der einheitlichen Artikelstruktur, kann automatisiert über Validierungstools geprüft werden. Somit ist – zumindest auf formaler Ebene – eine grundlegende, automatisierte Qualitätskontrolle möglich.

Variable Darstellung der Daten

Ein weiterer wesentlicher Vorteil, den die frühzeitige digitale Modellierung der Lexikonartikel mit sich bringt, besteht in der Flexibilität auf der Präsentationsebene der Daten. Diese Flexibilität ergibt sich aus der Trennung von Struktur und Darstellung, einem grundlegenden Konzept bei Markupsprachen. Durch die generische Auszeichnung von Dokumenten werden diese zunächst mit Information angereichert. Anhand dieser Informationen können die Dokumente dann mithilfe von *Stylesheets* (z. B. *Cascading Stylesheets*[19]) in beliebiger Form visuell dargestellt werden. Praktisch können so mehrere Stylesheets für ein XML-Dokument angelegt werden, um dieses jeweils für verschiedene Präsentationsszenarien zu optimieren. Mit der Transformationssprache *XSL* (*eXtensible Stylesheet Language*) können darüber hinaus beliebige Dokumentteile selektiert (z. B. alle 3-wertigen Verben aus dem Ahd.) und in eine beliebige Zieldarstellung übertragen werden. Im Falle der HSVW-Wörterbucherstellung könnte hier aus dem XML-Dokument einerseits eine spätere Printfassung (Serifenschrift, Seitenränder, A5-Format, etc.) generiert werden und andererseits eine web-optimierte Fassung mit Hyperlinks und dynamischen Sichten auf die einzelnen Informationen, d. h., die angezeigten Informationen unterscheiden sich je nach vorhergehender Suchanfrage oder können je nach Bedarf interaktiv ein- und ausgeblendet werden.

---

[19] Für mehr Informationen zu CSS vgl. die Erläuterungen des W3-Consortiums, online verfügbar unter http://www.w3.org/Style/CSS/Overview.en.html. Hinweise zur Verwendung von Stylesheets mit XML-Dokumenten finden sich in einer entsprechenden W3C-Empfehlung, online verfügbar unter http://www.w3.org/TR/xml-stylesheet/.

Durchsuchbarkeit der annotierten Daten

Der größte Vorteil einer elektronischen Repräsentation des Lexikons liegt zweifellos in der späteren Durchsuchbarkeit der annotierten Daten. Mithilfe der *XPath*-Syntax können beliebige Teilmengen in einem XML-kodierten Dokument selektiert und damit gesucht werden. Dabei wird im Wesentlichen ein Pfadausdruck formuliert, der die hierarchische Struktur des XML-Dokuments berücksichtigt. Um etwa aus einem Wörterbuch, dessen einzelne Artikel analog zur in Kap. 6 spezifizierten Dokumentgrammatik erstellt wurden, alle Satzglieder zu identifizieren, die als Temporalangabe annotiert wurden, ist folgender XPath-Ausdruck nötig:

```
/lexikon/artikel/semem/belegsaetze/satz/satzglied[funktion="A"
and semantik="temporal"]
```

Der Ausdruck navigiert, ausgehend vom Wurzelknoten „Lexikon“, durch alle hierarchisch darunterliegenden Kindelemente und selektiert dann all jene Elemente „Satzglied“, deren Kindelement „Funktion“ den Wert „A“ (Angabe) und deren Kindelement „Semantik“ den Wert „temporal“ aufweisen. Das Ergebnis der Anfrage ist eine Liste aller in den Belegsätzen entsprechend annotierten Satzglieder.

Mithilfe der XPath-Syntax können so je nach Forschungsfrage alle vorhandenen Informationen kombiniert und zu einer beliebig komplexen Suchanfrage verknüpft werden. So könnte beispielweise die morphosyntaktische Umsetzung auf Ebene der „Funktion“ für eine konkrete semantische Rolle (z. B. Agens) untersucht werden. Verknüpft mit den Informationen „Hyperlemma“ und Sprachstufe könnten zusätzlich diachrone Phänomene bzgl. der Entwicklung einer semantischen Rolle erfasst werden. Eine weitere diachrone Fragestellung könnte sich etwa mit der Veränderung der syntaktischen Umgebung eines Verbs im Verlauf der Sprachgeschichte beschäftigen. Hierzu werden für jede Sprachstufe die Art und Zahl der Ergänzungen erfasst, um beispielsweise systematische Effekte wie „Zurückdrängen des Genitivs in einem bestimmten Wortfeld“ sichtbar zu machen. Weiterhin kann in einem XML-kodierten Wörterbuch nach konkreten Elementabfolgen gesucht werden, z. B. nach Satzbauplänen (sowohl auf Ebene der Belegsatzanalysen als auch auf Ebene der allgemeinen Valenzanalyse des Verbs). Ebenfalls möglich ist die Untersuchung von Stellungsfeldern, d. h., über eine entsprechende Abfrage kann automatisch erfasst werden, ob z. B. für bestimmte Varietäten oder Sprachstufen Auffälligkeiten in der Besetzung des Vorfelds erkennbar sind.

Die im letzten Abschnitt vorgestellten Fragestellungen illustrieren, wie die verschiedenen elektronisch aufbereiteten Informationselemente eines Lexikonartikels in nahezu beliebiger Kombination abgefragt werden können. Es ergibt sich somit ein wertvolles Analyseinstrument, welches die systematische Identifikation von Valenzphänomenen und Mustern bei historischen Verben ermöglicht.

Literatur

Burghardt, Manuel (2012): Annotationsergonomie: Design-Empfehlungen für linguistische Annotationswerkzeuge. In: Information, Wissenschaft & Praxis, 63, 300–304.

Burghardt, Manuel (2014): Engineering Annotation Usability – Toward Usability Patterns for Linguistic Annotation Tools. Dissertation, Universität Regensburg.

Burghardt, Manuel/Wolff, Christian (2015): Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik. In: Book of Abstracts Workshop „Informatik und die Digital Humanities“. Leipzig.

Greule, Albrecht (1999): Syntaktisches Verbwörterbuch zu den althochdeutschen Texten des 9. Jahrhunderts. Frankfurt a. M. u. a. (= Regensburger Beiträge zur deutschen Sprach- und Literaturwissenschaft B 73).

Greule, Albrecht (2014): Diachrone Perspektiven im Historischen deutschen Valenzwörterbuch. In: Glottotheory 5, 57–63.

Gugerli, David (2009): Suchmaschinen. Die Welt als Datenbank. Frankfurt a. M.

Helbig, Gerhard/Schenkel, Wolfgang (1983): Wörterbuch zur Valenz und Distribution deutscher Verben. 7. Aufl. Tübingen.

Leech, Geoffrey (1997): Introducing Corpus Annotation. In: Roger Garside/Geoffrey Leech/Anthony McEnery (Eds.): Corpus Annotation. Linguistic Information from Computer Text Corpora. Harlow, Essex, 1–18.

Lobin, Henning (2004): Dokumentgrammatiken als Grundlagen von XML-Tools. In: Alexander Mehler/Henning Lobin (Hg.), 23–38.

Mehler, Alexander/Lobin, Henning (Hg.) (2004): Automatische Textanalyse: Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte. Wiesbaden.

Moretti, Franco (2007): Graphs, Maps, Trees: Abstract Models for Literary History. London.

Rapp, Rebecca/Reimann, Sandra (2016): *Gruoz enbieten* und *grüszen*. Ein Beitrag zum Desiderat eines syntaktisch-semantischen mittelhochdeutschen Verbwörterbuchs. In: Peter Ernst/Martina Werner (Hg.): Linguistische Pragmatik in historischen Bezügen. Berlin (= Lingua Historica Germanica 9), 253–271.

Schneider, Roman (2008): E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database. In: Roman Schneider/Bernhard Schröder (Hg.): Sprache und Datenverarbeitung. Duisburg, 35–46.

Schumacher, Helmut u. a. (2004): VALBU. Valenzwörterbuch deutscher Verben. Tübingen (= Studien zur Deutschen Sprache 31).

Sommerfeldt Karl-Ernst/Schreiber, Herbert (1996): Wörterbuch der Valenz etymologisch verwandter Wörter. Verben, Adjektive, Substantive. Tübingen.

Tarvainen, Kalevi (1987): Semantische Kasus im Deutschen unter praxisorientiertem Aspekt. In: Deutsch als Fremdsprache 24, 296–300.

Witt, Andreas (2004): Linguistische Informationsmodellierung mit XML. In: Alexander Mehler/Henning Lobin (Hg.), 39–54.