Tests mit philologisch-linguistischen Verfahren zur Erstellung der Meta-Lemmaliste

Version 2009-10-15
Arbeitspaket AP 5
verantwortlicher Partner: Universität Trier

TextGrid

Modulare Plattform für verteilte und kooperative wissenschaftliche Textdatenverarbeitung -
ein Community-Grid für die Geisteswissenschaften

Bundesministerium
für Bildung
und Forschung

Projekt: **TextGrid**
Teil des D-Grid Verbundes und der deutschen e-Science Initiative
BMBF Förderkennzeichen: 07TG01A-H
Laufzeit: Februar 2006 - Januar 2009

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren:
Stefan Büdenbender, Uni Trier
Werner Wegstein, Uni Würzburg
Andrea Rapp, Uni Trier

Inhaltverzeichnis

1 Ausgangslage

Vorbemerkungen

Der vorliegenden Report steht in engem Zusammenhang mit dem Projekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“, das maßgeblich von TextGrid-Partnern initiiert wurde und sich insofern auch auf Vorarbeiten im Rahmen des AP 5 stützt. Gerade mit dem Hinzustoßen der Würzburger Bioinformatik aber ergab sich in diesem Bereich eine Erweiterung des Spektrums, die die traditionellen Grenzen des Faches sprengt und insofern nur Gegenstand eines Ausblicks sein soll. Auch die abschließenden, im Rahmen von AP 5 zunächst sehr knapp und abstrakt gehaltenen konkreten Arbeitsvorschläge erfuhren im Zuge der Antragstellung eine Ausweitung und Präzisierung, die die ursprüngliche Fassung obsolet machten; in beiden Punkten soll im Folgenden deshalb auf den Antrag verwiesen werden.

1.1 Synchrone und diachrone Varianz im Schnittpunkt von Philologie und EDV

Synchrone wie diachrone Varianzen und Varietäten können als Regelzustand einer Sprache angesehen werden. Ihre Ermittlung, Dokumentation und Analyse ist von Anfang an Gegenstand der Sprachwissenschaften. Allerdings sind schon für die Beschreibung von Sprachdaten, etwa in Wörterbüchern, aber auch für die Analyse von Sprachkorpora feste Bezugspunkte notwendig, die es ermöglichen, Varianz abzubilden und damit zu ordnen. Ein bei der Ausdrucksseite der Sprache ansetzendes – semasiologisches – Konzept beschränkt zwar unvermeidlich die Reichweite des Bezugssystems, gewinnt durch die Fixierung auf die Ausdrucksseite dafür im Bereich der Empirie und Metrie, weil es sicherstellt, dass alle Sprachphänomene ausnahmslos in den Blick genommen werden.

Während gerade die ausdrucksseitige Varianz mittlerweile gut dokumentiert ist, handelt es sich hierbei oft um fragmentarisches Wissen, das, auf sehr enge Anwendungsbereiche bezogen, in zahlreichen, unverbundenen Spezialquellen vorgehalten wird[1]. Eine Bündelung und Zusammenführung stand bisher nicht im Fokus der klassischen Linguistik und Lexikographie bzw. entsprach nicht den vorherrschenden methodischen Ansätzen, obschon es sich hier zweifellos um ein zentrales Desiderat handelt.

Mit dem Einzug der EDV-Technologien in die Philologie und Linguistik haben sich die Voraussetzungen jedoch in mehrfacher Hinsicht geändert, sowohl was den Bedarf, als auch die Möglichkeiten einer Vereinheitlichung angeht.

Einerseits erfordert die maschinelle Analyse und Verarbeitung natürlichsprachlicher Texte eine weitestgehende Reduktion ausdrucksseitiger Varianz z.B. mittels lexikalischer Information oder Transformationsregeln. Andererseits rücken die Bemühungen rund um das Semantic Web wieder den

[1] Das Trierer Wörterbuchnetz vereint bereits 11 retrodigitalisierte Lexika, die einen historischen, dialektalen oder personalen (=Autorenwörterbücher) Fokus haben. Vgl. www.woerterbuchnetz.de und Kapitel 2.1. des vorliegenden Reports.

zweiten Aspekt von Sprache in den Mittelpunkt: ihre onomasiologischen Strukturen bzw. die Bedeutungsebene.

Diese, obwohl im herrschenden Paradigma untrennbar mit dem sprachlichen Ausdruck verbunden, bildet in der klassischen Sprachwissenschaft ein eigenes Untersuchungsfeld, und auch in der Computerlinguistik ist etwa das Verhältnis von Ontologien, die die inhaltlichen Beziehungen von Begriffen zueinander abbilden, zu deren konkreter sprachlicher Realisierung nach wie vor ein kritischer Punkt. Selbst die umfassendsten Wortnetze wie WordNet oder, für das Deutsche, GermaNet sind von ihrer lexikalischen Basis her im Wesentlichen auf eine Auswahlmenge der jeweiligen aktuellen Standardsprache beschränkt.

Während dies eine für kommerzielle Zwecke – etwa im „information retrieval“ – ausreichende Grundlage sein mag, bedeutet es für die Sprach- und Literaturwissenschaft eine wesentliche Beschränkung ihres Untersuchungsgebiets.

Primäre Aufgabe einer Meta-Lemmaliste, wie sie im Folgenden skizziert wird, wird es sein müssen, bereits vorhandene und künftige elektronische Quellen, die die deutsche Sprache in all ihrer Varianz dokumentieren, für einen einheitlichen Zugriff zu erschließen. Ausgehend von den Stichwortansätzen retrodigitalisierter Wörterbücher sowie aktuellen Referenzwortlisten aus der Corpuslinguistik einerseits, und andererseits Transformationsregeln, die nicht zuletzt mit Hilfe von Ansätzen aus der Bioinformatik zu modellieren sind, soll ein Wortnetz geschaffen werden, das unterschiedlichste Ebenen und Instanzen betrifft. So erleichtert es dem Philologen den Einstieg in Spezialquellen – Wörterbücher, aber auch historische oder dialektale Primärtexte – und unterstützt die maschinelle lexikalische Analyse natürlichsprachlicher Texte bis hin zur Lemmatisierung. Darüber hinaus kann es als Bindeglied zwischen Varietät und Standard zur „Andockstelle“ für weitere, bereits existierende standardsprachliche Wortnetze und Ontologien werden und somit einen großen, bisher ausgegrenzten Bereich der deutschen Literatur für Semantic Web-Technologien und darauf aufbauende onomasiologische Fragestellungen öffnen.

Die EDV erschließt in diesem Zusammenhang demnach nicht bloß neue Betätigungsfelder, sie ermöglicht neue Untersuchungsmethoden in einem interdisziplinären Kontext. Die Analyse und Systematisierung evolutionärer Wandlungsprozesse stellt gerade auch in zahlreichen naturwissenschaftlichen Disziplinen eine zentrale Herausforderung dar, an erster Stelle ist in diesem Zusammenhang die Bioinformatik zu nennen. Die Erforschung und Nutzung von Synergien, die die klassisch-philologischen Arbeitsmethoden ergänzen können, ist indes Gegenstand eines eigenen Projekts.

1.2 Semasiologie: Norm und Varianz

Der konzeptuelle Ansatz der Zusammenführung verschiedener Schreibvarianten einer Grundform bzw. eines Etymons steht – jeweils unter eigenen Leitaspekten – im Zentrum der Etymologie und der Dialektologie. Beide Disziplinen basieren auf der Grundannahme eines stetigen Sprachwandels, der sich überall dort, wo sprachliche Ausdrücke keiner strengen Normierung unterworfen sind, im täglichen Gebrauch unweigerlich einstellt, dabei aber immer auch einer gewissen Regelhaftigkeit unterworfen ist. Beide nehmen dabei in unterschiedlicher Weise Bezug auf einen sprachlichen Normzustand. Während die Dialektologie vor seinem Hintergrund die tatsächlich vorhandene Varianz

im Raum zu einem bestimmten Zeitpunkt dokumentiert[2], fragt die Etymologie, sofern sie sich nicht areal einschränken will, nach der zeitlichen Entwicklung eines übergreifenden Standards, der in sich auf jeder historischen Stufe indes schon eine Idealisierung darstellt.

Während die moderne Linguistik zu Recht die grundsätzliche Problematik normativer Ansätze hervorhebt, insofern sie oft mit impliziten Werturteilen bzw. einer unkritischen Vorstellung von „richtigem“ Sprachgebrauch einhergehen, liegt die Notwendigkeit stabiler, referenzierbarer Grundformen überall dort, wo sprachliche Ausdrücke über ihre äußere Form geordnet und systematisiert werden, jedoch auf der Hand.

Eine historisch-dialektale Meta-Lemmaliste wird insofern nicht auf ein semasiologisches Grundgerüst verzichten können, über das die areale Varianz innerhalb einer Sprachstufe auf einer Zeitachse verortet wird.

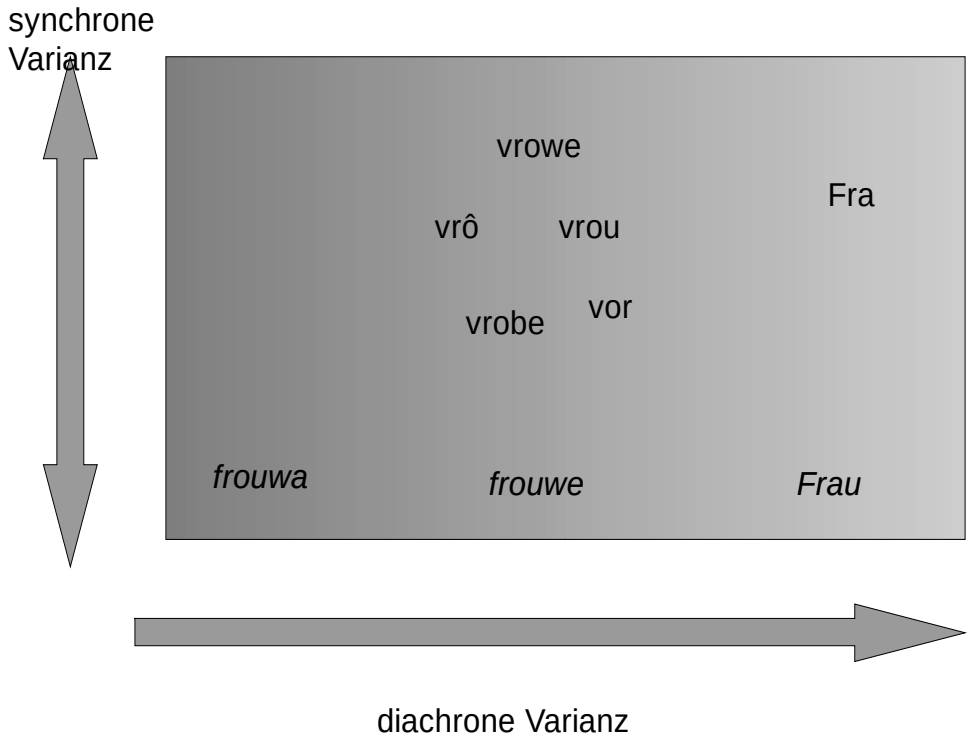

Abb. 1: semasiologisches Grundgerüst

Im Detail ist dabei eine Fülle von Problemfeldern zu bearbeiten, beginnend bei Wörtern, die keine Entsprechung (mehr) im neuhochdeutschen Standard haben (ausgestorbene Wörter, spezifische Dialektwörter mit regional begrenzter Reichweite, Fachsprachen etc.), und für die zur Verknüpfung mit den standardsprachlich üblichen Varietäten-Lemmata ein Konzept von „Pseudo-Meta-Lemmata“ entwickelt werden muss, bis hin zur Frage der Wortfestigkeit von Wortbildungen (z.B. im Bereich

[2] Wobei etwa auch die Stichwortansätze und Belege in klassischen Dialektwörterbüchern immer schon zu einem gewissen Grad Normierungen darstellen (s.u.) - im gedruckten Medium lässt sich nur eine begrenzte Granularität sinnvoll darstellen.

Komposition oder Konversion), dem Wortbegriff (z.B. Homonymie versus Polysemie) und dem Problemfeld Semantik. Als Beispiel für die Notwendigkeit solcher Pseudo-Meta-Lemmata kann etwa das ahd. Verb *jehan*, mhd. *jehen* 'sagen, sprechen, bekennen' dienen, das im Nhd. zwar in Wortbildungen wie nhd. *Beichte* oder *Gicht* noch enthalten, als eigenständiges Verb jedoch ausgestorben ist. Für solche Fälle müssen philologische „Varianz-Regeln“ definiert und Verfahren gefunden werden, die die Abbildung der Varietäten-Lemmata auf das Meta-Lemma ermöglichen.

Die Lemma-Ansätze der Stichworteinträge in Wörterbüchern sind varietätenspezifisch, also beispielsweise sprachstufenspezifisch (ahd. *frouwa*, mhd. *vrouwe*, nhd. Frau) oder dialektspezifisch (*Fra* im Lothringischen Wörterbuch) und repräsentieren in der Regel bereits eine gewisse Normierungsstufe, die sich signifikant von der „Textwirklichkeit“ mit einer weitaus höheren Varianz abhebt: für nhd. Frau z.B. *vrowe, vrobe, vrou, vrô, vor, vuor, ver, vir, vür froï, froïw* usw. (vgl. Abb. 1). Diese Varianz kann z.B. für phonologisch-graphonematische, morphologische und lexikalische Analysen des Wortschatzes von Texten, Wörterbüchern, aber auch Korpora, über eine Meta-Lemmaliste ausgeglichen werden, ohne dass dadurch das sprachliche Basismaterial verfälscht wird.

1.3 Onomasiologie

Dialektale und historische, auch sprachstadienübergreifende Lexika wie etwa das „Deutsche Wörterbuch“ bilden die Grundlage zahlreicher vor allem semasiologisch orientierter Untersuchungen. Fragen nach der inhaltlichen Seite sprachlicher Ausdrücke standen dabei insgesamt wesentlich seltener im Mittelpunkt als die nach ihrer formalen Entwicklung. Es darf jedoch bezweifelt werden, dass das einer mangelnden Ergiebigkeit des Gegenstands geschuldet ist; vielmehr scheint hier das Medium zu einem gewissen Grad das Forschungsinteresse zu präfigurieren.

Vor dem Hintergrund des Semantic Web, das auf eine inhaltliche Strukturierung elektronischer Ressourcen abzielt, ist dieses Verhältnis zumindest in Projekten, die im Bereich der EDV angesiedelt sind, neu zu bewerten. Wie eingangs bereits skizziert, stehen in Form von Thesauri und Wortnetzen, aber auch Taxonomien und Ontologien eine Reihe Ressourcen und Technologien zur Verfügung, die jedoch fast ausschließlich auf den neuhochdeutschen Standard fokussiert sind.

Im Kontext einer historisch-arealen Meta-Lemmaliste zeichnen sich hier in zwei Bereichen Berührungspunkte bzw. Überlappungen ab: einerseits ist es sinnvoll, existierende Ressourcen an die Liste „anzudocken“, andererseits können Methoden zur Beschreibung, aber auch zur Erschaffung solcher Ressourcen auf bisher unausgewertetes Rohmaterial angewendet werden.

Der neuhochdeutsche Standardansatz fungiert im ersten Fall als gemeinsames Drittes, über das die Verbindung zwischen beiden Bereichen hergestellt wird; doch erscheint es sinnvoll, onomasiologisch strukturierte Quellen wie etwa Thesauri und Wortnetze nicht von vornherein auf einen solchen, nachträglichen Anschluss einzuschränken. Unten (vgl. Abb.3) soll an einem komplexeren Beispiel gezeigt werden, wie sie bereits die Zuordnung von Schreibformen zum Standardansatz steuern können.

2 Meta-Lemmaliste in TextGrid

Wie unten auszuführen sein wird, bringen die TextGrid-Partner verschiedene Ressourcen ein, die quantitativ, qualitativ und in Bezug auf die sprachliche Vielfalt, die sie dokumentieren, ein äußerst ergiebiges Forschungsgebiet bilden.

Ziel des Arbeitspakets 5 ist es u. a., Tools und Konzepte zu entwickeln, die einerseits die Erschließung dieser Sprachdaten ermöglichen, andererseits das in ihnen hinterlegte Wissen für die Erschließung neuer Quellen nutzbar machen. Ein zentraler Ansatz sah dabei einen Linkeditor vor, der es ermöglichen sollte, die Stichwortansätze der digitalen Wörterbücher mit Hilfe von GermaNet zu einem diachronen Wortnetz zu verbinden, was sich jedoch aus lizenzrechtlichen Gründen nicht realisieren ließ. Während einzelne Aspekte durch andere Ressourcen abgefangen werden konnten (so soll etwa bei der Wörterbuchverknüpfung OpenThesaurus mit eingebunden werden), wurde der Bedarf nach einer offen zugänglichen Meta-Lemmaliste immer deutlicher, die in- aber auch außerhalb von TextGrid, beispielsweise in einem Semantic Web-Umfeld, einzusetzen wäre.

Ebenso offensichtlich war, dass die Schaffung einer solchen Meta-Lemmaliste sowohl vom Aufwand als auch vom interdisziplinären Ansatz her nur im Rahmen eines eigenständigen Projekts zu bewerkstelligen sein würde. Aufbauend auf einer Reihe gezielter Tests mit den vorhandenen Datenbeständen und einer Evaluierung klassischer wie auch interdisziplinärer Methoden im Rahmen des Arbeitspakets 5 wurde deshalb ein Projektantrag beim BMBF eingereicht; das Projekt startete am 1.10.2008.

Das eingangs bereits erwähnte Projekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“ (im Folgenden abgekürzt als „Wechselwirkungen“) ist dabei auf eine enge Kooperation mit TextGrid ausgelegt; einerseits werden die in AP5 entwickelten generischen Tools zur Recherche, Verknüpfung und Annotation von Wörterbucheinträgen eine wichtige Unterstützung darstellen, andererseits werden die zurückfließenden Ergebnisse eine substanzielle Bereicherung für TextGrid sein.

2.1 Ressourcen

Mit dem Kompetenzzentrum Trier, dem Institut für Deutsche Sprache Mannheim und der Universität Würzburg sind in TextGrid drei Institute vertreten, die über ebenso umfangreiche wie qualitativ hochwertige Datensammlungen sowie eine langjährige Erfahrungen im Umgang mit solch hochstrukturierten Daten verfügen.

Im Folgenden soll ein Überblick über die eingebrachten Daten und methodischen Kompetenzen gewährt werden, die die in TextGrid und dem Projekt „Wechselwirkungen“ vertretenen Partner einbringen:

Das Trierer Kompetenzzentrum hat in verschiedenen Projekten zur Digitalisierung und Erschließung von Referenzwerken und Primärquellen Modelle und Methoden entwickelt, die entscheidende Vorarbeiten für die Erstellung einer Meta-Lemmaliste liefern; neben den Wörterbuch-Daten selbst wurden Verfahren zur Vernetzung etabliert, die auch im Zentrum der Untersuchungen in AP5 standen. Im Gegensatz zu einer bloßen, unverbundenen Bereitstellung verschiedener Nachschlagewerke hat das Trierer Kompetenzzentrum im internationalen Vergleich erstmalig und mit großem Erfolg Wörterbuchverbünde mit multidirektionalen Verlinkungen entwickelt.[3]

[3] Vgl. Burch/Rapp: „Das Wörterbuch-Netz“.

Im Falle der „Mittelhochdeutschen Wörterbücher im Verbund“ wurde die multidirektionale Vernetzung anhand einer bereits jetzt schon vorhandenen standardisierten mittelhochdeutschen 'Meta-Lemmaliste' aller Stichwörter aus den vier eingebundenen Werken realisiert.

Auch im „Digitalen Verbund von Dialektwörterbüchern“ konnte durch eine reichhaltige, semasiologische wie onomasiologische Artikelverknüpfung innerhalb des Verbundes, aber auch mit dem DWB und den „Mittelhochdeutschen Wörterbüchern im Verbund“, bereits ein entscheidender Mehrwert erzielt werden. Hier ist der Bedarf nach einer Meta-Lemmaliste besonders augenfällig, da die betreffenden Dialekte einerseits durch reiche Synonymie und Heteronymie gekennzeichnet sind, die Wörterbücher andererseits aber in ihrer Systematik variieren: so sind die Lemmaansätze im Pfälzischen und Rheinischen Wörterbuch am hochdeutschen Standard orientiert, die des „Wörterbuchs der elsässischen Mundarten“ und des „Wörterbuchs der deutsch-lothringischen Mundarten“ jedoch dialektal.

Mit der digitalen Erstbearbeitung des „Deutschen Wörterbuchs von Jacob und Wilhelm Grimm“ (DWB) steht ein sprachstadienübergreifendes Wörterbuch zur Verfügung, das die zu schaffende Meta-Lemmaliste des Deutschen um das Frühneuhochdeutsche und Neuhochdeutsche erweitern wird. Demgegenüber bringen Autorenwörterbücher weitere Spezifizierungen: Hier steht das Goethe-Wörterbuch zur Verfügung, dessen elektronische Aufbereitung derzeit mit DFG-Förderung in Trier erfolgt und das mit den Einträgen im DWB verknüpft wird.

Anhand dieser Wörterbücher zu den verschiedenen Varietäten des Deutschen kann ein standardisierter Einstieg für effiziente Recherchen geschaffen werden, von dem aus sich die nicht standardisierten, heterogen organisierten und verstreuten Informationen verzweigen; heterogen im Hinblick auf synchrone wie diachrone Varianz. Dieses Konzept lässt sich nicht nur auf Wörterbücher anwenden, sondern auch auf die Erschließung großer Korpora übertragen.

Das Institut für Deutsche Sprache in Mannheim baut seit 1964 Korpora der deutschen Gegenwartssprache auf. Derzeit verfügt das Institut über die mit mehr als 3,2 Milliarden laufenden Textwörtern weltweit größte Sammlung deutschsprachiger Texte für die sprachwissenschaftliche Forschung – das Deutsche Referenzkorpus (DEREKO) und über eine Reihe von DEREKO-basierten Referenzwortlisten (DEREWO).

Auch die Universität Würzburg kann auf langjährige Erfahrungen und umfangreiche philologisch zuverlässige Datenbestände zurückgreifen. Dazu gehören die elektronische Edition des Campe-Wörterbuchs, die im Rahmen des TextGrid-Projekts geleistet wird (Wörterbuch der Deutschen Sprache. Veranstaltet und herausgegeben von Joachim Heinrich Campe. 5 Bdd. Braunschweig 1807-1811; Ergänzungsband 1813) sowie ein vollständiges Wörterbuch zu Klingemanns Roman „Nachtwachen”, verfasst unter dem Pseudonym Bonaventura (vgl. Arnold: „Klingemann”). Verfahren der Umkehrlexikographie wurden erstmals intensiv erprobt am Neuhochdeutschen Index zu den mittelhochdeutschen Wörterbüchern, der den ersten vollständigen Versuch einer Erschließung des mittelhochdeutschen Wortschatzes über die neuhochdeutschen Bedeutungsbeschreibungen darstellt (Erwin Koller/Werner Wegstein/Norbert Richard Wolf: Neuhochdeutscher Index zum mittelhochdeutschen Wortschatz, Stuttgart 1990).
Der Index erschließt den mhd. Wortschatz des mhd. Taschenwörterbuchs, das Matthias Lexer im Vorwort der letzten, von ihm überarbeiteten Auflage „als ein supplement und korrektiv des

[mittelhochdeutschen] Handwörterbuchs und im ganzen [...] auch als ein repertorium des dermaligen mittelhochdeutschen sprachschatzes“ kennzeichnet (Matthias Lexer: Mittelhochdeutsches Taschenwörterbuch, Leipzig 1885, Nachdruck Stuttgart 1992, S. XV). Die Daten der Druckfassung von 1990 wurden 2009 für die weitere Arbeit an dem Projekt auf dem Opus-Server der Universitätsbibliothek Würzburg öffentlich verfügbar gemacht (URN: urn:nbn:de:bvb:20-opus-35530; URL: http://www.opus-bayern.de/uni-wuerzburg/volltexte/2009/3553/). Das Erschließungskonzept des Umkehrwörterbuchs wurde gegenüber der Druckfassung von 1990 für die TextGrid-Nutzung grundlegend überarbeitet und auf eine verbesserte Nutzbarkeit getestet. Das nhd. Index-Lemma „Abgabe“ kann die Prinzipien veranschaulichen, nach denen im Index die mhd. Entsprechungen den neuhochdeutschen Lemmata zugeordnet sind:

Abgabe

[1] bërn, bëte, dienest, dienestgëlt, gâbe, geschôz, gewërf,_gewërft, hëlfe, ingëlt, nôtbëte, nôtstiure, phlëge, phlëgenisse, schatzunge, schoz, stiurunge, taz, ûfsatzunge, ûfsaz, zins · gevelle · lantkoste

[2] anval, *bercreht*, burcveste, collecte, dëchgëlt, dëhem, dëchtuom, forstrëht, fratz, gebürnisse, gerëhtecheit, gruntrëht, *hanse*, hantlôn, hantloese, holzloese, holzrëht, honecgëlt, huobegëlt, huoberëht, kamerschaz, kappengëlt, kappengülte, kappenzins, kirchengift, kirchloese, klagegëlt, klâstiure, kornungëlt, kramzol, küchenbëte, küchenstiure, *küchendienest*, kurmiete, lantrëht, lîpbëte, lôsunge, marcrëht, marketrëht, mêdeme, meisterrëht, muntschaz, münzgëlt, oblei, phahte, *râtschaz*, rëgelgëlt, rëgelphenninc, rîbegërste, rîbekorn, rouch, rouchval, schenkrëht, seilrëht, selderëht, slegeschaz, slageschaz, slahschaz, slahgelt, slozgëlt, slozrëht, stantgëlt, stalgëlt, stalmiete, statrëht, *stegegelt*, stëgrëht, stocrëht, swërtstiure, teverîe, turnloese, überdienest, überzins, ûfvart, ungëlt, umbegëlt, unpfliht, val, versenphenninc, vihestiure, *vischlêhen*, vridephenninc, vrideschaz, vrîrëht, vrônkost, vuorwîn, vürdinc, vürgedinge, vürvart, wagenleite, wahte, wahtgëlt, wahtphenninc, waltgëlt, waltrëht, watschar, wëgeloese, widemgëlt, wînban, wîngartstiure, wînkouf, wisegëlt, wîsôt, wîsunge, wîsoede, wochengëlt, wüestgëlt, zëhende, zëhente, zëhent, ziegelstiure, zinsphenninc, zol, zuoganc, zuoval · burgerrëht, *mit_stiure_unt_mit_bete*, riutegëlt, riute, riutzëhende.

Direkte mhd. Äquivalente des nhd. Lemmas werden in Kategorie [1] verzeichnet, wobei ein hochgestellter Punkt solche mhd. Belege markiert, die einer bestimmten Wortform des nhd. Lemmas entsprechen, im Beispiel etwa „gevelle“ mit der nhd. Bedeutungsangabe „abgaben“ im Plural, ähnlich adverbial verwendete Substantive (mhd. *abenthalben* = „am abend“) oder partizipiale Verbformen. Die zweite, durch [2] markierte Kategorie von mhd. Äquivalenten liefert mhd. Entsprechungen, bei denen das jeweilige nhd. Stichwort mit einer zusätzlichen attributiven Spezifizierung versehen ist. So erscheint z.B. mhd. *anval* „abgabe des erben für eine verleihung des hofes“ unter dem Stichwort „Abgabe“ in Kategorie [2] ebenso wie etwa mhd. *burcveste* „eine an die *burc* zu leistende abgabe“. Für den Aufbau der Metalemmaliste sind hier, abweichend von der gedruckten Ausgabe, auch die mhd. Äquivalente eingeordnet, in denen das nhd. Index-Lemma in einer morphologisch oder semantisch verwandten Form erscheint, also etwa in Komposita, Präfixbildungen oder Zusammenbildungen, im Falle von „Abgabe“ sind das die nhd. Index-Lemmata „aufenthalts-, fisch-, geld-, handels-, landes-, zwangs-abgabe“. Insgesamt handelt es bei dem mhd. Wortschatz der

Kategorie [2] im wesentlichen um mhd. Unterbegriffe zu einem nhd. Index-Lemma, das als Oberbegriff zu verstehen ist. Da in der lexikographischen Praxis Bedeutung begrifflich häufig in den Strukturen von *genus proximum* und *differentia specifica* definiert wird, bildet der Index über diese Kategorie zu vielen (nhd.) Stichwörtern wesentliche Bereiche des mhd. Wortfeldes ab und bietet damit eine tragfähige Grundlage für den Aufbau einer verfeinerten Ontologie zum Mhd. und die Verknüpfung von Metalemma-Strukturen über Sprachperioden des Deutschen hinweg.

Die mhd. Stichwörter aus dem Nachtrag zum Taschenwörterbuch von Pretzel wurden ausgesondert: *bercreht, hanse, küchendienest, râtschaz, stegegelt, vischlêhen,* zusammen mit der Phrase *mit stiure unt mit bete*. Denn davon stammen die Begriffe *bercreht¸ stegegelt* sowie ohne nhd. Bedeutungsangabe *râtschaz* und *vischlêhen* allesamt aus dem Lexerschen Handwörterbuch. Zu *küchendienest* wird nur die Bedeutungsangabe „= *küchenstiure*“ mitgeteilt und die Phrase *mit stiure unt mit bete* ist schon in Lexers Basiswörterbuch von Benecke/Müller/Zarncke für Hartmanns von Aue Roman „Der arme Heinrich“ nachgewiesen. Das Ergebnis der Tests innerhalb der Buchstabenstrecke A führt zu dem Schluss, das Indexkonzept des Umkehrwörterbuchs auf die Daten von Lexers Handwörterbuch anzuwenden und die Pretzelschen Nachträge vollständig zu auszuschließen, weil sie entweder aus dem Lexerschen Handwörterbuch übernommen sind oder, falls nicht, ohnehin nicht an Textbelegen verifiziert werden können. Diese Arbeit soll im Rahmen des BMBF-Projekts "Wechselwirkungen" weitergeführt werden, in das die Erstellung der Meta-Lemmaliste ausgegliedert wurde.

Ebenso wird die Kategorie [3] der Druckfassung zunächst ausgesondert, die unter dem neuhochdeutschen Lemma alle mhd. Äquivalente sammelt, zu deren Erklärung das nhd. Stichwort in einem sehr weiten Sinn in attributiver Funktion, also als eine Art von Bedeutungsmerkmal verwendet ist. Ob die Bearbeitung und Auswertung dieser Index-Kategorie soll zusätzlichen Aufschluss für die Metalemmaliste und die Ontologie liefert, soll erst in einem späteren Projektstadium untersucht werden.

Die Dialekt-Datenbank BayDat schließlich liefert Möglichkeiten für Tiefenanalysen in der Fläche: Es handelt sich um eine Oracle-Datenbank, die die Erhebungsdaten aus dem DFG-Projekt 'Bayerischer Sprachatlas' nachweist: Sie umfasst 1613 Orte in Bayern, die mit Fragebüchern von durchschnittlich 2.500 Fragen exploriert wurden. Für die unterfränkischen Daten liegen dazu auch Tonbandaufnahmen vor.

Einen weiteren zentralen Aspekt stellen in methodischer Hinsicht die gemeinsamen Arbeiten von Prof. Seipel und Prof. Wegstein dar, die sich auf der Basis des oben genannten Materials unterschiedlichen Problemen der weiteren Erschließung und Bearbeitung anhand eines inter- und transdisziplinären Ansatzes am Schnittpunkt von (Bio)Informatik und EDV-Philologie widmen.

Die bioinformatischen Analysen können auf öffentlich zugängliche Daten zugreifen. Für die Analyse der Varianz in einer Spezies kann auf dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) zurückgegriffen werden. Diese Datenbank enthält mehr als 34 Millionen Basenunterschiede für das menschliche Genom. Um Varianzen zwischen Genomen zu untersuchen, wird Ensembl verwendet werden (http://www.ensembl.org), die Standarddatenbank zur Speicherung von Daten über eukaryontische Genome. Die Analyse von Domänen wird auf Pfam (http://www.sanger.ac.uk/Software/Pfam/) und SMART (http://smart.embl-heidelberg.de) beruhen. Wichtig für das Projekt ist, dass alle Daten frei zur Verfügung stehen und vollständig heruntergeladen werden können.

2.2 Konzepte, Tests, Verfahren

Im Rahmen von AP5 galt es, im Zuge umfangreicher Tests und einer Evaluation der oben angeführten Ressourcen – Datengrundlagen wie Methodenwissen – ein einheitliches Konzept sowie prototypische Tools und Datenstrukturen für die Erstellung einer Meta-Lemmaliste zu entwicklen, die den in Kapitel 1 skizzierten Anforderungen gerecht wird.

Zu klären war einerseits, inwieweit bereits implementierte oder beschriebene Methoden das Projekt unterstützen können bzw. wo sie weiterentwickelt werden können und müssen.

2.2.1 Automatische Verknüpfung

Einen ersten zentralen Untersuchungspunkt bildeten die im Rahmen des Wörterbuchnetzes entwickelten Verfahren zur automatischen Verknüpfung von Lexikoneinträgen. Derartige Verknüpfungen sind geeignet, ein reiches Beziehungsnetz zwischen inhaltlich verwandten Einträgen zu schaffen, selbst wenn die zugehörigen Stichwortansätze etymologisch nicht verwandt sind. Eine Analyse der Ergebnisse zeigte, dass derartige Ansätze durchaus als zentraler Baustein für eine onomasiologische Vernetzung dienen können, gleichzeitig aber in methodischer Hinsicht an mehreren Stellen Optimierungspotenzial besteht. Des weiteren traten bei den Tests die Limitierungen derzeitiger Standard-Hardware deutlich zu Tage: die statistische Auswertung von über einer halben Million Wörterbucheinträge stellt Ansprüche an Rechenleistung und Speicherkapazität, die derzeit von einem Einzelplatzrechner nicht sinnvoll zu bewältigen sind. Die Tests wurden deshalb auf einer stark verkleinerten Datenmenge durchgeführt. Dies ist auch insofern hervorzuheben, als sich hier ein Szenario für ein Computinggrid in den Geisteswissenschaften ergibt: da die anfallenden Berechnungen gut parallelisierbar sind, ließe sich auf der Grundlage eines CPU-Grids eine Testumgebung aufbauen, in der sich Änderungen in den Verknüpfungsalgorithmen zeitnah nachverfolgen und auswerten ließen.

Ein besonderes Augenmerk bei der Analyse lag auf folgenden Punkten:

Stemming: Durch ein Stemming der Wörterbucheinträge (Definitions- und Belegteil) konnten im Rahmen der Tests signifikante Verbesserungen erzielt werden, da so eine wesentlich präzisere Gewichtung der Wortformen innerhalb der Vektoren möglich wurde, die die Grundlage für alle weiteren Berechnungen bilden. Stichprobenartige Untersuchungen weisen darüber hinaus darauf hin, dass durch die Zusammenführung verwandter Wörter innerhalb der Vektoren mittels Thesauri weitere Verbesserungen erzielt werden könnten, ein Ansatz, der im Rahmen des unten skizzierten Projekts weiter zu verfolgen sein wird.

Stoppwortlisten: Gerade bei kurzen Wörterbucheinträgen können systematische Eigenheiten des jeweiligen Lexikons (häufige Verweise auf allgemeine Quellen, Domänenangaben etc.) die statistische Auswertung verzerren. Individuell auf die Grundlage zugeschnittene Stoppwortlisten können dieses Problem weitestgehend eliminieren, besonders wenn sie mit einer Verfeinerung des Markups kombiniert werden.

Verfeinerung der XML-Auszeichnung: Eine feinere Granularität der Datengrundlage erlaubt es, bestimmte Teile der Artikelstruktur (etwa Definitions- und Belegteil) stärker zu gewichten. Besonders im Falle ausgedehnter Nestartikel ist eine stärkere Differenzierung von großer Bedeutung. Abgesehen von der Makrostruktur können so auch einzelne Elemente (z.B. lateinische Definitionen) individuell gewichtet werden.

Ein konkreter Schritt zum Ausbau der automatischen Vernetzung wurde mit der Entwicklung eines generischen Wörterbuch-Linkeditors unternommen, der derzeit in TextGrid-Lab integriert wird. Dieser Editor ermöglicht es dem Bearbeiter, die Verknüpfungsergebnisse systematisch zu durchsuchen und ggf. zu annotieren. Dabei können wahlweise semantische Relationen oder eine allgemeine Qualitätseinschätzung verzeichnet werden. Auch bei diesem Tool bietet sich die Einbindung weiterer externer Quellen wie z.B. OpenThesaurus an.

2.2.2 Umkehrlexikografie

Einen weiteren zentralen Baustein gerade auch zur Sprachstadien-übergreifenden Verknüpfung bildet die Umkehrlexikographie, bei der Stichwortansatz und Definition gegeneinander vertauscht werden. Im Rahmen des AP5 wurde ein solches prototypisches Umkehrlexikon aus den SGML-Daten des BMZ (Georg Friedrich Benecke, Wilhelm Müller und Friedrich Zarncke: Mittelhochdeutsches Wörterbuch [...], Stuttgart, 1990) erstellt und in die Wörterbuchsuche des TextGrid-Lab eingebunden. Dabei wurde eine Reduktion auf direkte Entsprechungen vorgenommen (Einwort-Definitionen) und auf eine weitere Aufbereitung der Wörterbuchdaten verzichtet, da mit dem „Neuhochdeutschen Index zu den mittelhochdeutschen Wörterbüchern“ (s.o.) bereits ein aufwendiges, philologisch geprüftes Umkehrlexikon zu den drei verfügbaren Mittelhochdeutschen WBB vorliegt, das nach den strukturellen Vorarbeiten nun jederzeit in das Lab eingebunden werden kann und auch bei der Erstellung der Meta-Lemmaliste eine wichtige Rolle spielen wird.

2.2.3 Transformationsregeln

Die Zuordnung etymologisch verwandter Stichwortansätze kann durch die Definition von Transformationsregeln und die Konstruktion vereinfachter Grundformen, die systematische Eigenheiten des Druckwerks ausgleichen, maßgeblich unterstützt werden. Im Kooperation mit dem Projekt „LexicoLux“, das u.a. eine Meta-Lemmaliste für die drei gedruckten luxemburgischen Wörterbücher erstellt, deren Entstehungszeitraum sich über mehr als ein Jahrhundert erstreckt, konnte ein Verfahren entwickelt werden, das einen großen Prozentsatz der Stichwortansätze zueinander in Beziehung setzt.

```
<meta2>agin</meta2><g> <meta1>A(n)-/a(n)-*-gin</meta1><type="abbr"><orig>A(n)-/a(n)-*-gin</orig>
<meta2>agin</meta2><m> <meta1>agin</meta1><type="lemma"><orig>a-gin</orig>
<meta2>agio</meta2><g> <meta1>Agio</meta1><type="lemma"><orig>Agio</orig>
<meta2>agiotage</meta2><u> <meta1>Agiotage</meta1><type="lemma"><orig>Agiotage</orig>
<meta2>agnes</meta2><g> <meta1>Agnes</meta1><type="lemma"><orig>Agnes</orig>
<meta2>agoen</meta2><g> <meta1>A(n)-/a(n)-*-goen</meta1><type="abbr"><orig>A(n)-/a(n)-*-goen</orig>
<meta2>agoen</meta2><g> <meta1>a(n)/A(n)--*-goen</meta1><type="abbr"><orig>a(n)/A(n)--*-goen</orig>
<meta2>agoen</meta2><m> <meta1>agoen</meta1><type="lemma"><orig>a-g&ocirc;en</orig>
<meta2>agoen</meta2><m> <meta1>agoen</meta1><type="lemma"><orig>&acirc;-g&ocirc;en</orig>
<meta2>agoen</meta2><u> <meta1>Agoen</meta1><type="lemma"><orig>A'goen</orig>
<meta2>agraffe</meta2><u> <meta1>Agraffe</meta1><type="lemma"><orig>Agraffe</orig>
<meta2>agräifen</meta2><g> <meta1>A(n)-/a(n)-*-gräifen</meta1><type="abbr"><orig>A(n)-/a(n)-*-gräifen</orig>
<meta2>agräifen</meta2><m> <meta1>agräifen</meta1><type="lemma"><orig>a-gr&Overline;ei&overline;fen</orig>
<meta2>agraire</meta2><u> <meta1>Agraire</meta1><type="lemma"><orig>Agraire</orig>
<meta2>agreabel</meta2><g> <meta1>agreabel</meta1><type="lemma"><orig>agreabel</orig>
<meta2>agreff</meta2><g> <meta1>A(n)-/a(n)-*-grëff</meta1><type="abbr"><orig>A(n)-/a(n)-*-grëff</orig>
<meta2>agreff</meta2><m> <meta1>Agrëff</meta1><type="lemma"><orig>A-gr&ouml;ff</orig>
<meta2>ägrett, kleng</meta2><g> <meta1>Ägrett, kleng</meta1><type="lemma"><orig>Ägrett, kleng</orig>
<meta2>agruewen</meta2><g> <meta1>A(n)-/a(n)-*-gruewen</meta1><type="abbr"><orig>A(n)-/a(n)-*-gruewen</orig>
<meta2>agruewen</meta2><m> <meta1>agruewen</meta1><type="lemma"><orig>a-gru&Super;e&super;wen</orig>
```

Abb. 2: Originalform und zwei unterschiedlich stark normierte (Pseudo)-Metaformen. Farbig hervorgehoben das Kürzel des Ursprungswörterbuchs.

Dabei kam eine Kombination von Transformationsregeln zum Einsatz, die individuell auf die Datengrundlage zugeschnitten war und u.a. die Vereinheitlichung bzw. Vereinfachung von Akzentcodierungen, die Auflösung stellungsbedingter Varianten (z.B. „Eifeler Regel“) und die Berücksichtigung bestimmter historisch bedingter Prozesse sprachlichen Wandels beeinhalteten. In einem größeren Projektrahmen, wie er unten skizziert wird, sollte vor allem letzterer Aspekt ausgebaut werden, indem die in der historischen Sprachwissenschaft ausführlich dokumentierten Gesetze des Lautwandels verstärkt einbezogen werden. Dabei ist es durchaus sinnvoll, auch Methoden zum Einsatz zu bringen, die über das Feld der traditionellen Philologien hinausgehen.

2.2.4 Bioinformatik

Die Verbindung von Sprachwissenschaft und Bioinformatik mag nicht nur dem Laien auf den ersten Blick wenig evident erscheinen, dennoch stehen beide Fächer schon seit längerem in einem fruchtbaren Austausch miteinander. So haben in der Vergangenheit zum Beispiel mit Hidden Markov Modellen (HMMs) und probabilistischen, kontextfreien Grammatiken (PCFGHs) linguistische Modelle erfolgreich in der Biologie Einzug gehalten, im Gegenzug hat sich die Anwendung phylogenetischer Methoden auf Sprachdaten als nutzbringend erwiesen. Auch die oben skizzierten Verfahren zur statistischen Auswertung und Vernetzung von Wörterbuchartikeln bedienen sich z.T. einer Methodik, wie sie auch in der Bioinformatik zur Anwendung kommt. Die strukturellen Gemeinsamkeiten (etwa das Verhältnis von Domänen und Proteinen auf der einen Seite zu Morphemen und Wörtern auf der anderen), die einen solchen Transfer erst möglich machen, können an dieser Stelle nicht im gebührenden Umfang ausgeführt werden. Sie versprechen ein großes Potenzial für den Einsatz moderner EDV-Technologie bzw. starke Synergieeffekte, sprengen aber vollkommen die Grenzen derjenigen germanistischen Fachteile, die klassisch an dem Konzept „Meta-Lemmaliste“ interessiert sind.

Hier sei auf dem Antrag zum Projekt „Wechselwirkungen“ verwiesen, in den die ausgiebige Erfahrungen eingeflossen sind, die die Würzburger Arbeitsgruppen von Prof. Wegstein und Prof. Seipel im Rahmen einer engen Zusammenarbeit in den vergangenen Jahren mit einem solchen Methodentransfer sammeln konnten.

3 Ausblick: Erstellung der Meta-Lemmaliste – Projektskizze

Im Rahmen des AP 5 wurden ursprünglich erste Arbeitschritte zur Erstellung einer Meta-Lemmaliste vorgeschlagen, die dann im Zuge der Antragstellung wesentlich detaillierter ausgearbeitet wurden. Ein kursorischer Überblick über die konkrete Arbeitsplanung des Projekts folgt unten, an dieser Stelle ist zunächst noch auf ein wesentliches Desiderat hinzuweisen:

Obwohl die Notwendigkeit und der zu erwartende Nutzen einer Meta-Lemmaliste innerhalb der Wissenschaftsgemeinde häufig betont wird und in diesem Zusammenhang auch relativ konkrete Vorstellungen von ihrer möglichen Gestalt geäußert werden, gibt es bisher keine theoretisch fundierte, abschließende Definition des Konzepts „Meta-Lemma“, geschweige denn ein entsprechendes Datenmodell. Im Rahmen des Projektantrags wurde ein erstes Verfahren vorgeschlagen, aus dem sich bereits eine grobe Struktur ableiten lässt:

Danach werden Stichwortansätze über (halb)automatische Verfahren (Transformationsregeln u.a.) auf einen nhd. Standard abgebildet; bei Bedarf kann eine Disambiguierung über den Abgleich des Kontextes erfolgen. (Definitionsteil – Referenzkorpus). Über die Standardform können dann weitere Quellen wie Thesauri und Wortnetze angedockt werden, auch hier soll ein zusätzlicher Abgleich geschehen.

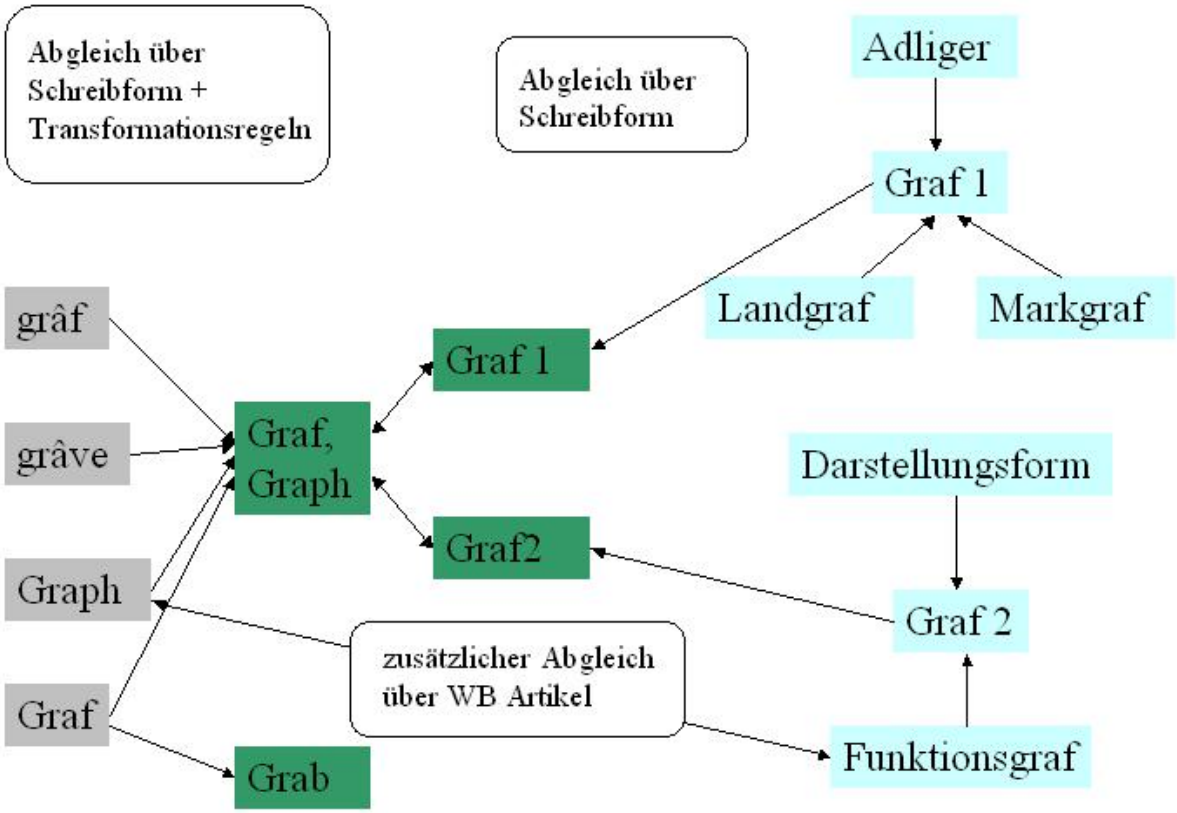

Abb.3: Datenstruktur Meta-Lemmaliste

Eine der ersten Aufgaben des Projekts „Wechselwirkungen“ wird es sein, diese Vorschläge zu evaluieren und ggf. auszubauen. Sodann gilt es, in einem interdisziplinären Diskurs das Konzept „Meta-Lemma“ in einer Form zu präzisieren, die es erlaubt, seine linguistische Komplexität auf eine Datenstruktur abzubilden, die den EDV-technischen Ansprüchen gerecht wird.

Diese Struktur wird die Grundlage für die folgenden Projektarbeiten bilden, die drei inhaltliche Schwerpunkte haben sollen:

- Erstellung einer Basis-Lemmaliste der neuhochdeutschen Standardsprache
- Erstellung Klassifizierter Varietäten-Lemmalisten
- Modellierung einer „Grammatik der Varianz“, Visualisierung, Gridifizierung

Für Details sei erneut auf den Antrag des Projekts „Wechselwirkungen“ verwiesen bzw. auf dessen Homepage unter http://www.sprache-und-genome.de/.

Anhang A: Bibliographie

Burch, Thomas und A. Rapp. Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In: Historisches Forum 10 (2007) Bd. 1, S. 607–627. URL: http://edoc.hu-berlin.de/histfor/10_I/PDF/Woerterbuecher_2007-10-I.pdf [Zugriff am 30.06.2009].

Das Trierer Wörterbuchnetz. URL: www.woerterbuchnetz.de [Zugriff am 30.06.2009].

Koller, Erwin, Werner Wegstein und Norbert Richard Wolf: Neuhochdeutscher Index zum mittelhochdeutschen Wortschatz. Stuttgart, 1990.

Projekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“. URL: www.sprache-und-genome.de [Zugriff am 15.10.2009].

Barbara Arnold: Lexikographische Studien zu August Klingemann (Diss. University of Exeter 2004).