INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

ISSN 1860-9422

2/2016

Im Auftrag des Instituts für Deutsche Sprache herausgegeben von
Hardarik Blühdorn, Mechthild Elstermann und Doris Stolberg

Luise Borek, Andrea Rapp (Hg.):

Varianz und Vielfalt interdisziplinär: Wörter und Strukturen

doi:10.14618/opal_02-2016

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Mitglied der Leibniz-Gemeinschaft

© 2016 IDS Mannheim – Alle Rechte vorbehalten

Das Werk einschließlich seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechts ist ohne Zustimmung der Copyright-Inhaber unzulässig und strafbar. Das zulässige Zitieren kleinerer Teile in einem eigenen selbstständigen Werk (§ 51 UrhG) erfordert stets die Angabe der Quelle (§ 63 UrhG) in einer geeigneten Form (§ 13 UrhG). Eine Verletzung des Urheberrechts kann Rechtsfolgen nach sich ziehen (§ 97 UrhG). Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen. Die zugänglichen Daten dürfen von den Nutzern also nur zu rein wissenschaftlichen Zwecken genutzt werden. Eine darüber hinausgehende Nutzung, gleich welcher Art, oder die Verarbeitung und Bearbeitung dieser Daten mit dem Zweck, sie anschließend selbst oder durch Dritte kommerziell zu nutzen, bedarf einer besonderen Genehmigung des IDS (Lizenz). Es ist nicht gestattet, Kopien der Textdateien auf externen Webservern zur Verfügung zu stellen oder Dritten auf sonstigem Wege zugänglich zu machen. Bei der Veröffentlichung von Forschungsergebnissen, in denen OPAL-Publikationen zitiert werden, bitten die Autoren und Herausgeber um eine entsprechende kollegiale Information an opal@ids-mannheim.de.

Inhalt

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Einleitung: Varianz und Vielfalt interdisziplinär: Wörter und Strukturen

Luise Borek und Andrea Rapp (*Technische Universität Darmstadt*)

Im Frühjahr 2007 publizierte das BMBF eine Ausschreibung mit dem Titel „Wechselwirkungen zwischen Natur- und Geisteswissenschaften“.[1] Zur gleichen Zeit führten die Gegebenheiten am Würzburger Uni-Campus und der Zufall zu einem Austausch zwischen Wissenschaftlern aus Bioinformatik, Informatik und EDV-Philologie, aus dem kurze Zeit später das interdisziplinäre Verbundprojekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“ entstand. Auch im Licht der zu der Zeit ansteigenden Zahl von Digital Humanities-Projekten konnte dieses Unterfangen als ein interdisziplinärer Sonderfall angesehen werden. Sind es für gewöhnlich Kooperationen zwischen einem geisteswissenschaftlichen Fach und der Informatik (oder vice versa), so war in dieser Konstellation von vornherein ein weiterer – und in diesem Kontext durchaus seltenerer – naturwissenschaftlicher Partner involviert. Das Projekt bot somit Gelegenheit zu einem anregenden, transdisziplinären Austausch über das Phänomen der Varianz. Unter der Annahme von Varianz als einem Regelzustand natürlicher Systeme, die synchronem und diachronem Wandel unterliegen, waren es daher die Hauptziele, Verfahren, Methoden und Algorithmen zu entwickeln und zu erproben, die Varianz erschließen und ordnen. Ausgehend von der strukturellen Ähnlichkeit des untersuchten Materials, sollte dabei der Einsatz von disziplinspezifischen Verfahren wechselseitig erprobt werden. Unter diesem Blickwinkel lässt sich die Aufgabe des Verbunds sehr passend beschreiben als Versuch, analoge Strukturen zwischen Sprache und Genomen zu identifizieren und zu analysieren und aus solchen Erkenntnissen einen Zugewinn über die eigene Fachwissenschaft hinaus zu erzielen. Die Fragestellung selbst ist nicht neu. Ihre Entstehungsgeschichte verdient jedoch, in diesem Rahmen festgehalten zu werden, gibt sie doch zugleich Aufschlüsse und Anregungen zu weiterem interdisziplinären Austausch.[2]

Der im 19. Jahrhundert für seine Stammbaumtheorie bekannt gewordene Sprachwissenschaftler August Schleicher veröffentlichte im Jahr 1863 einen dreißigseitigen Brief an seinen Kollegen Ernst Häckel unter dem Titel „Die Darwinsche Theorie und die Sprachwissenschaft. Offenes Sendschreiben an Herrn Dr. Ernst Häckel, a.o. Professor der Zoologie und Director des zoologischen Museums an der Universität Jena“ (Schleicher 1863). In der Einleitung dieses transdisziplinären Schreibens erklärt Schleicher, dass Häckel ihm „nicht eher Ruhe gelassen [habe], als bis [er] Darwins viel besprochenes Werk über die Entstehung der Arten im Thier- und Pflanzenreich durch natürliche Züchtung oder Erhaltung der vervollkommneten Rassen im Kampfe ums Dasein, nach der zweiten Auflage übersetzt von Bronn, Stuttgart 1860, gelesen hatte“ (Schleicher 1863, S. 3). Dafür dankt er ganz ausdrücklich und bringt Darwins Darlegungen und Ansichten in Verbindung mit dem eigenen Fach, der Sprachwissenschaft, und nimmt dabei Bezug auf die eigene Darstellung über „Die deutsche Sprache“ aus dem Jahr 1860:

> Von den sprachlichen Organismen gelten nämlich ähnliche Ansichten, wie sie Darwin von den lebenden Wesen überhaupt ausspricht, theils fast allgemein, theils habe ich zufällig im Jahre 1860, also in demselben Jahre, in welchem die deutsche Uebersetzung von Darwins Werk erschien, über den ‘Kampf

[1] Die Bekanntmachung der Förderrichtlinie ist online unter www.bmbf.de/foerderungen/7774.php einzusehen.

[2] Die folgende Darstellung baut auf den Abschlussbericht des Verbundprojekts auf und ist zum Teil aus diesem übernommen: Wegstein, Werner et al. (2012): Verbundprojekt Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen. Modellierung und Abbildung von Varianz in Sprache und Genomen, Teilprojekt: Bioinformatik und Informatik. Schlussbericht. Berichtszeitraum: 01.10.2008 - 31.12.2012. Zum weiteren Verlauf vgl. zudem Dörries (Hg.) (2002), bes. Richards (2002) und Suhr (2002).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

> ums Dasein', über das Erlöschen alter Formen, über die die grosse Ausbreitung und Differenzierung einzelner Arten auf sprachlichem Gebiete mich in einer Weise ausgesprochen, welche den Ausdruck abgerechnet, mit Darwins Ansichten in auffälliger Weise zusammen stimmt. (Schleicher 1863, S. 4)

Die Beobachtungen zu den jeweiligen Forschungsgegenständen weisen Ähnlichkeiten und sogar Übereinstimmungen auf, da sie auf der Grundannahme des lebendigen und der Veränderung unterworfenen Gegenstands beruhen. Auf S. 6 präzisiert Schleicher:

> Die Sprachen sind Naturorganismen, die, ohne vom Willen des Menschen bestimmbar zu sein, entstanden, nach bestimmten Gesetzen wuchsen und sich entwickelten und wiederum altern und absterben; auch ihnen ist jene Reihe von Erscheinungen eigen, die man unter dem Namen 'Leben' zu verstehen pflegt.[3]

In der Folge wird Schleichers Schrift 1869 als „Darwinism tested by the Science of Language" von Dr. Alex Bikkers ins Englische übersetzt und im Januar 1870 im ersten Band der naturwissenschaftlich geprägten Fachzeitschrift „Nature" von Max Müller rezensiert.[4] Er räumt zwar auch die Analogie des „struggle for life between languages" und des „struggle for life among the more or less favoured species in the animal and vegetable kingdoms" ein (Müller 1870, S. 257), sieht aber eine entschieden treffendere Analogie als die des „struggle for life among separate languages" im „struggle for life among words and grammatical forms which is going on in each language" (ebd.). Die Verlagerung auf diese konkretere und strukturelle Ebene des Sprachsystems entspricht zugleich dem Ansatz, dem das Wechselwirkungen-Projekt aus synchroner und diachroner Sicht nachging.

Die naturwissenschaftliche Reaktion auf Schleichers Brief und Müllers Rezension erfolgte 1871, als Darwin im ersten Band von „The descent of man and selection in relation to sex" beides wieder aufgreift (Darwin 1871). Im zweiten Kapitel „Comparison of the mental power of man and the lower animals" zitiert er im Abschnitt zu „Language" Schleichers Übersetzung (S. 54 und Anm. 34), setzt sich mit Müllers Rezension auseinander (S. 58) und erläutert seine Position:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same.[5] But we can trace the origin of many words further back than in the case of species, for we can perceive that they have arisen from the imitation of various sounds, as in alliterative poetry. We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. We have in both cases the reduplication of parts, the effects of long-continued use, and so forth. The frequent presence of rudiments, both in languages and in species, is still more remarkable.

Diese frühe, inzwischen über 140 Jahre alte Wechselwirkung zwischen Natur- und Geisteswissenschaften kann als Ausgangspunkt unseres oben beschriebenen Projektes angesehen werden. Wenn es strukturelle Parallelitäten zwischen Biologie und Sprache gibt, so könnten wir Methoden des einen Feldes nutzen, um Effekte des anderen Feldes zu erklären.

Die Analogie zwischen Sprache und Genetik wurde seit der Anfangszeit bereits des Öfteren thematisiert, im Rahmen unseres Verbundprojekts wurde sie aufgegriffen und unter neue Perspektiven gestellt. Hervorzuheben ist insbesondere, dass nunmehr erstmals die technischen Möglichkeiten und Ressourcen für empirische Untersuchungen ausreichend umfangreicher

[3] Schleicher (1863), S. 6.
[4] Müller, Max (1870): Darwinism tested by the Science of Language. Translated from the German of Professor August Schleicher. In: Nature 1, S. 256-259.
[5] Vgl. die interessanten Parallelen zwischen der Entwicklung von Arten und Sprachen, die Sir C. Lyell anstellt (Lyell 1863, Kap. xxiii).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Datenmengen gegeben sind. Die Ergebnisse bestätigen die Fruchtbarkeit des gewählten Ansatzes. So konnte u.a. nachgewiesen werden, dass sich Wörter aus der korpusgenerierten Basislemmaliste anders verhalten als lemmatisierte Wörter aus Wörterbüchern.

Die Beiträge des vorliegenden Bandes sind das Ergebnis eines interdisziplinären Workshops, der zum Abschluss des Projekts unter dem Titel „Varianz und Vielfalt interdisziplinär: Wörter und Strukturen“ im Dezember 2012 in Darmstadt stattfand.

Das Arbeitstreffen fasste Erkenntnisse und Erfahrungen aus der Untersuchung von „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen für die Modellierung und Abbildung von Varianz in Sprache und Genomen“ zusammen und diskutierte diese Ergebnisse mit den Fachwissenschaften (Computerlinguistik, Lexikografie, Informatik, Bioinformatik, u.v.m.). Ein Schwerpunkt lag hierbei auf elektronischen Wörterbüchern (retrodigitalisiert oder *born digital*), ihrer Heterogenität, der in ihnen dokumentierten Varianz sowie auf den Werkzeugen und Methoden, die zu ihrer Erschließung und Analyse dienen. Weitere sprachwissenschaftlich motivierte Themenbereiche umfassten daher z.B. die synchrone und diachrone Varianz, die quantitative Linguistik, Morphologie und Sprachwandelprozesse, Varianz in Wortfamilien wie auch die Erschließung von Varianz. Anschließend konnte das Phänomen der Varianz aus verschiedensten Perspektiven beleuchtet werden und ein Beitrag zur Konstituierung einer disziplinübergreifenden Abstraktionsebene geleistet werden.

Der vorliegende Band enthält einige der Vorträge und führt heterogene Forschungsgegenstände zusammen, die zwischen Lexikografie, Computerlinguistik, (historischer) Sprachwissenschaft und den digitalen Geisteswissenschaften transzendieren. Die Beiträge von Natalia Filatkina, Meike Meliss und Annette Klosa gründen sich dabei auf unterschiedlichen Typen von Wörterbüchern.

Filatkina stellt „Beobachtungen zu Varianz in (historischen) Wörterbüchern und Texten“ an, indem sie Typen von Varianz in polylexikalischen Strukturen untersucht. Dabei spielen sowohl die Diachronie als auch die Einflussnahme von Wörterbüchern auf den Forschungsgegenstand eine Rolle. Zudem werden methodische Konsequenzen des hohen Variationspotenzials reflektiert.

Meliss beleuchtet „Lexikalische Vielfalt und Varianz aus kontrastiver Perspektive“ und erläutert in ihrem Beitrag Überlegungen zu einem Produktionswörterbuch aus der Sicht des Deutschen und Spanischen. Dabei stellt die Vielfalt der sprachlichen Ausdrucksmittel eine Schwierigkeit für L2-Wörterbuchnutzer dar. Das Wörterbuch muss die lexikalische Ausdrucksvielfalt ebenso mitberücksichtigen wie die Bedeutungs- und Konstruktionsvarianz des Lemmas. Auch hier klingt an, dass Lösungsansätze für die durch Vielfalt komplexe lexikografische Dimension nur mit digitalen Methoden realisiert werden können.

Annette Klosa widmet ihren Beitrag über „Wortfamilien in Onlinewörterbüchern“ den synchronen Anforderungen der Thematik und skizziert Wortbildung als Ordnungsstruktur sowie deren Darstellung in Onlinewörterbüchern. Die untersuchte Varianz betrifft die Problematik von Schreibvarianten ebenso wie die morphologische Varianz.

Die anschließenden Beiträge von Heike Stadler, Werner Wegstein, Luise Borek und Dietmar Seipel widmen sich digitalen Verfahren, die Varianz-Phänomene in Ordnungsstrukturen abbildbar und analysierbar machen. Sie bieten damit Einblicke in die Ergebnisse des abge-

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

schlossenen Verbundprojekts und eröffnen gleichzeitig Anknüpfungspunkte für weitere Forschung.

Stadler und Wegstein betrachten in ihrem Beitrag „Korpusbasierte Wortfamilien-Lemmalisten und ihre TEI-Kodierung“ und knüpfen damit thematisch an den Beitrag von Annette Klosa an. Dabei klingt ein Vergleich von Wortschatzstruktur und XML-Struktur an und es wird eine Listenstruktur als ergänzendes Werkzeug für die Lexikografie vorgeschlagen, um der morphologischen Vielfalt gerecht werden zu können.

Borek thematisiert den Umgang mit dem Befund sprachlicher Vielfalt in retrodigitalisierten Wörterbüchern im Beitrag mit dem Titel „Die Metalemmaliste als Tool zum Erschließen von sprachlicher Varianz“. Die untersuchten Wörterbücher werden dabei als Datenbanken (historischer) sprachwissenschaftlicher Ressourcen aufgefasst. Die Metalemmaliste fungiert als Werkzeug und Grundlage für das Analysieren formaler und inhaltlicher Aspekte von Varianz.

Dietmar Seipel und Luise Borek skizzieren schließlich weitere informatische Methoden, indem sie „Ein halbautomatisches Werkzeug zum Alignieren von Wörtern“ vorstellen. Am Beispiel der mittelhochdeutschen Sprachstufe wird ein Alignmentverfahren als Verknüpfungswerkzeug des heterogenen Datenmaterials vorgestellt, das händische und automatische Verfahren kombiniert. Letzteres ist notwendig, da Varianz und Regelbasiertheit gemeinsam den Regelzustand einer natürlichen Sprache ausmachen.

Die Herausgeberinnen und Herausgeber danken allen, die an der Abschlusstagung als interessierte und aktive Beiträgerinnen und Beiträger beteiligt waren – von den anregenden Gesprächen und Kontakten profitieren sie noch heute. Annette Klosa ist für die Aufnahme in die Reihe OPAL zu danken, wo der Band einen optimalen Publikationsort gefunden hat. Allen, die uns bei der Fertigstellung beständig unterstützt haben, sei herzlich gedankt, namentlich Marc Adler, Franziska Horn, Celia Krause und Sandra Denzer.

Die BMBF-Ausschreibung „Wechselwirkungen zwischen Natur- und Geisteswissenschaften“ war ein förderpolitisches und wissenschaftliches Wagnis, in das sich viele Institutionen und Forschende – Etablierte wie Nachwuchs – begeistert gestürzt haben. Die Erfahrungen und Ergebnisse der beteiligten Forschungsprojekte wirken bei allen Beteiligten nach und haben neue Dialog-, Kooperations- und Forschungsperspektiven eröffnet. Den Verantwortlichen im BMBF mit den Betreuenden beim Projektträger im DLR sei für diesen Mut, diese Kreativität und diese Risikobereitschaft ganz besonders gedankt! Alle, die das Projekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“ durch ihre Mitarbeit begleitet und getragen haben, schließen wir in diesen herzlichen Dank ein![6]

[6] Namentlich sind dies: Cyril Belica, Agnes Brauer, Ludwig Eichinger, Michel Hartmann, Daniela Keller, Claudine Moulin, Rainer Perkuhn, Esther Ratsch, Christian Schneiker, Jörg Schultz, Dietmar Seipel, Heike Stadler, Florian Stefan und Werner Wegstein.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Literatur

Darwin, Charles (1871): The descent of man and selection in relation to sex. In two volumes. Bd. 1. London.

Dörries, Matthias (Hg.) (2002): Experimenting in tongues. Studies in science and language. Stanford.

Lyell, Charles (1863): The geological evidences of the antiquity of man with remarks on theories of the origin of species by variation. London.

Müller, Max (1870): Darwinism tested by the Science of Language. Translated from the German of Professor August Schleicher. In: Nature 1, S. 256-259.

Richards, Robert J. (2002): The linguistic creation of man: Charles Darwin, August Schleicher, Ernst Haeckel and the Missing Link in nineteenth-century Evolutionary Theory. In: Dörries, Matthias (Hg.), S. 21-48.

Schleicher, August (1863): Die Darwinsche Theorie und die Sprachwissenschaft. Offenes Sendschreiben an Herrn Dr. Ernst Häckel, a.o. Professor der Zoologie und Director des zoologischen Museums an der Universität Jena. Weimar.

Suhr, Stephanie (2002): Is the notion of language transferable to the genes? In: Dörries, Matthias (Hg.), S. 49-66.

Wegstein, Werner et al. (2012): Verbundprojekt: Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen, Teilprojekt: Bioinformatik und Informatik. Schlussbericht. Berichtszeitraum: 1.10.2008 - 31.12.2012. http://edok01.tib.uni-hannover.de/edoks/e01fb14/780793897.pdf (Stand: 11.3.2015).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Wie fest sind feste Strukturen?
Beobachtungen zu Varianz in (historischen) Wörterbüchern und Texten*

Natalia Filatkina (*Universität Trier*)

Der Beitrag widmet sich aus linguistischer Perspektive den Typen der Varianz im Bereich der polylexikalischen Strukturen, die synchron gesehen als mehr oder weniger fest gelten und von Sprecherinnen und Sprechern des Deutschen in einer bestimmten syntaktischen Form reproduziert werden. Gerade mit Blick auf das hohe Variationspotenzial solcher Strukturen in der Diachronie verwende ich in Anlehnung an die Erkenntnisse der Nachwuchsforschergruppe „Historische formelhafte Sprache und Traditionen des Formulierens (HiFoS)“ im Folgenden nicht den am gegenwartssprachlichen Material etablierten Begriff *Phraseologismus*, sondern spreche von *formelhaften Wendungen.*[1] In Kapitel 1 gehe ich theoretisch auf das Verhältnis von Festigkeit und Varianz im Bereich der formelhaften Wendungen ein. In den Kapiteln 2 und 3 zeige ich am Beispiel des Idioms *Perlen vor die Säue werfen*, dass formelhafte Wendungen immer Produkte des Sprachwandels und der vielfältigen Variation sind. Es wird sich dabei herausstellen, dass sowohl historische Wörterbücher als auch ältere Texte aufschlussreiche Auskünfte über das Variationspotential der formelhaften Wendungen geben, dabei aber ganz unterschiedliche Bilder von Varianz vermitteln. Kapitel 2 beschäftigt sich mit der Frage, wie (historische) Wörterbücher mit Varianz umgehen und inwiefern sie zur Entstehung der Festigkeit im Bereich der formelhaften Wendungen beitragen. Ausgehend von den Daten der HiFoS-Nachwuchsforschergruppe präsentiert hingegen Kapitel 3 die Varianz beim Idiom *Perlen vor die Säue werfen* in ihrer Entfaltung in älteren Texten und fragt danach, welche Bereiche der formelhaften Wendungen in der Diachronie von Varianz betroffen sind und welche Wechselwirkungen zwischen diesen Bereichen existieren. Es wird sich zeigen, dass unterschiedliche Typen der Varianz bei formelhaften Wendungen gehäuft zu einem Zeitpunkt in der Geschichte vorkommen und den ebenenübergreifenden Wandel bedingen, was für Einzellexeme weniger typisch ist. In Kapitel 4 diskutiere ich, welche methodischen Konsequenzen das hohe Variationspotenzial der formelhaften Wendungen für die Lexikographie mit sich bringt und stelle die Lösungsvorschläge der HiFoS-Nachwuchsforschergruppe vor.

1. Festigkeit vs. Varianz im Bereich der formelhaften Wendungen

Dass sich Sprachen wie auch biologische Naturorganismen verändern, wurde noch vor der Entstehung der Sprachwissenschaft und erst recht in ihrer ersten historisch-vergleichenden Phase im 19. Jahrhundert beobachtet. Auch wenn sich z.B. Hermann Paul von der absoluten Gleichsetzung des „Sprachorganismus“ (Paul 1995, S. 29) mit einem Naturorganismus distanziert, indem er die Sprachwissenschaft als Kulturwissenschaft profiliert und die Rolle der Gesellschaft und historischer Prozesse bei sprachlichen Entwicklungen hervorhebt, weist er auf Gemeinsamkeiten der sprachlichen und biologischen Veränderungen (ebd., S. 32), insbesondere im Bereich der Evolution und der Herausbildung der sprachlichen Vielfalt, hin:

> Es ist eine durch die vergleichende Sprachforschung zweifellos sicher gestellt Tatsache, dass sich vielfach aus einer im wesentlichen einheitlichen Sprache mehrere verschiedene Sprachen entwickelt haben, die ihrerseits auch nicht einheitlich geblieben sind, sondern sich in eine Reihe von Dialekten gespalten haben.

* Für eine kritische Durchsicht des Manuskripts bedanke ich mich bei Carina Hoff und Bernhard Ost.

[1] Zu theoretischen Unterschieden zwischen den beiden Begriffen vgl. Filatkina et al. (2009); Filatkina (2012, 2013a); Hanauska (2012).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

> Man sollte erwarten, dass sich bei der Betrachtung dieses Prozesses mehr als irgend wo anders die Analogieen [sic!] aus der Entwicklung der organischen Natur[2] aufdrängen müssten. [...] Hier in der Tat ist die Parallele innerhalb gewisser Grenzen eine berechtigte und lehrreiche (Paul 1995, S. 37).[3]

In die Nähe der Naturwissenschaften rücken die Sprachwissenschaft auch Junggrammatiker (im engeren Sinn), indem sie ihren Schwerpunkt auf die Aufdeckung der Gesetzmäßigkeiten der historischen Entwicklung indoeuropäischer Sprachen legen und den Gesetzesbegriff im Kontext einer „Organismusvorstellung" von Sprache definieren (Grimm 1819-1837/1967, Vorwort). Spätestens seit der Mitte des 20. Jahrhunderts ist auch bekannt, dass Variation als Prozess und Varianz als Resultat bei diesen Veränderungen die unabdingbaren treibenden Kräfte sind (Labov 2001). Allerdings standen diese beiden Erkenntnisse nicht immer im Zentrum des sprachwissenschaftlichen Interesses und waren in der Linguistik stark paradigmenabhängig (Gardt 1999). Noch seltener sind bis in die heutige Zeit hinein die Versuche, Gemeinsamkeiten und Unterschiede zwischen den biologischen Veränderungsprozessen und dem Sprachwandel systematisch und in Wechselwirkung aufzudecken.[4]

Variation und Varianz sind in der Natur der Sprache angelegt. Um ihre kommunikativen Ziele erfolgreich erreichen zu können, sind die Sprecherinnen und Sprecher einer Sprache im Sinne des Sinclair′schen *open choice principle* (Sinclair 1991, S. 109) relativ frei, was die Wahl der sprachlichen Mittel, ihre Verbindung miteinander sowie ihre Veränderung anbetrifft. Diese Freiheit ist besonders bei der Entstehung der Texte bzw. Diskurse aus Sätzen oder der Sätze aus Wörtern ausgeprägt. Eingeschränkter bis kaum möglich ist sie hingegen bei der Bildung der Wörter aus Morphemen. Der Erfolg eines kommunikativen Akts besteht nicht nur in der korrekten Verwendung einzelner sprachlicher Mittel, sondern in ihrer adäquaten Kombination miteinander und Variation mit Blick auf die situativen Kommunikationsbedingungen.

Unter den sprachlichen Einheiten gibt es komplexe polylexikalische Phänomene, die in der Kombinatorik der in ihrer Struktur vorkommenden Einzellexeme eingeschränkt sind. Im Gegenteil: Solche Einheiten ergeben nur dann Sinn, wenn sie in wiederkehrenden Kommunikationssituationen in ein und derselben Struktur nicht neu produziert, sondern nach vorgeformten Mustern reproduziert werden. Seitdem Sinclair (1991, S. 110) solche Einheiten mit dem Begriff *idiom choice principle* als den zweiten unabdingbaren Teil der menschlichen Kommunikation beschrieben hat, haben sie auch unter den Stichwörtern *idiomatische Prägung* (Feilke 1994), *lexical priming* (Hoey 2005), *formulaic language* (Wray 2002), *Muster* (Steyer 2011, 2013) oder auch *Phraseme/Phraseologismen* (Burger 2010) Beachtung gefunden und werden heutzutage im Rahmen der Konstruktionsgrammatik(en) „neu" diskutiert (Langacker 1987; Goldberg 1995). Ihre wichtige Funktion ergibt sich daraus, dass viele der in der Kommunikation auszuführenden Sprachhandlungen konventionalisierte, ritualisierte Kommunikationsformen sind. Formelhafte Wendungen gestalten das Formulieren ökonomischer und erleichtern das auf der *Common-Sense*-Kompetenz (Feilke 1996) beruhende Verstehen. Für das soziale Sprachhandeln ist somit neben Freiheit und Kreativität eine besondere Typik, Vorgeformtheit oder eben Festigkeit kennzeichnend, die die Traditionen des Formulierens einer Gesellschaft gestaltet, historisch erwachsen ist und sich deshalb in Raum und Zeit ändern kann.

---

2 Sperrung im Original [NF].

3 Vgl. auch Bopps Forderung „einer Naturbeschreibung der Sprache", die nach seiner Auffassung die Qualität der strukturellen Beschreibungen verbessert (Bopp 1816 [1975]).

4 Vgl. dazu wegweisend das „Sprache und Genome"-Projekt: www.sprache-und-genome.de.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bis vor kurzem wurde solchen Strukturen (insbesondere im Rahmen der klassischen Phraseologieforschung) jegliche Varianz wenn nicht komplett untersagt, dann doch zumindest in nur einem sehr eingeschränkten Maße attestiert. Die Festigkeit ihrer morphosyntaktischen Struktur und der Besetzung von lexikalischen Slots wurde zum definitorischen Merkmal solcher Einheiten erhoben. Oft wurde das Merkmal der syntaktischen Festigkeit an Idiomatizität im Sinne einer semantischen Festigkeit gekoppelt: Die Bedeutung einer Wendung ist fest, weil sie übertragen, d.h. nicht aus der Bedeutung der einzelnen Konstituenten in der Struktur dieser Wendung ableitbar, ist und nur beim Vorkommen dieser bestimmten Konstituenten zustande kommt. Je näher ein Phrasem am prototypischen Kern des phraseologischen Subsystems liegt, in desto höherem Maße wurde ihm das Merkmal der Festigkeit zugesprochen. Nach dieser nicht ganz unstrittigen Auffassung (Filatkina 2010; Burger 2012) sind idiomatische Ausdrücke (Idiome, Paarformeln, Sprichwörter u.a.) phraseologische Einheiten mit dem höchsten Festigkeitsgrad.

Teilweise ist diese Auffassung methodisch zu begründen: Zu Beginn der Phraseologieforschung, in der Phase ihrer Gegenstandsbestimmung und Begriffsfindung, bildeten vor allem idiomatische Print-Wörterbücher einiger moderner Sprachen die Materialgrundlage für phraseologische Studien. Eines der Prinzipien der modernen Lexikographie bzw. Phraseographie besteht in der Formulierung der so genannten Nennform, einer Art phraseologisches Lemma, das zwar in einigen Fällen die Varianz in der Struktur eines Phraseologismus mit dokumentiert, grundsätzlich aber die Abstrahierung von Varianz zum Ziel hat. Abbildung 1 mit dem Artikel „Perle“ aus dem Duden 11 „Redewendungen“ soll dies veranschaulichen:

Perle: **Perlen vor die Säue werfen** (ugs.): *etwas Wertvolles jmdm. anbieten, geben, der kein Verständnis dafür hat, es nicht zu würdigen weiß:* Dir Kaviar zu servieren heißt wirklich Perlen vor die Säue zu werfen!
◇ Die Wendung wurde durch die Bibel allgemein verbreitet, dort heißt es in Matthäus 7,6: »... eure Perlen sollt ihr nicht vor die Säue werfen, auf daß sie dieselbigen nicht zertreten mit ihren Füßen und sich wenden und euch zerreißen.« Die Wendung ist allerdings schon früher in der Literatur belegt. Da Schweine früher überwiegend mit Abfall gefüttert wurden, bringt die Wendung die völlige Unangemessenheit einer Handlung zum Ausdruck.
jmdm. fällt keine Perle/kein Stein aus der Krone: ↑Stein.

Abb. 1: Artikel „Perle“ im Duden 11 „Redewendungen“

Der typographisch fett unterlegte Eintrag *Perlen vor die Säue werfen* stellt solch eine Nennform dar, die im gegenwärtigen Deutsch nach Angaben des Wörterbuchs nur in dieser Form existiert. In der im Artikel angeführten einzigen Belegstelle ist die Wendung allerdings anders kontextualisiert und liefert somit keine Grundlage für die Formulierung einer abweichenden Nennform. Die Diskrepanz zwischen dem „phraseologischen Lemma“ und dem Beleg bleibt unkommentiert. Das sich am Ende des gleichen Artikels im Verweisteil befindende Idiom *jmdm. fällt keine Perle aus der Krone* ist hingegen mit der lexikalischen Variante *jmdm. fällt kein Stein aus der Krone* verzeichnet. Ob es zwischen den beiden Varianten Unterschiede im Gebrauch gibt bzw. ob sie Besonderheiten aufweisen, geht aus dem Artikel nicht hervor.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Mit der pragmatischen Wende der 70er Jahre des 20. Jahrhunderts, der Entwicklung der Gesprochene-Sprache-Forschung und erst recht durch die Korpus- und Computerlinguistik verlagerte sich der methodische Fokus auf Primärtexte/-daten. Die Analyse dieser Quellen führte zur Relativierung des Merkmals der Festigkeit: Auch wenn es nach wie vor zu Recht eine der wichtigsten Eigenschaften der formelhaften Wendungen beschreibt, hat es seinen absoluten Charakter verloren. Trotz der neuen Erkenntnisse fehlen aber umfassende Untersuchungen und folglich auch Darstellungen zur Varianz im Bereich der formelhaften Wendungen selbst in der wissenschaftlichen Sekundärliteratur, die den Status der Standardnachschlagewerke auf dem Gebiet der Variationslinguistik für sich beansprucht. Wesentliche Schritte in diese Richtung sind das abgeschlossene Projekt „Kollokationen im Wörterbuch“ an der Berlin-Brandenburgischen Akademie der Wissenschaften, das Projekt „Usuelle Wortverbindungen“ und die Sprichwortdatenbank deutsch, die beide am Institut für Deutsche Sprache in Mannheim verortet sind, aufeinander aufbauen und über die online-Plattform OWID zugänglich sind. Obwohl zurzeit in den beiden Ressourcen des IDS noch kein Wörterbuchartikel für das Idiom *Perlen für die Säue werfen* gefunden werden kann,[5] veranschaulichen die Einträge für andere formelhaften Wendungen (z.B. *an die große Glocke hängen*) neue Zugänge zur Abbildung der Varianz im modernen Deutsch: Der Wörterbuchartikel fängt zwar nach wie vor mit einer „Nennform“ an; im weiteren Verlauf des Artikels wird aber nicht diese Form kommentiert, sondern auf „typische Kontextmuster“ eingegangen. Hier können mit Blick auf die Varianz in Textkorpora auch mehrere Formen angeführt werden; sie bilden die Grundlage für die semantische Paraphrase und die Angaben zur pragmatischen Funktion. Auf neuen Möglichkeiten der Online-Lexikographie bauen ebenfalls die Projekte „Dynamics of Luxembourgish Phraseology (DoLPh)“ und „Online-Lexikon zur diachronen Phraseologie (OldPhras)“ für das Deutsche ab dem 18. Jahrhundert auf, die sowohl die lexikographischen Quellen als auch die Textüberlieferung berücksichtigen.

Der vorliegende Beitrag wird zeigen, dass Festigkeit im synchronen Sprachgebrauch nur als Produkt des Sprachwandels und der Varianz zu verstehen ist.[6] Aus der Vielfalt der formelhaften Wendungen greife ich das Idiom *Perlen vor die Säue werfen*[7] heraus, weil es im heutigen Sprachgebrauch ein seltenes Beispiel für eine tatsächlich in hohem Maße feste Struktur ist und in historischen Quellen im Gegensatz dazu ein großes Variationsspektrum aufweist. Mit diesem Beispiel soll veranschaulicht werden, dass eine der Herausforderungen, die bei der Untersuchung und Abbildung der Varianz im Bereich der formelhaften Wendungen zu berücksichtigen ist, darin besteht, dass unterschiedliche Typen von Varianz im Normal- und Regelfall gleichzeitig und gehäuft zu einem Zeitpunkt aufkommen und gegenseitig aufeinander wirken. Bedingt ist dies dadurch, dass formelhafte Wendungen oft Phänomene sind, die nicht einer sprachlichen Ebene zugeordnet werden können, sondern sich zwischen Lexik, Grammatik, Pragmatik und Diskurs bewegen. Das bedeutet für den Sprecher/die Sprecherin die gleichzeitige Kenntnis der lexikalischen, grammatischen, pragmatischen und diskurslinguistischen Regeln einer Sprache, um die formelhafte Wendung angemessen verwenden zu können. So muss er oder sie beim Gebrauch des Idioms *Perlen vor die Säue werfen* wissen, dass eben diese Konstituenten und nicht andere (etwa **Rubine unter die Ferkel legen*) vorkommen müssen, damit die übertragene Bedeutung ‘etwas Wertvolles vor jemandem vergeuden, der/die das nicht zu schätzen weiß’ aktualisiert werden kann. Die Substitution der verbalen Konstituente *Perlen vor die Säue schmeißen* und die Passivierung (*Hier werden musikalische Perlen vor die Säue geworfen)* sind ausgehend von den Analysen im DWDS-Korpus,

[5] Die Erweiterung der online veröffentlichten Daten ist laut der OWID-Homepage geplant.
[6] Das gilt uneingeschränkt auch für die selbst in ihrer Makrostruktur stark formalisierten Texte (Filatkina 2011).
[7] Vgl. die umfangreiche Literatur zu diesem Idiom, seinen ikonographischen und theologischen Traditionen sowie Entsprechungen in anderen Sprachen in Piirainen (2012, S. 231ff.) und Dobrovol’skij/Piirainen (2009, S. 35f.).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Deutschen Referenzkorpus DeReKo und im Leipziger Wortschatzportal aber zulässig. Einige grammatische Veränderungen an den einzelnen Konstituenten (etwa **eine Perle wurde unter die Sau geworfen*) zerstören ebenfalls die idiomatische Bedeutung, andere grammatische Varianten sind in den Korpora wiederum belegt, z.B. *Das ist doch Perlen vor die Säue geworfen!* und öfter sogar mit Auslassung der verbalen Konstituente: *Das ist/wäre Perlen vor die Säue!* Die Korpora zeigen, dass die zuletzt genannte Struktur heutzutage das präferierte strukturelle Muster bei diesem Idiom ist. Schließlich muss der Sprecher/die Sprecherin wissen, dass die Wendung pragmatisch gesehen im gegenwärtigen Deutsch eher ein Kommentar für Verschwendung, sinnlose Tätigkeit ist (*Pragmatik*), dass sie nicht in die gehobenen Stilregister gehört (*Stil*), dort aber unterschiedlich thematisch kontextualisiert werden kann (*Diskurs*: vorstellbar ist eine Kommunikationssituation zu einem beliebigen Thema, in der eine beliebige Tätigkeit mit diesem Idiom als Verschwendung charakterisiert werden kann). Die auf der Verwendung basierte Abbildung der Varianz bedeutet für den Lexikographen die Berücksichtigung eben all dieser Ebenen.

Beim Vergleich der Korpusbelege mit den Wörterbucheinträgen geht es mir nicht um die Aufdeckung der lexikographischen Mängel bei der Darstellung der formelhaften Wendungen (dieses Thema ist in der Phraseologieforschung nicht neu und gut beschrieben),[8] sondern vielmehr um die Tatsache, dass Wörterbücher und Korpora jeweils ein unterschiedliches Bild von Varianz bei formelhaften Wendungen vermitteln. Dazu bereits Burger (2000, S. 42):

> Wie empirische Untersuchungen zeigen, täuschen die Angaben in den Lexika, auch in den phraseologischen Spezialwörterbüchern, ein Bild vor, das der Sprachwirklichkeit vermutlich nicht entspricht. Die kodifizierten und damit als standardsprachlich tolerierten Varianten entsprechen bei weitem nicht der Vielfalt von Varianten, wie sie in gesprochener – und teilweise auch geschriebener Kommunikation vorkommen.

Das bedeutet nicht, dass Wörterbücher zur Untersuchung der Varianz im Bereich der formelhaften Wendungen nicht herangezogen werden können. Wie Kleine-Engel (2012) für das Luxemburgische und Dräger (2011, 2012) für das Deutsche nachweisen, gibt es Varianz durchaus auch in Wörterbüchern, nur ist sie anders motiviert. Vielmehr möchte ich mit dem vorliegenden Aufsatz aussagen, dass beide Quellentypen – historische Wörterbücher und Textkorpora – für die Untersuchung der phraseologischen Varianz unerlässlich sind. Wie in den Kapiteln 2 und 3 noch zu zeigen sein wird, verschärfen sich diese Unterschiede im diachronen Schnitt: Historische Wörterbücher und Texte unterscheiden sich nicht nur im Umgang mit Varianz, sondern auch in ihren Beständen, denn sie überliefern in unterschiedlicher Frequenz teilweise ganz unterschiedliche formelhafte Wendungen.

In weiteren Ausführungen konzentriere ich mich auf die nicht intendierte Varianz und lasse die bewussten Modifikationen, die für den Bereich der formelhaften Wendungen in einem ganz besonderen Ausmaß typisch sind (Hemmi 1994; Sabban 1998; Wotjak 1992), außer Acht, soweit die Abgrenzung im historischen Kontext möglich ist.[9]

[8] Vgl. stellvertretend Moon (2007) und für das Deutsche Müller/Kunkel-Razum (2007).
[9] Zu diesem Problem vgl. Filatkina (2012).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

2. Varianz und (historische) Wörterbücher

Bei der Untersuchung der Varianz beim Idiom *Perlen vor die Säue werfen* in den (historischen) Wörterbüchern beschränke ich mich exemplarisch auf diejenigen, die über das Trierer Wörterbuchnetz miteinander verbunden sind, und verwende im Folgenden für die Titel der Wörterbücher die im Verbund gängigen Abkürzungen. Die Ressourcen des Wörterbuchnetzes habe ich aus Gründen der thematischen Relevanz um weitere Quellen ergänzt:

- das „Wörterbuch der deutschen Sprache“ von Joachim Heinrich Campe,
- das „Wörterbuch der deutschen Sprache“ von Daniel Sanders,
- „Der Teutschen Weissheit“ von Friedrich Petri,
- das „Deutsche Sprichwörterbuch“ von Joachim Christian Blum und
- „Die deutschen Sprichwörter“ von Karl Simrock.

Es ist mir bewusst, dass es sich um konzeptionell sehr unterschiedliche Werke handelt, die nicht unmittelbar miteinander vergleichbar sind. Für die Fragestellung des vorliegenden Beitrags – Aufdeckung der Besonderheiten im Umgang mit Varianz bei formelhaften Wendungen in Wörterbüchern und Texten – können die Quellen aber durchaus allgemeine Tendenzen aufzeigen. Dabei beziehen sich meine Ausführungen ausschließlich auf das Idiom *Perlen vor die Säue werfen*. Gegenbeispiele, die in Wörterbüchern anders bearbeitet sind, lassen sich selbstverständlich auch finden, zahlreich sind sie aber nicht. Die allgemeinen Tendenzen in der Dokumentation der Varianz spiegelt das ausgewählte Beispiel trotz der eventuellen Gegenbeispiele gut wider.

Würde man anhand des Vorkommens des Idioms in Wörterbüchern Rückschlüsse auf seine Geläufigkeit ziehen, käme man zum Schluss, dass *Perlen vor die Säue werfen* nicht zu den verbreiteten Wendungen des Deutschen gehört. In vielen der im Wörterbuchnetz erfassten Quellen kommen weder dieses Idiom noch seine Varianten vor, so z.B. nicht in Lexer, BMZ, MHDBDB, FindeB, DRW, RhWb, PfWb, LoWb, LWb, LLU, WLM.

Die Besonderheiten, von denen unten noch die Rede sein wird, erklären sich aus der allgemeinen lexikographischen Praxis der historischen (und gegenwärtigen) Wörterbücher, die darin besteht, dass formelhafte Wendungen im deutlichen Unterschied zu Einzellexemen keinen Lemmastatus haben, sondern lediglich als veranschaulichende Beispiele für die Verwendung eines Lexems angeführt werden. Als „versteckte Informationen“ sind sie den Lexemen „beigegeben“, die nach Auffassung der Wörterbuchautoren sinntragende Konstituenten in der Struktur der Wendungen sind. So finden sich im DWB (Bd. 13, Sp. 1547-1555) unter *Perle* nach der rein auf die übertragene Bedeutung des Lemmas und nicht auf die Wendung bezogenen Paraphrase einige wenige Belegstellen von der ursprünglichen Bibelstelle bis hin zu Lessing. Diese Belegstellen enthalten das Idiom *Perlen vor die Säue werfen*, dienen aber der Veranschaulichung der übertragenen Bedeutung von *Perle*. Die Textstellen zeigen kursorisch, dass die zweite substantivische Konstituente *Säue* mit *Schweine* variieren kann. Auf diese Variation geht das Wörterbuch aber nicht weiter ein. Ähnlich verfahren auch die modernen Sprachstadienwörterbücher, z.B. das neue MWB: Es erweitert das Lemma *berle* um den aus dem Nhd. heraus gedachten Verweis „im Sprichw. ‘Perlen vor die Säue werfen’“ und führt unter Verweis auf TPMA 10, 318 eine Belegstelle aus „Dem Renner“ an, die aus Freidanks „Bescheidenheit“ übernommen wurde, ohne die Bedeutung der als Sprichwort bezeichneten Wendung zu beschreiben. Unter anderen Lemmata wie etwa *gimme* oder *margarite*, die als lexikalische Varianten in der Struktur der Wendungen vorkommen (vgl. unten Kap. 3 zu Varianz in älteren Texten), finden sich keine Belege.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Sollten Wörterbücher und Sammlungen das gleiche Idiom bzw. seine lexikalischen Varianten oder Synonyme im Sinne einer Mehrfachzuordnung an anderen Stellen verzeichnen, so lassen sie sich wegen der fehlenden einheitlichen Regeln zur Verzeichnung von formelhaften Wendungen in Wörterbüchern sowie zum Verweissystem nicht auf Anhieb finden, insbesondere wenn das Nachschlagewerk in digitaler Form nicht vorliegt.

Aus dieser lexikografischen Praxis im Umgang mit Wendungen ergeben sich folgende Konsequenzen für den Umgang mit Varianz:

Grammatische und lexikalische Varianz ist implizit mitdokumentiert, aber schwer zugänglich und oft zufällig

Auch wenn in anderen Wörterbüchern unser Idiom nach wie vor keinen Lemmastatus hat, sind diese etwas ausführlicher in der Dokumentation der lexikalischen Varianten. Im Wörterbuch von Daniel Sanders (Bd. 2, S. 515) ist das Lemma *Perle* im Abschnitt d) mit der Bedeutung ‘für etwas Köstliches, Wertvolles’ versehen. Dort ist als veranschaulichendes Beispiel auch der Beleg *Perlen vor die Säue werfen*, den Sanders selbst als Sprichwort bezeichnet, angeführt. Auch die lexikalischen Varianten *sein Perlen vor die Schweine, vor die Pfleglinge des verlorenen Sohnes werfen* sind jeweils mit Stellenangaben angeführt, allerdings ohne weitere Kommentare. Ferner finden sich hier auch weitere Beispiele bzw. Belege, über deren formelhaften Charakter anhand des Wörterbuchs aber keine Aussagen gemacht werden können, vgl. den folgenden Ausschnitt aus dem Wörterbuch:

> Perle**,** [...]
> — d) zur Bez. für etwas Köstliches, Werthvolles, z.B. sprchw.: Sein P-n vor die Säue (Matth. 7,6), vor die Schweine (Platen 2, 138), vor die Pfleglinge des verlorenen Sohnes (JG Müller Lind. 2, 196) werfen; Welche P. [welchen Schatz] warf ich hin! | welch Glück des Himmels hab ich weggeschleudert! Sch. 445a; Wirf nicht für eiteln Glantz und Flitterschein | die echte P. deines Werthes hin! 526a; [...] (D. Sanders, Wörterbuch der deutschen Sprache, Bd. 2, S. 515)

Auch im Wander sind unter dem Lemma *Perle* drei mhd. Textstellen und Entsprechungen aus anderen Sprachen zum Idiom verzeichnet. Auffällig ist, dass Wander den Beleg in einer heutzutage völlig unbekannten morphosyntaktischen Struktur und typologisch gesehen eher in Form eines Sprichworts *Man soll die Perlen nicht vor die Säue werfen* verzeichnet. Die von ihm gewählte morphosyntaktische Struktur und lexikalische Besetzung unterscheiden sich von den im Artikel angeführten Textstellen.

Aufgrund der äußerst seltenen Paraphrasen ist semantische Varianz nur in Ausnahmefällen untersuchbar

Als eine weitere Besonderheit, die auch alle oben angeführten Belegstellen veranschaulichen, stellt sich die Tatsache heraus, dass semantische Angaben zur Bedeutung der formelhaften Wendungen in historischen Wörterbüchern auch im 19. Jh. noch äußerst selten sind. Dies erschwert erheblich die Analyse der semantischen Varianz. Eine auffällige Ausnahme enthält Blums „Deutsches Sprichwörterbuch“ (Bd. 2, S. 71):

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

> 429.
> Man soll die Perlen nicht vor die Säue werfen.
> Wie, wenn man alle die schönen Lehren der Mässigkeit und Enthaltung, die auf Vernunft und Christenthum sich gründen, einer Gesellschaft von Trunkenen auch noch so beredt empfehlen; wie, wenn man Klopstocks Messias einer Studentengesellschaft in der Schenke vorlesen wollte?

Im Vergleich zu heute ist die hier verzeichnete Bedeutung beim Eintrag *Man soll die Perlen nicht vor die Säue werfen* enger auf die Nutzlosigkeit der Tugendlehren für Betrunkene fokussiert. Für semantisch ähnlich hält der Autor des Wörterbuchs diesen Eintrag mit seiner davor stehenden lexikalischen Variante *Wozu sollen der Kuh Muskaten?*, worauf der metasprachliche Kommentar *Aus dem Grunde gebietet ein andres Sprichwort* hindeutet:

> 428.
> Wozu sollen der Kuh Muskaten?
> Wozu Leckerbissen, dem, der sie nicht zu essen versteht, [...]
> Die schätzbarsten Güter, die edelsten und erhabensten Kenntnisse sind da überflüssig und übel angebracht, wo man nicht empfänglich für sie ist, und den Umständen nach eben itzt auch nicht seyn kann. Aus dem Grunde gebietet ein andres Sprichwort:
>
> 429.
> Man soll die Perlen nicht vor die Säue werfen. [...]
> (J. Chr. Blum, Deutsches Sprichwörterbuch, Bd. 2, S. 70f.)

Die Angaben zur Bedeutung im Wörterbuch lassen hier ein breiteres Verwendungsspektrum annehmen (vgl. *Güter*, *Kenntnisse*) als beim Eintrag 429: *Man soll die Perlen nicht vor die Säue werfen*. Diese Beobachtung muss allerdings eine Annahme bleiben, denn für die Existenz des Idioms in dieser Form in der textuellen Überlieferung aus der Zeitspanne 8.-17. Jh. findet sich in der HiFoS-Datenbank kein einziger Nachweis. Natürlich sind solche Sammlungen weder mit Blick auf ihren Aufbau noch ihre Zielsetzung als lexikographische Quellen im strengen Sinn zu betrachten: Im Vordergrund steht die möglichst umfassende Dokumentation der formelhaften Wendungen, oft durch die nationalpatriotische Überzeugung motiviert, dass die hohe Zahl solcher Wendungen den Reichtum der deutschen Sprache beweist.[10] Semantische Paraphrasierungen sind hier deshalb sekundär bzw. kommen gar nicht vor.

Eindruck der absoluten Festigkeit

In den allermeisten Fällen suggerieren die Sammlungen und Wörterbücher eher, dass das Idiom *Perlen vor die Säue werfen* in den älteren Sprachstufen des Deutschen über eine syntaktisch völlig feste Struktur verfügte. Meyers Großes Konversationslexion (hier ausnahmsweise mit dem Idiom im Lemmastatus), die Sammlung „Der Teutschen Weisheit“ von Friedrich Petri und „Die deutschen Sprichwörter“ Karl Simrocks verzeichnen das Idiom zwar in der Form, in der es in den gegenwärtigen Textkorpora und Wörterbüchern gar nicht vorkommt (*Perlen sol man nicht für die Sew werffen / Man soll die Perlen nicht für die Säue werfen*), aber die Form ist laut Angaben der Wörterbücher fest. Wie unten noch zu zeigen sein wird, vermitteln die Wörterbücher somit ein anderes Bild über die Variation im Bereich der formelhaften Wendungen. Damit sei nicht gesagt, dass dieses Bild ein falsches ist, aber es unterscheidet sich deutlich von der Varianz, die sich in den Texten entfaltet.

[10] Vgl. ähnliche Einstellungen und die zentrale Rolle der formelhaften Wendungen bereits in der barocken Spracharbeit in Hundt (2000); Filatkina (2009a). Zum frühen Mittelalter vgl. Filatkina/Hanauska (2011) und Filatkina et al. (2009).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Dies führt mich zur Annahme, dass Kodifizierung im Bereich der formelhaften Wendungen eine geringere Auswirkung auf die Reduzierung der Varianten und Entstehung der Festigkeit hat als z.B. im Bereich der Aussprache (Kohrt 1998), Rechtschreibung (Kohrt 1998) oder Grammatik (Werner 1998; Mattheier 2000). Zwar sehen Burger/Linke (1998, S. 747) in dem, „was wir heute als strukturelle 'Festigkeit' des Phraseologismus fassen, das Produkt der mehrhundertjährigen schriftsprachlichen (insbesondere lexikographischen) Normierung". Diese These muss aber m.E. auch heute mehr als 10 Jahre nach ihrer Entstehung eine unbeantwortete Forschungsfrage bleiben. Zieht man die neuesten Erkenntnisse zur Wirkung der barocken Grammatiker und Wörterbuchschreiber in Betracht (Bergmann 1982; Takada 1998), muss der Zusammenhang zwischen Festigkeit und Normierung – wenn überhaupt – als ein sehr junges Phänomen der gegenwartssprachlichen Print-Lexikographie betrachtet werden. Es wäre zu untersuchen, ob historisch nicht eher der Gebrauch einer Wendung in den großräumig über Jahrhunderte wirkenden Texten eine Rolle spielt, wie das etwa Hüpper/Topalovic/Elspaß (2002, S. 96) für Paarformeln in Eidestexten erarbeiten.

3. Varianz und Textkorpora

Im Weiteren stütze ich mich auf die Ergebnisse der HiFoS-Nachwuchsforschergruppe, die sich erstmalig die Dokumentation, Erfassung und Untersuchung der Dynamik und der Verfestigungsprozesse im Bereich der formelhaften Wendungen in deutschen Texten aus der Zeitspanne von ca. 750 bis ca. 1650 (mit Schwerpunkt in der ahd. Zeit) zum Ziel gesetzt hat. Die HiFoS-Datenbank besteht aus ca. 30.500 kommentierten Belegen; davon stammen ca. 9.630 aus den ahd., ca. 11.800 aus den mhd. und ca. 9.000 aus den fnhd. Texten (Stand: April 2013).

3.1 Varianz auf der Ebene der Form (Morphosyntax und lexikalische Besetzung)

Diachron gesehen befindet sich vor allem die Form einer formelhaften Wendung am meisten und am längsten in Bewegung, bevor sie – wenn überhaupt – zu einer festen Wendung wird.[11] Während gegenwärtig das ausgewählte Idiom als ein seltenes Beispiel für eine formelhafte Wendung mit einem geringen Variationspotenzial gelten kann, bei der alle drei Identifikationsmerkmale der Phraseologismen ohne Einschränkungen gelten (der Beleg ist polylexikalisch, es ist strukturell in sehr hohem Maße fest und es ist idiomatisch), stehen diesem Befund 33 Belegstellen aus den Texten vom 9. bis hin zum 16. Jh. gegenüber, in denen die Wendung kein einziges Mal in der gleichen morphosyntaktischen Struktur und lexikalischen Besetzung vorkommt. Da das Variationspotenzial der Wendung an einer anderen Stelle ausführlich beschrieben wurde (Filatkina 2013a), fasse ich hier die wichtigsten Entwicklungen zusammen.

Von Variation betroffen ist im Gegensatz zur Gegenwartssprache vor allem die erste substantivische Konstituente *Perlen*, die in den ältesten Belegen aus Heliand und Tatian, aber auch in den mittel- und frnhd. Texten als *meri griotun / merrigroze / margariten* 'Perle, Meeressand' vorkommt. Ab der mhd. Zeit ist die Substitution durch andere semantisch verwandte Lexeme möglich, so etwa durch *rôtez golt und edel gesteine*, *edel gestain*, *muschgot vnd negelein*, durch *schöne rosen* (in Kombination mit *Perlen*), durch das semantisch abstraktere *dat gude* und durch *diu gimme*. Möglich ist auch die Erweiterung durch das adjektivische Attribut *edel*. Nur in drei Belegen ist das Vorkommen anderer Konstituenten alleine durch den Reim be-

[11] Das scheint auch für das Englische zu stimmen, wie es Aurich (2012) für Sprichwörter überzeugend nahelegt. Für das Deutsche ab dem 18. Jahrhundert bezugnehmend auf Burger/Linke (1998) und Burger (2000) zuletzt ausführlich Dräger (2011, S. 93ff.).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

dingt und erlaubt somit keine Aussagen über das tatsächliche Variationspotenzial. Allerdings liefern sprichwörtliche Sammlungen, Chroniken, Predigten und andere Prosawerke zahlreiche weitere Belegstellen, bei denen das Reim-Argument nicht greift und die zusätzlich auch keine Übersetzungen aus dem Lateinischen sind. Sie stützen die Beobachtung, dass hier tatsächlich Varianz und keine okkasionelle Modifikation vorliegt – eine Beobachtung, die freilich durch das grundsätzliche Problem der sprachhistorischen Untersuchungen (es handelt sich *token*-mäßig um singuläre Belege) in ihrer Allgemeingültigkeit eingeschränkt bleibt.

Mit der Variation bei der Konstituente *Perlen* ist das Variationspotenzial noch nicht ausgeschöpft: Auch die zweite substantivische Konstituente *Säue* variiert zwischen *gifuinon* (ohne Artikel), *swîn*, *diu swîn*, *die schwein füez*, *Schweine und Hunde*, *mist* oder *die vercken*. Das gleiche gilt auch für die verbale Konstituente *werfen*, und selbst die Präposition *für* kann anders besetzt sein bzw. wegen des Vorkommens des präfigierten Verbs *vorwerfen* fehlen. All diese Beispiele veranschaulichen einen schwächeren Grad an Lexikalisierung in der Diachronie sowie synchron auf den historischen Sprachstufen des Deutschen. Sie zeigen ein Nebeneinander von Varianten mit unterschiedlicher Komponentenzahl und externer Valenz sowie die topologische, morphologische, grammatische und lexikalische Variation.[12] Mit Blick auf die korrekte Abbildung der lexikalischen Variation in einem Wörterbuch stellt sich die Frage danach, ob in der Tat Variation oder eher Synonymie vorliegt, als nicht trivial dar. Im Bereich der festen Strukturen bedarf sie weiterhin einer umfassenden theoretischen Fundierung.

Angesichts der anders als heute gelagerten syntaktischen Einbettung und der ausgeprägten semantisch-pragmatischen Komponente ‘Vermittlung der Moral, Belehrung’ ist es für die älteren Stufen schwierig, die Wendung als Idiom zu bezeichnen. Es ist eher ein Sprichwort, eine Sentenz. Sprichwörter, feste Phrasen und Sentenzen dienen in der Diachronie oft als Ausgangsbasen für zukünftige Idiome. Der Typenwechsel bildet einerseits gleichzeitig einen der Verfestigungswege der formelhaften Wendungen, andererseits kann er über den Abbau der morphosyntaktischen Struktur (wie etwa *bei Perlen vor die Säue*, nun auch ohne das Verb) zur Entstehung von Einzellexemen führen und ist somit als Mittel der Wortbildung zu betrachten.[13]

3.2 Semantische Varianz

Auch die Bedeutung des Idioms *Perlen vor die Säue werfen* verändert sich im Laufe der Geschichte. Dabei handelt es sich hier um die im 13. Jh. beginnende Bedeutungserweiterung in kleinen Schritten, die vom Übergang in ein anderes Stilregister und einer anderen Textsortenpräferenz begleitet ist. In den ältesten deutschen in der Bibeltradition stehenden Texten ist die Verwendung des Idioms stark durch seine Herkunft bestimmt:[14]

(1) Nolite dare sanctum canibus. | neque mittatis margaritas uestras | ante porcos.’ ne forte conculcent eas | pedibus suis. et conuersi | disrumpant uós

[12] Begrifflichkeiten nach Burger (2010), Barz (1992, S. 29) und Fleischer (1997, S. 205).

[13] Auf das Zusammenspiel zwischen Wortbildung und Phraseologie wurde in der Forschungsliteratur mehrfach hingewiesen, vgl. zuletzt Stein (2012). Barz (2007) unterscheidet z.B. die phrasembasierte Wortbildung *(Grundsteinlegung, Vieraugengespräch, Ohne-mich-Standpunkt*) und phrasembasierte Bedeutungsbildung (*Fettnapf*). Als Stichwort sei hier die Univerbierung oder auch die entgegengesetzte Tendenz der Entstehung von Phraseologismen aus Einzellexemen genannt (*radfahren >Rad fahren, Rad schlagen > radschlagen* oder der Zweifelsfall *brustschwimmen* und *Brust schwimmen*).

[14] Alle Belege werden im Folgenden nach der Datenbank der Nachwuchsforschergruppe HiFoS zitiert: Nach der Notation des Kontextes sind in Klammern die Angabe der Textstelle und die ID-Nummer angeführt, mit der der Beleg in der Datenbank versehen ist.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Nicuret heilagaz geban hunton | noh *nisentet íuuara merigrozza* | *furi súin*. min odouuan furtreten sie | mit iro fuozun Inti giuuentite | zibrehhent Iuuih (Tatian 72, 9 Mt; ID 12381)

(2) sithor mah | hie mid is lerun uuerthan helithon te | helpu sithor hie ina hlutteran uuet | Sundiono sicoran *ne sculun gi suinon* | *teforan iuuua meri griotun macon* | *eftha methmo gistriuni helag hals* | *meni* huand sia it an horo spurnat | suiliuuat it an sande niuuitun subres | gisceth fagarero fratoho Sulic sind | hier folc manag thia iuuua helag | uuord horean niuuilliat (Heliand 49r, 11; ID 5782)

Wie in der Bergpredigt ist bei Tatian und Heliand mit *Perlen* ausschließlich die Lehre Gottes, die heilige Lehre gemeint. Es heißt, dass man sie nicht an Menschen vergeuden soll, die ihren Wert nicht zu schätzen wissen, denn es wäre sinnlos. Nur diese Handlung wird als sinnlos betrachtet. Die Wendung ist – soweit man das anhand zweier erhaltener Belege beurteilen kann – seit Beginn der altdeutschen Überlieferung idiomatisch und in ihrem Bezug auf Gottes Lehre an religiöse Kontexte gebunden. Das letztere schließt natürlich den Gebrauch in profanen Kontexten auch im Ahd. prinzipiell nicht aus, allerdings findet sich im u.a. die gesamte ahd. Textüberlieferung umfassenden HiFoS-Korpus kein Beweis dafür. Die biblische Herkunft erlaubt nicht, die Wendung als umgangssprachlich zu bewerten; sie kommt eher in belehrenden, didaktischen Aussagen *Man soll nicht* [...] vor.

Zu Beginn des 13. Jhs. zeichnet sich der erste Bedeutungsschub ab: Im „Wigalois“ des Wirnt von Grafenberg steht die Wendung im Prolog und bezieht sich nicht mehr auf die Lehre Gottes, sondern auf die Lehre des Autors, die er mit seinem Werk vermitteln möchte:

(3) Si wellent daz daz iht witze sîn | swer *rôtez golt under diu swîn* | *werfe und edel* (*ge*)*steine* | des vreuwent si sich doch kleine | si wâren ie vür daz golt | der vil trüeben lachen holt | dâ bewellent si sich inne (Wirnt von Grafenberg, Wigalois, 1204, 75; ID 18019)

Im Prolog bittet der Autor um Nachsicht für sein Werk und dessen aufmerksame Aufnahme, denn er will es nicht umsonst schreiben: Es wäre genauso unklug, wie wenn man Perlen und Edelsteine unter die Schweine werfen würde. Eine weitere Bedeutungserweiterung erfährt die Wendung im „Renner“, also gegen Ende des 13. Jhs.:

(4) Der wirt beroubet ûf der strâzen | Sô diu sêle den lîp muoz lâzen | Daz er dort in sînes vater lant | Niht kumen tar âne schœn gewant | Frâz, hôchfart und gîtikeit | Brâhten uns von êrste in arbeit | Sô machte Kâîn durch nît und haz | Mit bluote sînes vater muoter naz | Do er Âbeln sînen bruoder sluoc | Der traz hât noch geverten genuoc | Des slangen rât und Êven tât | Brâhte alle die werlt in missetât | Des klaget meister Hugewitze | Daz zuht scham kunst und witze | Fleischlichem geluste entwîchen müezen | Und under der gîtikeit füezen | *Ligen als vor swînen edel gesteine*
(Hugo von Trimberg, Der Renner, 1290-1300, 6305; ID 18039)

Eine Bedeutungserweiterung liegt hier insofern vor, als unter Wertvollem nicht mehr eine Lehre verstanden wird, sondern menschliche Tugenden. Der Bezug auf Tugenden bleibt bis ins 15. Jh. hinein zentral: Bei Michel Beheim lesen wir z.B., dass ungeschlachte Menschen, die Bildung und Ehre ablehnen, mit Tieren wie Kalb, Esel oder Schwein verglichen und als Tore bezeichnet werden müssen. Wer sie ausbildet und erzieht, ist einem Toren gleich, der Schweinen Muskat und Nelken als teure, wertvolle Gewürze vorwirft:

(5) Mit diesem peispil da/ sing ich von groben läuten | die zuht vnd er verneuten/ Wer nu dieselben sein | Der namen nenn ich kein/ sie uinden sich wol selber | die vnuerschempten kelber/ selb werden offenbär | Zwar/ sicher es ist wour*/ Wer *würffet für die swein* | *muschgot vnd negelein*/ den gleich ich einem toren | für war es ist uerloren/ we*r* singet oder seit | Zuht er vn*d* hupischkeit vor manchem groben leffel | der seinen wüsten deffel / rihtet vff pös gespey
(M. Beheim, vor 1474-8, 230; ID 18031)

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bereits ab der Mitte des 14. Jhs. ist mit dem Wertvollen auch das Wissen gemeint. So schreibt Konrad von Megenberg in seinem „Buch der Natur“, dass das Wissen über Kräuter in einem „Straßenläufer“ – einem Buch, das der Allgemeinheit zugänglich ist, – geheim gehalten werden soll, weil es dafür zu schade ist:

(6) ez habent auch andreu | kräuter gar wunderleicheu werch sam patönigekraut und | eisenkraut, daz ze latein verbena haizt. iedoch schol man | in diu kniel decken in disem strâzenlaufær, wan ez waer | niht tugentleich getân, der die halichait für die hunt | würfe und der *daz edel gestain under der swein füez* | *würfe* : zwâr, daz wær unpilleich. ich waiz daz wol, daz | liebeu kint selten prôt handelnt, dâ reis den hunden | etwaz von und andern zuckern
(Konrad von Megenberg, Buch der Natur, 1348-1350, 380, 22; ID 18026)

Die heutige Bedeutung des Idioms *Perlen vor die Säue werfen* ist somit das Ergebnis einer Bedeutungserweiterung in kleinen Schritten, die ihre Anfänge im 13. Jh. hat. Die Bedeutungserweiterung geht einher mit den Änderungen in der Textsortenpräferenz.[15] Das ist ein weiterer Typ der Varianz im Bereich der formelhaften Sprache, der sich bei unserem Beispiel darin manifestiert, dass die Gebundenheit an religiöse Kontexte über den Verlust des semantischen Slots ‘Gottes Lehre’ zugunsten einer möglichen Verwendung in profanen Kontexten durchbrochen wird. Diese Trennung ist für das Mittelalter und die Frühe Neuzeit bekanntlich schwierig, aber in Bezug auf die angeführten Kontexte lässt sie sich m.E. durchführen (vgl. auch Kapitel 3.3 unten). Die Bedeutungserweiterung wird bei diesem Idiom aber auch durch den deutlichen Übergang vom gehobenen ins umgangssprachliche Stilregister begleitet, der nach den Ergebnissen des HiFoS-Projekts für die diachrone Dynamik der formelhaften Sprache genau in dieser Richtung prägend ist (Filatkina 2013a, 2013b). Die Bedeutungserweiterung ist möglicherweise durch die Verdunklung der kulturhistorischen Grundlage – den nächsten Typ der Varianz und des Wandels – zu erklären.

3.3 Varianz in der kulturhistorischen Grundlage

Unter kulturhistorischer Grundlage werden jene Aspekte der Kultur verstanden, die den bildlichen Hintergrund der formelhaften Wendung motivieren und/oder die Ausgangsbasis für die Entstehung der Metaphorizität bilden, z.B. Symbole der Kultur (wie in *sein Herz ausschütten*), Artefakte der materiellen (Alltags-)Kultur (*etwas auf dem Kerbholz haben*), stereotype Einstellungen und/oder Tabus (*Lange Haare, kurzer Verstand*), semiotisierte Gestik (*mit den Achseln/Schultern zucken*) usw. Eine Klassifikation solcher Phänomene wurde in Dobrovol’skij/Piirainen (2005) für moderne Sprachen vorgeschlagen; in HiFoS wurde sie ausgehend vom historischem Material um einige Domänen ergänzt bzw. gekürzt. So verfügt auch das Idiom *Perlen vor die Säue werfen* über solch eine kulturhistorische Grundlage. Sie besteht in der Intertextualität und konkret darin, dass das Idiom aus der Bergpredigt des Matthäusevangeliums, die Jahrhunderte hindurch bis in die allerjüngste Zeit einen herausragenden, äußerst bekannten Text darstellte, stammt: *Neque mittatis margaritas vestras ante porcos ne forte conculcent pedibus suis* (Matthäus 7,6 Vulgata).[16] Die Theorie des bildlichen Lexikons (*Conventional Figurative Language Theory*) geht von der Annahme aus, dass die bildliche Grund-

[15] Noch deutlicher lässt sich der Wandel in der Textsortenpräferenz beim Übergang eines phraseologischen Terminus aus der Fachsprache in die allgemeine Lexik beobachten.

[16] Das Vorkommen der Konstituente *Perlen* für *margaritas* hat die Forschung auf eine Tradition der byzantischen Kirche zurückgeführt, in der das heilige Brot, als kleine Brocken zerkrümelt, *margaritas* genannt wurde und das Neugriechische Perlen und Brotkrümel immer noch mit demselben Begriff bezeichnet (Röhrich 2004, Bd. 2, S. 1148). So wäre die Bibelstelle sinngemäß zu interpretieren als ‘Wirf nicht den Hunden das geheiligte Fleisch und den Schweinen das geheiligte Brot vor’. Zu Missverständnissen bzw. falschen Übersetzungen von griech. *μαργαρίτης* ‘Perle’ über lat. *margarīta* ‘Perle’ durch die mittelalterliche volkssprachliche *Margerite* als Blume vgl. auch Piirainen (2012, S. 231) und Mokienko (2011).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

lage ein wichtiges Element des Inhaltsplans figurativer Einheiten ist und ihre Auswirkungen auf die Verwendung der Idiome, d.h. auf ihre lexikalisierte figurative Bedeutung, hat (Dobrovol'skij/Piirainen 2009, S. 183). Die Analyse des Wandels beim Idiom *Perlen vor die Säue werfen* erlaubt die Vermutung, dass diese Annahme auch für die Diachronie der Entwicklung gilt und selbst dann greift, wenn die kulturhistorische Grundlage in der Intertextualität (und nicht nur in der synchron motivierbaren, auf der bildlichen Vorstellung basierenden wörtlichen Lesart) besteht. Wie oben gezeigt, hat die biblische Herkunft die ursprüngliche Verwendung des Belegs nur in Bezug auf die Lehre Gottes motiviert. Vermutlich sind sich die meisten (besonders kulturhistorisch nicht gebildeten) Muttersprachlerinnen und Muttersprachler des Deutschen der biblischen Herkunft nicht mehr bewusst. Hier findet also ein Wandel statt, der im Verblassen der kulturhistorischen Grundlage besteht und der m.E. eine Reihe von Konsequenzen mit sich zieht. Zum einen kann die Bedeutungserweiterung im Verblassen der Motivationsgrundlage ihren Ursprung haben: Je weniger bewusst die biblische Herkunft ist, desto uneingeschränkter ist der Kreis der Gebrauchskontexte. Es ist unwahrscheinlich, dass die These über das verblasste Bewusstsein der biblischen Herkunft für das Mittelalter und die frühe Neuzeit genauso wie für heute gelten kann. Aber wir beobachten seit dem 13. Jh. den Verlust des Bezugs auf die Lehre Gottes und den Übergang der Wendung aus den Texten in zahlreiche Kunstwerke (etwa das Simultangemälde von Pieter Bruegel dem Älteren „Der Blaue Mantel", 1559, oder sein aus 12 Tafeln zusammengefügtes Werk „Zwölf Sprichwörter", 1558, um nur zwei der bekanntesten zu nennen), die didaktische Funktionen in einem nicht unbedingt theologischen Kontext bzw. gar keine didaktischen Funktionen hatten und auf denen nicht die Perlen, sondern die Blumen (Margeriten oder Rosen) geworfen werden – wohl basierend auf den Missverständnissen der griechischen und lateinischen biblischen Vorlagen in den Volkssprachen (vgl. Anmerkung 16 und die Abbildungen 2 und 3). Dieser Übergang kann den wohl viel später einsetzenden Wandel im Bewusstsein beeinflusst haben.

Abb. 2: Pieter Bruegel, d.Ä., Ausschnitt aus dem „Blauen Mantel", 1559, Staatliche Museen Berlin, Stiftung Preußischer Kulturbesitz.

Abb. 3: Pieter Bruegel, d.Ä., ein Tafelbild aus den „Zwölf Sprichwörtern", 1558, Museum Mayer van den Bergh, Antwerpen.

Zum zweiten kann der Wandel der kulturhistorischen Grundlage Auswirkungen auf die Form der Wendung gehabt haben. Oben wurde erwähnt, dass es im diachronen Schnitt 33 Beleg-

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

stellen mit dieser Wendung gibt. Einige davon sind lexikalische Varianten, in denen die Konstituente *Perlen* durch andere semantisch verwandte substituiert war, z.B. *Edelsteine* oder *gimme*. Nimmt man an, dass diese Art von Substitution keine okkasionelle Modifikation ist, so ist sie nur möglich, wenn die kulturhistorische Grundlage opak wird.

Die dargestellte Abhängigkeit der unterschiedlichen Varianztypen voneinander ist durch die „mehrgliedrige Struktur“ (Burger 2000, S. 35) der formelhaften Wendungen als Zeichen der sekundären Nomination bedingt. Grafisch lässt sie sich wie folgt darstellen:

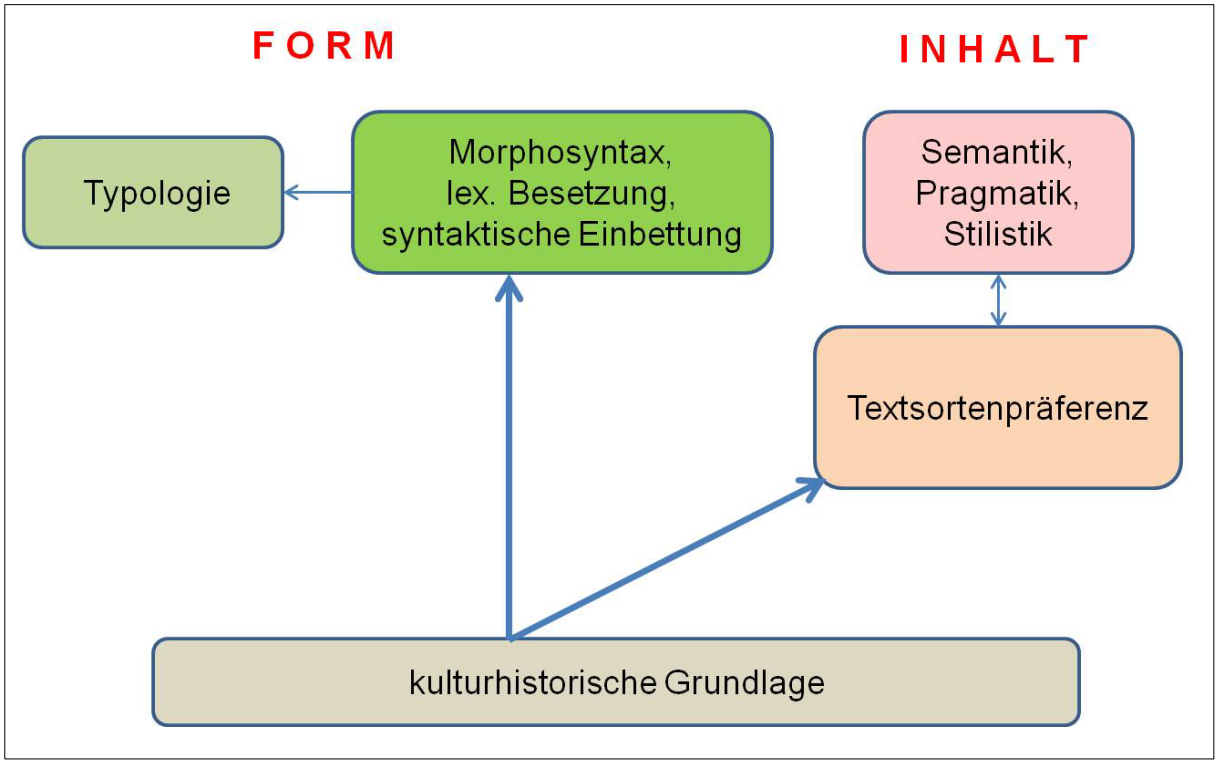


Abb. 4: Die Abhängigkeit einzelner Varianztypen voneinander

4. Umgang mit Varianz in der HiFoS-Datenbank

Die Vielfalt und Komplexität der Varianz in ihrer Entfaltung in historischen Texten stand im Mittelpunkt der Analysen in der Nachwuchsforschergruppe „Historische formelhafte Sprache und Traditionen des Formulierens (HiFoS)“ an der Universität Trier. Aufgrund dieser Komplexität und insbesondere mit Blick darauf, dass für formelhafte Wendungen in den mhd. und fnhd. Texten systematische Erhebungen nach wie vor fehlen, bestand eines der wichtigsten methodischen Postulate des HiFoS-Projekts darin, jede Variante einer formelhaften Wendung zunächst einzeln als eigenständigen Beleg in der Datenbank zu erfassen und im Kontext zu kommentieren. Abbildung 5 veranschaulicht den kommentierten Tatian-Beleg mit der ältesten überlieferten Form der formelhaften Wendung *Perlen vor die Säue werfen*:

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Abb. 5: Die Erfassungsmaske „Beleg-Kontext“ in der HiFoS-Datenbank

Die Abbildung macht einen grundsätzlichen Unterschied zur gegenwärtigen lexikographischen Praxis deutlich, bei der in idiomatischen Wörterbüchern eine Nennform definiert wird, deren Varianten durch Schrägstriche, eckige Klammern oder in Belegkontexten im gleichen Wörterbuchartikel mit angeführt werden (vgl. Abb. 1 am Anfang des vorliegenden Beitrags). In die HiFoS-Datenbank werden die Belege so aufgenommen, wie sie in Kontexten tatsächlich vorkommen; es werden keine übergeordneten Nennformen formuliert, weil es für historische Sprachstufen nur bei ausreichender Menge hochvariabler Daten möglich ist, die gegenwärtig noch nicht erreicht wurde. Dieses Vorgehen erlaubt eine detaillierte Beschreibung der Varianz bei jedem Beleg auf allen Ebenen in der Synchronie und auf allen historischen Sprachstufen des Deutschen.

Es ermöglicht aber auch die Zusammenführung der einzelnen in der Datenbank zunächst isoliert stehenden Varianten, um zu einem kompletten Bild der Varianz zu gelangen. Die Zusammenführung bzw. Bündelung der Varianten ist in der Datenbank auf zweifache Art und Weise möglich. Zum einen können insbesondere lexikalische und grammatische Varianten manuell[17] über die Metalemmata gebündelt werden, die bei der Kommentierung des Konstituentenbestandes für jede Konstituente einer Wendung in der Regel auf Nhd. formuliert werden. Abbildung 6 veranschaulicht die Metalemmata für den Konstituentenbestand des ältesten überlieferten Tatian-Belegs:

[17] Zu Vorteilen und Nachteilen des manuellen Belegverwaltungstools vgl. ausführlich Filatkina (2009b).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

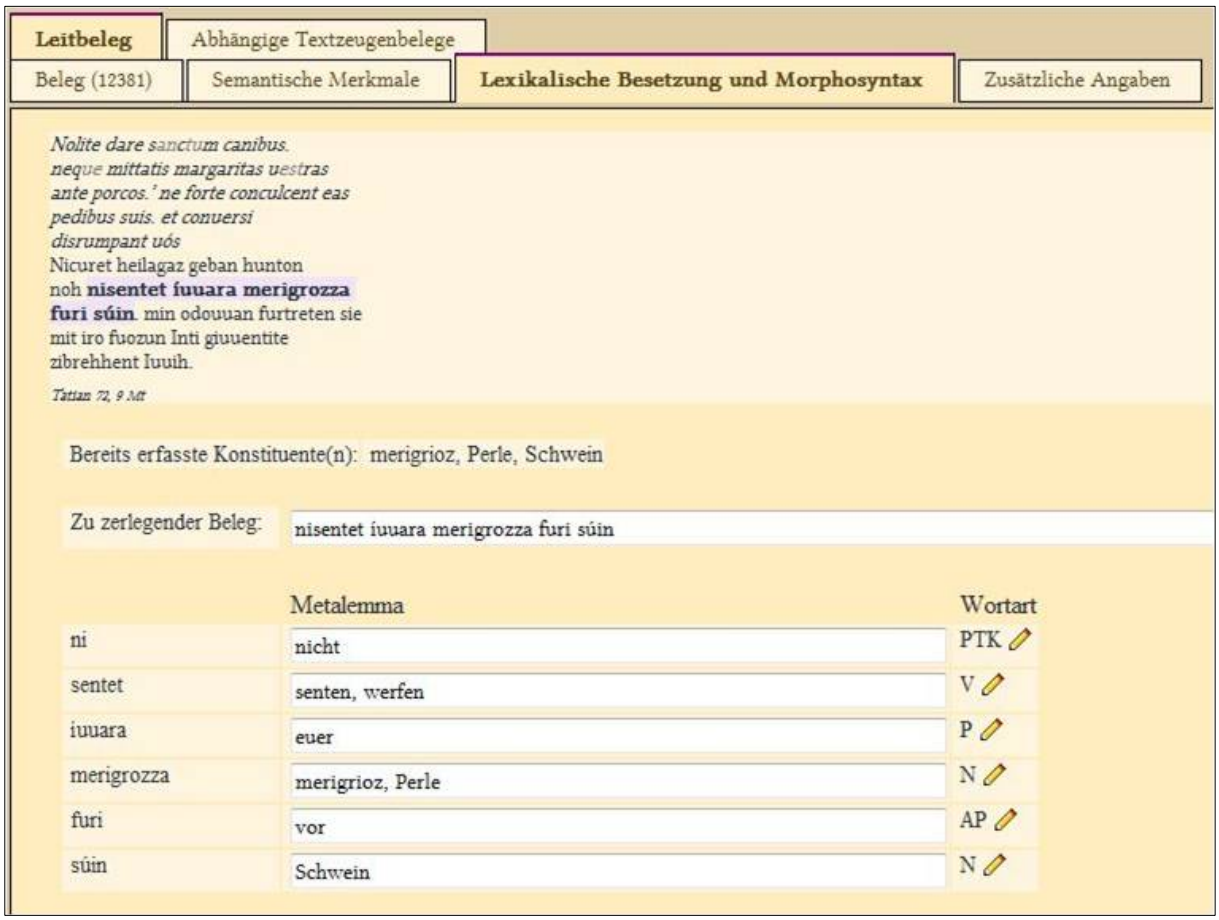


Abb. 6: Der Tatian-Beleg für *Perlen für die Säue werfen* mit Metalemmata

Die Abbildung weist aber auch gleichzeitig auf die Schwierigkeiten bei diesem Verfahren hin: Das Nhd. als Bezugssystem für die Formulierung der Metalemmata ist weniger hilfreich im Fall des Bedeutungswandels bei einzelnen Konstituenten (vgl. ahd. *senten* – nhd. 'werfen') sowie in den Fällen, in denen phonetisch und grafisch ähnliche nhd. Entsprechungen fehlen (vgl. ahd. *merigrioz* – nhd. *Perle*). In solchen Fällen wurden den nhd. Lemmata die Lemmata der jeweiligen Sprachstufen in der Rechtschreibung vorangestellt, in der sie in den Referenzwerken für die jeweilige Sprachstufe vorkommen.[18] Über die Angabe der Metalemmata in einer Suchmaske gelangt man zu einem Referenzkorpus bzw. Subkorpus mit Belegen und den dazugehörigen Metadaten, die nun zu engerer Analyse der Varianz zusammengefasst und als Gruppe für einen unbegrenzten Zeitraum gespeichert werden können.

Um die Bearbeiter bei der manuellen Erstellung der Beleggruppen und somit auch bei der Untersuchung der Varianz zu unterstützen, wurde zum zweiten ein Programm implementiert, das automatisch Ähnlichkeiten zwischen Belegen berechnet (Dostert 2009). Es zerlegt die zu vergleichenden Attribute (Belege) in Bi-Gramme und bestimmt die Ähnlichkeit über den Anteil der übereinstimmenden Bi-Gramme, vgl. Abbildung 7.

[18] Für das Ahd. wurde aufgrund seiner Abgeschlossenheit das „Althochdeutsche Wörterbuch" Rudolf Schützeichels gewählt.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

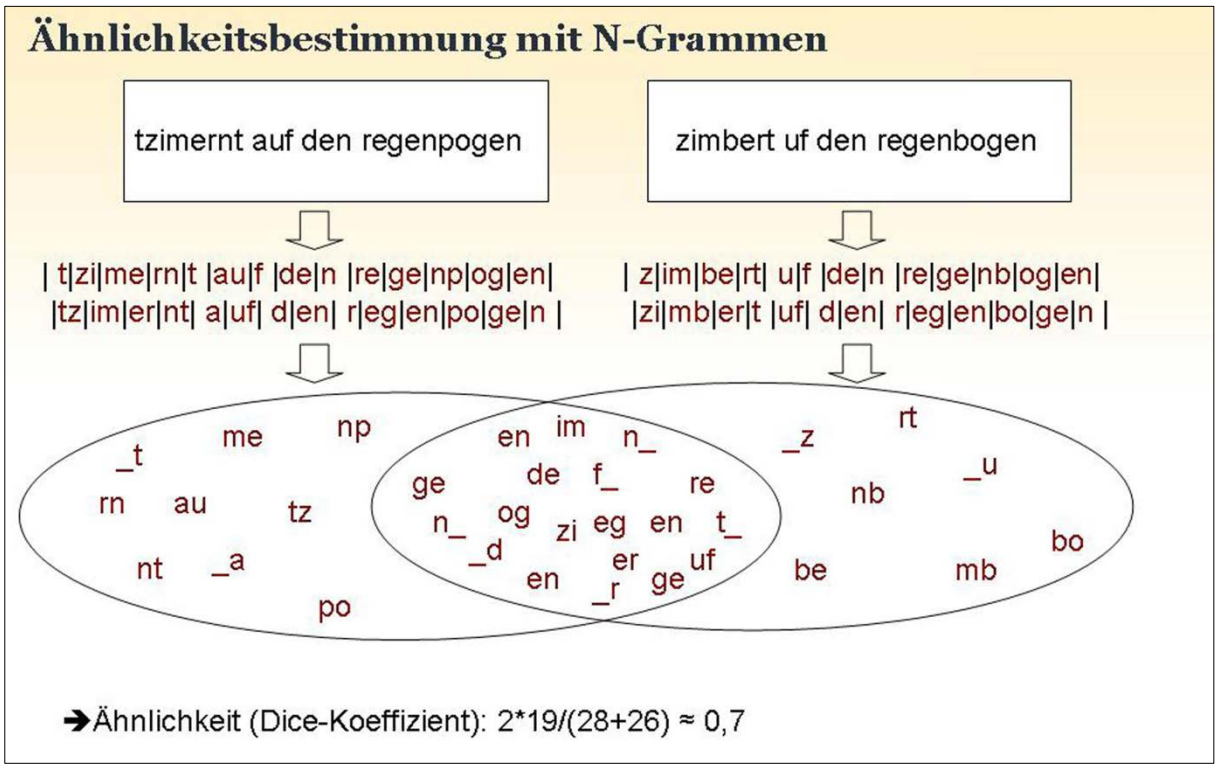


Abb. 7: Ähnlichkeitsbestimmung mit Bi-Grammen in der HiFoS-Datenbank

Es wird hierbei davon ausgegangen, dass zwei Belege umso ähnlicher sind, je mehr die Werte der einzelnen Attribute übereinstimmen. Das Programm erhält als Eingabe einen Referenzbeleg und berechnet paarweise die Ähnlichkeit zwischen diesem und allen in der Datenbank gespeicherten Belegen, deren orthographische Varianz (im Hintergrund für das Tool und nicht in den Primärdaten!) normalisiert wurde. Als Ergebnis liegt eine absteigend nach dem Grad der Ähnlichkeit sortierte Liste der Belege mit den höchsten Übereinstimmungen vor. Die betrachteten Attribute können der jeweiligen Fragestellung entsprechend individuell ausgewählt und unterschiedlich stark gewichtet werden.

Laut Dostert (2009) ist der Einsatz des n-Gramm-Verfahrens bei der Bündelung der semantischen Varianten nicht zielführend. Die Abbildung der semantischen Variation ist in der HiFoS-Datenbank gegenwärtig ausschließlich manuell möglich, über eine im Projekt entwickelte Ontologie. Abbildung 8 veranschaulicht einen Ausschnitt aus der Ontologie sowie das Feld „Zielkonzept“ in den Erfassungsmasken der Datenbank, über welches die Ontologie mit der Datenbank verknüpft ist. Die Ontologie ist nicht gleichzusetzen mit der Metalemmaliste: Während sich die erstere auf die Inhaltsebene der formelhaften Wendungen und die Konzepte bezieht, die mit ihrer Hilfe versprachlicht werden, enthält die letztere die Konstituenten in der Struktur der Wendungen und zielt somit auf ihre Ausdrucksseite ab.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

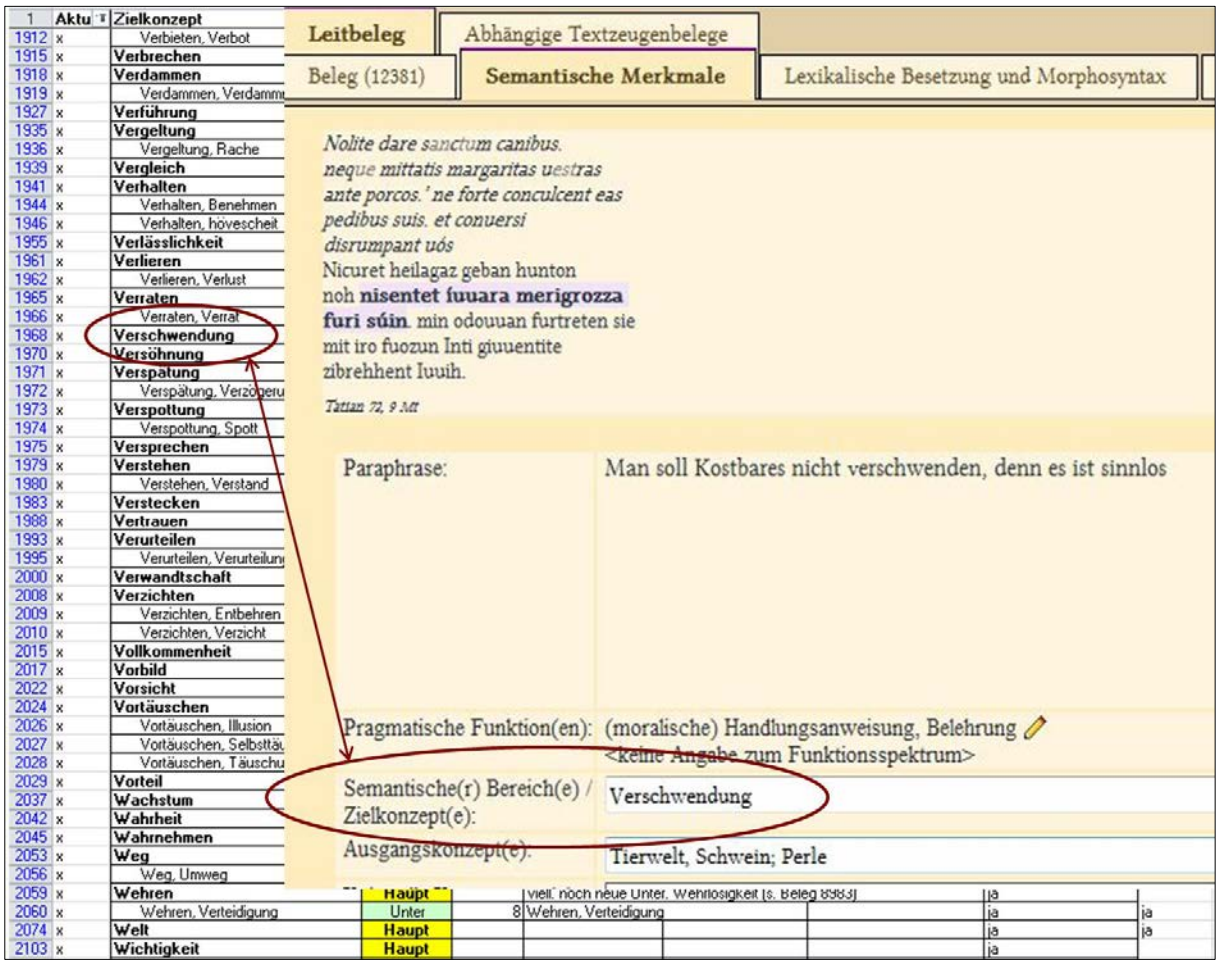


Abb. 8: Die Bündelung der semantischen Varianten in der HiFoS-Datenbank und die Ontologie des Projekts

5. Anstelle einer Zusammenfassung

Der vorliegende Beitrag hat gezeigt, dass die Formelhaftigkeit der „Strukturen“ und ihre Festigkeit Produkte der vielfältigen und ineinander greifenden diachronen Prozesse der Variation und des Wandels sind. Die Nachwuchsforschergruppe HiFoS an der Universität Trier stellte den ersten theoretisch wie methodisch neuen Versuch dar, diese Varianz in der Diachronie systematisch zu erfassen und in ihren Wechselwirkungen zu beschreiben. Sowohl historische Wörterbücher als auch historische Texte erweisen sich dabei als ergiebige Quellen für die Untersuchung der Varianz, sie vermitteln allerdings ganz unterschiedliche Bilder davon. Diese Diskrepanz wirft die Frage auf, ob der Kodifizierung im Bereich der formelhaften Wendungen beim Abbau der Varianz in der Diachronie die gleiche Rolle wie in der Aussprache, Rechtschreibung oder Grammatik zukommt. Die vorläufigen Ergebnisse des HiFoS-Projekts erlauben eher den Einfluss anderer Faktoren anzunehmen. Der Unterschied zwischen der Abbildung der Varianz in Wörterbüchern und ihrer Entfaltung in Texten lässt sich für die Gegenwart ebenfalls feststellen, allerdings ist er eher durch die Verfahren der modernen Print-Lexikographie bedingt. Die im Beitrag dargestellten neuen Ansätze der korpusbasierten (Online-)Mehrwortlexikographie lassen darauf hoffen, dass dieser Unterschied sich zugunsten der real vorkommenden Varianz beheben lässt – ein Programm, von dessen Realisierung in Synchronie wie Diachronie die derzeitige Forschung noch weit entfernt ist.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

6. Literatur

Primärquellen

Duden 11 = DUDEN (2008): Redewendungen. Wörterbuch der deutschen Idiomatik. 3., neu bearb. und aktual. Aufl. (= Der Duden 11). Mannheim.

Blum, Joachim Christian (1739/1790 [1990]): Deutsches Sprichwörterbuch. 2 Bde. in einem Band. Leipzig. [Nachdruck der Ausgabe Leipzig 1780 und 1782. Hildesheim 1990].

Campe, Joachim Heinrich (1807 [1969]): Wörterbuch der Deutschen Sprache. 5 Bde. Braunschweig. [Nachdruck hrsg. v. Henne, Helmut. Hildesheim 1969].

Petri, Friedrich (1604/1605 [1983]): Der Teutschen Weissheit / Das ist: Außerlesen kurtze / sinnreiche / lehrhafte und sittige Sprüche und Sprichwörter in schönen Reimen oder schlecht ohn Reim. Hamburg. [Faksimiledruck hrsg. v. Mieder, Wolfgang. Bern/Frankfurt a.M. 1983].

Sanders, Daniel (1860 [1969]): Wörterbuch der Deutschen Sprache mit Belegen von Luther bis auf die Gegenwart. 3 Bde. Leipzig. [Nachdruck Hildesheim 1969].

Simrock, Karl J. (1846): Die deutschen Sprichwörter. Frankfurt a. M.

Trierer Wörterbuchnetz. http://woerterbuchnetz.de/.

HiFoS Datenbank = Datenbank des Projekts „Historische formelhafte Sprache und Traditionen des Formulierens (HiFoS)“. www.hifos.uni-trier.de.

Röhrich, Lutz (2004): Lexikon der sprichwörtlichen Redensarten. 3 Bde. 7. Aufl. Darmstadt.

Schützeichel, Rudolf (2006): Althochdeutsches Wörterbuch. 6., überarb. und um Glossen erw. Aufl. Tübingen.

Sekundärliteratur

Aurich, Claudia (2012): Proverb structure in the history of English: Stability and change. A corpus-based study. Baltmannsweiler.

Barz, Irmhild (1992): Phraseologische Varianten. Begriff und Probleme. In: Földes, Csaba (Hg.): Deutsche Phraseologie in Sprachsystem und Sprachverwendung. Wien, S. 25-47.

Barz, Irmhild (2007): Wortbildung und Phraseologie. In: Burger, Harald et al. (Hg.): Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung. 1. Halbbd. Berlin/New York, S. 27-36.

Bergmann, Rolf (1982): Zum Anteil der Grammatiker an der Normierung der neuhochdeutschen Schriftsprache. In: Sprachwissenschaft 7, S. 261-281.

Besch, Werner et al. (Hg.) (1998): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Bd. 1. 2., vollst. neu bearb. und erw. Aufl. Berlin/New York.

Bopp, Franz (1816 [1975]): Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache. Hg. und mit Vorerinnerungen begleitet von Karl Joseph Windischmann. [Nachdruck Hildesheim/New York 1975].

Burger, Harald (2000): Konzepte von „Variation“ in der Phraseologie. In: Häcki Buhofer, Annelies (Hg.): Vom Umgang mit sprachlicher Variation. Soziolinguistik, Dialektologie, Methoden und Wissenschaftsgeschichte. Tübingen/Basel, S. 35-51.

Burger, Harald et al. (Hg.) (2007): Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung. 2. Halbbd. Berlin/New York, S. 909-918.

Burger, Harald (2010): Phraseologie. Eine Einführung am Beispiel des Deutschen. 4., überarb. Aufl. Berlin.

Burger, Harald (2012): Alte und neue Fragen, alte und neue Methoden der historischen Phraseologie. In: Filatkina et al. (Hg.), S. 1-20.

Burger, Harald/Linke, Angelika (1998): Historische Phraseologie. In: Besch et al. (Hg.), S. 743-755.

Dobrovol’skij, Dmitrij/Piirainen, Elisabeth (2005): Figurative language. Cross-cultural and cross-linguistic perspectives. Amsterdam u.a.

Dobrovol’skij, Dmitrij/Piirainen, Elisabeth (2009): Zur Theorie der Phraseologie. Kognitive und kulturelle Aspekte. Tübingen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Dostert, Heiko (2009): Ähnlichkeitssuche in sprachhistorischen hochvariablen Daten. Eine Studie am Beispiel des Belegkorpus der Nachwuchsforschergruppe „Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)“. Diplomarbeit, Universität Trier.

Dräger, Marcel (2011): Der phraseologische Wandel und seine lexikographische Erfassung. Konzept des „Online-Lexikons zur diachronen Phraseologie (OldPhras)“. Diss., Universität Freiburg i.Br.

Dräger, Marcel (2012): Plädoyer für eine diachrone Perspektive in der Phraseographie. In: Filatkina et al. (Hg.), S. 193-226.

Feilke, Helmut (1994): *Common sense*-Kompetenz. Überlegungen zu einer Theorie „sympathischen“ und „natürlichen“ Meinens und Verstehens. Frankfurt a. M.

Feilke, Helmut (1996): Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik. Frankfurt a.M.

Filatkina, Natalia (2009a): *Und es duencket einem noch/wann man euch ansiehet/daß ihr Sand in den Augen habt*. Phraseologismen in ausgewählten historischen Grammatiken des Deutschen. In: Földes, Csaba (Hg.): Phraseologie disziplinär und interdisziplinär. Tübingen, S. 15-31.

Filatkina, Natalia (2009b): Historische formelhafte Sprache als „harte Nuss“ der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt. In: Linguistik online 39, 3, S. 75-95. www.linguistik-online.de/39_09/filatkina.html (Stand: 5.5.2015).

Filatkina, Natalia (2010): Phraseologie der germanischen Sprachen kontrastiv: Geschichte, Ergebnisse und Perspektiven. In: Dammel, Antje/Kürschner, Sebastian/Nübling, Damaris (Hg.): Kontrastive Germanische Linguistik. 1. Teilbd. (= Germanistische Linguistik 206-209.1). Hildesheim, S. 275-309.

Filatkina, Natalia (2011): Variation im Bereich der formelhaften Wendungen am Beispiel der Luxemburger Rechnungsbücher (1388-1500). In: Elspaß, Stephan (Hg.): Sprachvariation und Sprachwandel in der Stadt der Frühen Neuzeit. Heidelberg, S. 79-95.

Filatkina, Natalia (2012): *Wann wer beschreibt der welte stat | der muoß wol sagen wie es gat.* Manifestation, functions and dynamics of formulaic patterns in Thomas Murner’s “Schelmenzunfft” revisited. In: Filatkina et al. (Hg.), S. 21-44.

Filatkina, Natalia (2013a): Wandel im Bereich der historischen formelhaften Sprache und seine Reflexe im Neuhochdeutschen: Eine neue Perspektive für moderne Sprachwandeltheorien. In: Vogel, Petra (Hg.): Sprachwandel im Neuhochdeutschen. Berlin/New York, S. 34-51.

Filatkina, Natalia (2013b): *Wehre auch der Teutschen Jugend zu vielen guten ersprießlich / wan die Teutschen Sprichwoerter recht bey zeiten beygebracht und erklaeret wuerden.* Formelhafte Wendungen im mittelalterlichen und frühneuzeitlichen Sprachunterricht. In: Kášová, Martina (Hg.): Wege zu Sprache und Literatur. Festschrift anlässlich des 70. Geburtstages von Ladislav Sisák. Prešov, S. 65-94.

Filatkina, Natalia/Hanauska, Monika (2011): Wissensstrukturierung und Wissensvermittlung durch Routineformeln: Am Beispiel ausgewählter althochdeutscher Texte. In: Yearbook of Phraseology 1, S. 45-71.

Filatkina, Natalia et al. (2009): Formelhafte Sprache im schulischen Unterricht im Mittelalter: Am Beispiel der so genannten „Sprichwörter“ in den Schriften Notkers des Deutschen von St. Gallen. In: Sprachwissenschaft 34, 4, S. 341-397.

Filatkina, Natalia et al. (Hg.) (2012): Aspekte der historischen Phraseologie und Phraseographie. Heidelberg.

Fleischer, Wolfgang (1997): Phraseologie der deutschen Gegenwartssprache. 2., durchges. und erg. Aufl. Tübingen.

Gardt, Andreas (1999): Geschichte der Sprachwissenschaft in Deutschland vom Mittelalter bis ins 20. Jahrhundert. Berlin/New York.

Goldberg, Adele E. (1995): Constructions. A construction grammar approach to argument structure. Chicago.

Grimm, Jacob (1819-1837 [1967]): Deutsche Grammatik. [Nachdruck der Ausgabe von 1870. Hildesheim 1967].

Hanauska, Monika (2012): *Hystoria dye is eyn gezuyge der zijt*. Untersuchungen zur pragmatischen Formelhaftigkeit in der volkssprachigen Kölner Stadthistoriographie des Spätmittelalters. Diss. Universität Trier.

Hemmi, Andrea (1994): *Es muss wirksam werben, wer nicht will verderben*. Kontrastive Analyse von Phraseologismen in Anzeigen-, Radio- und Fernsehwerbung. Bern.

Hoey, Michael (2005): Lexical priming. A new theory of words and language. London/New York.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Hundt, Markus (2000): „Spracharbeit“ im 17. Jahrhundert. Studien zu Georg Philipp Harsdörffer, Justus Georg Schottelius und Christian Gueintz. Berlin/New York.

Hüpper, Dagmar/Topalovic, Elvira/Elspaß, Stephan (2002): Zur Entstehung und Entwicklung von Paarformeln im Deutschen. In: Piirainen, Elisabeth/Piirainen, Ilpo Tapani (Hg.): Phraseologie in Raum und Zeit. Akten der 10. Tagung des Westfälischen Arbeitskreises „Phraseologie/Parömiologie“, Münster 2001. Baltmannsweiler, S. 77-99.

Kleine-Engel, Ane (2012): Some arguments for a historical approach to phraseology in not (fully) standardized languages. In: Filatkina et al. (Hg.), S. 127-146.

Kohrt, Manfred (1998): Historische Phonologie und Graphematik. In: Besch et al. (Hg.), S. 551-572.

Labov, William (2001): Principles of linguistic change. Bd. 1-3. Oxford.

Langacker, Ronald W. (1987): Foundations of cognitive grammar. Stanford.

Mattheier, Klaus (2000): Die Herausbildung neuzeitlicher Schriftsprachen. In: Besch, Werner et al. (Hg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Bd. 2. 2., vollständig neu bearb. und erw. Aufl. Berlin/New York, S. 1085-1107.

Mokienko, Valerij M. (2011): Biblismen als Quellen der Europäisierung nationaler Phraseologismen und Sprichwörter. In: Pamies, Antonio/Dobrovol’skij, Dmitrij (Hg.): Linguo-cultural Competence and Phraseological Motivation. Baltmannsweiler, S. 91-100.

Moon, Rosamund (2007): Phraseology in general monolingual dictionaries. In: Burger et al. (Hg.), S. 909-918.

Müller, Peter O./Kunkel-Razum, Kathrin (2007): Phraseographie des Deutschen. In: Burger et al. (Hg.), S. 939-949.

Paul, Hermann (1995): Prinzipien der Sprachgeschichte. 10., unveränd. Aufl. Tübingen.

Piirainen, Elisabeth (2012): Widespread idioms in Europe and beyond. Towards a lexicon of common figurative units. New York.

Sabban, Annette (1998): Okkasionelle Variationen sprachlicher Schematismen. Eine Analyse französischer und deutscher Presse- und Werbetexte. Tübingen.

Sinclair, John (1991): Corpus, concordance, collocation. Oxford.

Stein, Stephan (2012): Phraseologie und Wortbildung des Deutschen. Ein Vergleich von Äpfeln mit Birnen? In: Prinz, Michael/Richter-Vapaatalo, Ulrike (Hg.): Idiome, Konstruktionen, „verblümte Rede“. Beiträge zur Geschichte der germanistischen Phraseologieforschung. Stuttgart, S. 225-240.

Steyer, Kathrin (2011): Von der sprachlichen Oberfläche zum Muster: Zur qualitativen Interpretation syntagmatischer Profile. In: Travaux Neuchâtelois de Linguistique 55, S. 219-239.

Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht. Tübingen.

Takada, Hiroyuki (1998): Grammatik und Sprachwirklichkeit von 1640-1700. Zur Rolle deutscher Grammatiker im schriftsprachlichen Ausgleichsprozess. Tübingen.

Werner, Otmar (1998): Historische Morphologie. In: Besch et al. (Hg.), S. 572-596.

Wray, Alison (2002): Formulaic language and the lexicon. Cambridge u.a.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Lexikalische Vielfalt und Varianz aus kontrastiver Perspektive

Überlegungen zu einem Produktionswörterbuch aus der Sicht des Deutschen und Spanischen

Meike Meliss (*Universidad de Santiago de Compostela*)

1. Vorbemerkungen

Der Beitrag steht in direkter Verbindung mit dem Forschungsprojekt DICONALE, welches sich mit der Erstellung eines onomasiologisch-konzeptuell orientierten Produktions-Wörterbuches mit bilateralem Online-Zugang für das Sprachenpaar Deutsch-Spanisch beschäftigt.[1] Es sollen die Relevanz der adäquaten Behandlung lexikalischer Vielfalt und Varianz in zweisprachigen lexikographischen Werken für fremdsprachige Produktionssituationen thematisiert und neue Vorschläge, die im Rahmen des besagten Projekts ausgearbeitet wurden, zur Diskussion gestellt werden. Ausgangspunkt ist die L2-Produktionssituation in Deutsch bzw. Spanisch als Fremdsprache (DaF/ELe) und die Beobachtung, dass die gängigen ein- und zweisprachigen Lernerwörterbücher (LWB) für fremdsprachige Produktionssituationen nur ungenügende Information für den fremdsprachigen Wortfindungs- und Wortauswahlprozess bereitstellen (Honnef-Becker 2002) und auch für die adäquate Benutzung im konkreten fremdsprachigen Kontext zu wenig Information zur Bedeutungsdisambiguierung bei Polysemie, zu den konkreten morpho-syntaktischen Realisierungsmöglichkeiten, zu Kollokationen und zu alternierenden Ausdrucksformen anbieten. Der L2-Wörterbuchbenutzer muss für seinen fremdsprachigen Produktionszweck aus der Vielfalt der sprachlichen Ausdrucksmittel auswählen und dabei zusätzlich die mögliche Bedeutungs- und Konstruktionsvarianz eines Lemmas berücksichtigen. Im Allgemeinen ist die Wörterbuchinformation für den L2-Findungs- und Auswahlprozess zu wenig miteinander vernetzt, aber auch für die Anwendung in einer konkreten L2-Kommunikationssituation nur unzureichend strukturiert (Meliss 2013, 2014a, 2015b, 2015c, 2016b). Zwar bieten einsprachige syntagmatische Spezialwörterbücher[2] konkrete Informationen zu verschiedenen Aspekten der Kombinatorik an, die für den L2-Produktionsprozess nutzbar gemacht werden können. Auch die einsprachige Lernerlexikographie hat u.a. im DaF-Bereich in der letzen Dekade durch die konsequente Lesartdisambiguierung mit Hilfe der Angabe von Strukturformeln (Dentschewa 2006; Gouws 1998; Schafroth 2002) und der Einbeziehung von Information zu Kollokationen und häufigen Verbindungen (Köster/Neubauer 2002; Lehr 1998) für die sprachliche Produktionssituation große Leistung getan. Leider hat aber die zweisprachige Lernerlexikographie in diesem Sinne noch kaum Fortschritte vorzuweisen (Model 2010). Auf konkrete Schwierigkeiten bezüglich der Wortfindung und Wortauswahl bei der Produktion eines (fremdsprachigen) Textes hat die ein- und zweisprachige lexikographische Forschung bis jetzt nur völlig ungenügend reagiert, obwohl diese Aktivitäten der eigentlichen sprachlichen Produktion zunächst vorausgehen.

Das geplante Online-Wörterbuch DICONALE will genau diesen Defiziten entgegentreten und ausgehend von einer onomasiologisch-konzeptuellen Perspektive für den Benutzer im fremd-

[1] Dieser Beitrag ist im Rahmen der durch Drittmittel geförderten Forschungsprojekte DICONALE-estudios (Xunta de Galicia: IN.CI.TE: 10PXIB204 188 PR), DICONALE-online (MINECO-FEDER: FFI2012-32658), COMBIDIGLEX (MINECO-FEDER: FFI2015-64476-P) und in Verbindung mit dem lexikographischen Netzwerk RELEX (Xunta de Galicia/FEDER: CN2012/290 und R/2014/042) an der USC entstanden.

[2] Vgl. dazu u.a. aktuelle Kollokationswörterbücher (z.B. Quasthoff 2011; DUDEN Stilwörterbuch 2010) und neuere Valenzwörterbücher (VALBU: Schumacher et al. 2004) für das Deutsche. In der Lexikographie der spanischen Sprache beschäftigen sich, wenn nicht ausschließlich, dann doch in besonderem Grade die Wörterbücher von Seco et al. (DEA: 1999), Bosque (2004) und Cuervo (1998) mit syntaktischen Konstruktionsmustern.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

sprachigen Kontext Informationen zu lexikalischer Vielfalt und Varianz beider Sprachen im Vergleich durch einen bilateralen Zugriff systematisch bereitstellen.[3] Der vorliegende Beitrag beschäftigt sich daher aus einer synchronen Perspektive sowohl mit der lexikalischen Ausdrucksvielfalt im semiotischen Sinne der *eins-zu-viele-Relation* zwischen Inhalts– und Ausdrucksebene als auch mit der Varianz als Bedeutungs- und struktureller Konstruktionsvarianz. Dabei wird angestrebt, beide Begriffe für die Erstellung eines onomasiologisch-konzeptuell orientierten Produktionswörterbuches neu zu perspektivieren. In dem hier zu untersuchenden Sprachenpaar stehen folglich der Begriff der Vielfalt in Zusammenhang mit der Benennungs- oder Bezeichnungsfrage eines mentalen Konzepts und der Begriff der Varianz in Zusammenhang einerseits mit der Bedeutungspolysemie und andererseits mit den unterschiedlichen morpho-syntaktischen und funktionalen Realisierungsmöglichkeiten und semantischen Füllungen der einzelnen Mitspieler eines verbalen Szenarios (Engelberg 2010, 2015, erscheint; Engelberg et al. 2011, 2012).

Die zunächst einzelsprachig formulierte Problemdarstellung verbindet sich durch die anwendungsorientierte DaF-/ELe-Perspektive mit sprachvergleichenden, kontrastiven Überlegungen. Die exemplarische Darstellung einiger Fälle in Verbindung mit der konkreten fremdsprachigen Produktionssituation und der damit verbundenen kritischen Auswertung der durchgeführten Rechercheergebnisse ein- und zweisprachiger (Lerner-)Wörterbücher sollen als Plädoyer für die Erstellung spezieller Produktionswörterbücher unter besonderer Berücksichtigung der Wortfindungs- und Wortauswahlproblematik im zweisprachigen Kontext verstanden werden. Dabei stützen sich die Überlegungen im Weiteren exemplarisch auf die verbalen Lexikalisierungsmöglichkeiten der auditiven SINNESWAHRNEHMUNG. Die empirische Grundlage für den Sprachvergleich bilden Pressetexte aus dem Deutschen Referenzkorpus (DeReKo) und dem entsprechenden spanischen Referenzkorpus der Real Academia Española (CREA).[4] Folgende vier Problemkomplexe sollen behandelt werden:

1) Das Benennungsproblem: Suchen und Finden bei lexikalischer Ausdrucksvielfalt
2) Auswahl aus der lexikalischen Vielfalt
3) Ausdrucksvarianz in der Kombinatorik
4) Ausdrucksvielfalt und Varianz bei der fremdsprachigen Äquivalenzsuche.

2. Untersuchung – Analyse

2.1 Das Benennungsproblem: Suchen und Finden bei lexikalischer Ausdrucksvielfalt

Welche sprachlichen Mittel stehen in beiden Sprachen zum Ausdruck der SINNESWAHRNEHMUNG – AUDITION zur Verfügung? Welches sprachliche Mittel soll aus der Vielfalt der einzelsprachlichen Ausdrucksmittel für die konkrete L2-Kommunikationssituation ausgewählt werden? Bieten die lexikographischen Nachschlagewerke ausreichende Information für die systematische Abgrenzung bedeutungsähnlicher Lexeme? Werden die einzelnen Lesarten eines Lemmas ausreichend semantisch und syntagmatisch disambiguiert, um bei einer Vielzahl von Ausdrucksmöglichkeiten die Auswahl zu steuern? Erhalten wir neben der distinktiven Bedeutungsinformation auch Informationen zu den einzelnen Mitspielern des Verbszenarios? Diesen und ähnlichen Fragen soll hier aus der Perspektive des hispanophonen DaF-Lerners nachgegangen werden.

[3] Zu den theoretischen Grundlagen von DICONALE vgl. Meliss (2014b), González Ribao/Meliss (2015), Meliss/Sánchez Hernández (2015).
[4] Zu den methodologischen Grundlagen der korpusgestützten Analysen von DICONALE vgl. González Ribao (2015), Meliss (2015a).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bei der Suche nach Benennungen für bestimmte konzeptuelle Einheiten in einer konkreten Kommunikationssituation kann methodologisch unterschiedlich verfahren werden. Klassische onomasiologische Wörterbücher wie das von Dornseiff/Wiegand/Quasthoff (2004) und Wehrle/Eggers (1993) für das Deutsche oder das ideologische Wörterbuch von Casares (2001) für das Spanische geben zwar einen ersten Überblick über die Vielfalt der möglichen Ausdrucksmittel, sichern aber nicht den korrekten und adäquaten Gebrauch in der fremdsprachigen Produktion ab, da die angebotenen Wortlisten[5] selten konkrete Gebrauchsinformationen liefern. Außerdem handelt es sich um Spezialwörterbücher, die in der hier im Fokus stehenden Benutzersituation kaum vorliegen. Paradigmatisch orientierte Spezialwörterbücher, die ebenfalls eine onomasiologisch motivierte Suchanfrage erlauben, zeigen in dem klassischen Printformat, aber auch in den entsprechenden Modulen zahlreicher lexikographischer Online-Portale,[6] ähnliche Probleme bezüglich der hier relevanten Produktionssituation auf. Es handelt sich größtenteils nur um Wortlisten (siehe Abb. 1-3), die zwar für Produktionssituationen unterschiedliche Ausdrucksmittel zur Verfügung stellen, aber für eine korrekte Anwendung im fremdsprachigen Kontext der weiteren detaillierten Informationsbeschreibung bedürfen.[7] Über den Online-Zugang wird die Informationserweiterung durch Verlinkung mit weiteren Informationsmodulen der einzelnen lexikographischen Portale oder mit kooperierenden Portalen[8] ermöglicht und stellt daher gegenüber den klassischen, paradigmatisch orientierten Printwörterbüchern einen wichtigen Vorteil dar. Online-Wörterbücher, die auf ausführlichen Studien zu umfangreicheren lexikalisch-semantischen Paradigmen beruhen, liegen bis jetzt allerdings nur sehr vereinzelt vor (Meliss 2005, 2006).[9]

Abb. 1: Paradigmatische Sinnrelationen: Informationsmodul aus DWDS zu dem Eintrag *zuhören* (links); Ausschnitt aus dem Online-Portal canooNet zu *zuhören* (rechts).

[5] Vgl. dazu auch Hausmann (1990a, 1990b) und Engelberg/Lemnitzer (2009).

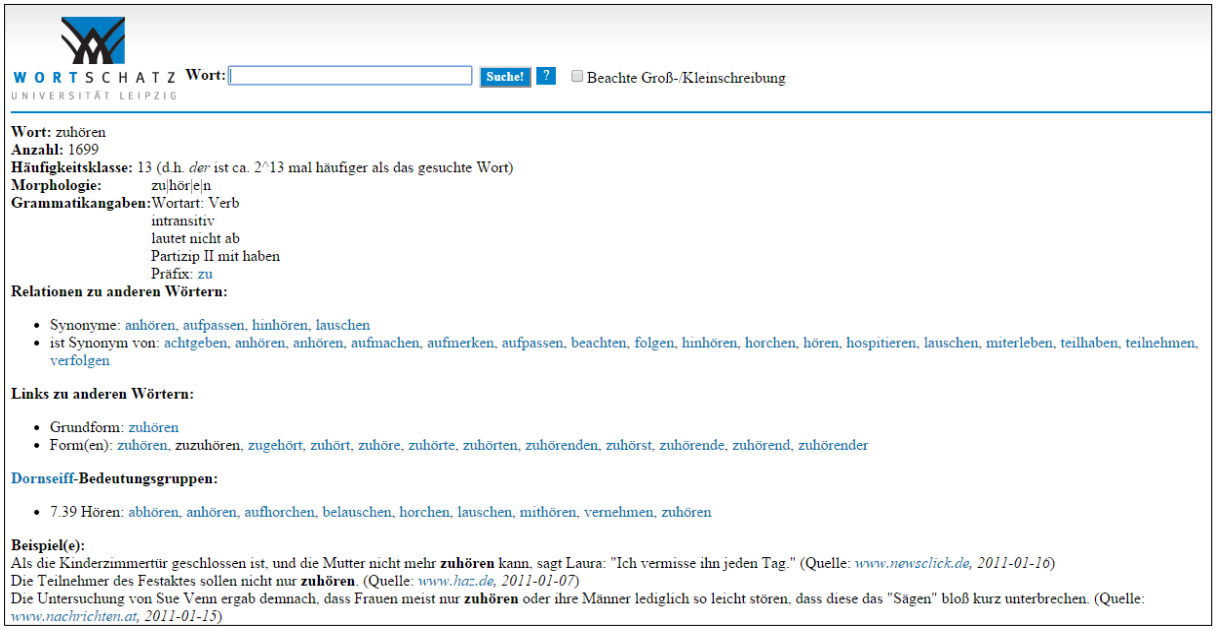
[6] Neben den klassischen Printversionen dieses Wörterbuchtyps kann man über verschiedene Portale auch *online* auf Module zu paradigmatischer Bedeutungsinformation zurückgreifen (z.B. für das Deutsche: *elexiko*, DWDS, Wortschatz Uni-Leipzig, canooNet, woxikon, Wörterbuch für Synonyme etc.; für das Spanische: diccionarios.com: „diccionario de sinónimos“, WordReference etc.).

[7] Die angebotenen Lexeme bedürfen einer konsequenteren Anordnung nach konzeptuellen und semantisch relevanten Kriterien, wie sie z.B. *elexiko* oder in Ansätzen canooNet (Abb. 1b) vorschlagen.

[8] *elexiko* verlinkt mit canooNet, Wortschatz Uni-Leipzig verlinkt mit der elektronischen Version des onomasiologischen Wörterbuches von Dornseiff, DWDS verlinkt mit OpenThesaurus. Das spanische Portal WordReference verlinkt die Information mit den lexikographischen Materialien des Verlages Espasa-Calpe sowie mit dem DRAE bzw. DLE und diccionarios.com verlinkt mit den Materialien der Verlage Vox und Larousse.

[9] Hier ist besonders das elektronische Wörterbuch der Kommunikationsverben (KV-online) zu erwähnen. Dieses Wörterbuch mit Online-Zugriff über OWID basiert auf den Studien von Harras et al. (2004) und Harras/Proost/Winkler (2007) (vgl. Proost/Müller-Spitzer 2014).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.



WORTSCHATZ UNIVERSITÄT LEIPZIG Wort: Suche! ? Beachte Groß-/Kleinschreibung

Wort: zuhören
Anzahl: 1699
Häufigkeitsklasse: 13 (d.h. *der* ist ca. 2^13 mal häufiger als das gesuchte Wort)
Morphologie: zu|hör|e|n
Grammatikangaben: Wortart: Verb
intransitiv
lautet nicht ab
Partizip II mit haben
Präfix: zu

Relationen zu anderen Wörtern:

- Synonyme: anhören, aufpassen, hinhören, lauschen
- ist Synonym von: achtgeben, anhören, anhören, aufmachen, aufmerken, aufpassen, beachten, folgen, hinhören, horchen, hören, hospitieren, lauschen, miterleben, teilhaben, teilnehmen, verfolgen

Links zu anderen Wörtern:

- Grundform: zuhören
- Form(en): zuhören, zuzuhören, zugehört, zuhört, zuhöre, zuhörte, zuhörten, zuhörenden, zuhörst, zuhörende, zuhörend, zuhörender

Dornseiff-Bedeutungsgruppen:

- 7.39 Hören: abhören, anhören, aufhorchen, belauschen, horchen, lauschen, mithören, vernehmen, zuhören

Beispiel(e):
Als die Kinderzimmertür geschlossen ist, und die Mutter nicht mehr **zuhören** kann, sagt Laura: "Ich vermisse ihn jeden Tag." (Quelle: *www.newsclick.de, 2011-01-16*)
Die Teilnehmer des Festaktes sollen nicht nur **zuhören**. (Quelle: *www.haz.de, 2011-01-07*)
Die Untersuchung von Sue Venn ergab demnach, dass Frauen meist nur **zuhören** oder ihre Männer lediglich so leicht stören, dass diese das "Sägen" bloß kurz unterbrechen. (Quelle: *www.nachrichten.at, 2011-01-15*)

Abb. 2: Paradigmatische Sinnrelationen: Ausschnitt aus dem Portal Wortschatz Uni-Leipzig zu dem Eintrag *zuhören* mit Verlinkung zu der Information aus Dornseiff;

escuchar
tr.
1 atender*.
CITA: «El que escucha no pone en ejercicio más que el sentido del oído. El que atiende observa los gestos y los movimientos. El primer verbo se aplica al ruido de las cosas inanimadas, pero no el segundo.» ? **José Joaquín de Mora** ? «Se escucha para oír bien lo que se dice. Se atiende para comprender bien lo que se oye. El primero [escuchar] representa una operación inmediata del oído, el segundo [atender] una operación del ánimo. El que oye bien al predicador, atiende, está atento al sermón, no se distrae, para no perder nada de él. El que está lejos, escucha para poder oír. Para escuchar se evita el ruido; para atender se evita la distracción.» ? **José López de la Huerta**
2 atender, tomar en consideración, dar oídos, hacer caso. COMENTARIO: Son sinónimos en la acepción de *escuchar* una proposición, los dictados de la conciencia, los avisos de un amigo.
© Vox, marca registrada por Larousse Editorial

oír LISTEN: ESPAÑA
definición | en inglés | en francés | conjugar | en contexto | imágenes
Diccionario de sinónimos y antónimos © 2005 Espasa-Calpe:
oír
- escuchar, percibir, atender, notar, sentir, advertir

'oír' also found in these entries:
atender - entender - entenderse - enterar - enterarse - escuchar - notar - percibir

Abb. 3: Paradigmatische Sinnrelationen: *escuchar* aus dem Portal diccionarios.com („diccionario de sinónimos“) (links); *oír* aus dem Portal WordReference (rechts).

Der Versuch, eine semasiologisch und onomasiologisch orientierte Suchanfrage miteinander zu kombinieren und für den L2-Bereich nutzbar zu machen, ist erstmals von Kempcke et al. (1999) für den DaF-Bereich vorgelegt worden. Diese semasiologisch-onomasiologische Vernetzung wird allerdings nur ungenügend genutzt und weist klare Grenzen in der Umsetzbarkeit im Rahmen eines Printwörterbuches auf (Roelcke 2002). Hinweise zu paradigmatischen Bedeutungsrelationen sind laut Lernerwörterbuchforschung aber gerade bei der Textproduktion für eine stilistische Ausdrucksvarianz von großem Nutzen (Wolski 2002; González Ribao/Proost 2015) und in diesem Sinne weist schon Fuentes Morán (1997, S. 84) für den zweisprachigen, spanisch-deutschen Kontext darauf hin, dass ein Wörterbuchbenutzer für die freie fremdsprachige Produktion ein Nachschlagewerk brauche, das genügend paradigmatische Bedeutungsinformationen anbietet.

DICONALE greift die aufgezeigten lexikographischen Defizite auf und hat sich zum Ziel gesetzt, für die *Lexem-Auffindung* (Suchen + Finden) und die anschließende *Lexem-Auswahl* verschiedene hierarchisch strukturierte Such- und Auswahlmöglichkeiten, die in Verbindung mit konzeptuell-paradigmatisch und szenenorientierter Information stehen, anzubieten. Die entsprechende lexikologische Informationsstrukturierung und ihre Vernetzung erfolgt in un-

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

terschiedlichen Modulen[10] und auf vier Stufen.[11] Auf der lexikographischen Mikrostrukturebene ist die Information dem entsprechend ebenfalls in verschiedenen ein- und zweisprachigen Modulen abrufbar. Im Fall der allgemeinen Lexikalisierung von AUDITION ergibt sich durch verschiedene onomasiologisch-konzeptuell geleitete Methoden für das Deutsche und Spanische zunächst eine offene Liste mit einer Reihe von ein- und mehrteiligen Lemmata zum Ausdruck von AUDITIVER WAHRNEHMUNG (Abb. 4).

Deutsch	**Lemmata**	**Spanisch**
	Einwortlemmata	
hören, horchen, lauschen ...	**Simplizia**	***auscultar, escuchar, oír ...***
ab/an/hin/mit/zu ... –***hören*** sich an/um/ver ... –***hören*** ab/auf/hin/zu ... –***horchen*** sich um ... –***horchen*** be***lauschen*** ***...***	**Präfigierungen**	
jemandem sein Ohr leihen ganz Ohr sein die Ohren spitzen ...	**Mehrwortlemmata**	*abrir los oídos* *aguzar los oídos* *aplicar el oído* *ser todo oídos* ...

Abb. 4: **Stufe 1:** Lexikographische Makrostruktur: konzeptuelle Strukturierung in Felder und Subfelder. Offene Liste der möglichen lexikalischen Ausdrucksmittel zu AUDITION im Deutschen und Spanischen.

Eine Subklassifizierung der AUDITIVEN WAHRNEHMUNG auf Grund von auf Grund von unterschiedlichen szenenorientierten Spezifizierungen erlaubt eine Zuordnung der einzelnen Lexeme zu mindestens fünf konzeptuell definierten Subfeldern (Abb. 5).

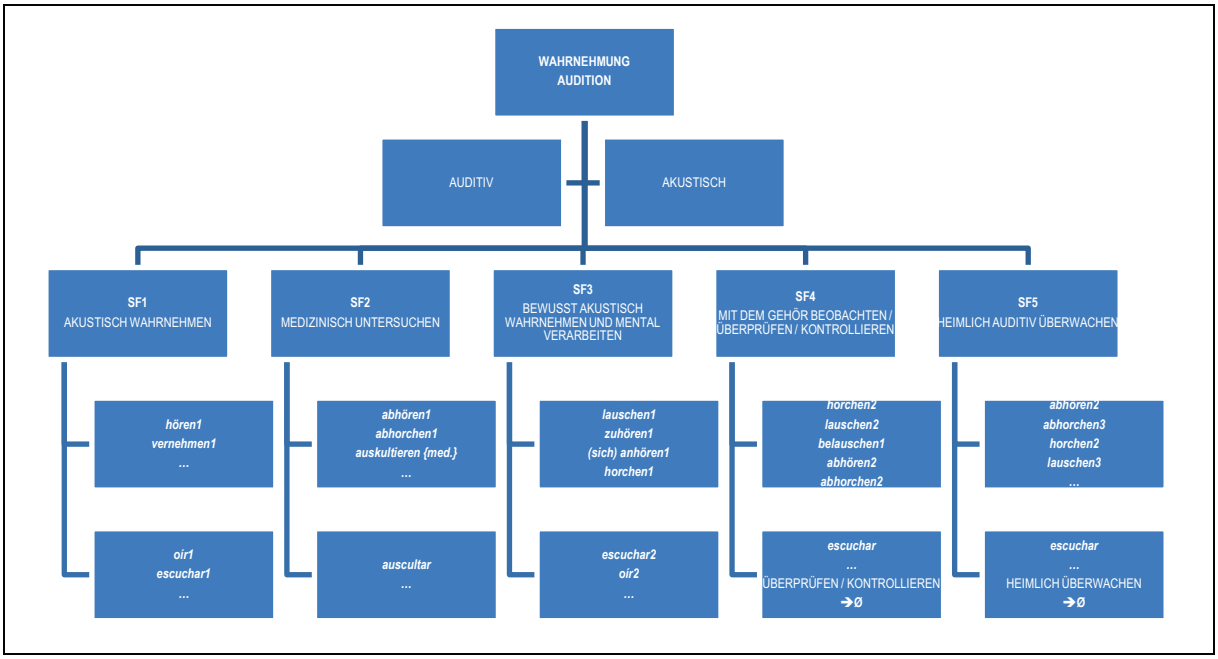


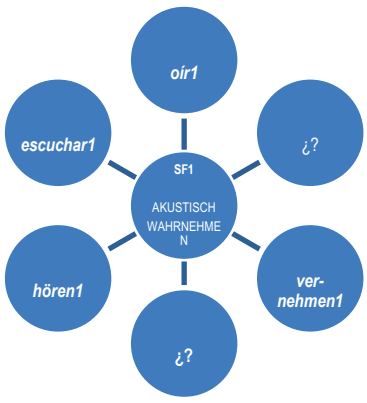
Abb. 5: **Stufe 1:** Strukturierung der lexikalischen Ausdrucksmittel zu AUDITION im Deutschen und Spanischen: konzeptuell geleitete Klassifizierung in Subfelder (SF) und Zuordnung der Lexeme.

[10] Die Beschreibung in Modulen erfolgt zunächst einzelsprachig: **Modul 0**: Information zur Form; **Modul 1**: Bedeutungsbeschreibung (semantisch-distinktive Bedeutungskomponenten, paradigmatische Sinnrelationen etc.); **Modul 2**: Kombinatorik: M2.1.: Argumentstruktur und Spezifizierung der Argumente; M2.2.: morpho-syntaktisch-funktionale Information: Funktion und Realisierungsmöglichkeiten der Argumente; M2.3.: kategoriale Bedeutung, Kollokationen und semantische Füllung, Kookkurrenzprofil (cfr. CCDB: Belica 2007); **Modul 3**: pragmatische Information: Register, Stil und textsortenspezifische Aspekte.

[11] **Stufe 1**: konzeptuelle Makrostruktur: Felder + Subfelder; **Stufe 2**: Mikrostrukturelle Information zu den einzelnen Lemmata in verschiedenen Beschreibungsmodulen; **Stufe 3**: Mediostrukturelle Information: interne paradigmatische und syntagmatische Strukturen der einzelnen Subfelder; **Stufe 4**: Kontrastive Information auf mikro- und mediostruktureller Ebene;

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Durch die konzeptuell geleitete Subklassifizierung erhält der Benutzer für seine entsprechende Produktionssituation ein übersichtliches Angebot von möglichen Lexemen. Die Angabe von exemplarischen Belegbeispielen (Abb. 6a-e)[12] und der Hinweis auf mögliche feldübergreifende Hyperonyme (markiert mit dem Symbol ⌂) in Verbindung mit einer weiteren konzeptuell gesteuerten, semantischen (vgl. Abb. 6c: *lauschen1, horchen1*: KONZENTRIERT etc.) und stilistischen (vgl. Abb. 6b: *auskultieren* {med.}) Spezifizierung einzelner Lexeme dienen dazu, aus der Vielfalt der lexikalischen Benennungsmöglichkeiten auszuwählen.



Deutsch: ⌂ *wahrnehmen* [+Wahrnehmumg] [+/−auditiv]

(1.1.$_{dt}$) Ein Anwohner ***hörte*** das Klirren und rief die Polizei. (BRZ06/OKT.04252 Braunschw. Z., 10.10.2006)

(1.2.$_{dt}$) Auf einem umgefallenen Baumstamm harrend, ***vernimmt*** das lauschende Ohr des Naturfreundes die liebliche Geräuschkulisse der gefiederten Waldbewohner. (RHZ98/APR.34923 RZ, 21.04.1998)

Spanisch: ⌂ *percibir* [+Wahrnehmung] [+/−auditiv]

(1.1.$_{sp}$) Desde nuestra casa ***oíamos*** las músicas de tus fiestas. (CREA)

(1.2.$_{sp}$) De madrugada ***escuchó*** un ruido extraño en el exterior; [...]. (CREA)

Abb. 6a: Subfeld 1 (SF1) mit der Zuordnung einiger Lexeme und entsprechender Belegbeispiele in beiden Sprachen.

Die Erfassung lexikalischer Lücken (markiert durch das Symbol Ø), wie sie z.B. im Fall von **SF4** und **SF5** für das Spanische vorliegen (vgl. Anhang: Abb. 6d+6e) wird so besonders eindeutig ermöglicht. Eine in dieser Form paradigmatisch vernetzte Information zu den einzelnen Elementen eines konzeptuell orientierten Feldes und dessen Subfelder muss für den Produktionsprozess mit Information zum Bedeutungs- und Konstruktionspotenzial erweitert werden. Die diesbezüglich auftretenden einzelsprachigen Probleme zu Vielfalt und Varianz sollen in den folgenden Abschnitten exemplarisch für das Deutsche anhand ausgewählter Fälle kurz erörtert werden (vgl. Abschnitte 2.2 und 2.3).

2.2 Auswahl aus der lexikalischen Vielfalt

Wie oben dargestellt werden konnte (vgl. Abb. 6a-d), existieren zur Benennung bestimmter konzeptueller und szenenorientierter Einheiten in den meisten Fällen mehrere lexikalische Ausdrucksmöglichkeiten, aus denen ausgewählt werden muss. In Textproduktionssituationen, in denen bei der Existenz von sprachlicher Ausdrucksvielfalt ein hoher Disambiguierungsbedarf zwecks Lexemauswahl herrscht, stehen zwar die herkömmlichen Wörterbücher und spezifische Lernerwörterbücher[13] zur Verfügung, aber die angebotenen Informationen sind für den fremdsprachigen Produktionsprozess nicht immer ausreichend (vgl. Engelberg 2010). Die Notwendigkeit, die konzeptuell geleitete Such- und Auffindungsaktivität von Benennungsmöglichkeiten zusammen mit adäquaten Auswahlkriterien bei Polysemie in der Lernerlexikographie ernster zu nehmen, soll anhand des Eintrages zu *lauschen* aus LGWB-DaF (Abb. 7) dargestellt werden. Beide Lesarten von *lauschen* werden durch semantische Merkmale wie

[12] Die Abbildungen 6b-e befinden sich im Anhang.

[13] Für den DaF-Bereich existieren neben dem Langenscheidt Großwörterbuch DaF (2015) u.a. folgende weitere Lernerwörterbücher: Kempcke, PONS-DaF (print: 2015 und online), Duden-DaF, Wahrig-DaF. Ein freier Online-Zugriff existiert auch über das Duden-online-Portal. Für den ELe-Bereich sind u.a. das Diccionario de Salamanca (DSA), das ELe-Lernerwörterbuch (DEE) und dessen online-Zugriff über das Verlagsportal SM bzw. das Wörterbuch CLAVE: CLAVE-online und das „Diccionario de Alcalá“ (DA) von VOX, welches *online* über das Portal diccionarios.com begrenzt frei konsultiert werden kann, zu nennen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

[+auditiv] [+konzentriert] beschrieben, dienen aber nicht zur gegenseitigen Bedeutungsdifferenzierung und einer konzeptuell geleiteten Subklassifizierung. Die Bemühung um eine Lesartdisambiguierung mit Hilfe der Angabe unterschiedlicher Strukturmuster hat vor allem in der Lernerlexikographie die Bedeutungsdisambiguierung durch distinktive Bedeutungsmerkmale in den Hintergrund gedrängt. So wird eher selten explizit auf bedeutungsrelevante, distinktive Merkmale hingewiesen, obwohl sie zusammen mit der Information zu den jeweiligen Strukturmustern bei der Bedeutungs- und Lesartdisambiguierung eine entscheidende Rolle spielen (Konerding 1998; Wolski 2002; Meyer/Wiegand 2000).

lau·schen; *lauschte, hat gelauscht*; *[Vi]* **1.** sich stark konzentrieren, damit man etwas hört ≈ horchen: *an der Tür lauschen* **2.** ***jemandem/etwas lauschen*** jemandem/etwas konzentriert zuhören: *dem Gesang der Vögel lauschen* [...] © Langenscheidt KG, Berlin und München	***lauschen1:*** ≈ *hören* ≈ *horchen* [+konzentriert] → [+bewusst]? → [+mental verarbeiten]? → [+heimlich]? → [+überwachend?] → [+kontrollierend]? ***lauschen 2:*** ≈ *zuhören* [+konzentriert]

Abb. 7a: Ausschnitt[14] zu *lauschen1*[15] *und lauschen2*[16] aus Langenscheidt Großwörterbuch DaF-digital und Systematisierung der Information.

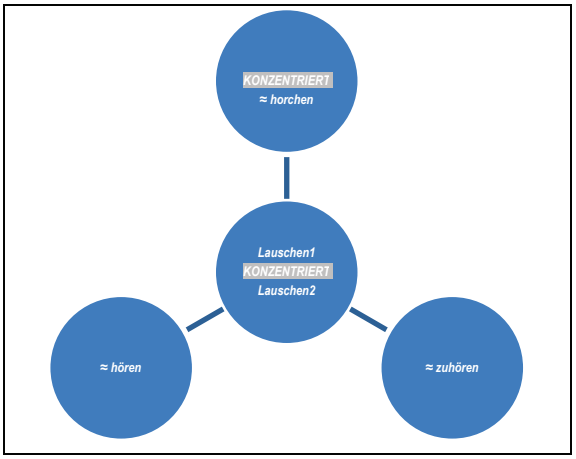


Abb. 7b: Darstellung verschiedener paradigmatischer Sinnrelationen zu bedeutungsähnlichen Lexemen mit Zuordnung von distinktiven semantischen Merkmalen.

Erst semantisch-distinktive Merkmale wie [+heimlich] [+überwachend] [+kontrollierend] etc. so wie es z.B. DWDS (Abb. 8) aber auch Duden online für *lauschen* anbieten, tragen im Zusammenspiel mit den entsprechenden Informationen zu den unterschiedlichen Strukturmustern zur Lesartdisambiguierung bei.

[14] Die farbliche Markierung des Wörterbuch-Eintrags erfolgte durch die Autorin dieses Beitrages.
[15] In diesem Sinne Teil von **SF3** und bedeutungsähnlich mit *zuhören* (vgl. Abb. 6c).
[16] In diesem Sinne Teil von **SF4** (vgl. Abb. 6d) und bedeutungsähnlich mit *hören* und *horchen* aber auch Teil von **SF5** und bedeutungsähnlich mit *horchen*, *belauschen*, *abhören* etc. (vgl. Abb. 6e).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

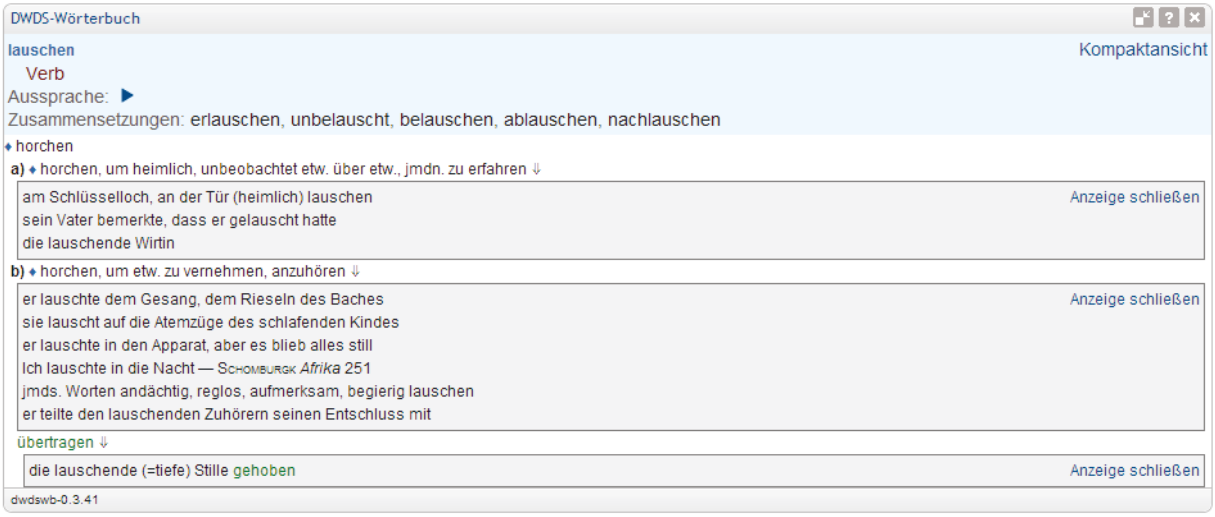


Abb. 8: Ausschnitt zu *lauschen* aus DWDS.

			Modul 1							Modul 2					
[+Wahrnehmung] [+auditiv] [+akustisch]	**Sub-feld**	**Beleg-beisp.**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
lauschen1 ⌂ *zuhören*	**SF3**		+	+	+										
Jemand (A1) *lauscht* jemandem (A2) **SBP[17] <s d>**		**3.6.dt**								**s**	**d**				
Jemand (A1) *lauscht* etwas (A3) **SBP <s d>**		**3.7.dt** **3.8.dt** **3.9.dt**								**s**		**d**			
lauschen2 ⌂ *beobachten, überprüfen, kontrollieren* [+/-Wahrnehmung] [+/-auditiv]	**SF4**		+	+		+	(+)								
Jemand (A1) *lauscht* etwas / auf etwas (A4) **SBP <s d/prp$_{auf}$>**		**4.7.dt** **4.8.dt** **4.9.dt**								**s**			**d / prp**		
Jemand (A1) *lauscht* in eine bestimmte Richtung (A5) **SBP <s adv$_{dir}$>**		**4.10.dt**								**s**				**adv**	
lauschen3 ⌂ *überwachen, kontrollieren, spionieren* [+/-Wahrnehmung]	**SF5**		+	+			+	+	+						
Jemand (A1) *lauscht* an einem bestimmten Ort (A6) **SBP <s adv$_{lok}$>**		**5.4.dt** **5.5.dt**								**s**					**adv**
semantisch-distinktive Merkmale			**S1**: [+bewusst], **S2**: [+konzentriert], **S3**: [+mental verarbeitend], **S4**: [+beobachtend], **S5**: [+heimlich], **S6**: [+überwachend], **S7**: [+kontrollierend];												
Komplemente:[18]			**s** = Subjekt, **d** = Dativkomplement, **prp** = Präpositivkomplement, **adv** = Adverbialkomplement (**dir**: direktiv, **lok**: lokal)												
Argumente mit kategorial-semantischer Beschreibung[19] [...] und Deskriptoren zur weiteren Spezifizierung			**A1**: WAHRNEHMER [+belebt]; **A2:** LAUTVERURSACHER: [+hum] Redner, Sänger etc.; **A3:** WAHRGENOMMENE VERBALE HANDLUNG [+intell] Rede, Vortrag, Lesung, Gesang etc.; **A4**: WAHRGENOMMENE GERÄUSCHE [+inmat, +auditiv] Geräusch, Krach, etc.; **A5:** GERÄUSCHQUELLE [+mat] Ort, Medium etc.; **A6:** WAHRNEHMUNGSORT [+lok] an der Tür etc.;												

Abb. 9: **Stufe 2:** Ausschnitt aus der lexikographischen Mikrostrukturbeschreibung zu *lauschen*: Lesarten von *lauschen* und entsprechende Information aus den Modulen 1 + 2

[17] SBP = Satzbauplan

[18] Die Subklassifizierung der Komplementtypen erfolgt in Anlehnung an Engel (2004), wobei die Metasprache an unsere kontrastiven Bedürfnisse angepasst wurde.

[19] Die Merkmale zur kategorial-semantischen Beschreibung der Argumente erfolgt in Anlehnung an Engel (2004).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Für das hier konkret analysierte Beispiel ergibt sich nach einer umfangreichen Beleganalyse neben den in LGWDaF erwähnten Lesarten zu *lauschen* (vgl. Abb. 6c: **SF3,** Abb. 6e: **SF5**) noch eine weitere Lesart in Verbindung mit dem Subfeld 4 (vgl. Abb. 6d: **SF4**). Die unterschiedlichen Szenarien, die sich aus der Beteiligung der verschiedenartigen Mitspieler (Argumente) und ihrer unterschiedlichen morpho-syntaktischen und funktionalen Realisierungen ergeben, unterstützen die Notwendigkeit einer Zuordnung zu drei verschiedenen, zunächst konzeptuell geleiteten, lexikalisch-semantischen Subfeldern (Abb. 9).

AUDITIVE WAHRNEHMUNG [+Wahrnehmung] [+akustisch]						
SF3	**Modul 1**	**Modul 2**				
[+bewusst] [+mental verarbeitend]	**dist. sem. Merkmale**	**Belegbsp.**	**A1**	**A2**	**A3**	**Supplemente[20]**
zuhören1						**Wie? mod**
ASTM1 Jemand (A1) *hört* jemandem (A2) *zu* **SBP <s d>**		**3.5.**$_{dt}$	**s** [+hum]	**d** [+hum]		...
ASTM2 Jemand (A1) *hört* etwas (A3) *zu* **SBP <s d>**		**3.6.**$_{dt}$ **3.11.**$_{dt}$	**s** [+hum]		**d** [+intell] Predigt, Ausführungen, Rede, Konzert, ... **SF**: was	...
(sich) anhören1	[+genau]					
ASTM2 Jemand (A1) *hört* (sich) etwas (A3) *an* **SBP <s a>**		**3.3.**$_{dt}$ **3.4.**$_{dt}$ **3.13.**$_{dt}$	**s** [+hum]		**a** [+intell] Probleme, Musik, ... **SF: was**	...
lauschen1	[+konzentriert]					
ASTM1 Jemand (A1) *lauscht* jemandem (A2) **SBP <s d>**		**3.6.**$_{dt}$	**s** [+hum]	**d** [+hum]		...
ASTM2 Jemand (A1) *lauscht* etwas (A3) **SBP <s d>**		**3.7.**$_{dt}$ **3.8.**$_{dt}$ **3.9.**$_{dt}$ **3.14.**$_{dt}$	**s** [+hum]		**d** [+intell] Krimis, Erklärungen, Gesang, ... **SF**: **was**	...
hören2						
ASTM1 Jemand (A1) *hört* jemanden (A2) **SBP <s a>**		**3.1.**$_{dt}$	**s** [+hum]	**a** [+hum] Musiker, Professor, ...		...
ASTM2 Jemand (A1) *hört* etwas (A3) **SBP <s a>**		**3.2.**$_{dt}$ **3.12.**$_{dt}$	**s** [+hum]		**a** [+intell] Geschichten, Meldungen, Rede, Ansprache, Begründung, Berichte, Musik, ... **SF: was**	..
horchen1	[+konzentriert]					
ASTM2 Jemand (A1) horcht etwas (A3)[21] **SBP <s d>**		**3.10.**$_{dt}$ **3.15.**$_{dt}$ **3.16.**$_{dt}$	**s** [+hum]		**d** [+intell] Rede, Predigt, ... **SF**: **was**	...

Abb. 10: Stufe 3: Lexeme aus SF3 und entsprechende (Teil)information aus den Modulen 1+2

Nur eine Kombination der unterschiedlichen lexikalischen Ebenen, wie sie in modularer Form in DICONALE vorgesehen ist, hilft dem Benutzer, die Bedeutungs- und Konstruktionsdisambiguierung zwischen polysemen Lesarten (*lauschen1, lauschen2, lauschen3*) zu erfassen und die eine oder andere Lesart in konkreten Produktionssituationen anzuwenden. Ausgehend von

[20] Die Erweiterung der Information durch korpusgestützte, quantitativ berechnete Daten zu typisch und/oder häufig auftretenden Supplementen scheint sinnvoll und ist in dem Modul 2 von DICONALE ebenfalls vorgesehen.

[21] Vereinzelt auch „*zuhorchen*".

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

einer onomasiologisch-konzeptuellen Perspektive dient eine dient eine schematisch-modulare Erfassung und die anschließende modulare Darstellung der verschiedenen lexikologischen Beschreibungsparameter (Abb. 9) aber hauptsächlich der Differenzierung bedeutungsähnlicher Lexeme innerhalb eines lexikalisch-semantischen Subfeldes und bietet somit in Produktionssituationen ein adäquates Informationsangebot, das bei der Auswahl aus der Vielfalt unabdingbar ist. Exemplarisch wird dies für die verschiedenen Ausdrucksmöglichkeiten und deren Ausdrucksformen zur Lexikalisierung des konkreten Konzepts: BEWUSST AKUSTISCH WAHRNEHMEN UND MENTAL VERARBEITEN (*hören2* / *horchen1* / *lauschen1* / (sich) *anhören1* / *zuhören1*) dargestellt (Abb. 10: **SF3**). Evident werden u.a. folgende Beobachtungen: (i) nicht alle Lexeme dieses Feldes realisieren das Argument A2, (ii) das Argument 3 kann satzförmig v.a. durch einen indirekten Fragesatz realisiert werden, (iii) bei semantisch sehr ähnlichen Lexemen (*lauschen1, horchen1*) werden die jeweiligen Distributionsbeschränkungen der unterschiedlichen Ebenen deutlich.

2.3 Ausdrucksvarianz in der Kombinatorik: morpho-syntaktische Realisierungsformen und semantische Füllung

Nach einem erfolgreichen Auffindungs- und Auswahlprozess ergeben sich für die fremdsprachige Produktionssituation weitere Fragestellungen, die hauptsächlich mit dem korrekten und adäquaten Gebrauch in Verbindung stehen. Bezüglich des Konstruktionspotenzials bei Verben fällt innerhalb eines Subfeldes die mögliche Varianz bezüglich der morpho-syntaktisch-funktionalen Information und der semantischen Füllung in Verbindung mit den einzelnen Argumenten auf. Die konsultierten Lernerwörterbücher bieten dazu wenig bis keine Information an, obwohl dies für Produktionszwecke nicht nur im fremdsprachigen Kontext von großem Nutzen ist (Wolf 2001). Exemplarisch soll anhand von einigen Beispielen die Reichweite dieser Information für den fremdsprachigen Produktionsprozess dargestellt werden. Eine schematische Gegenüberstellung der unterschiedlichen Informationen zu jedem Lexem eines (Sub)feldes dient zur kontextadäquaten Auswahl aus der lexikalischen Vielfalt und Varianz (Abb. 10) und macht die teilweise morpho-syntaktisch unterschiedlichen Realisierungsmöglichkeiten innerhalb eines Subfeldes sichtbar. Die exemplarisch realisierte Analyse und schematische Darstellung einiger Elemente von **SF3** zeigt, dass die Argumente A2 und A3 als Akkusativkomplemente bei (*sich*) *anhören* und *hören* (3.1.-3.4.) oder als Dativkomplemente bei *zuhören*, *lauschen* und *horchen* (3.5.-3.11.) realisiert werden (Abb. 10: **SF3**). Alternanzen zwischen der einen oder anderen morpho-syntaktischen Realisierungsform ein und desselben Arguments sind u.a. in **SF4** (Abb. 6d) bei *lauschen2* (4.8.1. und 4.8.2.) sichtbar. Einige Argumente der unterschiedlichen Subfelder können außerdem satzförmig realisiert werden. So werden die Argumente **A3** und **A4** von *hören1* (**SF1**) hauptsächlich durch Subjunktorsätze mit „dass“ und durch Infinitivsätze[22] gebildet (1.3. und 1.4.), während bei den Lexemen von Subfeld 3 (**SF3**) die belegten satzförmig realisierten Argumente vor allem durch einen indirekten Fragesatz – eingeleitet mit „was“ – realisiert sind (3.12.-3.16.).

(1.3. $_{dt}$) Bevor sie um die Ecke biegt, drückt sie den Knopf mit der Klingel – damit jeder *hört*, **dass** sie kommt. (BRZ05/OKT.04859 Braunschw. Z., 1.10.2005).

(1.4. $_{dt}$) Seine Freundin *hörte* ihn röcheln, fühlte immer wieder seinen Puls. (BRZ09/AUG.12385 Braunschw. Z., 27.8.2009)

(3.12. $_{dt}$) Er marschiert von Haustür zu Haustür und *hört zu*, **was** ihm die Leute erzählen. (RHZ96/AUG.03591 RZ, 7.8.1996)

(3.13. $_{dt}$) Frei nach dem Grundsatz „Erst mal *hören*, **was** die Zeugen wissen“, verlegte sich das Muskelpaket und Vater von drei Kindern aufs Schweigen. (RHZ03/JUL.12417 RZ, 16.07.2003)

[22] Auf die hier vorliegende AcI-Konstruktion kann hier nicht näher eingegangen werden.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

(3.14. $_{dt}$) Oberbürgermeister Dr. Schulte-Wissermann, CDU-Ratsmitglied Anne Schumann-Dreyer und SPD-Ratsmitglied Jochem Bröhl *hören sich an*, **was** die kleinsten Schängel zu sagen haben. (RHZ98/JUL.07221 RZ, 15.07.1998)

(3.15. $_{dt}$) Die etwa 80 anwesenden Freisinnigen verstummen, schauen auf und *lauschen*, **was** der Präsident der FDP Region Rorschach, Michael Haas, zu sagen hat. (A12/JAN.02812 St. Galler Tagbl., 13.01.2012, S. 37)

(3.16. $_{dt}$) „Ich wollte mal *horchen,* **was** da so vorgebracht wird“, begründete Preuß seine Teilnahme. (BRZ06/OKT.14858 Braunschw. Z., 30.10.2006)

Das kombinatorische Informationsangebot zu unterschiedlichen semantischen Füllungen der einzelnen Argumente eines Lexems ist aus semasiologischer Perspektive relevant und trägt zusammen mit anderen Informationen zu der Lesartdisambiguierung bei, wie exemplarisch an *abhören* gezeigt werden kann (Abb. 11). Andererseits bietet diese Art lexikologischer Daten für eine konzeptuell-onomasiologisch orientierte Perspektive die notwendige Gebrauchsinformation für die Auswahl aus der Vielfalt von lexikalischen Mitteln (Abb. 10).

			A1	A2	A3	A4
abhören1	**SF2**	2.1.$_{dt}$ 2.2.$_{dt}$	[+hum] Arzt	[+hum] [+zool] Patient	-	[+mat] Organ/Körper
abhören2	**SF4**	4.3.$_{dt}$	[+hum]		Gespräche, (Audio)aufnahmen, ...	[+mat] Geräte: Tonband, Anrufbeantworter ...
abhören3	**SF5**	5.11.$_{dt}$ 5.12.$_{dt}$ 5.13.$_{dt}$	[+hum] Polizei	[+hum]	Gespräche, Telefonate, ...	[+mat] Telefon ...

Abb. 11: **Stufe 2:** Ausschnitt aus der Mikrostrukturanalyse zu verschiedenen Lesarten von *abhören* und Information zu den semantischen Füllungen der einzelnen Argumente.

2.4 Ausdrucksvielfalt bei fremdsprachiger Äquivalenz

Im Gegensatz zu den in Abschnitten 2.1-2.3 beschriebenen Ausgangssituationen, in denen der L2 Benutzer direkt für seine Produktionszwecke Ausdrucksmöglichkeiten in der L2 sucht, ist es auch denkbar, dass man zunächst von der Muttersprache ausgeht und eine äquivalente Benennungsmöglichkeit in der L2 durch Übersetzungsäquivalenz sucht (MS:L1 → FS:L2). In diesem Fall wird von den dazu konsultierten zweisprachigen Wörterbüchern erwartet, dass sie die unterschiedlichen Lesarten zunächst in der Ausgangssprache MS:L1 sorgfältig semantisch und syntaktisch disambiguieren, um dann zu den möglichen zielsprachigen Entsprechungen die Information bezüglich des Konstruktionspotenzials etc. für den angebrachten Gebrauch anzugeben. Die zweisprachigen Äquivalenzrelationen zeigen häufig eine *eins-zu-viele*-Relation auf, wobei der Benutzer aus der Vielfalt der möglichen Äquivalente aussuchen muss (Wiegand 2005). Die exemplarische Analyse dreier zweisprachiger Wörterbücher für das Sprachenpaar Deutsch-Spanisch[23] ergibt folgende Resultate für die Richtung Spanisch (MS: L1) → Deutsch (FS: L2)[24] und eine Produktionssituation, die wie folgt skizziert werden kann: „Suche nach Ausdrucksmitteln in der ZS Deutsch zu *escuchar* im Sinne von AUDITIV WAHRNEHMEN UND MENTAL VERARBEITEN (→ SF3) und genauer im Sinne von „prestar atención a la palabra“ (LGWBsp-dt) oder „prestar atención a lo que se oye“ (DRAE) mit dem Argumentstrukturmuster ⎪A1 A3⎪)“.

[23] LHWB, PONS online-WB, LEO

[24] Eine intensive Auseinandersetzung mit den Kriterien zur Lemmatisierung, Auswahl der Einträge, Disambiguierung der Lesarten, Benutzerforschung etc. für die zweisprachige Lexikographie liegt zwar in Ansätzen vor (Haensch/Omeñaca 2004, Hausmann 1991, Hausmann/Werner 1991, Werner 1998, Meyer/Wiegand 2000, Korhonen (Hg.) 2001, Fuentes Morán et al. (Hg.) 2009, Model 2010), muss aber mit den neuesten Erkenntnissen theoretisch erweitert und speziell für das Sprachenpaar Deutsch-Spanisch verstärkt umgesetzt werden.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bei der Konsultation von LGWBsp-dt zu dem Eintrag *escuchar* erfolgt die Lesartdisambiguierung mithilfe der auf Spanisch angegebenen Bedeutungsparaphrase und der Angabe des Strukturmusters (*escuchar a alguien*) (Abb. 12). Den einzelnen Lesarten werden verschiedene, nicht weiter semantisch differenzierte Entsprechungen in der Zielsprache (ZS) mit Information zu dem Strukturmuster zugeordnet (*jemanden (an)hören* (ac)).[25] Der Bedeutungsunterschied der einen oder anderen Entsprechung in der deutschen ZS wird nicht expliziert. So kann ein DaF-Lerner mit der angegebenen Information nicht zwischen den Entsprechungen „jemanden (an)hören" und „jemanden belauschen" unterscheiden, was in bestimmten Kontexten, in denen *belauschen* verwendet wird, zu Missverständnissen führen kann (vgl. SF5). Das Online-Wörterbuch von PONS (Abb. 13a) bietet hinsichtlich der Disambiguierung der Ausgangssprache (AS) Information durch Angabe von distinktiven Bedeutungsmerkmalen („en secreto" → Ü: 'heimlich') bzw. Hinweise zu typischen Verbindungen oder Kollokationen (*escuchar* una conversación telefónica / un concierto → Ü: 'ein Telefongespräch / ein Konzert anhören'), führt aber zu wenig Information über die zielsprachige Kombinatorik an. So kann der Benutzer zwar erfahren, dass *escucharII.1* mit dem Merkmal „en secreto" (→ Ü: 'heimlich') die Entsprechung *belauschen* hat, erhält aber keine weitere Information über die mögliche Kombinatorik (wer kann was/wen belauschen?). Das Wörterbuch LEO (Abb. 13b) hingegen legt vor allem Wert auf diese kombinatorische Information, wobei die Bedeutungsdisambiguierung der AS und ZS in den Hintergrund gestellt wird. Es lässt sich zusammenfassend sagen, dass der Benutzer für die beschriebene Ausgangssituation in keinem der drei zweisprachigen Wörterbücher ausreichend Informationen für seine Kommunikationszwecke in der fremdsprachigen ZS erhält, da die Information zu unsystematisch, zu ungenau und teilweise auch widersprüchlich dargeboten wird.

Wörterbucheintrag

escuchar
I *v/t*
1. *(prestar atención a la palabra)* ***escuchar a alg*** jemanden **(an)hören** *(ac)*, jemandem **zuhören** *(dat)*; jemanden **belauschen** *(ac)*; ***escuchar la radio*** Radio **hören**
2. *(dar oído)* **erhören** *(ac)*, **Gehör schenken** *(dat)*; auf *jemanden o etwas* **hören**
II *v/i*
(zu)hören; *TEL* **mithören**; ***¡escucha!*** hör mal!, pass auf!
[...]
© Langenscheidt KG, Berlin und München

Spanisch →	**Deutsch**
I. transitiv	
1. prestar atención a la palabra ***escuchar a alg***	jemanden **(an)hören** *(ac)*
	~~jemandem **zuhören** *(dat)*~~
	jemanden **belauschen** *(ac)*
escuchar la radio	Radio **hören**
2.dar oído	…
II.intransitiv	(zu)hören
[...]	*TEL* **mithören**

Abb. 12: Ausschnitt zu *escuchar* aus LGWBsp-dt und Systematisierung der Information

[25] „ac" steht für sp. „acusativo"/Akkusativ.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

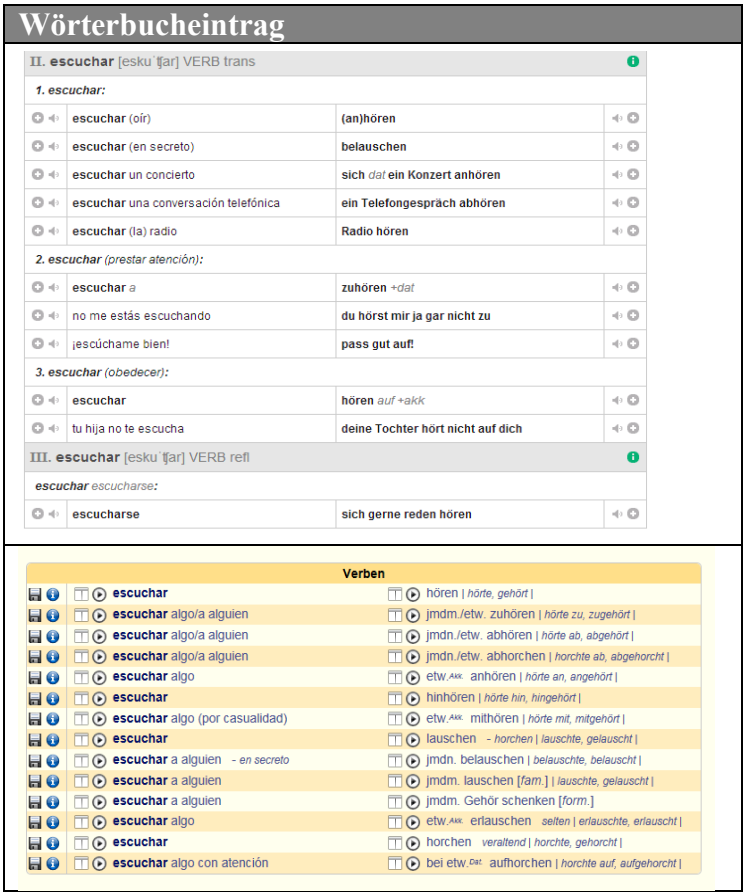


Abb. 13: Ausschnitte zu *escuchar* aus PONS-O und zu *escuchar* aus LEO

Eine für die fremdsprachige Produktion ausgerichtete Information in zweisprachigen Wörterbüchern sollte daher anschaulicher und benutzerfreundlicher die verschiedenen Entsprechungsmöglichkeiten zusammen mit den weiteren lexikologischen Informationen für die AS und ZS im Kontrast darbieten. Dafür erweist sich das mehrstufig modularisierte Informationsangebot, welches für DICONALE vorgeschlagen wird, als nützlich und soll exemplarisch für *escuchar2* aus dem Subfeld **SF3** dargestellt werden (Abb. 13). Für die möglichen lexikalischen Elemente des Subfeldes, die das Argumentstrukturmuster │A1 A3│(ASTM2)[26] aufweisen, ergeben sich mindestens fünf Entsprechungen ((*sich*) *anhören1*, *zuhören1*, *lauschen1*, *hören2*, *horchen2*), während für das ASTM1 │A1 A2│nur drei in dem SF3 aufgeführten Elemente als Entsprechungen möglich sind (*zuhören1*, *lauschen1*, *hören2*), da für (*sich*) *anhören1* und *horchen1* keine Belege in dieser Lesart mit A2 registriert werden konnten (vgl. Abb. 10). Zur Auswahl der einen oder anderen Entsprechung benötigt der Benutzer ausführliche Informationen bezüglich der einzelnen Argumentstrukturen und der jeweiligen Restriktionen, wie sie schon in Abb. 10 für die Feldmitglieder des Deutschen von SF3 unter einzelsprachlichem Aspekt angedeutet wurden. Hier soll exemplarisch verdeutlicht werden, dass die kontrastiv relevanten Aspekte gerade auf dieser Ebene zum Tragen kommen und entscheidend für die Auswahl der einen oder anderen Entsprechung sind. Relevante Divergenzen liegen hier u.a. auf der Ebene der morpho-syntaktischen Realisierung der Komplemente (SBP: sp.: <s dc> ≠ dt.: <s d> / <s a>) und auf der semantischen Ebene (durch eine mehr oder weniger differenzierte semantische Spezifikation). Genauere Unterschiede auf der Ebene der semantischen Füllungen und der Art und Häufigkeit der möglichen satzförmigen Realisierungen ha-

[26] Jedem Subfeld lassen sich verschiedene Argumentstrukturmuster (ASTM) zuordnen; so können in SF3 mindestens zwei ASTM angenommen werden: ASTM1: │A1 A2│, ASTM2: │A1 A3│ (Abb. 10).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

ben sich in diesem Fall nicht belegen lassen, sind aber durch genaue quantitative korpusgestützte Auswertungen zu überprüfen.[27]

Beschreibungsstufe 4: Kontrastiv

SF3 AKUSTISCH WAHRNEHMEN UND BEWUSST MENTAL VERARBEITEN: „zuhören-Paradigma“ • ASTM2 | A1 A3 |				Modul 2	
[+Wahrnehmung] [+akustisch] [+bewusst] [+mental verarbeitend]	Mögliche Entsprechungen	dist. sem. Merkmale	Belegbsp.	A1 Wahrnehmer	A3 Wahrgenommene verbale Handlung
escuchar2					
Alguien (A1) escucha algo (A3) **SBP <s (cd)*>**			**3.1.**$_{sp}$ **3.3.**$_{sp}$ **- 3.9.**$_{sp}$	**s** [+hum]	**cd** [+intell] historia, debate, música, noticias … **SF**: (lo) que
Alguien (A1) oye algo (A3) **SBP <s (cd) >**		≠			
→	***zuhören1***				
	Jemand (A1) *hört* etwas (A3) *zu* **SBP <s d>**		**3.6.**$_{dt}$ **3.11.**$_{dt}$	≠	
→	***(sich) anhören1***	[+genau]			
	Jemand (A1) *hört* (sich) etwas (A3) *an* **SBP <s a>**		**3.3.**$_{dt}$ **3.4.**$_{dt}$ **3.13.**$_{dt}$	≠	
→	***lauschen1***	[+konzentriert]			
	Jemand (A1) *lauscht* etwas (A3) **SBP <s d>**		**3.7.**$_{dt}$ **3.8.**$_{dt}$ **3.9.**$_{dt}$ **3.14.**$_{dt}$	≠	
→	***hören***				
	Jemand (A1) *hört* etwas (A3) **SBP <s a>**		**3.2.**$_{dt}$ **3.12.**$_{dt}$	≠	
→	***horchen1***	[+konzentriert]			
	Jemand (A1) *horcht* etwas (A3) **SBP <s d>**		**3.10.**$_{dt}$ **3.15.**$_{dt}$	≠	

* Für das Spanische werden u.a. folgende Siglen verwendet: **cd** = direktes Komplement, **ci** = indirektes Komplement;

sp. Belegbeispiele

(3.3.$_{sp}$) Y son muchos los curiosos que quieren ***escuchar*** esta historia. (El Mundo – Su vivienda (Suplemento), 17/01/2003)

(3.4.$_{sp}$) Allí, mientras ***escuchaba*** las noticias por televisión, se quedó impresionada cuando una locutora narraba con frialdad el siguiente suceso [...]: (El Diario Vasco, 31/01/2001)

(3.5.$_{sp}$) Como la música cubana ha sido la novedad, la gente la ***escucha*** sin parar. (El País, 24/11/2004)

(3.6.$_{sp}$) En la tribuna de invitados ***escucharon*** el debate el secretario general de UGT [...] (El Mundo, 20/11/2002)

(3.7.$_{sp}$) Desde los escaños los escolares ***escuchaban*** en silencio aunque con evidentes signos de inquietud. (Faro de Vigo, 22/11/2002)

(3.8.$_{sp}$) Pero antes del espectáculo, cuando el autor se ofrece a responder a las preguntas, se nota que el público ***escucha*** lo que Miller dice [...]. (El Mundo, 24/08/1994)

(3.9.$_{sp}$) Nunca está de más ***escuchar*** lo que los otros tienen que decir. (ABC Cultural, 08/11/1991)

Abb. 14a und 14b: **Stufe 4:** Ausschnitt aus der kontrastiven Analyse Spanisch → Deutsch für Elemente aus **SF3** mit ASTM2.

[27] In diesem Rahmen konnte nur der einseitigen Perspektive Spanisch → Deutsch nachgegangen werden. DICONALE ist aber als bilaterales Wörterbuch konzipiert und zieht beide Richtungen gleichermaßen in Betracht.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

3. Schlussfolgerungen – Ausblick

Es konnte aufgezeigt werden, dass lexikalische Ausdrucksvielfalt und Varianz für fremdsprachige Produktionssituationen bisher lexikographisch zu wenig beachtet wurden. Im Rahmen von Forschungsarbeiten zur Erstellung eines konzeptuell orientierten zweisprachig bilateralen Produktionswörterbuches für das Sprachenpaar Deutsch-Spanisch (DICONALE) konnten einige Lösungsvorschläge skizziert werden, die allerdings wegen ihrer Komplexität nur über die Digitalisierung der Information und deren Vernetzung und Verlinkung einem zukünftigen Benutzer zugänglich gemacht werden können. In diesem Sinne knüpfen die vorausgehenden Überlegungen, Analysen und Vorschläge an neue lexikographische Dimensionen an, die sich für die Online-Lexikographie[28] seit knapp zwei Jahrzehnten darbieten. Neue Wörterbuchkonzepte entstehen, die den Benutzerbedürfnissen entsprechend Onomasiologie und Semasiologie, Paradigmatik und Syntagmatik, Form und Inhalt, Einsprachig- und Mehrsprachigkeit etc. miteinander verbinden. Die verschiedenen Zugangsmöglichkeiten zu einem breiten Informationsspektrum, aus dem individuell ausgewählt werden kann, gestalten sich daher ebenfalls vielseitig. In diesem Sinne schließe ich mich im Einverständnis mit Tarp folgender Meinung an:

> When a deep-rooted millenarian culture practice like lexicography passes from one medium to another, one would expect such a gigantic step to be more than a mere change of platform and that it would also involve improvements in terms of quality which in the case of lexicography can be translated into quicker, more accurate and personalized satisfaction of corresponding user needs. (Tarp 2012, S. 253)

4. Literatur

4.1 Zitierte (elektronische) Wörterbücher, Wörterbuchportale, elektronische Korpora
(alle Links zuletzt aufgerufen am 12.04.2016)

Belica, Cyril (2007): Kookkurrenzdatenbank CCDB – V3. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. http://corpora.ids-mannheim.de/ccdb/.

Bosque, Ignacio (2004): Redes. Diccionario combinatorio del español contemporáneo. Madrid.

CanooNet. Portal: Deutsche Wörterbücher und Grammatik. www.canoo.net/.

Casares, Julio (2001): Diccionario ideológico de la lengua española. 3. Aufl. Barcelona.

CLAVE-online: Diccionario de uso del español actual. Portal SM-Diccionarios. http://clave.smdiccionarios.com/app.php.

COSMAS: Corpus Search, Management and Analysis System. Institut für Deutsche Sprache, Mannheim. www.ids-mannheim.de/cosmas2/uebersicht.html.

CREA: Corpus de referencia del español actual. Real Academia Española. http://corpus.rae.es/creanet.html.

Cuervo, Rufino J. (1998): Diccionario de construcción y régimen de la lengua castellana. Continuado y editado por Instituto Caro y Cuervo. Barcelona.

DA: Diccionario de Alcalá (1995): Diccionario para la Enseñanza de la Lengua Española. Alcalá de Henares.

DEA: Seco, Manuel et al. (1999): Diccionario del español actual. 2 Bde. Madrid.

DEE: Diccionario de Español para Extranjeros (2002). Madrid.

DeReKo: Deutsches Referenzkorpus des Instituts für Deutsche Sprache Mannheim. Abrufbar über das Korpusrecherche und -analysesystem COSMAS II (Corpus Search, Management and Analysis System). https://cosmas2.ids-mannheim.de/cosmas2-web/.

Diccionarios.com. Portal Vox & Larousse. www.diccionario.com/.

[28] Zu unterschiedlichen Aspekten der existierenden elektronischen Ressourcen und Plädoyers zu zukünftigen elektronischen Werken verweise ich u.a. auf die Studien von Abel (2008), Chuchuy/Moreno (2002), Haß/Schmitz (2010), Haß (Hg.) (2005), Kemmer (2010), Klein/Geyken (2010), Klosa (Hg.) (2008), Klosa (2009), Mann (2010), Meliss (2016a), Müller-Spitzer/Engelberg (2011), Pöll (2010), Storrer (2010), Tarp (2012), Torres del Rey (2009).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

DICONALE: Diccionario conceptual del alemán y del español. Konzeptuell orientiertes Wörterbuch: Deutsch-Spanisch. USC. www.usc.es/gl/proxectos/diconale/aleman/.

DLE (2014): Diccionario de la Lengua Española. 23. Aufl. Madrid.

DLE: Diccionario de la Lengua española. Madrid. http://dle.rae.es/.

Dornseiff, Franz/Wiegand, Herbert E./Quasthoff, Uwe (2004): Der deutsche Wortschatz nach Sachgruppen. 8., neubearb. Aufl. Berlin. (print und elektronisch).

DSA: Diccionario de Salamanca – español para extranjeros (2006). Madrid.

DUDEN (2010): Stilwörterbuch. 9., völlig neu bearb. Aufl. (= Der Duden 2). Mannheim.

DUDEN-DaF (2010): Deutsch als Fremdsprache. Standardwörterbuch. 2., neu bearb. und erw. Aufl. Mannheim.

DUDEN-online Portal. www.duden.de/.

DWDS: Digitales Wörterbuch der deutschen Sprache. www.dwds.de/.

elexiko. Online-Wörterbuch der deutschen Gegenwartssprache. Institut für Deutsche Sprache, Mannheim. www.owid.de/wb/elexiko/start.html (Stand: 1.3.2013).

E-VALBU: Valenzwörterbuch online. Institut für Deutsche Sprache, Mannheim. http://hypermedia.ids-mannheim.de/evalbu/index.html.

Kempcke, Günter et al. (1999): Wörterbuch Deutsch als Fremdsprache. Berlin.

KV-online: Online Nachschlagewerk zu Kommunikationsverben. Institut für Deutsche Sprache, Mannheim. www.owid.de/docs/komvb/start.jsp.

LEO. Zweisprachiges Wörterbuchportal. www.leo.org/.

LGWB-DaF: Götz, Dieter (2015): Langenscheidt Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Berlin/München.

LHWBe: Langenscheidts Handwörterbuch Spanisch (2006). Spanisch – Deutsch (LHWBe-SD) / Deutsch – Spanisch (LHWBe-DS): elektronische Fassung. Berlin/München.

OpenThesaurus: Ein freies deutsches Wörterbuch für Synonyme. www.openthesaurus.de/.

OWID: Online-Wortschatz-Informationssystem Deutsch. Institut für Deutsche Sprache, Mannheim. www.owid.de/.

PONS-O: Das Sprachenportal. http://de.pons.eu/.

PONS-DaF (2015): Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Stuttgart.

Quasthoff, Uwe (2011): Wörterbuch der Kollokationen im Deutschen. Berlin.

Schumacher, Helmut et al. (2004): VALBU – Valenzwörterbuch deutscher Verben. (= Studien zur Deutschen Sprache 31). Tübingen.

Wahrig-DaF: Großwörterbuch Deutsch als Fremdsprache (2008). Berlin.

Wehrle, Hugo/Eggers, Hans (1993): Deutscher Wortschatz: ein Wegweiser zum treffenden Ausdruck. 17. Aufl. Stuttgart.

WordReference: Diccionario de Español, Inglés Francés y Portugues. www.wordreference.com/es/.

Wörterbuch für Synonyme: www.synonym.de.

Wortschatz Universität Leipzig. http://wortschatz.uni-leipzig.de/.

Woxikon: Online-Lexikon. www.woxikon.de.

4.2 Zitierte Literatur

Abel, Andrea (2008): ELDIT (Elektronisches Lernerwörterbuch Deutsch-Italienisch) und *elexiko*: Ein Vergleich. In: Klosa (Hg.), S. 175-189.

Abel, Andrea/Vettori, Chiara/Ralli, Natascia (Hg) (2014): Proceedings of the 16th EURALEX International Congress: The User in Focus. Bozen.

Chuchuy, Claudio/Moreno, Antonio (2002): Diccionarios españoles en formato electrónico. In: Fuentes Morán/ Werner (Hg.), S. 89-108.

Dentschewa, Emilia (2006): DaF-Wörterbücher im Vergleich: Ein Plädoyer für „Strukturformeln. In: Dimova, Ana et al. (Hg.): Zweisprachige Lexikographie und Deutsch als Fremdsprache. Hildesheim, S. 113-128.

Domínguez Vázquez, María José et al. (Hg.) (2015): Lexicografía de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva. Bd. 2. Berlin u.a.

Engel, Urich (2004): Deutsche Grammatik. Neubearbeitung. München.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Engelberg, Stefan (2010): Die lexikographische Behandlung von Argumentstrukturvarianten in Valenz- und Lernerwörterbüchern. In: Fischer, Klaus/Fobbe, Eilika/Schierholz, Stefan (Hg.): Valenz und Deutsch als Fremdsprache. Frankfurt a.M., S. 113-141.

Engelberg, Stefan (erscheint): The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In: Boas, Hans/Ziem, Alexander (Hg.): Constructional approaches to argument structure in German. Boston/Berlin.

Engelberg, Stefan (2015): Gespaltene Stimulus-Argumente bei Psych-Verben. Quantitative Verteilungsdaten als Indikator für die Dynamik sprachlichen Wissens über Argumentstrukturen. In: Engelberg, Stefan/Meliss, Meike/Proost, Kristel/Winkler, Edeltraud (Hg.): Argumentstruktur zwischen Valenz und Konstruktion. (= Studien zur Deutschen Sprache 68). Tübingen, S. 469-491.

Engelberg, Stefan et al. (2011): Argumentstrukturmuster als Konstruktionen? Identität – Verwandtschaft – Idiosynkrasien. In: Engelberg, Stefan/Holler, Anke/Proost, Kristel (Hg.) (2011): Sprachliches Wissen zwischen Lexikon und Grammatik. Jahrbuch 2010 des Instituts für Deutsche Sprache. Berlin, S. 71-112.

Engelberg, Stefan et al. (2012): Argument structures and text genre: Cross-corpus evaluation of the distributional characteristics of argument structure realizations. In: Heid, Ulrich (Hg.): Corpora and lexicography. (= Lexicographica 28). Berlin, S. 13-48.

Engelberg, Stefan/Lemnitzer, Lothar (2009): Lexikographie und Wörterbuchbenutzung. 4. Aufl. Tübingen.

Fuentes Morán, María Teresa (1997): Gramática en la lexicografía bilingüe. Morfología y sintaxis en diccionarios español-alemán desde el punto de vista del germanohablante. Tübingen.

Fuentes Morán, María Teresa/Modal, Benedikt (Hg.) (2009): Investigaciones sobre lexicografía bilingüe. Granada.

Fuentes Morán, María Teresa/Werner, Reinhold (Hg.) (1998): Lexicografías iberrománicas: problemas, propuestas y proyectos. Frankfurt a.M.

González Ribao, Vanessa (2015): Sobre algunos conflictos en la 'pre'-lexicografía: La selección de corpus para la elaboración de un diccionario conceptual alemán-español. In: Domínguez Vázquez et al. (Hg.), S. 247-269.

González Ribao, Vanessa/Meliss, Meike (2015): Vorschläge zur Ausarbeitung eines onomasiologisch-konzeptuell orientierten Produktionswörterbuches im zweisprachigen Lernerkontext: Deutsch-Spanisch. In: Calañas Continente, José A. et al. (Hg.): Die Wörterbücher des Deutschen. (= Studien zur Linguistik des Deutschen – Spanische Akzente 2). Frankfurt a.M., S. 109-136.

González Ribao, Vanessa/Proost, Kristel (2015): El campo léxico al servicio de la lexicografía: Un análisis contrastivo en torno a algunos campos de los verbos de comuncación en alemán y español. In: Domínguez Vázquez et al. (Hg.), S. 223-245.

Gouws, Rufus H. (1998): Das System der sogenannten Strukturformeln in Langenscheidts Grosswörterbuch Deutsch als Fremdsprache. Eine kritische Übersicht. In: Wiegand (Hg.), S. 63-76.

Haensch, Günther/Omeñaca, Carlos (2004): Los diccionarios del español en el siglo XXI. 2. Aufl. Salamanca.

Harras, Gisela et al. (2004): Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. (= Schriften des Instituts für Deutsche Sprache 10.1). Berlin/New York.

Harras, Gisela/Proost, Kristel/Winkler, Edeltraud (2007): Handbuch deutscher Kommunikationsverben. Teil 2: Lexikalische Strukturen. (= Schriften des Instituts für Deutsche Sprache 10.2). Berlin/New York.

Haß, Ulrike (Hg.) (2005): Grundfragen der elektronischen Lexikographie. *elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York.

Haß, Ulrike/Schmitz, Ulrich (2010): Einleitung: Lexikographie im Internet 2010. In: Haß/Schmitz (Hg.), S. 1-18.

Haß, Ulrike/Schmitz, Ulrich (Hg.) (2010): Lexikographie im Internet 2010. (= Lexicographica 26). Berlin/New York.

Hausmann, Franz J. (1990a): Das Synonymenwörterbuch. Die kumulative Synonymik? In: Hausmann et al. (Hg.), S. 1076-1081.

Hausmann, Franz J. (1990b): Das Antonymenwörterbuch. In: Hausmann et al. (Hg.), S. 1081-1083.

Hausmann, Franz J. (1991): Die zweisprachige Lexikographie Spanisch – Deutsch / Deutsch – Spanisch. In: Hausmann et al. (Hg.), S. 2987-2991.

Hausmann, Franz J./Werner, Reinhold (1991): Spezifische Bauteile und Strukturen zweisprachiger Wörterbücher: Eine Übersicht. In: Hausmann et al. (Hg.), S. 2729-2770.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Hausmann, Franz J. et al. (Hg.) (1990): Dictionaries. An international encyclopedia of lexicography. 2. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.2). Berlin/New York.

Hausmann, Franz J. et al. (Hg.) (1991): Dictionaries. An international encyclopedia of lexicography. 3. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.3). Berlin/New York.

Honnef-Becker, Irmgard (2002): Die Benutzung des DE GRUYTER WÖRTERBUCHES DEUTSCH ALS FREMDSPRACHE in Situationen der Textproduktion. In: Wiegand (Hg.), S. 623-646.

Kemmer, Katharina (2010): Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien. (= OPAL 2/2010). Mannheim. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2010-2.pdf.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Haß/Schmitz (Hg.), S. 79-96.

Klosa, Annette (2009): Modern German dictionaries and their impact on linguistic research. In: Bruti, Silvia et al. (Hg.): Perspectives on lexicography in Italy and Europe. Cambridge, S. 175-199.

Klosa, Annette (Hg.) (2008): Lexikographische Portale im Internet. (= OPAL. Sonderheft 1/2008). Mannheim. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf.

Konerding, Klaus P. (1998): Die semantischen Angaben in Langenscheidts Grosswörterbuch Deutsch als Fremdsprache. In: Wiegand (Hg.), S. 107-143.

Korhonen, Jarmo (Hg.) (2001): Von der mono- zur bilingualen Lexikographie für das Deutsche. Frankfurt a.M.

Köster, Lutz/Neubauer, Fritz (2002): Kollokationen und Kompetenzbeispiele im DE GRUYTER WÖRTERBUCH DEUTSCH ALS FREMDSPRACHE. In: Wiegand (Hg.), S. 283-310.

Lehr, Andrea (1998): Kollokationen in Langenscheidts Grosswörterbuch Deutsch als Fremdsprache. In: Wiegand (Hg.), S. 256-281.

Mann, Manfred (2010): Internet-Wörterbücher am Ende der „Nulljahre“. In: Haß/Schmitz (Hg.), S. 19-45.

Meliss, Meike (2005): Recursos lingüísticos alemanes relaticos a ‘GERÄUSCH’ y sus posibles correspondencias en español. Un estudio lexicológico modular-integrativo. Frankfurt a.M.

Meliss, Meike (2006): Kontrastive Wortfeldstudie für das Sprachenpaar Deutsch-Spanisch am Beispiel der Verben für GERÄUSCH. In: Wolf, Dietrich et al. (Hg.): Lexikalische Semantik und Korpuslinguistik. Tübingen, S. 141-167.

Meliss, Meike (2013): Das zweisprachige Wörterbuch im bilateralen deutsch-spanischen Kontext. Alte und neue Wege. In: Domínguez Vázquez, María José (Hg.): Trends in der deutsch-spanischen Lexikographie. (= Spanische Akzente. Studien zur Linguistik des Deutschen 1). Frankfurt a.M., S 61-87.

Meliss, Meike (2014a): Vorüberlegungen zu einem zweisprachigen Produktionslernerwörterbuch für das Sprachenpaar DaF und ELE. In: Reimann, Daniel (Hg.): Kontrastive Linguistik und Fremdsprachendidaktik. Iberoromanisch – Deutsch. (= Romanische Fremdsprachenforschung und Unterrichtsentwicklung 2). Tübingen, S. 113-137.

Meliss, Meike (2014b): Die fremdsprachige Produktionssituation im Fokus eines onomasiologisch-konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches für das Sprachenpaar Deutsch - Spanisch: Theoretische und methodologische Grundlagen von DICONALE. In: Abel/Vettori/Ralli (Hg.), S. 1119-1134.

Meliss, Meike (2015a): Argumentstrukturen, Valenz und Konstruktionen: Eine korpusbasierte Studie deutscher und spanischer „Geruchsverben“ im Kontrast. In: Engelberg, Stefan et al. (Hg.): Argumentstruktur zwischen Valenz und Konstruktion. (= Studien zur Deutschen Sprache 68). Tübingen, S. 317-339.

Meliss, Meike (2015b): Zum Kombinationspotenzial von Verben in einsprachigen DaF-Lernerwörterbüchern: Kritische Bestandsaufnahme – Neue Anforderungen. In: Zeitschrift für Deutsch als Fremdsprache 20, 1, S. 14-27.

Meliss, Meike (2015c): Was suchen und finden Lerner des Deutschen als Fremdsprache in aktuellen Wörterbüchern? Auswertung einer Umfrage und Anforderungen an eine aktuelle DaF-Lernerlexikographie. In: Roelcke, Thorsten (Hg.): Wörterbücher für Deutsch als Fremdsprache- Probleme und Perspektiven. Themenreihe: Info DaF 42,4/2015, S. 401-432.

Meliss, Meike (2016a): Wie viele und welche bilingualen Online-Wörterbücher brauchen wir für den DaF-Bereich? Erstellung eines Kriterienrasters und erste Bestandsaufnahme aus der Sicht eines hispanophonen Lernerkontextes. In: Cuartero Otal, Juan et al. (Hg.): Querschnitt durch die deutsche Sprache aus spanischer Sicht. Perspektiven der Kontrastiven Linguistik. Berlin, S. 187-210.

Meliss, Meike (2016b): Gesprochene Sprache in DaF-Lernerwörterbüchern. In: Handwerker, Brigitte et al. (Hg.): Gesprochene Fremdsprache Deutsch. Baltmannsweiler, S. 179-199.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Meliss, Meike/Sánchez Hernández, Paloma (2015): Theoretical and methodological foundations of the DICONALE project: A conceptual dictionary of German and Spanish. In: Silvestre, João Paulo et al. (Hg.): Dicionários que não existem. Lissabon, S. 163-179.

Meyer, Meike/Wiegand, Herbert E. (2000): Gemischt-semiintegrierte Mikrostrukturen für deutsch-spanische Printwörterbücher. In: Wiegand, Herbert E. (Hg.): Studien zur zweisprachigen Lexikographie mit Deutsch V. (= Germanistische Linguistik 151-152). Hildesheim, S. 87-171.

Model, Benedikt (2010): Syntagmatik im zweisprachigen Wörterbuch. Berlin.

Müller-Spitzer, Caroline/Engelberg, Stefan (2011): Elektronische Lexikographie zwischen Grammatik und Lexikon. In: Engelberg et al. (Hg.), S. 559-572.

Pöll, Bernhard (2010): Internetlexikographie der iberomanischen Sprachen. In: Haß/Schmitz (Hg.), S. 169-173.

Proost, Kristel/Müller-Spitzer, Caroline (2014): Degrees of synonymity as the basis of a network for German communication verbs in the Online Reference Work Kommunikationsverben in OWID. In: Abel/Vettori/Ralli (Hg.), S. 1171-1186.

Roelcke, Thorsten (2002): Das Verhältnis der semasiologischen und onomasiologischen Angaben im DE GRUYTER WÖRTERBUCH DEUTSCH ALS FREMDSPRACHE. In: Wiegand (Hg.), S. 201-244.

Schafroth, Elmar (2002): Die Grammatik der Verben im DE GRUYTER WÖRTERBUCH DEUTSCH ALS FREMDSPRACHE. In: Wiegand (Hg.), S. 57-74.

Storrer, Angelika (2010): Deutsche Internet-Wörterbücher: Ein Überblick. In: Haß/Schmitz (Hg.), S. 154-164.

Tarp, Sven (2012): Online dictionaries: Today and tomorrow. In: Heid, Ulrich (Hg.): Corpora and lexicography. (= Lexicographica 28). Berlin, S. 253-267.

Torres del Rey, Jesús (2009): Diccionarios electrónicos bilingües: Nuevas posibilidades de futuro. In: Fuentes Morán/Modal (Hg.), S. 29-79.

Werner, Reinhold (1998): La selección de lemas en los diccionarios español-alemán y alemán-español o ¿un diccionario de qué lengua es un diccionario de las lenguas española y alemana? In: Fuentes Morán/Werner (Hg.), S. 139-156.

Wiegand, Herbert E. (2005): Äquivalentpräsentation und Wörterbuchfunktion in zweisprachigen Printwörterbüchern. Mit einem Seitenblick auf die so genannte „moderne lexikographische Funktionslehre“. In: Igla, Birgit/Petkov, Pavel/Wiegand, Herbert E. (Hg.): Kontrastive Lexikologie und zweisprachige Lexikographie. Hildesheim, S. 1-38.

Wiegand, Herbert E. (Hg.) (1998): Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand von „Langenscheidts Großwörterbuch Deutsch als Fremdsprache“. Tübingen.

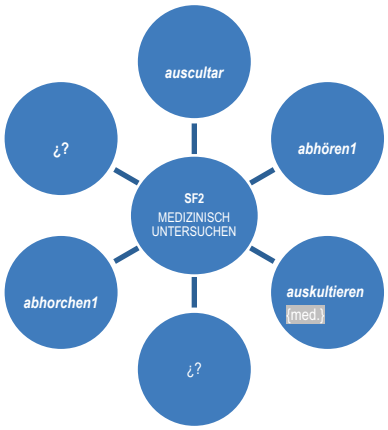
Wiegand, Herbert E. (Hg.) (2002): Perspektiven der pädagogischen Lexikographie des Deutschen II. Untersuchungen anhand des „de Gruyter Wörterbuchs Deutsch als Fremdsprache“. Tübingen.

Wolf, Norbert R. (2001): Kollokationen und semantische Valenz im einsprachigen Wörterbuch. In: Korhonen (Hg.), S. 153-162.

Wolski, Werner (2002): Das DE GRUYTER WÖRTERBUCH DEUTSCH ALS FREMDSPRACHE und LANGENSCHEIDTS GROßWÖRTERBUCH DEUTSCH ALS FREMDSPRACHE. Ein Vergleich im Hinblick auf die Semantik. In: Wiegand (Hg.), S. 3-34.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Anhang: Abbildungen 6b-e



Deutsch: ⌂ *untersuchen* [−Wahrnehmung]

(2.1.$_{dt}$) Er ***hört*** den Brustkorb ***ab***, sieht in den Rachen, begutachtet die Ohren. (NUN07/MAR.02161 NN, 19.03.2007)

(2.2.$_{dt}$) Ulrich Eggert ***hört*** im Transporter einem Patienten aus der Männerunterkunft die Herztöne ***ab.*** (HAZ09/MAR.01732 HAZ, 10.03.2009, S. 15)

(2.3.$_{dt}$) Wie ein Arzt seine Patienten ***abhorcht***, röntgt oder verbindet, wird Kindern anhand ihrer Teddybären gezeigt. (BRZ05/SEP.16312 Braunschw. Z., 28.09.2005)

(2.4.$_{dt}$) ***Auskultiert*** man die Lunge direkt nach dem Aufstehen der Patienten nach langer Liegezeit, hört man [...] ein leises Knistern. (COSMAS)

[…]

Spanisch: ⌂ *examinar* [−Wahrnehmung]

(2.1.$_{sp}$) El médico le ***auscultó*** el pecho y la espalda, en silencio. (CREA)

[...]

Abb. 6b: **Stufe 1:** Subfeld 2 (SF2) mit der Zuordnung einiger Lexeme und entsprechender Belegbeispiele in beiden Sprachen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Deutsch: ⌂ *zuhören* [+Wahrnehmung] [+auditiv]

(3.1.$_{dt}$) 15-jährig ***hörte*** er den Jazzpianisten Eroll Garner, spielte nach, was er ***hörte***. (BRZ10/APR.13294 Braunschw. Z.,30.04.2010)

(3.2.$_{dt}$) Sie ***hörte*** die Aussagen der verängstigten Kinder. (RHZ05/JUN.16629 RZ, 15.06.2005)

(3.3.$_{dt}$) Wenn ich das über einen andern Künstler lesen würde, würde ich dessen Musik nie ***anhören.*** (A97/OKT.28423 St. Galler Tagblatt, 07.10.1997)

(3.4.$_{dt}$) Sie ***hören sich*** die Probleme ***an,*** die den Kindern auf den Nägeln brennen [...]. (R97/SEP.72386 Frankf. Rundschau, 15.09.1997, S. 4)

(3.5.$_{dt}$) [...] sie hat mir geduldig ***zugehört*** und geantwortet. (BRZ08/JUN.06305 Braunschw. Z., 12.06.2008)

(3.6.) Mit großem Interesse hatten die Pflegekräfte [...], dem Vortrag ***zugehört*** [...]. M03/FEB.11701 Mannh. Morgen, 22.02.2003

(3.7.$_{dt}$) Am Ziel angekommen, ***lauschten*** die Kinder aufmerksam der Piratin Lilli [...]. (A12/JUN.09072 St. Galler Tagbl., 20.06.2012, S. 41)

(3.8.$_{dt}$) Aber das Publikum ***lauschte*** an diesem Abend nicht nur den spannenden Krimis. (RHZ05/SEP.18942 RZ, 15.09.2005

(3.9.$_{dt}$) Jago, das Aas, ***lauscht*** hinter Säulenfluchten den honorigen Erklärungen Othellos vor dem Rat von Venedig. (UN93/JUN.01879 NN, 26.06.1993, S. 22)

(3.10.$_{dt}$) In dem bis auf den letzten Platz gefüllten Gotteshaus ***lauschten*** mehr als 2000 Besucher andächtig den Chorgesängen [...]. (L98/NOV.16306 Berliner Morgenpost, 01.11.1998, S. 9)

(3.11.$_{dt}$) Man setzt sich, ***horcht*** der Feldpredigt und geniesst die Sonne. (A00/JUN.38103 St. Galler Tagblatt, 02.06.2000)

[…]

Spanisch: ⌂ *escuchar* [+Wahrnehmung] [+auditiv]

(3.1.$_{sp}$) „Cuando ***escuché*** a Bob Dylan tuve la sensación de que me iba a dedicar a la música“. (CREA)

(3.2.$_{sp}$) Esta noche, escribió, mientras ***oíamos*** el Otello de Verdi en el palco de sus distinguidos padres, he sentido la tentación de inclinarme hacia delante y besar sus hombros. (CREA)

[…]

Abb. 6c: **Stufe 1:** Subfeld 3 (SF3) mit der Zuordnung einiger Lexeme und entsprechender Belegbeispiele in beiden Sprachen[29]

[29] Es werden in diesem Feld keine Kommunikationsverben (z.B. jemandem die Vokabeln *abhören* (abfragen), jemanden *verhören*, die Zeugen *hören*, (*sich*) den Angeklagten *anhören* etc.) aufgenommen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Deutsch: ⌂ *beobachten, überprüfen, kontrollieren* [+/–Wahrnehmung] [+/–auditiv]

(4.1.$_{dt}$) Er ***hörte*** auf den regelmäßigen Atem des Tieres, [...]. (NUZ03/SEP.00686 NZ, 06.09.2003)

(4.2.$_{dt}$) [...] er ***hört*** in die Kompositionen hinein, formt sie [...]. (O94/APR.33863 Neue Kronen-Ztg., 13.04.1994, S. 20)

(4.3.$_{dt}$) Er ***hört*** seine Mailbox ***ab*** [...]. (RHZ09/SEP.14382 RZ, 16.09.2009)

(4.4.$_{dt}$) [...] blieb noch einige Sekunden über sie gebeugt und ***horchte*** auf die tiefen, regelmäßigen Atemzüge. (GR1/TL1.12150 Weyden: Träume, 1990 [S. 23])

(4.5.$_{dt}$) Die Zeiten, in denen ein Auto-Mechaniker mit seinem fachlich geschulten Ohr kurz in den Motorraum ***horcht*** [...] sind heute größtenteils vorbei. (BRZ06/APR.12848 Braunschw. Z., 26.04.2006)

(4.6.$_{dt}$) Mit Geophonen ***horchen*** Bundesherrsoldaten den Boden ***ab*** - bisher erfolglos. (O98/JUL.72449 Neue Kronen-Ztg., 25.07.1998, S. 8)

(4.7.$_{dt}$) Um Amphibien zu zählen, zieht er am Abend los und ***lauscht*** auf die Rufe der Tiere. (A12/MAI.12895 St. Galler Tagbl., 30.05.2012, S. 26)

(4.8.1.$_{dt}$) Moritz schloß die Augen und ***lauschte*** auf die Atemzüge_des Alten. (BRZ07/APR.00154 Braunschw. Z., 28.04.2007)

(4.8.2.$_{dt}$) Dann lag sie da und ***lauschte*** den Atemzügen_ihrer Mutter. (M03/APR.27702 Mannh. Morgen, 28.04.2003)

(4.9.$_{dt}$) Er blieb mitten im Zimmer stehen und ***lauschte*** auf den Hund, der da unten bellte, [...] (BRZ06/NOV.07368 Braunschw. Z., 14.11.2006)

(4.10.$_{dt}$) Auf einer vorgegebenen Route wandernd, ***lauscht*** sie in den Wald. (A12/MAI.12993 St. Galler Tagbl., 30.05.2012, S. 39)

(4.11.$_{dt}$) Der Verein [...] bietet im März einen Erlebnistag am Weiher für Kinder und Jugendliche oder beobachtet und ***belauscht*** im April Vögel in der Region. (A97/SEP.24729 St. Galler Tagblatt, 17.09.1997)

[…]

Spanisch: ⌂ *observar, comprobar, controlar, sentir, revisar, registrar* [+/–Wahrnehmung] [+/–auditiv]

(4.1.$_{sp}$) Es una enfermedad muy común a lo largo del viaje, y es contagiosa, incluso al acercarse y ***sentir*** la respiración de otro […]. (CREA)

(4.2.$_{sp}$) Necesita hablar para desalojar la angustia y recomponer todos los días que transcurrieron hasta que ocurrió lo que podía no haber sucedido si una noche, al volver a casa, hubiera ***escuchado*** la única llamada que ***registraba*** su contestador. (CREA)

[…]

Abb. 6d: **Stufe 1:** Subfeld 4 (SF4) mit der Zuordnung einiger Lexeme und entsprechender Belegbeispiele im Deutschen. Das Spanische lexikalisiert die hier relevanten Konzepte nicht und muss daher auf andere sprachliche Ausdrucksmittel zurückgreifen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Deutsch: ⌂ *überwachen, kontrollieren, spionieren* [−Wahrnehmung]

(5.1.$_{dt}$) Damals ***horchte*** die Bundespolizei unter ihrem berühmt-berüchtigten Direktor Edgar Hoover ständig US-Bürger und Politiker illegal ***ab.*** (P93/NOV.37089 Die Presse, 19.11.1993)

(5.2.$_{dt}$) Wolfsburger Rauschgift-Großhändler muss 3 Jahre 6 Monate in Haft – Beamten ***horchten*** Telefonate ***ab.*** (BRZ10/JAN.08580 Braunschw. Z., 21.01.2010)

(5.3.$_{dt}$) Aus sicherer Entfernung wurde das Flugzeug mit Spezialgeräten ***abgehorcht,*** die wie die raffinierten „James-Bond-Waffen" der Einheit auf geheime Sonderbestellung von der französischen Industrie angefertigt worden sind. (NUN94/DEZ.02222 NN, 28.12.1994, S. 3)

(5.4.$_{dt}$) Tim ***lauschte*** an ihrer Tür – ob sie vielleicht verreist war? (M02/DEZ.97379 Mannh. Morgen, 28.12.2002)

(5.5.$_{dt}$) Der präpubertäre Lüdecke ***lauschte*** an der verschlossenen Tür seiner zehn Jahre älteren Schwester. Mit ihrem Freund diskutierte sie Orte und Themen der Zeit. (RHZ03/APR.02569 RZ, 03.04.2003)

(5.6.$_{dt}$) Sie stieg die Treppe hinauf und ***horchte*** an ihrer Tür: alles still. (BRZ06/NOV.08431 Braunschw. Z., 16.11.2006)

(5.7.$_{dt}$) Der Heimatdienst-Mann erklärte John, er könne am Telefon nicht offen sprechen, da er das Gefühl habe, seine Kollegen ***belauschten*** ihn. (L98/MAI.02387 Berliner Morgenpost, 28.05.1998, S. 31)

(5.8.$_{dt}$) Bis einer der drei zufällig eine Unterhaltung der beiden anderen ***belauscht*** und unangenehme Dinge über sein Privatleben erfährt. (M12/JAN.03051 Mannh. Morgen, 12.01.2012, S. 28)

(5.9.$_{dt}$) Die Polizei ***belauschte*** die Telefone der Bande, verwanzte ihre Autos [...]. (D HMP11/APR.00482 MOPO, 06.04.2011, S. 13)

(5.10.$_{dt}$) Im Vorfeld ***belauschten*** die Ermittler hunderte Telefongespräche. (I00/JUL.42625 Tiroler Tagesztg., 22.07.2000)

(5.11.$_{dt}$) Polizei ***hört*** Anwaltsgespräche ***ab***. (SWR 4M08/NOV.87862 Mannh. Morgen, 12.11.2008, S. 22)

(5.12.$_{dt}$) Während zweier Jahre habe die Politische Polizei seine Post geöffnet, das Telefon ***abgehört*** und versucht, Freunde zu überreden, ihn zu bespitzeln (E97/APR.08712 Zürcher Tagesanzeiger, 11.04.1997, S. 9)

(5.13.$_{dt}$) Ihm selbst folgten Sicherheitsbeamte auf Schritt und Tritt, ***hörten*** ihn ***ab*** und schüchterten seine Interviewpartner ein. (NUN05/DEZ.00312 NN, 03.12.2005)

[...]

Spanisch: ⌂ *controlar, espiar, interceptar, pinchar* ... [−Wahrnehmung]

(5.1.$_{sp}$) [...] los servicios de inteligencia ***interceptaron*** una comunicación radiofónica en la que el jefe del bloque Caribe de las FARC, Iván Márquez, hablaba sobre el frustrado atentado. (CREA)

(5.2.$_{sp}$) Durante la última etapa de las negociaciones entre estadounidenses y panameños, los primeros ***espiaron*** con micrófonos las conversaciones de los segundos […]. (CREA)

(5.3.$_{sp}$) Los jueces anulan la condena a dos „narcos" porque les ***pincharon*** el teléfono ilegalmente. (CREA)

(5.4.$_{sp}$) [...] los propios servicios de inteligencia del Estado ***pincharon*** y grabaron conversaciones del Rey sobre el Consejo de Europa [...]. (CREA)

[...]

Abb. 6e: **Stufe 1:** Subfeld 5 (SF5) mit der Zuordnung einiger Lexeme und entsprechender Belegbeispiele im Deutschen. Das Spanische lexikalisiert die hier relevanten Konzepte nicht und muss daher auf andere sprachliche Ausdrucksmittel zurückgreifen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Wortfamilien im Onlinewörterbuch

Annette Klosa (*Institut für Deutsche Sprache, Mannheim*)

1. Einleitung

Unter einer Wortfamilie wird – zunächst noch eher unscharf – die Menge an komplexen Wörtern bzw. Lexemen verstanden, die den gleichen Wortstamm bzw. das gleiche Grundmorphem enthalten und die somit als Gruppe zusammengehören. In diachroner Sicht gehören zu einer Wortfamilie genauer alle Wortbildungen, „die in ihrer Struktur über ein etymologisch identisches Grundmorphem (Kernlexem) verfügen, vgl. *zieh*(*en*), *Ziehung*, *ausziehen*, *Zug*, *Regionalzug*“ (Fleischer/Barz 2012, S. 99), sodass bei dieser Definition zu einer Wortfamilie auch Wörter gehören, „die in der Gegenwartssprache formal und semantisch nicht mehr ohne Weiteres an das Kernlexem anzuschließen und auch untereinander nur noch bei diachroner Betrachtung als verwandt erkennbar sind“ (ebd.). Bei synchroner Betrachtung gehören zu einer Wortfamilie hingegen nur solche Lexeme, die hinsichtlich ihrer Bildung und Bedeutung durchsichtig sind und von gegenwärtigen Sprechern zu einem Grundmorphem bzw. Kernlexem zugeordnet werden können. Diesen Ansatz verfolgt etwa Augst (1998) in seinem Wortfamilienwörterbuch. Splett (zuletzt 2013, S. 120-124) geht davon aus, dass bei einer lexikographischen Erfassung von Wortfamilien weniger die etymologische Zusammengehörigkeit von Lexemen, als vielmehr die synchron erkennbaren Motivationsbeziehungen zwischen gebildeten Wörtern zu berücksichtigen und zu beschreiben sind. Im Kontext des vorliegenden Beitrags, der sich mit der Frage der Darstellung von Wortfamilien in Onlinewörterbüchern beschäftigt, wird ebenfalls ein synchrones Verständnis der Wortfamilie vorausgesetzt. Zu einer Wortfamilie gehören hier nur solche Bildungen, deren lexikalische Bedeutung sich aus der Bedeutung des Kernlexems sowie der weiteren Bildungsbestandteile erschließen lässt.

Die Frage der Behandlung von Wortfamilien im allgemeinsprachigen Onlinewörterbuch steht im weiteren Kontext der Frage nach der Behandlung von Wortbildung im Wörterbuch allgemein. Zunächst ist festzuhalten, dass in alphabetisch angeordneten (gedruckten) Bedeutungswörterbüchern die einer Wortfamilie zugehörigen Lexeme nicht immer an der gleichen Stelle im Wörterbuch stehen. Beispielsweise finden sich Präfixbildungen zu einem Kernlexem oder Komposita mit dem Kernlexem als Grundwort an alphabetisch vom Kernlexem entfernter Stelle. Traditionelle Wörterbücher enthalten außerdem keine explizite Information dazu, zu welcher Wortfamilie bzw. zu welchem Kernlexem ein Stichwort gehört. Dies hängt auch damit zusammen, dass in allgemeinsprachigen Bedeutungswörterbüchern häufig sehr unsystematische Angaben dazu gemacht werden, wie komplexe Stichwörter gebildet sind.[1] Ebenso fehlen häufig Angaben zu Wortbildungsmitteln, und nur teilweise werden zu einem Stichwort gebildete Wörter in den Wortartikeln verzeichnet.[2]

Es stellt sich vor diesem Hintergrund die Frage, wie in allgemeinsprachigen Onlinewörterbüchern (und anderen elektronischen Wörterbüchern) trotzdem Wortfamilien sichtbar gemacht

[1] Vgl. hierzu genauer Klosa (2005, S. 141-143).
[2] Einen ausführlichen Überblick über die lexikographische Praxis sowohl in gedruckten wie in elektronischen Wörterbüchern gibt Ulsamer (2013b).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

und auch als Zugriffsstruktur genutzt werden können.[3] Dies gilt umso mehr, als Wortbildungsregeln nicht nur das Lexikon einer Sprache erweitern, wenn sie angewendet werden, sondern dieses auch strukturieren (vgl. ten Hacken 2013, S. 167). Die Wörter eines Lexikons, die im allgemeinsprachigen (gedruckten) Wörterbuch notwendigerweise nach dem Alphabet geordnet stehen, sind in Wirklichkeit jedoch „miteinander in vielen unterschiedlichen Weisen verbunden [...], darunter auch durch Wortbildungsregeln“ (ebd.). Diesen auch semantischen Zusammenhang innerhalb einer Wortfamilie u.a. mithilfe „einer multidimensionalen Präsentationsweise“ (Eichinger 2013, S. 84) aufzuzeigen, könnte Aufgabe von Onlinewörterbüchern sein, die generell nicht an die alphabetische Anordnung der Stichwörter gebunden sind, sondern in denen der Zugriff auf die Einträge bzw. Gruppen von Einträgen über Suchanfragen nach Zeichenfolgen oder nach bestimmten, auch semantischen Kriterien geschieht. De Caluwe (2013, S. 105) spricht in diesem Kontext auch vom „onomasiological aspect of word formation“, den es zu nutzen gelte, damit die Wörterbuchbenutzer tatsächlich das im Wörterbuch finden können, was sie suchen.

Die folgenden Überlegungen vor diesem Hintergrund beziehen sich insbesondere auf *elexiko*, das elektronische, lexikalisch-lexikologische und korpusbasierte Informationssystem des Instituts für Deutsche Sprache Mannheim,[4] das als Prototyp eines einsprachigen Bedeutungswörterbuches im Internet[5] gelten soll.

2. Wortbildung in *elexiko*

Mit *elexiko* entsteht ein korpusgestützt erarbeitetes Onlinewörterbuch zum Gegenwartsdeutschen,[6] das über die gesamte, etwa 300.000 Einträge umfassende Stichwortliste Angaben zu Rechtschreibung und Worttrennung enthält und automatisch ausgewählte Korpusbeispiele sowie Hyperlinks zu weiteren grammatischen und semantischen Informationen zum Stichwort bietet. In ausgewählten Wortschatzbereichen werden die Stichwörter auch vollständig lexikographisch in ihrer Bedeutung und Verwendung beschrieben. Illustrationen und gesprochensprachliche Belege ergänzen das Angebot. *elexiko* erscheint innerhalb des Wörterbuchportals OWID,[7] in dem weitere, am IDS erarbeitete wissenschaftliche Wörterbücher (z. B. Neologismenwörterbuch, Sprichwörterbuch, Diskurswörterbücher) publiziert werden. Diese Wörterbücher sind untereinander vernetzt und werden über verschiedene Suchmöglichkeiten gemeinsam erschlossen.[8]

Entgegen der in traditionellen Printwörterbüchern üblichen, eher unsystematischen Behandlung von Wortbildung werden in *elexiko* alle durch Wortbildung entstandenen Lemmata in der Art ihrer Gebildetheit analysiert. Die Bestandteile der Bildung werden im Wortartikel verzeichnet und online, soweit sie selbst Stichwort in *elexiko* sind, als Hyperlink zu den entsprechenden Einträgen dargestellt (vgl. Abb. 1).

---

[3] Splett (2013, S. 117) versteht unter einem elektronischen Wörterbuch hingegen generell „ein elektronisches Wortfamilienwörterbuch“, da nur der neue Typ des Wortfamilienwörterbuches „die Wortschatzstrukturen in angemessener Weise explizit darzulegen“ (ebd., S. 120) vermöge.

[4] Vgl. www.elexiko.de.

[5] Zu *elexiko* als Bedeutungswörterbuch vgl. Klosa (2011a).

[6] Zur Konzeption von *elexiko* vgl. Haß (Hg.) (2005). Zur praktischen Umsetzung dieser Konzeption vgl. Klosa (Hg.) (2011).

[7] Vgl. www.owid.de.

[8] Vgl. Müller-Spitzer (2011).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

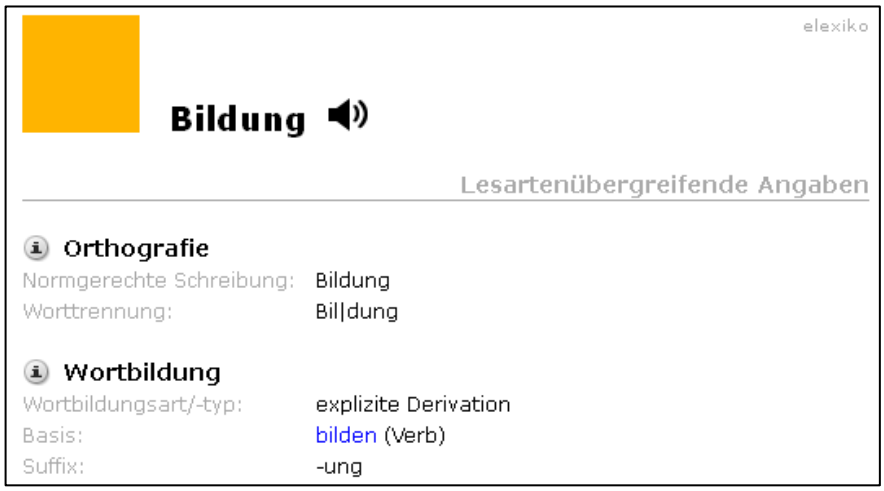


Abb. 1: Wortbildungsangaben zum Stichwort *Bildung* in *elexiko*

Die Wortbildungsangabe kann auch bezogen auf Einzelbedeutungen (in *elexiko* „Lesarten“) erfolgen, wenn verschiedene Bedeutungspunkte durch unterschiedliche Wortbildungsprozesse entstanden sind. Im Fall des Stichwortes *Bild* ist etwa die Lesart ‚künstlerische Darstellung‘ als Konversion zum Verb *bilden* zu bestimmen (vgl. Abb. 2), eine weitere Lesart ‚Spiegelung‘ aber als Kurzwort zum Kompositum *Spiegelbild*.

Abb. 2: Lesartenbezogene Wortbildungsangabe zur Lesart ‘künstlerische Darstellung’ im Stichwort *Bild* in *elexiko*

Bei Simplizia, die oft Kernlexeme von Wortfamilien sind, werden die (automatisch ermittelten) Wortbildungsprodukte[9] eingetragen. Öffnet man das Bildschirmfenster mit den Angaben zu Wortbildungsprodukten, werden diese in Komposita, Derivate und andere Wortbildungsprodukte unterteilt und nach dem Alphabet oder ihrer Frequenz im zugrundeliegenden Korpus sortierbar angezeigt (vgl. Abb. 3). Die Gestaltung der in eine Bildung eingegangenen

[9] Zur Ermittlung und Präsentation der Wortbildungsprodukte vgl. genauer Ulsamer (2013a). Zum Einsatz von Morphologie-Tools zur automatischen Wortformenanalyse vgl. Simon (2013).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Bestandteile als Hyperlinks einerseits und die Erfassung der Wortbildungsprodukte zu einem Stichwort als Hyperlinks andererseits ermöglicht es den Benutzern von *elexiko*, sich innerhalb einer Wortfamilie kreisförmig zu bewegen. In den in Abbildungen 1 und 2 gezeigten Fällen gelangt man vom expliziten Derivat *Bildung* bzw. der Konversion *Bild* über die Hyperlinks zum Verb *bilden*, das den lexikalischen Kern der Wortfamilie bildet. Unter den hier eingetragenen Wortbildungsprodukten findet sich beispielsweise das Nomenderivat *Bildung* wieder (vgl. Abb. 3), dessen Wortartikel, in dem wiederum *bilden* als Ableitungsbasis verzeichnet ist, man über den Hyperlink erreichen kann.

Abb. 3: Wortbildungsprodukte zum Stichwort *bilden* in *elexiko*

Die Darstellung der Wortbildung erfolgt in *elexiko*, wie die Beispiele zeigen, nach wie vor bezogen auf das Einzelwort. So sind auch die Wortfamilien derzeit nicht explizit als solche erkennbar. Insgesamt ist die Erfassung der Wortbildungsbeziehungen in *elexiko* unvollständig, was zum einen am noch geringen Bestand redaktionell bearbeiteter Stichwörter liegt, zum anderen an den Grenzen automatischer Verfahren (die Konversion *Bild* ist in den Wortbildungsprodukten zum Stichwort *bilden* beispielsweise nicht erfasst worden). Ein Ausbau der Angaben zur Gebildetheit der Stichwörter mithilfe automatisch ermittelter Informationen ist geplant,[10] kann dann allerdings nur bezogen auf ein Lemma, nicht bezogen auf einzelne Lesarten erfolgen. Der gleichen Beschränkung unterliegen die automatisch ermittelten Wortbildungsprodukte, die nicht nach Lesarten sortiert sind. Wünschenswert wäre auf jeden Fall eine Kombination automatischer und redaktioneller Verfahren, um Fehler korrigieren, Lücken füllen und vor allem – wo nötig – auf einzelne Lesarten bezogene oder nach Lesarten sortierte Angaben machen zu können.[11] Dies wäre auch deshalb wünschenswert, da nur so

[10] Vgl. Klosa (2011b, S. 155) und Klosa (Hg.) (2011, S. 18).

[11] Dies ist allerdings aufgrund personeller und organisatorischer Bedingungen im Projekt nicht zu leisten.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

die im folgenden Abschnitt beschriebenen Recherchemöglichkeiten in *elexiko* bzw. dem Wörterbuchportal OWID ihr volles Potenzial entfalten können.

3. Recherchemöglichkeiten in *elexiko* bzw. in OWID

Innerhalb von *elexiko* können Nutzer nicht nur einfach nach Stichwörtern suchen, sondern die sogenannte „Erweiterte Suche“ bietet die Möglichkeit, nach Gruppen von Stichwörtern zu suchen, die durch ein gemeinsames Merkmal oder eine Kombination verschiedener gemeinsamer Merkmale verbunden sind. Ein hier zur Verfügung stehendes Auswahlkriterium ist die Art der Wortgebildetheit der Stichwörter,[12] d.h., man kann etwa nach allen Stichwörtern suchen, die als Konversion bestimmt sind oder bei denen eine der Lesarten als Konversion analysiert wird (vgl. Abb. 4). Dieses Kriterium kann (wie in Abbildung 4 gezeigt) mit weiteren kombiniert werden, z.B. einem bestimmten Buchstabenbereich oder mit einer bestimmten Wortart.

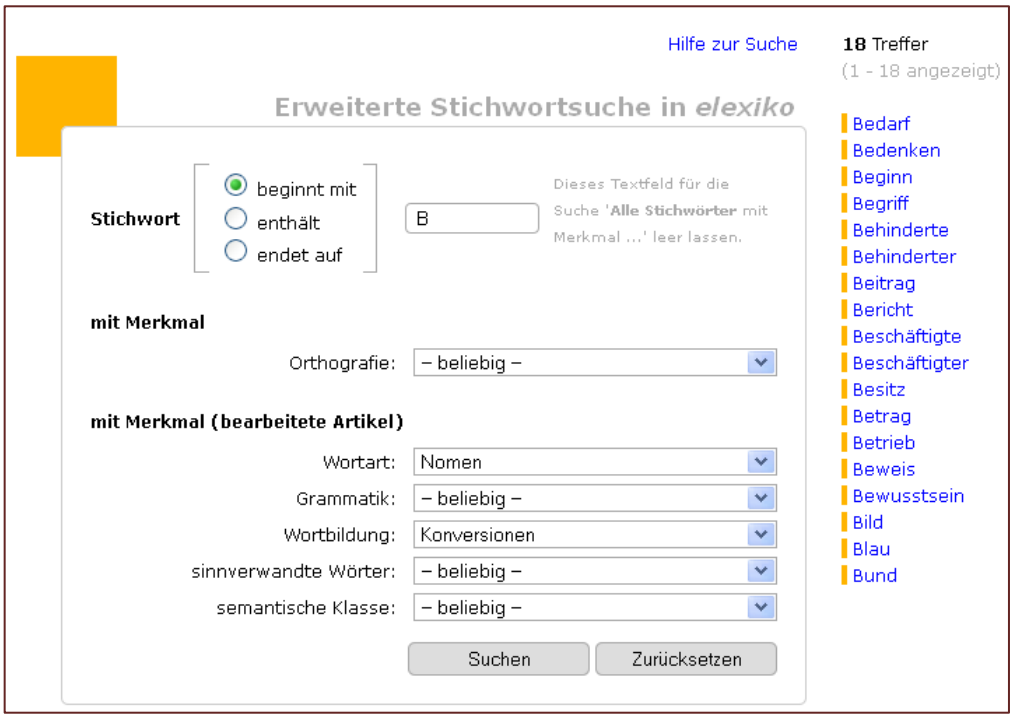


Abb. 4: Erweiterte Suche in *elexiko* nach allen Nomen mit dem Anfangsbuchstaben B, die Konversionen sind, und Anzeige der gefundenen Stichwörter in der rechten Bildspalte.

Die einfache Suche nach einem Stichwort bzw. nach einer Zeichenfolge im Wörterbuchportal OWID führt zu einer Ergebnispräsentation, die nicht nur die Stichworttreffer in verschiedenen Wörterbüchern innerhalb des Portals enthält, sondern daneben den Zugriff auf Stichwörter ermöglicht, die mit der gesuchten Zeichenfolge beginnen bzw. enden oder diese enthalten (vgl. Abb. 5).

[12] Soweit dies überhaupt relevant ist – Simplizia können über diese Suche natürlich nicht gefunden werden.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Abb. 5: Ergebnis für die Suche in OWID nach dem Stichwort *Bild* bzw. der Zeichenfolge *bild*

Über dieses Suchergebnis ist folglich der Zugriff auf Bildungen mit dem gesuchten Stichwort (z.B. *Bild*) als Bestimmungswort (z.B. *Bildgießerei*, *Bildgröße*, *Bildgrund*, *Bildgruppe*) bzw. als Grundwort (z.B. *Standbild*, *Stehbild*, *Steinbild*, *Stereobild*) in Komposita möglich, eingeschränkt auch der Zugriff auf weitere Bildungen (z. B. *bildhaft*, *ausbilden*, *Einbildung*, *Vorbild*). Es ist allerdings zu beachten, dass die Suche im derzeitigen Zustand keine morphologische, sondern tatsächlich eine rein zeichenbasierte Suche ist. Sie kann also nur eingeschränkt zur Suche nach Wortfamilien mit einem bestimmten Grundmorphem oder Kernlexem genutzt werden, da unter den Ergebnissen auch falsche Treffer enthalten sein können.

Die gezeigte Suche in OWID ermöglicht daneben, morphologische Varianten zu einem Stichwort zu finden. Allerdings zeigt sich auch hier, dass aufgrund der Tatsache, dass es sich nicht um eine morphologisch basierte Suche handelt, nicht alle Varianten gefunden werden können. Beispielsweise ergibt die Suche nach der Zeichenfolge bzw. dem Stichwort *dunkel* keine Hinweise auf die morphologische Varianten *dunkl* (vgl. Abb. 6). Sucht man nach der Zeichenfolge *dunkl*, lässt das Suchergebnis immerhin erkennen, dass dies in Bildungen eine morphologische Variante zu *dunkel* sein kann (vgl. Abb. 7).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

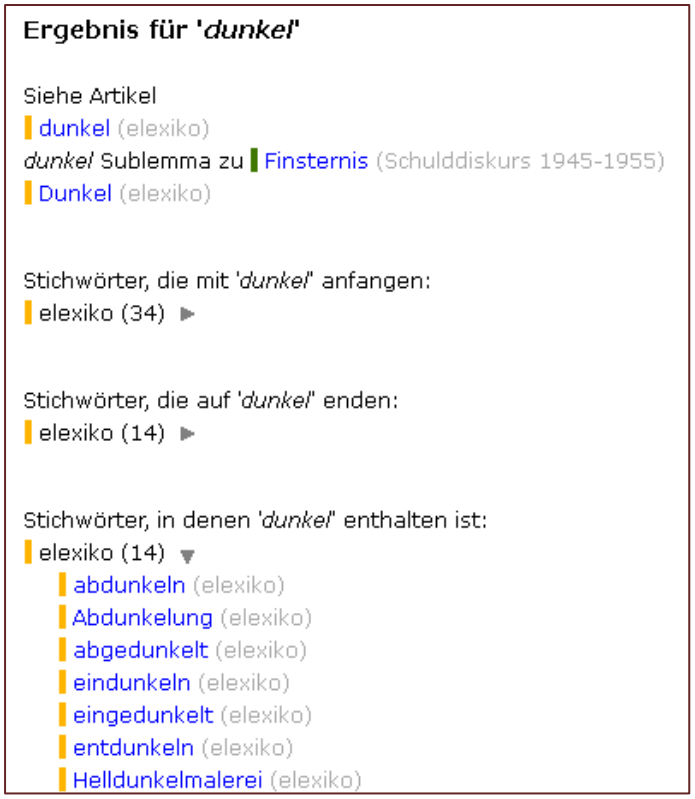


Abb. 6: Keine morphologische Varianten im OWID-Suchergebnis aller Stichwörter, die mit *dunkel* beginnen

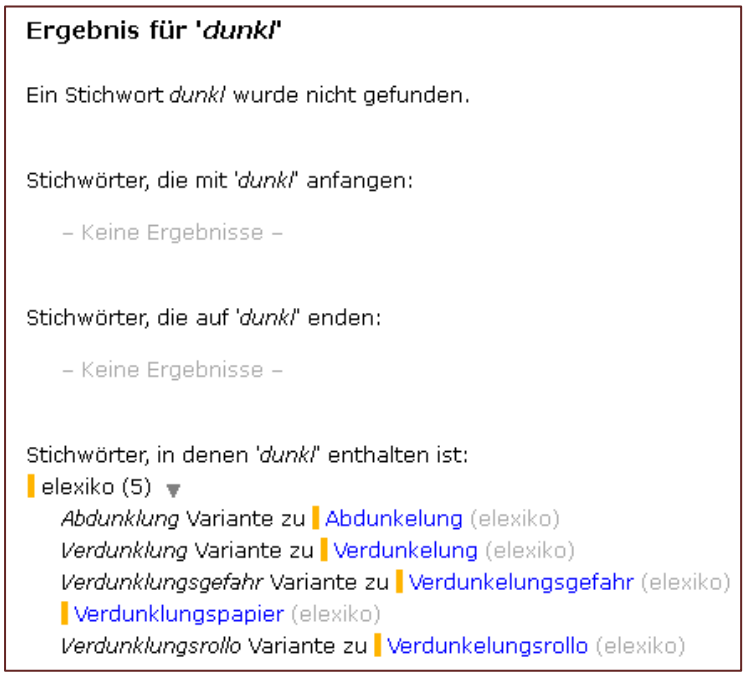


Abb. 7: Morphologische Varianten im OWID-Suchergebnis aller Stichwörter, die mit *dunkl* beginnen

Als Fazit zu den Recherchemöglichkeiten in *elexiko* bzw. OWID ist festzuhalten, dass es derzeit keine Suche gibt, welche alle Stichwörter, die zur gleichen Wortfamilie gehören, zuverlässig findet. Es gibt außerdem keine Möglichkeit, nach allen denjenigen Stichwörtern zu suchen, die als Kernlexem einer Wortfamilie zu bestimmen sind. Dies liegt allerdings auch daran, dass die entsprechenden Wortartikel keine lexikographische Angabe enthalten, die für

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

solch eine Suchanfrage genutzt werden könnte. Allerdings könnte zumindest eine Liste all derjenigen Stichwörter angeboten werden,[13] zu denen Wortbildungsprodukte vorliegen. Ob solch eine Liste neben dem formbasierten, semasiologischen Zugriff auf die Wortartikel in *elexiko* tatsächlich einen von den Motivationsbeziehungen ausgehenden, onomasiologischen Zugriff auf Wortartikel schaffen könnte, ist nicht eindeutig zu beurteilen. Jedenfalls wäre es dem elektronischen Medium angemessen und wohl im Sinne der Nutzer, wenn die Recherchemöglichkeiten in *elexiko* bzw. OWID weiter ausgebaut und nicht zuletzt auf morphologische Analysen basiert würden, damit beispielsweise auch morphologische Varianz bei den Suchanfragen berücksichtigt wird.

4. Meinungen von Wörterbuchbenutzern zu Wortbildungsangaben

Mit Fragen der Einschätzung und Bewertung bestimmter lexikographischer Inhalte und Gestaltungsmöglichkeiten in allgemeinsprachigen Onlinewörterbüchern durch Wörterbuchbenutzer haben sich einige Benutzungsstudien befasst, die am IDS im Rahmen des drittmittelfinanzierten Projektes „Benutzeradaptive Zugänge und Vernetzungen in *elexiko* (BZV*elexiko*)“[14] zwischen 2009 und 2011 durchgeführt wurden. Mithilfe von Onlinefragebögen, die befragende und experimentelle Elemente enthielten, wurden relativ große Probandengruppen (Sprachwissenschaftler, Studierende der Sprachwissenschaften, Übersetzer, Deutschlehrer, DaF-Lehrer und DaF-Lerner sowie Laien) erreicht. Zwei der Studien beschäftigten sich gezielt mit *elexiko* und hiermit vergleichbaren einsprachigen, auf die Beschreibung von Bedeutung und Verwendung konzentrierten Onlinewörterbüchern.[15] Untersucht wurde u.a. die Bewertung der Nützlichkeit einzelner lexikographischer Angabebereiche durch die Probanden, die erwarteten Informationen pro Angabebereich, der gewünschte Ausbau der Rechercheмöglichkeiten sowie die Bewertung unterschiedlicher Ansichten für lexikographische Angaben.

Im Ranking nach der Wichtigkeit der Angabebereiche in *elexiko* nimmt die Wortbildung einen relativ gesehen niedrigen Rang ein (vgl. Abb. 8), wobei nicht übersehen werden darf, dass alle im Onlinefragebogen überprüften Angabebereiche in der Bewertung ihrer Wichtigkeit sehr dicht beieinander liegen.[16]

[13] Solch eine Liste wäre z.B. vergleichbar mit der Liste aller illustrierten Wortartikel, die im Menüpunkt „Wortartikel“ in www.elexiko.de angeboten wird. Der Einsatz von Listen, über die bestimmte Gruppen von Wörtern erschlossen werden können, ist auch an anderer Stelle im Wörterbuchportal OWID erprobt (z.B. Liste der phraseologischen Neologismen im Neologismenwörterbuch unter dem Menüpunkt „Wortartikel“).

[14] Vgl. hierzu www.benutzungsforschung.de.

[15] Vgl. Klosa/Koplenig/Töpel (2011) sowie Klosa/Töpel/Koplenig (2012) zu einem Teil der Ergebnisse. Eine vollständige Dokumentation der Ergebnisse aller Studien im Projekt BZV*elexiko* liegt mit Müller-Spitzer (Hg.) (2014) vor.

[16] Vgl. hierzu ausführlicher Töpel (2013).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

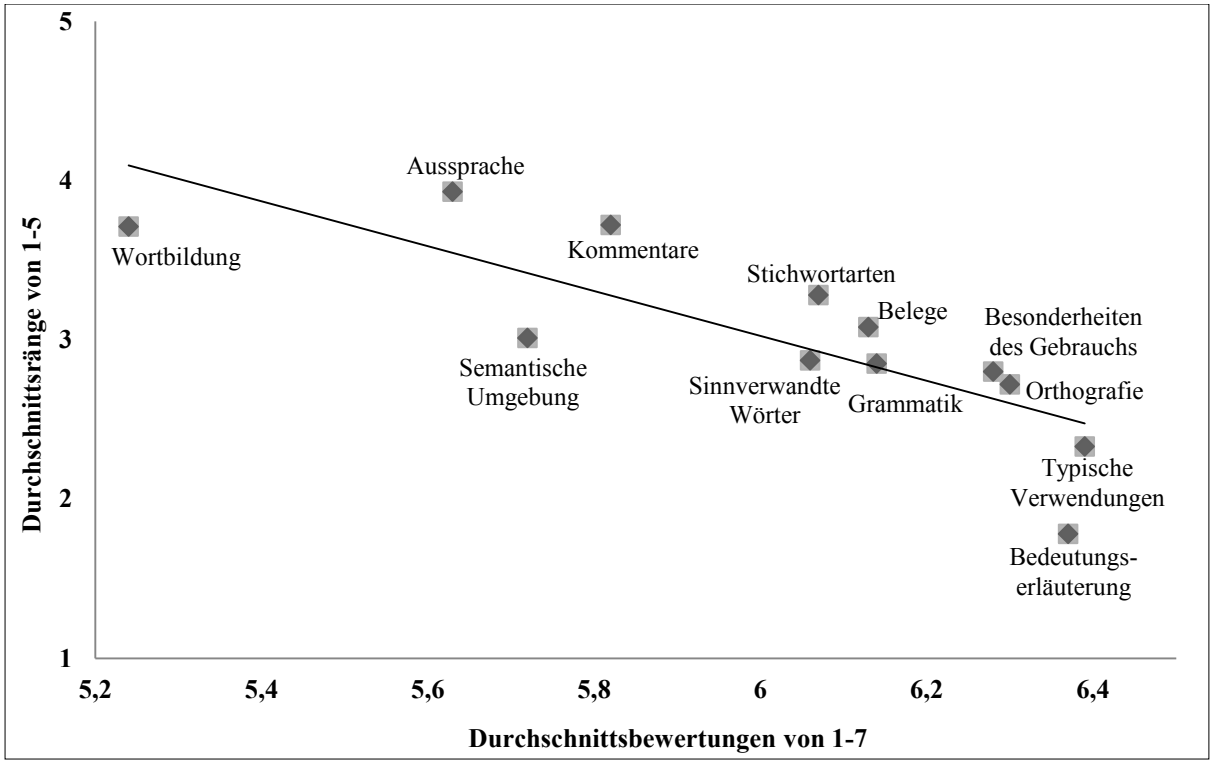


Abb. 8: Die Wichtigkeit der Angabebereiche in Onlinewörterbüchern wie *elexiko* (Bewertung und Rangfolge)

Bei der Frage danach, welche Angaben die Probanden im Angabebereich „Wortbildung“ als besonders nützlich erachten, kam die folgende Rangfolge zustande (Mehrfachnennungen waren möglich):

- Verlinkung zu den Bestandteilen eines Stichwortes bzw. zu den Wortbildungsprodukten (63,67% der Fälle),
- Erfassung und Beschreibung einzelner Wortbildungsmittel (62,95%),
- Information dazu, welche anderen Wörter zu einem Stichwort gebildet sind (62,59%),
- Analyse der Gebildetheit des Stichwortes (60,43%),
- Grafik, die das Stichwort mit allen zugehörigen Bildungen (z.B. in Form eines Wortnetzes) zeigt (38,49%).

Eine grafische Darstellung von Wortbildungszusammenhängen (bzw. Wortfamilien) wird also als weniger nützlich bewertet als die Angaben zur Gebildetheit des Stichwortes, zu Wortbildungsprodukten und Wortbildungsmitteln. Den Nutzen der Verlinkung zwischen Kernlexemen und Wortbildungsprodukten erkennen die Probanden klar. In der Befragung wurde nicht überprüft, für wie nützlich die Probanden die Möglichkeit, über Wortfamilien auf einzelne Stichwörter zugreifen zu können, eingeschätzt hätten. Allerdings wurde generell ein Ausbau der Recherchemöglichkeiten positiv bewertet.[17]

[17] Ein Ergebnis der Studie war, dass mehr erweiterte Suchmöglichkeiten in einem Onlinewörterbuch wie *elexiko* für die an der Umfrage Teilnehmenden wichtig und wünschenswert sind: Auf einer siebenstufigen Skala von *überhaupt nicht wichtig/wünschenswert* bis *sehr wichtig/wünschenswert* lag der Mittelwert der Antworten bei 5,86 (vgl. Klosa/Koplenig/Töpel 2014, S. 347).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Ob Wörterbuchbenutzer tatsächlich, wie Splett (2013) meint, an der Verdeutlichung der Wortschatzstrukturen in einem elektronischen Wortfamilienwörterbuch interessiert sind, müsste in weiteren Benutzungsstudien untersucht werden. Da Onlinewörterbücher bei entsprechender Modellierung der Datenstruktur und Erfassung der Angaben grundsätzlich nicht auf eine Darstellungsweise beschränkt sind, könnte hierin auch untersucht werden, wie und in welchen Benutzungssituationen die Wörterbuchbenutzer mit einem online (zumindest bei den Zugriffsmöglichkeiten kombinierten) Angebot aus (klassischem) semasiologischen Bedeutungswörterbuch und (neuerem) Wortfamilienangebot umgehen würden.[18]

5. Ausblick

Voraussetzung für eine solche Erweiterung der Recherchemöglichkeiten in *elexiko* (wie in vergleichbaren Onlinewörterbüchern) ist zunächst die vollständige und korrekte morphologische Analyse aller Stichwörter. Diese muss immer dann, wo es inhaltlich nötig ist, auf der Ebene der Lesarten erfolgen (dies ist in *elexiko* derzeit nur bei der Analyse der Gebildetheit von redaktionell bearbeiteten Stichwörtern der Fall, nicht aber bei der Angabe der Wortbildungsprodukte). Außerdem ist wünschenswert, die erweiterten Suchmöglichkeiten nicht nur um inhaltliche Kriterien auszubauen, sondern auch die Abfrage intuitiver zu gestalten sowie die Ergebnisdarstellung zu verbessern.

Hierzu gibt es Vorarbeiten im Kontext des Wörterbuchportals OWID. In der interaktiven[19] Suche nach Stichwörtern, die auf bestimmte Art und Weise gebildet sind (vgl. Abb. 9), wird versucht, sich von einem klassischen Suchformular, in dem Kriterien aus Drop-down-Menüs ausgewählt werden müssen (siehe oben Abb. 4), zu lösen.

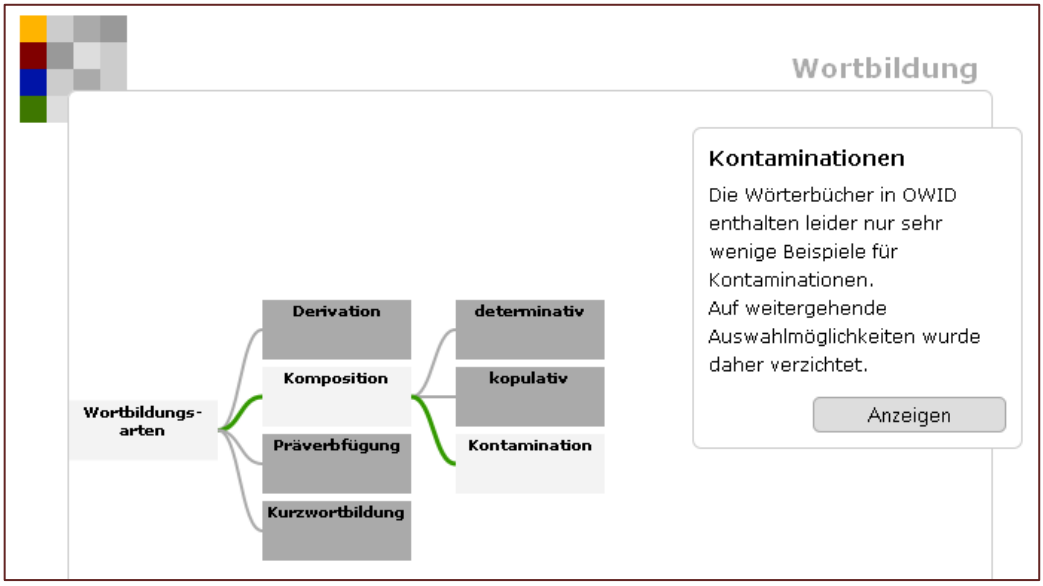


Abb. 9: Interaktive Suche nach allen Stichwörtern, die nach einer bestimmten Wortbildungsart gebildet wurden, aus OWID-intern

[18] Vgl. hierzu auch De Caluwe (2013, S. 115).

[19] Die Interaktivität der Suche kann auf Papier nicht gezeigt werden. In der Anwendung selbst erscheinen beim Klicken auf eines der Felder die jeweils nächsten Felder (beim Klicken auf das Feld „Wortbildungsarten“ öffnen sich beispielsweise die Felder „Derivation“, „Komposition“ usw., beim Klicken auf das Feld „Komposition“ öffnen sich die Felder „determinativ“, „kopulativ“ usw.).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Ebenso lässt sich die Darstellung der Suchergebnisse verbessern, indem nicht einfach Listen mit den gefundenen Stichwörtern gezeigt werden (siehe oben Abb. 4), sondern die Stichwörter direkt mit ihren Bildungsbestandteilen präsentiert werden (vgl. Abb. 10).[20]

Kontamination	Bestandteil		Bestandteil	
	Form	Typ	Form	Typ
Arabellion	arabisch	Adjektiv	Rebellion	Nomen
Besserwessi	Besserwisser	Nomen	Wessi	Nomen
denglisch	deutsch	Adjektiv	englisch	Adjektiv
Mechatronik	Mechanik	Nomen	Elektronik	Nomen
Naturlaub	Natur	Nomen	Urlaub	Nomen
Ostalgie	Osten	Nomen	Nostalgie	Nomen
Sidebag	side ('Seite')	engl.	Airbag	Nomen
Telematik	Telekommunikation	Nomen	Informatik	Nomen
Westalgie	Westen	Nomen	Nostalgie	Nomen
Wossi	Wessi	Nomen	Ossi	Nomen

Abb. 10: Verbesserte Darstellung der Suchergebnisse nach allen Stichwörtern in OWID-Wörterbüchern, die durch Kontamination entstanden sind

Eine interaktive Erkundung von Wortbildungsbestandteilen und -produkten ermöglicht daneben ein im Projekt BZV*elexiko* entwickeltes Visualisierungsverfahren (vgl. Meyer/Müller-Spitzer 2013, S. 270-277), das Wortbildungsbeziehungen (denen man in *elexiko* derzeit nur durch Folgen der Hyperlinks auf die Spur kommen kann, vgl. Abschnitt 2) in einem mithilfe von farbigen Kästchen sehr übersichtlich gestalteten und mithilfe der −/+-Symbole veränderbaren Baumgraphen darstellt (vgl. Abb. 11).

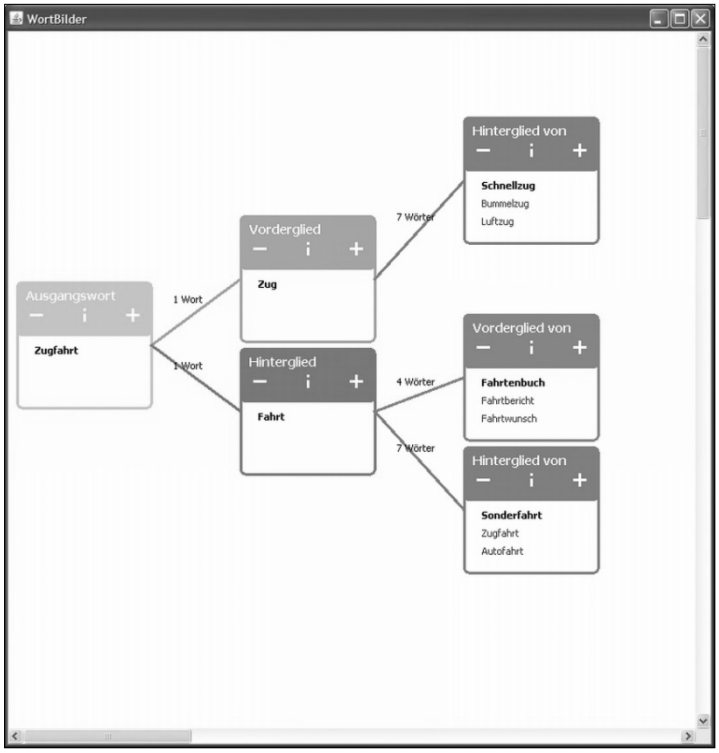


Abb. 11: Graphendarstellung von Wortbildungsbeziehungen am Beispiel *Zugfahrt* nach Meyer/Müller-Spitzer (2013, S. 273)

[20] Die gezeigte Suchmöglichkeit sowie die verbesserte Ergebnisdarstellung wurden von Frank Michaelis und Carolin Müller-Spitzer entwickelt.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

(Statische) Baumgraphen werden auch im Onlinewörterbuch Canoo.net genutzt, um Wortfamilien darzustellen (vgl. Abb. 12). Anders als bei dem im Projekt BZV*elexiko* entwickelten Tool kann die Darstellung hier bei umfangreichen Wortfamilien allerdings sehr unübersichtlich werden, weil sich die Nutzer durch die zum Teil vertikal stark aufgefächerte Darstellung scrollen müssen.[21]

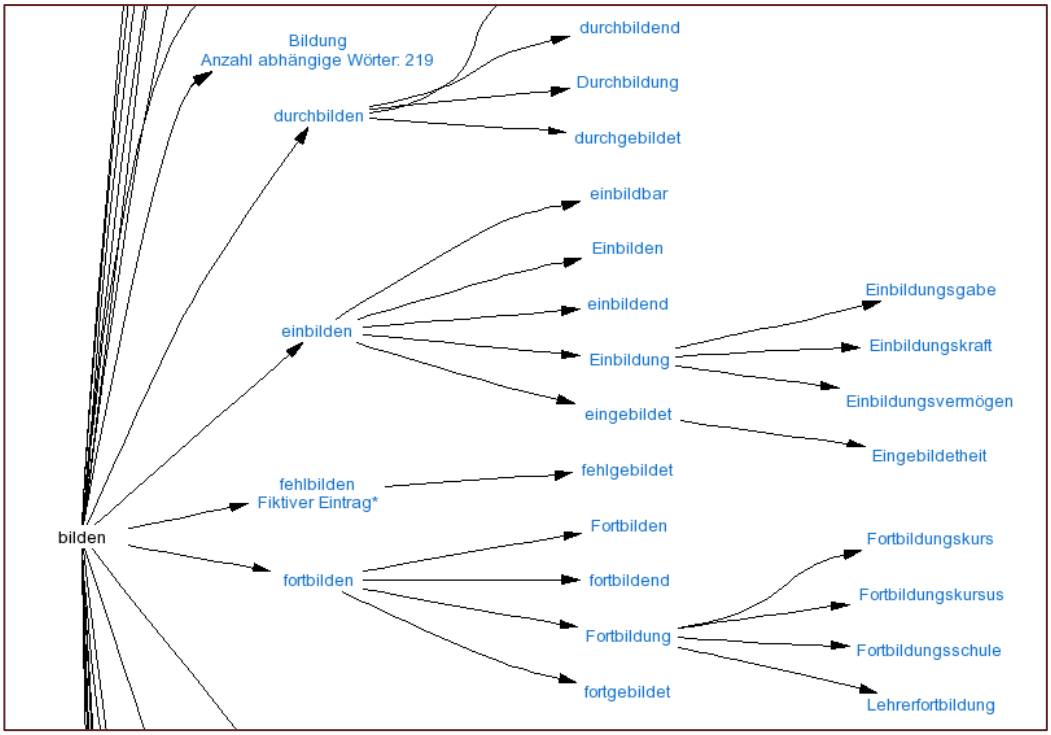


Abb. 12: Ausschnitt aus der Darstellung der Wortfamilie zu *bilden* in Canoo.net

Zur Visualisierung von Wortbildungsbeziehungen könnten schließlich netzartige Graphen genutzt werden, wie sie im Projekt *elexiko* (intern) zur Visualisierung von paradigmatischen Relationen eingesetzt werden.[22] Abbildung 13 zeigt die Synonyme zum Stichwort *Bild* in Verkleinerung, um zu illustrieren, dass solche netzartigen Graphen auch bei einer Fülle von Angaben übersichtlich gestaltet werden können.

[21] Zu Problemen mit dieser Visualisierungsmöglichkeit vgl. auch Meyer/Müller-Spitzer (2013, S. 262-267).

[22] Das entsprechende Tool, das auch der Verwaltung der Vernetzungen zwischen *elexiko*-Wortartikeln dient, wurde von Peter Meyer entwickelt (vgl. Meyer/Müller-Spitzer 2010). Die Graphen können in diesem Tool mithilfe verschiedener Auswahlmöglichkeiten je nach Bedarf sehr differenziert (z.B. nach Tiefe des Graphen oder Art der Visualisierung [organisch – radial]) gestaltet werden.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

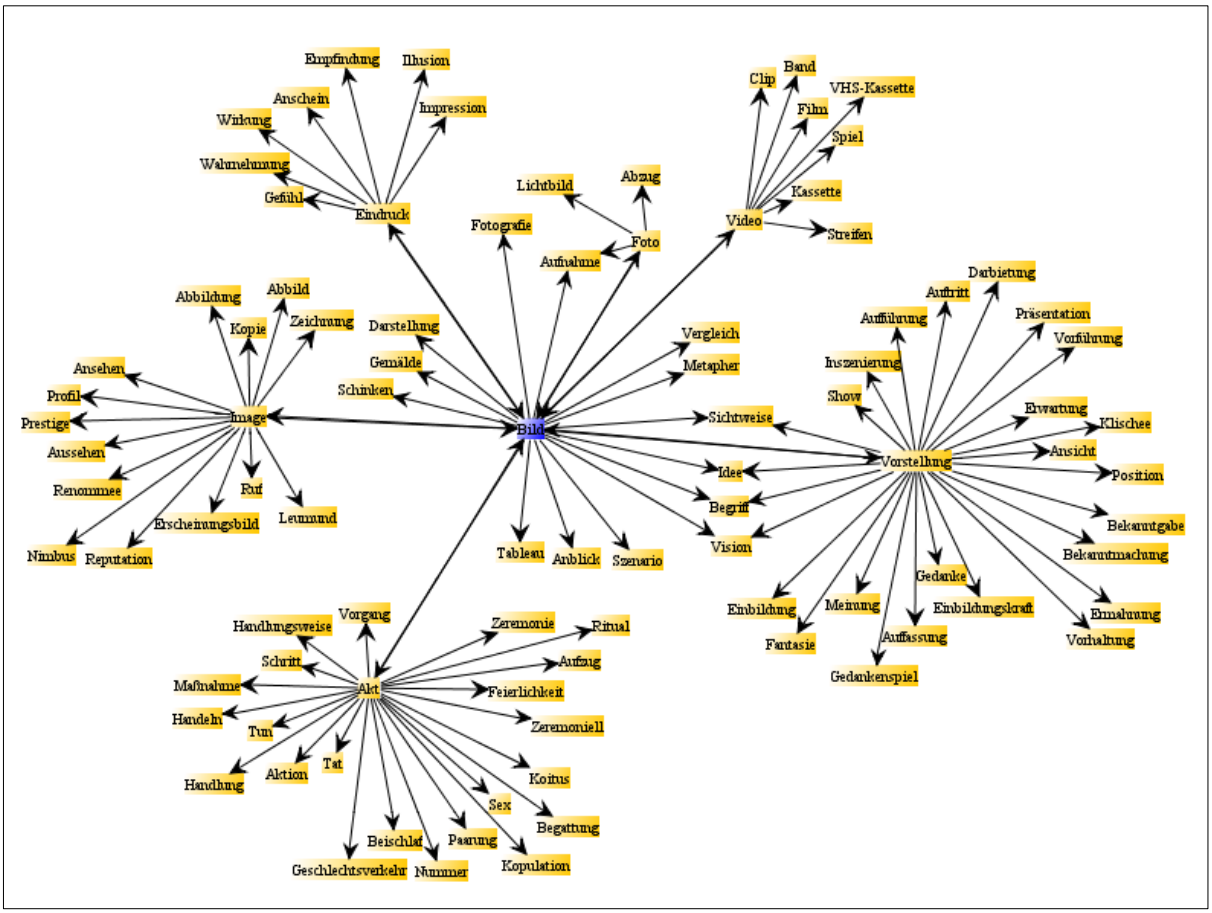


Abb. 13 Netzartige Darstellung von Synonymen zum Stichwort *Bild*, die für die Visualisierung von Wortbildungszusammenhängen adaptiert werden könnte

Eine Herausforderung für die Sichtbarmachung von Wortfamilien in Onlinewörterbüchern wird zukünftig sein, die gezeigten Visualisierungsmöglichkeiten zu adaptieren und ggf. weiterzuentwickeln. Zu prüfen ist außerdem, ob solche Visualisierungen zum Zugriff auf die Wortfamilien bzw. zugehörige Stichwörter genutzt werden können. Wörterbuchbenutzer würden dann nicht mehr nur über Eingabe von Suchwörtern oder über erweiterte Suchen, sondern auch über Klicken auf in solchen Graphen gezeigte Stichwörter bzw. Stichwortgruppen zu diesen gelangen. Möglicherweise ließe sich durch diese Verdeutlichung der Wortschatzstrukturen die Akzeptanz von Wortbildungsangaben in allgemeinsprachigen Onlinewörterbüchern (wie *elexiko*) verbessern.

6. Elektronische Ressourcen

Canoo.net. www.canoo.net (Stand: 4.4.2013).

elexiko (2003ff.): In: OWID (2008ff.). www.elexiko.de (Stand: 3.4.2013).

OWID: Online-Wortschatz-Informationssystem Deutsch (2008ff.), hrsg. v. Institut für Deutsche Sprache, Mannheim. www.owid.de (Stand: 4.4.2013).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

7. Literatur

Augst, Gerhard (1998): Wortfamilienwörterbuch der deutschen Gegenwartssprache. Tübingen.

De Caluwe, Johan (2013): Dictionary entries as windows on the onomasiological aspects of word formation. In: Klosa (Hg.), S. 105-116.

Eichinger, Ludwig M. (2013): Wortbildung im Wörterbuch. Aus der Sicht eines Grammatikers. In: Klosa (Hg.), S. 63-86.

ten Hacken, Pius (2013): Wortbildung in elektronischen Lernerwörterbüchern. In: Klosa (Hg.), S. 157-174.

Haß, Ulrike (Hg.) (2005): Grundfragen der elektronischen Lexikografie. *elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York.

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4., völlig neu bearb. Aufl. Berlin/Boston.

Klosa, Annette (2005): Wortbildung. In: Haß (Hg.), S. 141-162.

Klosa, Annette (2011a): *elexiko* – ein Bedeutungswörterbuch zwischen Tradition und Fortschritt. In: Sprachwissenschaft 36, 2/3, S. 275-306.

Klosa, Annette (2011b): Korpusgestützte Angaben zu Grammatik und Wortbildung. In: Klosa (Hg.), S. 145-156.

Klosa, Annette (Hg.) (2011): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen.

Klosa, Annette (Hg.) (2013): Wortbildung im elektronischen Wörterbuch. Tübingen.

Klosa, Annette/Koplenig, Alexander/Töpel, Antje (2011): Benutzerwünsche und Meinungen zu einer optimierten Wörterbuchpräsentation. Ergebnisse einer Onlinebefragung zu *elexiko*. (= OPAL 3/2011). Mannheim. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2011-3.pdf (Stand: 3.4.2013).

Klosa, Annette/Koplenig, Alexander/Töpel, Antje (2014): Benutzerwünsche und Benutzermeinungen zu dem monolingualen deutschen Onlinewörterbuch *elexiko*. In: Müller-Spitzer (Hg.), S. 281-384.

Klosa, Annette/Töpel, Antje/Koplenig, Alexander (2012): Zur Funktion und Rezeption von Belegen. Ergebnisse einer Benutzungsstudie zum Onlinewörterbuch *elexiko*. In: Sprachwissenschaft 37, 1, S. 93-123.

Meyer, Peter/Müller-Spitzer, Carolin (2010): Consistency of sense relations in a lexicographic context. In: Barbu Mititelu, Verginica/Pekar, Viktor/Barbu, Eduard (Hg.): Proceedings of the workshop „Semantic Relations. Theory and Applications“, 18th May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta. www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf (Stand: 3.4.2013).

Meyer, Peter/Müller-Spitzer, Carolin (2013): Überlegungen zur Visualisierung von Wortbildung in elektronischen Wörterbüchern. In: Klosa (Hg.), S. 255-279.

Müller-Spitzer, Carolin (2011): Der Aufbau einer maßgeschneiderten XML-basierten Modellierung für ein Wörterbuchnetz. In: Klosa, Annette/Müller-Spitzer, Carolin (Hg): Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. Mannheim, S. 37-51. http://pub.ids-mannheim.de/laufend/opal/opal11-2.html (Stand: 3.4.2013).

Müller-Spitzer, Carolin (Hg.) (2014): Using online dictionaries. (= Lexicographica. Series Maior 145). Berlin/New York.

Simon, Christian (2013): Finite-State-basierte Morphologie-Tools und ihre Stärken und Schwächen bei der maschinellen Wortbildungsanalyse. In: Klosa (Hg.), S. 217-233.

Splett, Jochen (2013): Grundlegende Bemerkungen zu einem auf einer pragmatischen Sprachtheorie fußenden Wortfamilienwörterbuch als legitimem Ort einer integrierten Wortbildung. In: Klosa (Hg.), S. 117-129.

Töpel, Antje (2013): Die Wortbildungsangaben im Online-Wörterbuch und wie Nutzer sie beurteilen – eine Umfrage zu *elexiko*. In: Klosa (Hg.), S. 197-214.

Ulsamer, Sabina (2013a): Chancen und Probleme bei der automatischen Ermittlung von Wortbildungsprodukten für *elexiko* und bei ihrer Präsentation. In: Klosa (Hg.), S. 235-254.

Ulsamer, Sabina (2013b): Wortbildung in Wörterbüchern – Zwischen Anspruch und Wirklichkeit. In: Klosa (Hg.), S. 13-59.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Korpusbasierte Wortfamilien-Lemmalisten und ihre TEI-Kodierung

Heike Stadler (*Universität Hildesheim*), *Werner Wegstein* (*Universität Würzburg*)

Im Rahmen des Forschungsverbunds „Wechselwirkungen“ dienen korpusgenerierte, frequenzsortierte Lemmalisten als Basislemmalisten für den Wortschatz der schriftlichen deutschen Gegenwartssprache. Sie dienen zugleich als Referenzdaten für die Zuordnung von Lemmata aus zeitlich und regional markierten Wörterbüchern und ihre Verknüpfung. Die Listen enthalten für jedes Lemma Angaben zur Wortklasse (part-of-speech, POS) und zur Frequenz. Die Lemmalisten wurden automatisch aus dem mehrfach linguistisch annotierten Deutschen Referenzkorpus DeReKo (http://www1.ids-mannheim.de/kl/projekte/korpora/) des Instituts für Deutsche Sprache erstellt. Zum Zeitpunkt der Erstellung bestand DeReKo aus circa fünf Milliarden Wortformen, die sich überwiegend auf Zeitungstexte und zu einem geringen Anteil auf literarische Texte und Gebrauchstexte verteilten. Die automatische Generierung von Lemmalisten aus Korpora setzt eine linguistische Annotation der Wortformen im Korpus voraus, die Zuordnung einer das Lexem kennzeichnenden Grundform (Lemmatisierung) und Wortart (POS-Tag). Für die Zusammenstellung der Wortfamilien-Lemmalisten wurden die Lemma- und POS-Annotationen des Lemmatisierers *glemm* (Belica 1994) und der POS-Tagger *Machinese Phrase Tagger* (www.connexor.com/nlplib/?q=demo/mpt) und *TreeTagger* (Schmid 1994, 1995) verwendet.[1]

Jochen Splett hat sein „Deutsches Wortfamilienwörterbuch“ (Splett 2009) aus rund 160.000 Stichwörtern des achtbändigen DUDEN-Wörterbuchs (Duden 1993-1995) erarbeitet, mit dem Ziel der – wie der Untertitel erläutert – „Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes“. „Das Wörterbuch besteht aus zwei Teilen. Der erste umfasst den Teil des zugrunde gelegten Wortschatzes, der zu Wortfamilien zusammengestellt werden kann [...] 8264 Wortfamilien [...] alphabetisch nach dem Kernwort angeordnet [...] Der zweite Teil „Einzeleinträge“ besteht aus den alphabetisch angeordneten Einzelwörtern, die keiner Wortfamilie eingegliedert werden können“ (Splett 2009, Bd. 1, S. XIX). Anhang 1 bietet einen Ausschnitt aus dem Wortartikel „Gabel“ von Splett, der insgesamt 54 Einträge zum Wortfamilienkern „GABEL“ nachweist. Die Konzeption des Wörterbuchs und sein Artikelaufbau werden in der Einleitung (S. XI-XXX) ausführlich erläutert unter Rückgriff auf Spletts „Althochdeutsches Wörterbuch“ (Splett 1993) und weitere Vorarbeiten.[2]

Wir haben mit diesem Beitrag den Versuch unternommen, zu prüfen, inwieweit Spletts am Althochdeutschen und an einem Wörterbuch zur deutschen Gegenwartssprache erprobte Beschreibungskonzepte für Wortschatzstrukturen auch auf korpusgenerierte Lemmalisten, kodiert nach TEI-Richtlinien, anwendbar sind und ob Spletts Arbeiten für eine Strukturgeschichte des deutschen Wortschatzes auch in eine nachhaltige XML-Struktur auf der Grundlage der TEI-Richtlinien und internationaler Standards konvertierbar sind. Wir verwenden dazu Daten zum Wortfamilienkern „GABEL“.

[1] Die Eigenschaften der aus den linguistischen Annotationen der einzelnen Lemmatisierer und POS-Tagger generierten separaten Lemmalisten werden in Stadler (2014) dargestellt.

[2] Splett (1993, S. XI-XIII); ferner Hundsnurscher (1985, 1987); Herbermann (1981a, 1981b).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Aus dem Deutschen Referenzkorpus können ca. 1.300 lexikographisch validierte Lemmata extrahiert werden, die das Morphem *gabel* oder eines der daraus abgeleiteten Morpheme *gabl* oder *gäbel* enthalten.[3] Mit regulären Ausdrücken und POS-Spezifikationen können alle lexikographisch validierten Lemmata aus DEREKO automatisch den Kategorien der Klassifikation von Splett zur Strukturierung der Einträge einer Wortfamilie zugeordnet werden. Die meisten Lemmata gehören zur Kategorie „1. Simplex bzw. Flexionstyp", ein kleiner Teil wird durch Derivation gebildet und befindet sich in der Kategorie „2. Suffixbildung". Beide Kategorien werden durch POS- und Flexions-Merkmale feiner unterteilt. In Tabelle 1 wird die prozentuale Verteilung der Lemmata auf die einzelnen Kategorien gezeigt. Die Kategorien, für die in DEREKO kein Lemma vertreten ist, sind auch im Wortartikel „GABEL" von Spletts „Deutschem Wortfamilienwörterbuch" unbesetzt.

	1	Simplex bzw. Flexionstyp		2	Suffixbildung	
starkes Verb	1.0	–		-	–	
schwaches Verb	1.1	1,2 %	aufgabeln	2.1	–	
Nomen	1.2	86,8 %	Gabelflug	2.2	6,5%	Waldweggabelung
Adjektiv	1.3	4,6%	gabelzart	2.3	0,8%	achtgabelig
Adverb (u. übrige Wortarten)	1.4	0,1%	ausgabelnd	2.4	–	

Tab. 1: Prozentuale Verteilung der Lemmata der Wortfamilien-Lemmaliste auf die einzelnen Kategorien der Wortfamilie *gabel*

Die korpusbasierte Wortfamilien-Lemmaliste bildet die Lemmata auf Spletts Wortfamilienstruktur ab: „target" verweist auf die betreffende Wortfamilie im Wortfamilienwörterbuch, die beiden Endziffern von „target" beziehen sich auf die in Tab. 1 dargestellte Gliederung der Wortfamilien. Als „item-id" fungiert die laufende Nummer, „lemma" enthält das Stichwort, „pos" steht für die Wortklasse (part-of-speech) mit den Kategorien NN für Substantiv, V für Verb, ADJ für Adjektiv und ADV für Adverb. Der Wert „num" enthält die Häufigkeitsklasse. Häufigkeitsklassen sind ein von Korpus- und Annotationsspezifika weitgehend unabhängiges Assoziationsmaß, dessen Werte über das Verhältnis des untersuchten Lemmas zum häufigsten Lemma des Korpus bestimmt werden (vgl. DEREWO – Deutsche Referenzwortlisten). Zu Beginn der Lemmaliste stehen die frequenten Lemmata, denen niedrige Häufigkeitsklassen zugeordnet sind. Innerhalb einer Häufigkeitsklasse ist die Wortfamilien-Lemmaliste absteigend nach den absoluten Frequenzen und innerhalb einer absoluten Frequenz absteigend alphabetisch sortiert.

[3] In der unbereinigten Wortfamilien-Lemmaliste, die alle Lemmata enthält, die mit den regulären Ausdrücken 'G|gabel', 'G|gabl' or 'G|gäbel' extrahiert werden, befinden sich mehrere hundert Lemmata, in denen sich die Buchstabenfolgen auf mehrere Morpheme verteilen (*<Eingabe><leiste>, <Mega><block>*), oder die Eigennamen sind (*Gabelentz*). Diese Lemmata werden automatisch über POS-Tags und Listen sowie per Hand ausgefiltert.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

```
Header: type="lemma list" target="#jsdw-GABEL"
item id="1" lemma="Gabel" pos="NN" num="15" target="jsdw-GABEL-1.2"
item id="2" lemma="Gabelstapler" pos="NN" num="17" target="jsdw-GABEL-1.2"
item id="3" lemma="gabeln" pos="V" num="17" target="jsdw-GABEL-1.1"
item id="4" lemma="Stimmgabel" pos="NN" num="18" target="jsdw-GABEL-1.2"
item id="5" lemma="Mistgabel" pos="NN" num="18" target="jsdw-GABEL-1.2"
item id="6" lemma="Gabelung" pos="NN" num="19" target="jsdw-GABEL-2.2"
item id="7" lemma="Weggabelung" pos="NN" num="19" target="jsdw-GABEL-2.2"
item id="8" lemma="Heugabel" pos="NN" num="19" target="jsdw-GABEL-1.2"
item id="9" lemma="aufgabeln" pos="V" num="19" target="jsdw-GABEL-1.1"
item id="10" lemma="Astgabel" pos="NN" num="19" target="jsdw-GABEL-1.2"
item id="11" lemma="Gabelstaplerfahrer" pos="NN" num="20" target="jsdw-GABEL-1.2"
item id="12" lemma="Federgabel" pos="NN" num="20" target="jsdw-GABEL-1.2"
item id="13" lemma="gegabelt" pos="ADJ" num="20" target="jsdw-GABEL-1.3"
item id="14" lemma="Telegabel" pos="NN" num="21" target="jsdw-GABEL-1.2"
item id="15" lemma="Teleskopgabel" pos="NN" num="21" target="jsdw-GABEL-1.2"
item id="16" lemma="Gabelfrühstück" pos="NN" num="21" target="jsdw-GABEL-1.2"
item id="17" lemma="Gabelbock" pos="NN" num="21" target="jsdw-GABEL-1.2"
item id="18" lemma="Kuchengabel" pos="NN" num="21" target="jsdw-GABEL-1.2"
item id="19" lemma="Straßengabelung" pos="NN" num="21" target="jsdw-GABEL-2.2"
item id="20" lemma="gabelig" pos="ADJ" num="22" target="jsdw-GABEL-2.3"
item id="21" lemma="Vorderradgabel" pos="NN" num="22" target="jsdw-GABEL-1.2"
item id="22" lemma="gabelförmig" pos="ADJ" num="22" target="jsdw-GABEL-1.3"
item id="23" lemma="Marschgabel" pos="NN" num="22" target="jsdw-GABEL-1.2"
...
item id="1315" lemma="Zündgabel" pos="NN" num="29" target="jsdw-GABEL-1.2"
item id="1316" lemma="Zungengabel" pos="NN" num="29" target="jsdw-GABEL-1.2"
item id="1317" lemma="Zweigabeltechnik" pos="NN" num="29" target="jsdw-GABEL-1.2"
item id="1318" lemma="zweigegabelt" pos="ADJ" num="29" target="jsdw-GABEL-1.3"
item id="1319" lemma="Zweizackgabel" pos="NN" num="29" target="jsdw-GABEL-1.2"
item id="1320" lemma="Zwischengabelung" pos="NN" num="29" target="jsdw-GABEL-2.2"
```

Abb. 1: Auszug von Anfang und Ende der frequenzsortierten Lemmaliste der Wortfamilie *gabel* (aus DEREWO)

Für die weitere Bearbeitung der korpusbasierten Häufigkeitsdaten bietet sich an, die Liste in eine TEI-konforme Elementstruktur zu transferieren, die es erlaubt, die Korrektheit des Strukturgerüsts zu überprüfen und mit etwas mehr Aufwand auch die Attributwerte zu kontrollieren und teils auch die Datenkonsistenz. An einem kurzen Auszug von Abbildung 1 wollen wir zeigen, wie eine solche TEI-konforme Wortfamilien-Lemmaliste aussehen könnte.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="tei_all.rng" type="application/xml"
schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="de">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>Encoding a DeReWo word-list
                            with items of the word-family "GABEL"</title>
                <author>Heike Stadler, Werner Wegstein</author>
            </titleStmt>
            <publicationStmt>
                <p>for test purposes only</p>
            </publicationStmt>
            <sourceDesc>
                <biblStruct xml:lang="de" type="word-list">
                    <monogr><title>DeReWo</title>
                        <imprint>
                            <pubPlace>Mannheim</pubPlace>
                            <publisher>Institut für deutsche Sprache</publisher>
                            <date>2013</date>
                        </imprint>
                    </monogr>
                </biblStruct>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
```

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

```
<text xml:lang="de">
  <body>
   <div>
    <list>
            <head>Auszug aus der korpusgenerierten und frequenzsortierten
                  DeReWo-Lemmaliste zur Wortfamilie 'GABEL'</head>

     <item xml:id="derewo-l-0001-gabel" ana="#jsdw-GABEL">Gabel
            <abbr type="pos" >NN</abbr><num type="jsdw-rank">1.2</num>
            <measure type="freqGroup"><num type="cardinal">15</num></measure></item>
     <item xml:id="derewo-l-0002-gabelstapler" ana="#jsdw-GABEL">Gabelstapler
            <abbr type="pos">NN</abbr><num type="jsdw-rank">1.2</num>
            <measure type="freqGroup"><num type="cardinal">15</num></measure></item>
     <item xml:id="derewo-l-0003-gabeln" ana="#jsdw-GABEL">gabeln
            <abbr type="pos">V</abbr><num type="jsdw-rank">1.1</num>
            <measure type="freqGroup"><num type="cardinal">17</num></measure></item>
     <item xml:id="derewo-l-0004-stimmgabel" ana="#jsdw-GABEL">Stimmgabel
            <abbr type="pos">NN</abbr><num type="jsdw-rank">1.2</num>
            <measure type="freqGroup"><num type="cardinal">18</num></measure></item>
     <item xml:id="derewo-l-0005-mistgabel" ana="#jsdw-GABEL">Mistgabel
            <abbr type="pos" >NN</abbr><num type="jsdw-rank">1.2</num>
            <measure type="freqGroup"><num type="cardinal">18</num></measure></item>
                                                              <!-- ...  -->
     <item xml:id="derewo-l-1320-zwischengabelung" ana="#jsdw-GABEL">Zwischengabelung
            <abbr type="pos" >NN</abbr><num type="jsdw-rank">2.2</num>
            <measure type="freqGroup"><num type="cardinal">29</num></measure></item>
    </list>
   </div>
  </body>
</text>
</TEI>
```

Abb. 2: Beispieleinträge für Lemmata mit einer TEI-konformen Elementstruktur

Ohne das Transkript im Detail zu erläutern, sei darauf hingewiesen, dass durch die Beschränkung der Kodierung auf eine Listenstruktur mit den TEI-Elementen `<list>` und `<item>` aus dem TEI-Modul „core“ (‘Elements common to all TEI documents’) Elemente ausgeschlossen sind, die zu anderen Textkategorien gehören, z.B. zu dem TEI-Modul „dictionaries“ und die hier auch nicht zur Anwendung kommen sollten, da es sich bei einer Wortliste eben nicht um ein Wörterbuch handelt. Das Attribut `xml:id` sorgt dafür, dass der Datensatz nur einmal mit dieser Kennzeichnung in der Wortliste auftauchen darf, das Attribut `ana` verknüpft den Listeneintrag mit der Wortfamilie „Gabel“ in Jochen Spletts „Deutschem Wortfamilienwörterbuch“, das Attribut `type="jsdw-rank"` zum Element `<num>` kennzeichnet Spletts Ordnungsbezifferung für Wortfamilieneinträge und das Attribut `type="freqGroup"` von `<measure>` markiert die Zahl in dem folgenden `<num>`-Element als Angabe der Häufigkeitsklasse.

Die Fülle von mehr oder weniger Text in und zwischen spitzen Klammern und das auch noch in verschiedenen Farben, die sich die Programmierer des XML-Editors „oxygen“, den wir hierfür benutzten, ausgedacht haben, mag auf den ersten Blick verwirren. Man sollte sich davon aber nicht irritieren lassen, denn die auf diese Weise kodierten Daten sind in dem XML-Strukturgerüst sehr gut aufgehoben. Man kann obendrein die Elementstrukturen, die man nicht permanent sehen will, mit einfachen Mitteln ausblenden, so dass nur noch der Text sichtbar bleibt, ohne dass die Strukturkennzeichnungen verloren gehen, und schließlich kann man mit passenden Werkzeugen, wie sie die TEI-Community zur Verfügung stellt, den Text in eine Fülle andere Formate konvertieren, ohne die ursprüngliche TEI-Kodierung aufgeben zu müssen.

Von hier aus gesehen liegt es nahe, zu überlegen, ob man die Analysearbeit an den korpusgenerierten und frequenzsortierten Wortlistendaten nicht auch auf der Basis von TEI-Kodierungen weiterführt, angelehnt an Strukturen die die ISO Norm 24613 „Lexical Markup

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Framework“ empfiehlt. Hierzu noch zwei Beispiele. Das erste Beispiel beschreibt mit „Gabel“ den Kern der Wortfamilie. Die ersten drei Zeilen sind mit der Listenstruktur identisch. Mit dem Element `<seg>` werden nun noch zwei weitere Informationsstränge integriert: das erste `<seg>`-Element beschreibt das morphologische Muster, in diesem Fall „Gabel“ als Basismorphem (Root), das zweite enthält die Komponenten, in diesem Fall das Wort (kodiert mit dem Element `<w>`), das identisch ist mit dem Basismorphem „Gabel“ und einem Verweis auf dessen Position in Spletts „Deutschem Wortfamilienwörterbuch“.

```
<item xml:id="derewo-l-0001-gabel" ana="#jsdw-GABEL">Gabel
    <abbr type="POS" >NN</abbr><num type="jsdw-rank">1.2</num>
     <measure type="freqGroup"><num type="cardinal">15</num></measure>
                        <seg type="morphological_pattern">(Root)</seg>
                        <seg type="components">
                            <w>
                                <m type="Root" ana="#jsdw-GABEL-1.2-Gabel">Gabel</m>
                            </w>
                        </seg>
                </item>
```

Das zweite Beispiel ist etwas komplexer, denn es zeigt im ersten `<seg>`-Element ein Beispiel für Spletts Kodierung von Wortstrukturenformeln durch (wS)(wS). „w“ steht für „Kernwort (Wurzel)“ und „S“ für „Substantiv“ (Splett 2009, Bd. 1, S. XXIV). In dem zweiten `<seg>`-Element werden die Kompenenten jeweils dem Element `<w>` für „Wort“ bzw. `<m>` für „Morphem“ zugeordnet und der Wortbildungstyp (Kompositum/compound) im Attribut des Wortrahmenelements festgehalten. Die `ana`-Attribute im `<m>`-Element verweisen auf die Positionierung der Kernwörter „Ast“ und „Gabel“ in Spletts Wortfamilienwörterbuch. Über solche robuste Strukturen wird die Untersuchung des Vernetzungsgrads der Wortfamilienkerne über die gesamte Breite des Wortbestands zuverlässig möglich.

```
<item xml:id="derewo-l-0010-astgabel">Astgabel
         <abbr type="POS">NN</abbr><num type="jsdw-rank">1.2</num>
         <measure type="freqGroup"><num type="cardinal">19</num></measure>

         <seg type="morphological_pattern" ana="#jsdw1-struct">(wS)(wS)</seg>
         <seg type="components">
                 <w type="compound">
                         <w>
                                 <m type="Root" ana="#jsdw-AST-1.2-Ast">Ast</m>
                          </w>
                          <w>
                                 <m type="Root" ana="#jsdw-GABEL-1.2-Gabel">gabel</m>
                           </w>
                  </w>
          </seg>
</item>
```

In Anlehnung an das Lexical Markup Framework LMF (ISO Norm 24613)

Die Verfügbarkeit digitaler korpusgenerierter Lemmalisten mit Frequenzangaben, Verweisen auf die Wortfamilienstruktur und präzisen morphologischen Angaben ermöglicht die detaillierte Analyse der deutschen Wortstrukturen anhand von Daten aus dem aktuellen Sprachgebrauch. Sprachliche Neuschöpfungen unterliegen neben den kombinatorischen Möglichkeiten der Wortbildungsregeln soziolinguistischen, linguistischen und pragmatischen Beschränkungen (Quirk et al. 1985, S. 1531). Die Existenz eines bestimmten Wortes in einer Sprache ist zu einem gewissen Grad idiosynkratisch: eine bestimmte Wortbildungsregel wird priorisiert (*Gänsebraten* vs. **Rehebraten*), die Verteilung von Fugenelementen (*Verkehrsteuer* vs. *Verkehrszeichen*) sowie die Mitglieder von Komposita (*Knoblauchzehe* vs. **Knoblauchzahn* – spanisch *diente de ajo*) sind allein aufgrund der Wortbildungsregeln nicht immer eindeutig determiniert. Die Wortfamilien-Lemmalisten bieten mit der TEI-Kodierung eine

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Datenbasis, über die sich Wortstrukturen und auch Wortbildungsprozesse dynamisch darstellen lassen. Über die morphologisch und morphosyntaktisch motivierte Einteilung der Wortfamilienstruktur sind die im tatsächlichen Sprachgebrauch vorkommenden Wortformen einer Kategorie einfach zu finden. Über die Frequenzangaben kann der Usus bei der Erstellung eines Wörterbuches wie dem Duden, auf dessen Einträgen das Wortfamilien-Wörterbuch basiert, mit korpusgenerierten Daten verglichen werden.

Bleibt abschließend noch die Frage: Kann auch Jochen Spletts „Deutsches Wortfamilienwörterbuch“ für die intensivere Arbeit an einer Strukturgeschichte des deutschen Wortschatzes in solche XML-Strukturen transformiert werden? Nach unseren Test können wir diese Frage ohne jeden Zweifel bejahen. Uns scheint nach einigem Testen eine Transformation in eine valide TEI-Kodierung auf der Basis von XML und weiteren internationalen Standards möglich. Sie erfordert zwar erheblichen Aufwand, dafür wären die Daten aber für die wissenschaftliche Arbeit langfristig gesichert. Spletts Arbeit wäre es wert, und es würde sich ohne Zweifel lohnen.

Elektronische Ressourcen

Die Internetadressen für Forschungsliteratur, Elektronische Ressourcen, Dokumentationen und Projekte beziehen sich auf den Stand Juni 2014.

DeReKo: Deutsches Referenzkorpus. www.ids-mannheim.de/kl/projekte/korpora/.

DeReWo: Korpusbasierte Grund-/Wortformenlisten. www.ids-mannheim.de/kl/projekte/methoden/derewo.html.

Machinese Phrase Tagger Demo. www.connexor.com/nlplib/?q=demo/mpt.

TreeTagger Download. www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.html.

Literatur

Belica, Cyril (1994): A German Lemmatizer. Final Report MLAP93-21/WP2. Luxemburg. http://www1.ids-mannheim.de/fileadmin/kl/dokumente/glemmrep.pdf.

Belica, Cyril et al. (2011): The morphosyntactic annotation of DeReKo: Interpretation, opportunities, and pitfalls. In: Konopka, Marek et al. (Hg.): Grammatik und Korpora. Dritte Internationale Konferenz. Mannheim, 22.-24.9.2009. Tübingen, S. 451-469.

DeReKo (2011): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2011-I (Release vom 29.03.2011). Mannheim. www.ids-mannheim.de/kl/projekte/korpora/archiv.html.

Duden (1993-1995): Das große Wörterbuch der deutschen Sprache in acht Bänden. 2. völlig neu bearb. und stark erw. Aufl. Mannheim u.a.

Herbermann, Clemens-Peter (1981a): Wort, Basis, Lexem und die Grenze zwischen Lexikon und Grammatik. München.

Herbermann, Clemens-Peter (1981b): Moderne und antike Etymologie. In: Zeitschrift für vergleichende Sprachforschung 95, S. 22-48.

Hundsnurscher, Franz (1985): Wortfamilienforschung als Grundlage einer Bedeutungsgeschichte des deutschen Wortschatzes. In: Stötzel, Georg (Hg.): Germanistik – Forschungsstand und Perspektiven. Teil 1. Berlin, S. 116-123.

Hundsnurscher, Franz (1987): Probleme des sprachstufenübergreifenden Wortschatzvergleichs. In: Bergmann, Rolf et al. (Hg.): Althochdeutsch. Bd. 2: Wörter und Namen. Forschungsgeschichte. Heidelberg, S. 1012-1024.

Perkuhn, Rainer et al. (2012): DeReWo: Korpusbasierte Grundformenliste. http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip.

Quirk, Randolph et al. (1985): A comprehensive grammar of the English Language. London/New York.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Romary, Laurent/Wegstein, Werner (2012): Consistent modeling of heterogeneous lexical structures. In: Journal of the Text Encoding Initiative 3: TEI and Linguistics. http://jtei.revues.org/540.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, UK. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT Workshop, Dublin. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf.

Splett, Jochen (1993): Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. Bd. 1, Teil 1: Einleitung, Wortfamilien A-L. Berlin/New York.

Splett, Jochen (2009): Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 18 Bde. Berlin/New York. (Eintrag „Gabel“, Bd. 4, S. 235f.).

Stadler, Heike (2014): Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora. In: OPAL 5/2014. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2014-5.pdf.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Anhang

Jochen Splett (2009): Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 18 Bde. Berlin/New York. (Eintrag „Gabel“, Bd. 4, S. 235f.)

GABEL [1] vgl. **GAFFEL**

1.1 **gabeln** sw.V. (wS)V/ **1. sich g.** 'sich von einem Punkt aus teilen u. gabelförmig verzweigen, auseinander streben' **2.** 'etw. mit der Gabel (2) auf- oder abladen, aufnehmen (u. irgendwohin befördern)' **3.** selten: 'etw. mit der Gabel (1) aufspießen (u. irgendwohin befördern)' **4.** selten: 'mit der Gabel (1) essen'

auf- p((wS)V)/ **1.** salopp: 'jmdn. irgendwo zufällig kennen lernen u. eine private, dienstliche Beziehung anknüpfen' **2.** '(Heu o.Ä.) mit der Gabel (2) aufnehmen u. aufladen'

1.2 **Gabel** F. wS/ **1.** 'Essgerät mit zwei oder mehr Zinken, das beim Essen zum Zerlegen, zum Aufnehmen oder Vorlegen von Speisen dient' **2.** 'Gerät mit zwei oder mehr Zinken u. langem Stiel, das in der Landwirtschaft bes. zum Auf- u. Abladen von Heu, Mist o.Ä. gebraucht wird' **3.** 'Stelle, an der ein Weg oder eine Straße sich in zwei spitzwinkelig auseinander gehende Wege oder Straßen teilt' **4.** 'Teil des Telefons, auf den der Hörer aufgelegt oder in den er eingehängt wird' **5.** 'gabelähnlicher Teil des Fahrrads, in den das Rad eingehängt ist' **6.** 'Astgabel' **7.** 'Gabeldeichsel' **8.** Jägerspr.: 'Gehörn oder Geweih mit nur zwei Enden' **9.** Schach: 'Angriff eines Bauern gegen zwei feindliche Figuren durch einen Zug'

Ast- (wS)(wS)/ 'Stelle, an der ein Ast abzweigt'
Austern- / 'kleine Gabel mit verstärkten Zinken zum Öffnen von Austern'
Dessert- / 'kleine Gabel für Kuchen oder Dessert'
Fisch- / 'Gabel, die zum Fischbesteck gehört'
-Tele-fon- ([wF][wF]S)(wS)/ 'Teil des Telefons, auf den der Hörer aufgelegt oder in den er eingehängt wird'
Fondue- ([wF+]S)(wS)/ 'langstielige Gabel mit zwei Zinken zum Fondueessen'
Grabe- (wV)(wS)/ 'vierzinkige Gabel zum Umgraben des Bodens'
Heu- (wS)(wS)/ 'landwirtschaftliches Gerät mit langem Stiel u. drei oder vier Zinken zum Aufheben o.Ä. des Heus'
Kuchen- / 'kleine Gabel mit drei Zinken, mit der Kuchen, bes. Torte, gegessen wird'
-Vor-lege- (p((wV)V))(wS)/ 'Gabel zum Vorlegen (4)'
Mist- (wS)(wS)/ 'Gerät mit langem Stiel u. drei oder vier Zinken zum Auf-, Abladen von Mist'
Ofen- / landsch.: 'Schürhaken'
Rad- / 'gabelähnlicher Teil des Fahrrads, in den das Rad eingehängt ist'
Hinter- (((wAD)A)(wS))(wS)/ 'Gabel (5) des Hinterrads'
Vorder- ((wA)(wS))(wS)/ 'Gabel (5) des Vorderrads'
Rüben- (wS)(wS)/ 'Gabel (2) zum Ernten von Rüben'
Ruder- / 'Dolle'
Sardinen- ([wF]sS)(wS)/ 'kleine Gabel zum Vorlegen von Ölsardinen'
Schoss- [(wV)S](wS)/ Landw.: 'Gabel (2) mit eng stehenden, am Ende verdickten Zinken zum Aufnehmen von Kartoffeln, Rüben o.Ä.' [← nach schießen im Sinne von schieben, gleiten lassen] [2]
Stimm- ((wS)V)(wS)/ Musik: 'mit Griff versehener Gegenstand aus Stahl in länglicher U-Form, mit dem man durch Anschlagen (8) eine bestimmte Tonhöhe, bes. die des Kammertons, erzeugen kann'
Tranchier- ((wS)sV)(wS)/ 'große Gabel mit Griff u. zwei festen langen Zinken (u. einem aufklappbaren Bügel als Handschutz) zum Tranchieren'
Weg- (wS)(wS)/ 'Weggabelung'
-Ge-wehr- (p((wV)S)S)(wS)/ früher: 'gabelförmiger Stock zum Auflegen des Gewehrs beim Schießen'

Gabel-antilope F. (wS)(wS)/ 'Gabelbock (1)'
-arbeit / 'Gabelhäkelei'
-bein N. / Zool.: 'gabelförmig zusammengewachsener Schlüsselbeinknochen der Vögel'
-bissen M. (wS)((wV)S)/ 'kleines, zusammengerolltes Stück Heringsfilet in pikanter Marinade'
-bock (wS)(wS)/ **1.** 'der Antilope ähnliches, in der Prärie Nordamerikas beheimatetes Tier mit gegabeltem, hirschgeweihähnlichem Gehörn' **2.** Jägerspr.: 'Rehbock, dessen Stangen nur je zwei Enden haben; Gabler'
-deichsel F. / 'aus zwei Stangen bestehende Deichsel, zwischen die ein einzelnes Zugtier eingespannt wird'
-frühstück N. (wS)((wA)(wS))/ veraltend: 'bei besonderen (festlichen) Anlässen eingenommenes zweites Frühstück am späten Vormittag, bei dem zu alkoholischen Getränken pikant zubereitete kalte Speisen gereicht werden'
-griff M. (wS)((wV)S)/ Musik: 'Grifftechnik bei bestimmten Blasinstrumenten, wobei Zeige- u. Ringfinger je ein Grifffloch bedecken u. der Mittelfinger erhoben bleibt'
-häkelei F. (wS)(((wS)sV)sS)/ Handarb.: 'mit einer Hakelnadel u. einer U-förmig gebogenen, großen Nadel ausgeführte Häkelarbeit'
-hirsch M. (wS)(wS)/ Jägerspr.: 'Rothirsch, dessen Geweihhälften nur je zwei spitze Enden haben'
-huhn N. / Jägerspr.: 'junges Rebhuhn mit gegabeltem Schwanz'

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

	Gabel-klavier	(wS) (wS) /	'heute nicht mehr gebräuchliches Tasteninstrument, bei dem Stimmgabeln anstelle von Saiten zum Klingen gebracht werden'
	-knochen M.	/	Zool.: 'Gabelbein'
	-kreuz N.	/	'besondere Art eines Kreuzes in der Form eines Ypsilon'
	-mücke F.	/	'in tropischen u. südeuropäischen Ländern vorkommende Stechmücke (die Malaria überträgt)' [← nach den gabelförmigen Kiefertastern]
	-schlüssel M.	(wS) ((wV) sS) /	'flacher Schraubenschlüssel mit gabelförmiger Öffnung'
	-stapler	((wS)+((wS)V)) sS/	'kleines, motorgetriebenes Fahrzeug, das an seiner Vorderseite mit einer gabelähnlichen Vorrichtung zum Aufnehmen u. Verladen oder Stapeln von Stückgut ausgestattet ist'
	-staplerfahrer	(((wS)+((wS)V)) sS) ((wV) sS) /	'Fahrer eines Gabelstaplers'
	-stütze F.	(wS) ((wV) S) /	'gegabelte Stütze'
	-weihe	(wS) (wS) /	'Milan mit an der Oberseite rotbraunem Gefieder; Roter Milan'
	-wender M.	(wS) ((wV) sS) /	'landwirtschaftliche Maschine, die zum Heuwenden dient'
	gabel-artig Adj.	((wS)+(wS)) sA/	'in der Art einer Gabel, ähnlich wie eine Gabel (1-3)'
	-förmig	/	'von einem Punkt aus in zwei Richtungen auseinander strebend, sich in zwei Arme teilend, in der Form einer Gabel (1) ähnlich'
2.2	**Gäbel-chen** N.	(wS) sS/	'kleine Gabel'
	Gabl-er M.	(wS) sS/	'Gabelbock, -hirsch'
	Gabel-ung F.	((wS)V) sS/	1. 'das Sichgabeln' 2. 'Stelle, an der sich etw. gabelt'
	Weg-	(wS) (((wS)V) sS) /	'Gabelung eines Weges'
2.3	**gabel-ig** Adj.	(wS) sA ((wS)V) sA/	selten: 'sich gabelnd, gabelt, in Form einer Gabel'

[1] KLUGE/SEEBOLD, 325; PFEIFER, Etym.Wb., 493.
[2] GRIMM IX, 1598; 1602.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Die Metalemmaliste als Tool zum Erschließen von sprachlicher Varianz

Luise Borek (*Technische Universität Darmstadt*)

1. Einleitung

In Systemen verschiedenster Disziplinen bedeutet das Vorkommen von Varianz eher einen Regelzustand als dass es eine Ausnahme darstellt. Innerhalb des kollaborativen Forschungsprojekts „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen*"*, gefördert vom BMBF von 2008 bis 2012, wurde dieses natürliche Phänomen anhand von fachspezifischen Daten auf methodischer Ebene interdisziplinär untersucht.

Eines der heterogenen Merkmale von Sprache stellt die in ihr vorkommende diachrone wie synchrone Varianz dar. Während die Bioinformatik für die Analyse von Varianz auf verschiedene Genomdatenbanken, wie z.B. *Ensembl* (siehe Flicek et al. 2012), zurückgreifen kann, musste für das Einnehmen der linguistischen Perspektive zunächst eine äquivalente Datenmenge erschlossen werden. Als Quelle der sprachlichen Vielfalt dienen die über das Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren an der Universität Trier elektronisch zur Verfügung gestellten, retrodigitalisierten historischen Wörterbücher. Mit der Lexikografie wird hier auf eine Disziplin zurückgegriffen, die in ihrer Methodologie als Schnittstelle zwischen Philologie und Informationswissenschaft angesehen werden kann: „In brief, the dictionary, in theory a systematic relational database, with ordered records and recurrent fields, may in human practice be as variable as the lexicon it sets out to describe." (Wooldridge 2004).

Um ihrer Aufgabe der Dokumentation des Wortschatzes gerecht werden zu können, wird ein lexikografischer Prozess etabliert, in dem Kriterien aufgestellt und ein systematisches Vorgehen angelegt werden. Es ist daher wenig verwunderlich, dass die ersten, heute als *Digital Humanities* angesehenen Arbeiten aus diesem Umfeld stammen.[1]

Mit dem Konzept der Metalemmaliste soll hier ein Werkzeug vorgestellt werden, das die Heterogenität der betrachteten Wörterbücher und des in ihnen verzeichneten Wortschatzes – auf formaler wie inhaltlicher Ebene – überbrückt und somit die enthaltene Vielfalt erschließbar macht. Sie stellt ein nutzerfreundliches Hilfsmittel für Wörterbuchnutzer dar, indem sie die im Material enthaltene Vielfalt leichter zugänglich macht. Gleichzeitig bildet sie eine flexible Möglichkeit zur Visualisierung retrodigitalisierter Wörterbücher und bietet für die Verknüpfung mit weiteren Informationssystemen eine Andockstelle.

Während aktuelle Online-Wörterbücher theoretisch tagesaktuell angepasst werden können und Möglichkeiten bieten, den Bedarf von Nutzern oder die Interaktion mit und zwischen diesen in die Konzeption mit einfließen zu lassen, sind die Möglichkeiten bei retrodigitalisierten Wörterbüchern in diesem Bereich relativ eingeschränkt. Vielmehr geht es hier um die werkgetreue Überführung der ursprünglichen Printausgaben in einen digitalen Repräsentanten, der seinen Charakter auch im jüngeren Medium widerspiegelt. Der Mehrwert zeigt sich nicht durch die reine Verfügbarkeit des digitalen Artefakts, sondern liegt vielmehr in dessen

[1] Vgl. Hockey (2004). Hier insbesondere das Arbeiten mit Konkordanzen (Busa).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Aufbereitung und Visualisierung. Bleiben die Inhalte weitgehend unverändert,[2] lassen sich so auch ohne Eingriffe in den Text Maßnahmen treffen, die Intentionen der Wörterbücher aufgreifen und nutzerfreundlich umsetzen. Darunter fallen z.B. das Realisieren bereits enthaltener Verweise mittels Hyperlinks, vielfältige Suchoptionen und eine ansprechende Gestaltung der Oberfläche im Browser (*Graphical User Interface*).

In diesem Artikel wird die Konzeption der Metalemmaliste beschrieben. Hierfür wird zunächst die Grundidee anhand der verwendeten Daten erläutert. Die Einbettung in den Projektkontext steht dabei im Hintergrund und wird nur am Rande beleuchtet, um ggf. Maßnahmen zu erläutern, die durch diesen bedingt sind. Anschließend wird das sich hieraus ergebende formale Konzept dargelegt sowie die angewandten Zuordnungsverfahren beschrieben. In einem Ausblick werden Möglichkeiten zur Integration und Nachnutzung skizziert.

2. Metalemmaliste

2.1 Erschließen sprachlicher Vielfalt

Wörterbücher dienen der Dokumentation des Wortschatzes. Traditionell müssen in der Lexikografie eine Vielzahl von Entscheidungen getroffen werden, die die Aufnahme und Ausführlichkeit der Behandlung einzelner Wörter betreffen. Die Heterogenität der Sprache und ihr ständiger Wandel stellen dabei Herausforderungen dar, denen auf verschiedene Arten begegnet werden kann. Das Einhalten der zugrunde gelegten Konzepte ist dabei von verschiedenen Faktoren abhängig, wie z.B. der Dauer des Vorhabens, der Anzahl der Bearbeiter sowie ihrer Kontinuität.

Mit der Metalemmaliste wird ein Werkzeug zur Verfügung gestellt, das die konzeptuelle und sprachliche Heterogenität von Wörterbüchern in ihrer elektronischen Verwendung überwindet, ihre Vielfalt erschließbar macht und eine flexible Suche ohne besondere Vorkenntnisse ermöglicht. Dabei erfolgt eine Vernetzung der Lemmata auf rein semasiologischer Ebene. Für mögliche regional oder diachron bedingte semantische Unterschiede wird auf die hinterlegten Wörterbuchartikel verwiesen.

2.2 Beschreibung der Datengrundlage

Die sprachwissenschaftlich-lexikografische Datengrundlage setzt sich zum Teil aus einer heterogenen Sammlung retrodigitalisierter, historischer Wörterbücher zusammen, die am *Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften* über das Trierer Wörterbuchnetz öffentlich zur Verfügung stehen (www.woerterbuchnetz.de).[3] Für die Metalemmaliste wurden daraus folgende Wörterbücher untersucht: „Deutsches Wörterbuch“ von Jacob und Wilhelm Grimm, Adelungs „Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart“, „Wörterbuch der elsässischen Mundarten“, „Wörterbuch der deutsch-lothringischen Mundarten“, „Pfälzisches Wörterbuch“, „Rheinisches Wörterbuch“ und die mittelhochdeutschen Wörterbücher von Matthias Lexer und Benecke/Müller/Zarncke (BMZ).

[2] Ausnahmen bilden Korrekturen von Tippfehlern etc., die stillschweigend vorgenommen und/oder gesondert kenntlich gemacht werden können.
[3] Das Wörterbuchnetz ist ein Angebot der Universität Trier. Weitere Informationen finden sich in Burch/Rapp (2007).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Abb. 1: Die Suchmaske des Trierer Wörterbuchnetzes mit der Eingabe *Pferd*.

Ergänzend konnte mit dem an der Universität Würzburg digital aufbereiteten „Wörterbuch der Deutschen Sprache“ von Joachim Heinrich Campe (Campe 1807-1811) und dem „Wörterbuch der deutschen Gegenwartssprache“ (WDG, Klappenbach/Steinitz 1961-1977), digitalisiert durch die Berlin-Brandenburgische Akademie der Wissenschaften,[4] auf weitere einschlägige Wörterbücher des Deutschen zurückgegriffen werden. Die betrachteten Wörterbücher bilden verschiedene historische wie regionale Facetten der deutschen Sprache ab und ergeben somit eine ausreichend repräsentative Datenbank, anhand derer Methoden und Algorithmen zum Erschließen und Untersuchen von Varianz entwickelt und erprobt werden. Sämtliche Wörterbücher liegen in SGML/XML und nach TEI kodiert vor. Das Vorgehen bei der Auszeichnung sowie deren Tiefe ist dabei aufgrund der verschiedenen Entstehungskontexte einerseits und Unterschieden in der lexikografischen Ausrichtung andererseits heterogen. In diesem Artikel liegt der Fokus auf Aspekten, die aus philologischer Perspektive für das Erstellen der Metalemmaliste relevant sind.

Bereits ein Blick auf die verschiedenen Arten des Lemmaansatzes genügt, um die zu überwindende Heterogenität zu verdeutlichen. Die Dialektwörterbücher (Rheinisches Wörterbuch, Pfälzisches Wörterbuch) verzeichnen weitgehend normalisierte Lemmata, während ein Idiotikon wie das Wörterbuch der deutsch-lothringischen Mundarten eine an die Aussprache angelehnte Verschlagwortung verfolgt. Die Suche nach dem nhd. Substantiv *Pferd* über die globale Suchmaske des Trierer Wörterbuchnetzes führt z.B. zu Ergebnissen im Rheinischen und Pfälzischen Wörterbuch, ergibt jedoch keinen Treffer für das Lothringische. Der entsprechende Eintrag *Perd* ist nur mithilfe von Vorkenntnissen gezielt anwählbar. Gleiches gilt für das Auffinden von Lemmata früherer Sprachstufen, z.B. des Mittelhochdeutschen, die ebenfalls keiner (modernen) orthografischen Normierung unterliegen und daher in ihrer Vielfalt in den Wörterbüchern verbucht und – unabhängig von der Art des Mediums – nur Kundigen zugänglich sind. Auch die Funktion als Belegwörterbuch spielt hier eine Rolle, indem auch Varianten verzeichnet werden, die von der Leitvariante abweichen, gesondert gelistet werden und auf die Leitvariante verweisen, unter dieser aber nicht zwangsläufig auch erwähnt werden.[5] Kandidaten wie *phærît*, *pfard* oder *pherch* sind zwar untereinander im Verbund der mittelhochdeutschen Wörterbücher weitgehend vernetzt (vgl. Burch/Fournier/Gärtner 2000), werden bei der obigen Suche aber dennoch nicht berücksichtigt. Vor diesem Hintergrund ist der Einsatz der Metalemmaliste für die verbesserte Erschließung des in den Wörterbüchern enthaltenen Wortschatzes eine interessante Anwendung.

[4] Für die Rolle des WDG im Kontext des DWDS Projektes siehe die dortige Beschreibung der Ressourcen, Kap. 2.2 unter www.dwds.de/ressourcen/woerterbuecher/#part_22 (Stand: 1.9.2014).

[5] Solche Fälle werden im Trierer Wörterbuchnetz als sog. Automatische Rückverweise behandelt und werden beim Gebrauch des Online-Wörterbuchs in einer parallelen Ansicht angezeigt.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Das Erschließen dieser Vielfalt und das Entwickeln eines Tools, das die Vernetzungsdichte der Wörterbücher untereinander weiter ausbaut, ist die Hauptaufgabe der Metalemmaliste. Die Ansprüche zum Erstellen der virtuellen Metaebene erfordern eine große Datenmenge, die den allgemeinen, gegenwartssprachlichen Gebrauch widerspiegelt. Das jüngste der betrachteten Wörterbücher, das WDG, ist mit seinem Umfang von ca. 60.000 Lemmata[6] nicht umfangreich genug, um die Vielfalt aller anderen Wörterbücher auf sich abbilden zu können.

Für die Analyse der Varianz wie auch für die intuitive Nutzung über einen standardsprachlichen Zugang sind Daten erforderlich, die entsprechende referentielle Kriterien erfüllen. Mit einem Umfang von über 5 Milliarden Textwörtern ist das am Institut für Deutsche Sprache aufgebaute Deutsche Referenzkorpus, DEREKO, die umfassendste Sammlung elektronischer Korpora des Deutschen. Zur Gegenüberstellung mit den Lemmalisten der retrodigitalisierten Wörterbücher wurden aus dem Korpus verschiedene Lemmalisten generiert; als primäre Referenz fungierte dabei die 250.000 Lemmata umfassende Basislemmaliste des Neuhochdeutschen (siehe Stadler 2013).

2.3 Struktur der Metalemmaliste

Die Metalemmaliste ist als dynamisches Clusternetzwerk angelegt, in dem semasiologisch identische Lemmata um eine neuhochdeutsche Referenz, das Metalemma, gebündelt werden. Die Notwendigkeit eines standardneuhochdeutschen Metalemmas ergibt sich aus zwei Faktoren. Zum einen bedarf es für die Untersuchung von Varianz einer Referenz, damit die eingesetzten Verfahren Vielfalt als solche erkennen können, und zum anderen ist ein intuitiver Zugang für den Gebrauch gewährleistet. Ein Vorgehen, das z.B. von germanischen Stämmen ausgeht, wäre reizvoll, ist aber in Ermangelung des nicht vorhandenen Datenmaterials und in Hinblick auf die Vorteile bei der Nutzung als Tool über das Projekt hinaus nicht umsetzbar. Eine spätere Anreicherung oder Umgestaltung der aufgebauten Struktur wäre hier jederzeit möglich.

Das vorerst restriktiv semasiologisch angelegte Vorgehen erlaubt eine klare Abgrenzung für die erste Version der Metalemmaliste, die im Folgenden verdichtet und ergänzt werden kann. Über die Ebene der Metalemmata stehen Anknüpfungspunkte für den Anschluss weiterer Informationssysteme zur Verfügung.

Unter dem Metalemma *Pferd* finden sich sämtliche semasiologische Varianten verbucht, die die angeschlossenen Wörterbücher als Lemmata führen:

[6] Auch mit Hinzunahme der Zusammensetzungen stehen den (dann 90.000) Lemmata des WDG ca. 300.000 Grimm'sche Lemmata gegenüber.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

```
<metalemma id="201" value="Pferd" pos="nn" source="bll">
    <lemma corpus="bll" corpus_id="1816" value="Pferd" frequency_class="11" />
    <lemma corpus="rhwb" value="Perd" id="RP01745"/>
    <lemma corpus="rhwb" value="Pferd" id="RP02470"/>
    <lemma corpus="dwb" value="Pferd" id="GP03285"/>
    <lemma corpus="elswb" value="Pfërd" id="EP02749"/>
    <lemma corpus="lexer" value="pfard" id="LP00457"/>
    <lemma corpus="lexer" value="phert" id="LP00636"/>
    <lemma corpus="lexer" value="phercht" id="LP00610"/>
    <lemma corpus="lexer" value="pherfrit" id="LP00616"/>
    <lemma corpus="lexer" value="pherift" id="LP00619"/>
    <lemma corpus="lothrwb" value="Perd" id="CB00558"/>
    <lemma corpus="bmz" value="phærît" id="BP00248"/>
    <lemma corpus="bmz" value="phert" id="BP00346"/>
    <lemma corpus="pfwb" value="Pferd" id="PB02803"/>
    <lemma corpus="findebuch" value="phert" id="FP00235"/>
</metalemma>
```

Abb. 2: Prototypischer XML-Ausschnitt aus der Arbeitsversion der Metalemmaliste zum Eintrag *Pferd*.

Für jedes im Metalemma integrierte Wörterbuchlemma ist die Herkunft (*corpus*), die Schriftform (*value*) sowie die eindeutige Wörterbuch-ID (*id*) gespeichert.[7] Das Metalemma führt zusätzlich die für alle geltende Wortklasse (*pos*) und wird in der Regel aus der Basislemmaliste gespeist (*source*=bll),[8] dem Basislemma ist als weitere Information eine Häufigkeitsangabe beigegeben (*frequency class*), die es aus der Basislemmaliste erbt.

Das Vorkommen von *Pferd* in der Basislemmaliste belegt seinen Platz im öffentlichen Sprachgebrauch und legitimiert seine zusätzliche Überführung in den Status eines Metalemmas, das die Clusterbildung ermöglicht. Der Klon aus dem Standardneuhochdeutschen nimmt hierbei ordnungsstrukturell eine Sonderrolle ein, da es sich bei ihm nicht – wie bei seinen Äquivalenten auf der gleichen Strukturebene – um ein Wörterbuch-Lemma handelt. Dieser Umstand darf jedoch nicht als wertendes Strukturmerkmal missverstanden werden. Der Mehrwert dieses Vorgehens liegt in der Funktion als Referenz zum Identifizieren von Varianz und entsprechenden Zuordnungsmöglichkeiten sowie in seiner Fähigkeit, den Wortschatz zu ordnen und damit einfachere Zugriffsmöglichkeiten zu liefern. Mit dem Einfügen von Relationen zwischen einzelnen Metalemmata können leicht Informationen zu weiteren Beziehungen hergestellt werden (z.B. kann das Metalemma als Schnittstelle verwendet werden, um *Pferd* mit *Klepper* oder *Schimmel* in Beziehung zu setzen).

2.4 Zuordnungsverfahren

Für die Generierung der Metalemmaliste werden verschiedene Verfahren miteinander kombiniert, auf die im Folgenden näher eingegangen wird. Zum einen wird dabei auf vorhandene Strukturen und Informationen der behandelten Wörterbücher zurückgegriffen, zum anderen werden die dort ermittelten Zugehörigkeiten mittels computergestützter Verfahren zum Vergleichen von Lemmalisten ergänzt. Die Information, welches Verfahren die Zugehörigkeit zu einem Metalemma (in eine Metalemma-Umgebung) ermittelt hat, wird in Form eines Attributes dem zugeordneten Lemma beigefügt, um Transparenz zu schaffen.

[7] Hinzu kommen weitere, wörterbuchspezifische Informationen wie Sprachstufe, Dialektraum und eingetragene Querverweise. Diese sind aus Gründen der Übersichtlichkeit hier nicht abgebildet.
[8] In Sonderfällen, in denen dies nicht möglich ist, stehen Pseudometalemmata an seiner Stelle.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

2.4.1 In den Wörterbüchern enthaltene Verweisstrukturen

Eine Möglichkeit für die elektronische Aufbereitung und Visualisierung von retrodigitalisierten Wörterbüchern, die im Trierer Wörterbuchnetz umgesetzt ist und hier nachgenutzt wird, besteht in der Auswertung und Realisierung von Verweisen, die in den Wörterbuchartikeln vermerkt sind. Auf diese Art kann das vergleichsweise junge Pfälzische Wörterbuch in seinem Artikel zu *Pferd* auf die entsprechenden Artikel im Lothringischen, Elsässischen und Rheinischen Wörterbuch verweisen. Im Trierer Wörterbuchnetz werden dem Anwender diese Links mit entsprechendem Herkunftshinweis zur Verweisart (in diesem Fall: „Im Wörterbuch eingetragene Verweise“) angezeigt. Im Umkehrschluss wurden diese Bezüge auch für die entgegengesetzte Richtung eingefügt („Automatisch erzeugte Rückverweise“) und tragen so zu einer erweiterten Funktionalität im Vergleich zum Printmedium bei (vgl. Hildenbrandt 2011, S. 34).

Für den Eingang in die Metalemmaliste müssen die eingetragenen Verweise zunächst mit automatischen oder halbautomatischen Verfahren überprüft werden, da die hinterlegten Verweise nicht auf das Semasiologische beschränkt sind. So finden sich in obigem Beispiel *Pferd* Zusammensetzungen (*Steckenpferd*, *Tanzpferd*) ebenso wie ein onomasiologischer Bezug (*Vollblut*) und ein sich aus dem Kontext ergebender Querverweis (*Mönch*) im gleichen Wörterbuch. Das Ineinandergreifen der Zuordnungsverfahren trägt dazu bei, dass solche, für die Metalemmaliste nicht relevanten (nicht in semasiologischem Verhältnis stehenden) Kandidaten, herausgefiltert werden können. Das Alignment-Verfahren, das aufbauend auf fachspezifischer Information und einer evaluierten Grundmenge stringbasiert wörterbuchübergreifend etymologisch gleiche Partner identifiziert (vgl. den Beitrag von Seipel/Borek in diesem Band), wird hier umgekehrt eingesetzt. Ein schlechtes Ergebnis bei den Ähnlichkeitswerten weist auf die Unwahrscheinlichkeit einer semasiologischen Identität hin. Auch die Zugehörigkeit zu unterschiedlichen Wortklassen bildet ein Ausschlusskriterium für die Zuordnung zu einem Metalemma (Cluster).

Neben den direkt in den Wörterbüchern erfolgten Bezügen werden aus den Angaben zur Etymologie, so diese behandelt wird, Informationen für eine weitere Verdichtung der Metalemmaliste gewonnen. An dieser Stelle sei auf die Auswertung des Deutschen Wörterbuches im Hinblick auf die darin angeführten mittelhochdeutschen Formen verwiesen (siehe Seipel/ Borek in diesem Band), die den Anschluss an die mittelhochdeutschen Wörterbücher im Verbund erlaubt. Hier finden sich weitere Informationen zur Erschließung und dem daran gekoppelten Alignment-Verfahren.

2.4.2 Ergänzende Verfahren zum Vergleich von Lemmalisten

Unabhängig von den eingetragenen Verweisen und weiteren in den Wörterbüchern enthaltenen Angaben, wurden die Lemmastrecken der verschiedenen Wörterbücher betrachtet und aufeinander abgebildet. Dabei wurden einzelne Lemmata als Strings behandelt und wörterbuchübergreifend miteinander verglichen, um Entsprechungen auffinden und diese miteinander in Bezug setzen zu können. Je nach Vergleichsansetzung wurden für diesen Zweck philologische Kriterienkataloge erstellt, um abweichende Konzepte in der Orthografie oder durch Lautwandel bedingte Veränderungen zu kompensieren. Beispiele hierfür sind die unterschiedlichen Lautrepräsentationen *th* zu *t* und *sz* für *ß* beim Vergleich der Lemmastrecke des DWB mit der Basislemmaliste oder das automatische Vollziehen von Lautgesetzen vom Mittelhochdeutschen zum Neuhochdeutschen: z.B. *û* zu *au* und *uo* zu *u*. Damit dieses Vorgehen

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

ausgeschöpft werden kann, muss es eine vielschichtige Kombinatorik erlauben: Angenommene Veränderungen müssen dabei iterativ ausprobiert werden, um anschließend überprüfen zu können, ob das so Generierte zum Auffinden von weiteren Kandidaten geführt hat.

Ein ähnliches Verfahren kam bei der Verknüpfung der Luxemburgischen Wörterbücher erfolgreich zur Anwendung (vgl. Büdenbender 2011). Die vorgenommenen Zuordnungsregeln können nur zum Teil als allgemeingültig angesehen werden (Monophthongierung, Diphthongierung), da sie spezifisch auf die Lemmalisten angepasst sind und auf die im lexikografischen Prozess getroffenen Normierungen der Lemmaansätze reagieren. Gleichzeitig stehen die verschiedenen Zuordnungsverfahren in unmittelbarem Zusammenhang, indem ein ermittelter Kandidat zu weiteren Verweisen etc. führen kann:

Lemma (Lexer)	Verweislemma (Lexer)	Lemma (DWB)	Verweislemma (DWB)
apfel	öpfel, epfel, [epfelmost]	Apfel	Öpfel
phert	pfard, pfärft, phärit	Pferd	–
sûber	sûver, sûfer, souber	sauber	–
ûf hoeren	–	aufhören	–

Abb. 3: Übersicht zur Interaktion von Zuordnungen auf Lemmaebene mittels entsprechender in den Artikeln enthaltener Verweise.

Die Tabelle zeigt Treffer, die durch das Mapping der Lemmastrecke des Mittelhochdeutschen Wörterbuchs (Lexer) mit der des DWB ermittelt wurden. Zusätzlich werden die Verweislemmata aufgeführt, die über das einfache Verknüpfen auf Lemmaebene automatisch hinzugewonnen werden, für die Metalemmaliste aber nur berücksichtigt werden, wenn es sich um rein semasiologische Verweise handelt.

Der Ansatz des stringbasierten Alignment-Verfahrens, das ausführlicher in dem Beitrag von Seipel/Borek (in diesem Band) beschrieben wird, ist zwar in seiner Form des Vergleichens ähnlich – weist aber entscheidende Unterschiede auf, die zu einer Parallelisierung beider Verfahren geführt haben, die es anschließend erlaubt, die Ergebnisse aufeinander zu beziehen, um ihre jeweils individuellen Vorteile ausnutzen zu können. Während im oben skizzierten Mapping-Verfahren über ein gezieltes Ersetzen ganze Lautgruppen und funktionale Morpheme in veränderter Form experimentell gesucht werden können, erlaubt es das Alignment, auch nicht primär in den Regelkatalogen berücksichtigte Phänomene aufgrund bloßer Wahrscheinlichkeiten aufzufinden.

2.5 Qualitätssicherung

Der Aufbau der Metalemmaliste erfolgt durch eine Kombination von händischen Verfahren (z.B. das Erstellen von Evaluierungsmengen) mit automatisierten Abläufen. Zu ersteren gehören das Aufstellen und Formalisieren von Lautregeln bzw. Mappings orthografischer Varianten, das Erstellen von Evaluierungsmengen zum Ermitteln von Wahrscheinlichkeiten für das Alignment-Verfahren sowie ein ständiges kritisches Überprüfen der Ergebnisse. Die computergestützten Verfahren greifen auf manuelle Vorarbeiten und Einstellungen zurück, setzen diese um und werden anschließend wiederum (stichprobenartig) händisch überprüft und nötigenfalls optimiert. Darunter fallen Vergleiche extrahierter Lemmalisten, die entweder durch das gezielte Austauschen von Zeichen oder durch Alignierungsverfahren erfolgen. Das Formalisieren vorhandener sprachwissenschaftlicher Regeln ist nur bis zu einem gewissen Grad

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

produktiv. In einem natürlichen, dem Wandel unterworfenen System greift keine Regel zu einhundert Prozent. „Ausnahmen“ hingegen lassen sich kaum formalisieren, da ein solches Vorgehen zur Folge hätte, dass die Ausnahmeregel auch auf eigentlich „regelmäßige“ Fälle angewendet würde und so zu einer möglichen Fehlerquelle avancierte.

Im Folgenden werden die Maßnahmen zum Umgang mit dieser Dialektik geschildert, die zur Optimierung und Transparenz des Prozesses beitragen sollen.

Zunächst einmal bietet der Rückgriff auf das fundierte und lexikografisch aufgearbeitete Material eine belastbare Datengrundlage, auf die bei den Operationalisierungen zurückgegriffen werden kann. Die Auswertung der annotierten Informationen zu einzelnen Lemmata ermöglicht eine präzisere Zuordnung beim Mapping. Während das mit Zeichenersatzregeln operierende Verfahren zunächst unabhängig von der Wortklassenzugehörigkeit nach geeigneten Partnern sucht und diese erst bei einem erzielten Treffer als abgleichende Probe herangezogen wird, gilt beim Alignment ein identisches *pos* von vornherein als Grundvoraussetzung für eine Zuordnung. Damit die Wortklassenangabe wörterbuchübergreifend ausgewertet werden kann, bedarf es wiederum verschiedener, händisch angelegter Auffangkategorien, die unterschiedliche Benennungen der Wörterbücher und Einordnungen kompensieren:

```
regel(X, Y) :-
   X = [v.],
   Y = [verb].

regel(X, Y) :-
   X = [vb.],
   Y = [verb].

regel(X, Y) :-
   X = ["schwaches  v|Rest],
   Y = [verb].
```

Das Beispiel zeigt einen Ausschnitt aus dem Wortklassen-Mapping für die Zuordnung von Verben aus dem DWB,[9] das zuverlässig Angaben zum part-of-speech aufweist, die sich jedoch in ihrer Systematik und Granularität stark unterscheiden (vgl. *v.*, *verb*, *vb.* etc.). Die Formalisierung mittels PROLOG erlaubt ein ökonomisches, leicht zu vermittelndes Vorgehen. Neben der Vereinheitlichung varianter Terminologie und Kürzel wurden funktionale, in Wörterbüchern traditionell eher wenig repräsentierte Wortklassen zum Teil zusammengefasst, um ihr Auffinden zu erleichtern. Lemmata, denen die *pos*-Information fehlt, wurden unabhängig von dieser Restriktion völlig ausgenommen.

Die Vorab-Beschränkung der Wortklasse beim Alignment führt nicht nur zum Ausschluss von vornherein falschen Kandidaten, sondern gleichzeitig zu einer erheblichen Reduktion des Rechenaufwands. Eine weitere Schranke dieser Art bildet die Vorgabe von bestimmten Zeichen im Anlaut, die für einen Ausgangswert überhaupt infrage kommen. Ergänzend zu der bloßen Auflistung möglicher Zeichengruppen wurden für das Alignment vom Mittelhochdeutschen zum Neuhochdeutschen zusätzlich verschiedene, die Position innerhalb der Zeichenkette berücksichtigende Strings definiert und als „Regeln“ in XML hinterlegt:

```
<rule position="head" sgnA="sl" sgnB="schl" />
<rule position="tail" sgnA="nusse" sgnB="nis" />
```

Diese beiden Regeln für den Anlaut (`position="head"`) und Auslaut (`position="tail"`) gelten für die Überführung eines mittelhochdeutschen Lemmas (`sgnA`) in eine neuhochdeut-

[9] Insgesamt wurden 53 Regeln für die Wortklassen innerhalb des DWB erfasst. Sie dienen dazu, die varianten Wortklassen-Beschreibungen in sechs vereinheitlichende Kategorien umzuleiten, die anschließend verwendet werden können, um eine korrekte Zuordnung zu unterstützen. Ein Informationsverlust oder Eingriff in die Textdaten erfolgt dabei nicht.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

sche Form (`sgnB`). Durch den Rückgriff auf bekannte lautgeschichtliche Abläufe wird der Tendenz des Alignments, bei einer abweichenden Zeichenanzahl schlechtere Ergebnisse zu berechnen, entgegengewirkt.

Wird während des Rechenablaufs ein bestimmter Wert überschritten – ist die Ähnlichkeit also sehr gering – wird der Ansatz nicht weiterverfolgt und der Prozess abgebrochen. Umgekehrt kann dieser Grenzwert aufgrund der im Alignment erlangten Erfahrungswerte als Indikator für Zuordnungen dienen, die nicht der semasiologischen Konzeption der Metalemmaliste entsprechen. Je nachdem wie aussagekräftig die errechnete Wahrscheinlichkeit ist, kann eine vorgeschlagene Zuordnung völlig ignoriert werden oder eine Markierung zur weiteren Überprüfung durch einen Bearbeiter bekommen. Auf diese Art kann eine gezielte Evaluierung von Zweifelsfällen erfolgen. Aufgrund der Vielfalt des Materials und der großen Datenmenge kann die manuelle Überprüfung darüber hinaus nur stichprobenartig erfolgen. Diesem Umstand wird mit Transparenz begegnet: die Herkunft eines ermittelten Links wird offenbart, indem jedem Link die Information zu seiner Entstehungsart beigefügt wird. Ein von einem Bearbeiter manuell angelegter oder überprüfter Link ist dabei verlässlicher einzuschätzen als ein rein automatisch generierter.

3. Anwendung und Perspektiven

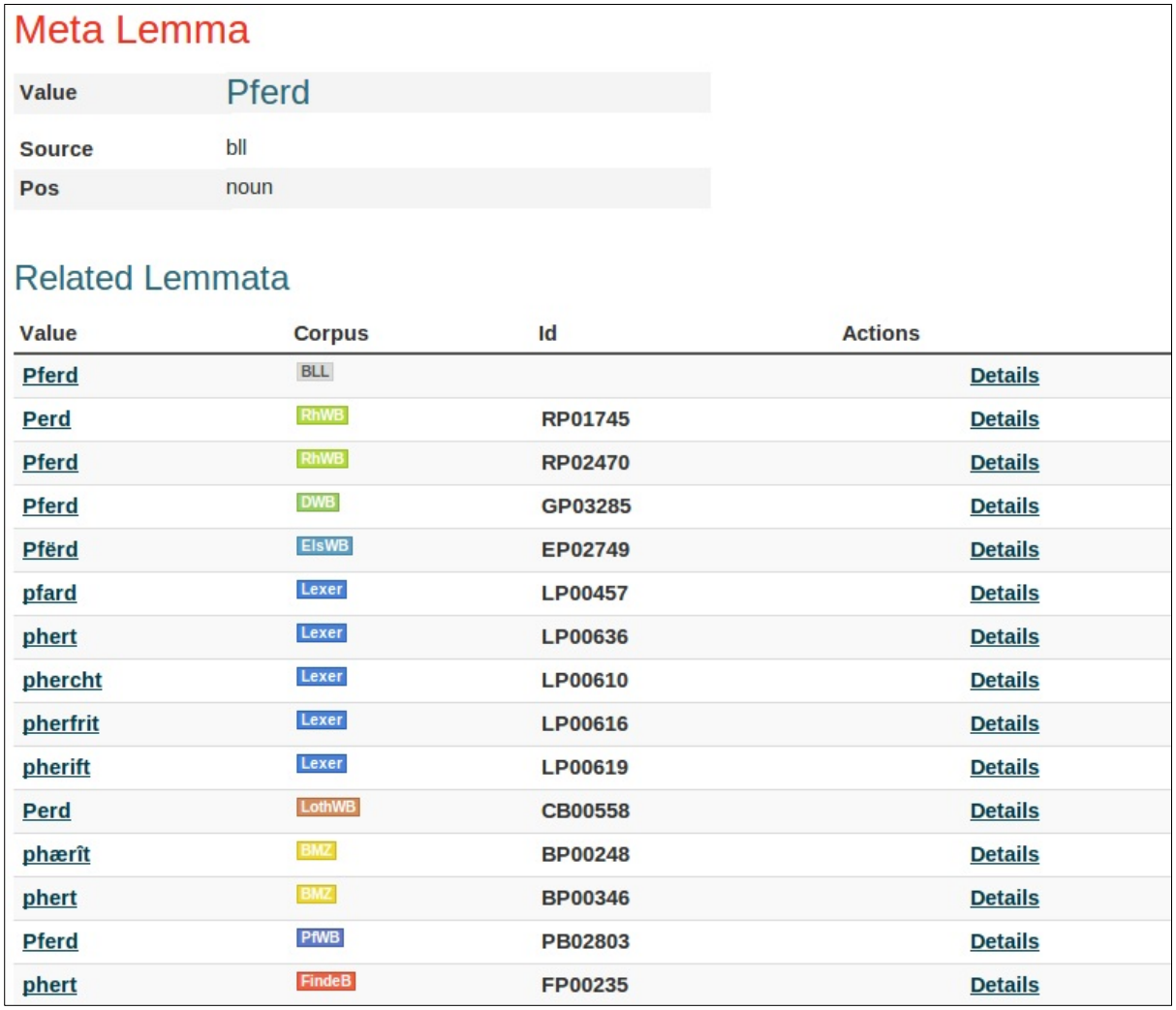
Als Werkzeug, das nach semasiologischen Kriterien identische Lemmata gruppiert, gibt es verschiedene Anwendungsszenarien für die Metalemmaliste. Neben der Funktion als strukturierte Datengrundlage für die Untersuchung sprachlicher Vielfalt in dem erwähnten Wechselwirkungen-Projekt, lassen sich zwei Haupteinsatzbereiche identifizieren, die im Folgenden exemplarisch skizziert werden. Das erste Beispiel beschreibt die Metalemmaliste als eigenen Forschungsbereich, indem es die für sie eingerichtete Website vorstellt. In einem zweiten Beispiel wird anschließend vorgeführt, wie sich die Metalemmaliste als Verknüpfungstool in ein bestehendes Wörterbuchprojekt einbetten ließe.

3.1 Die Metalemmaliste als Web-Service[10]

Während die Metalemmaliste bei der Anwendung als Verknüpfungstool eher unsichtbar im Hintergrund wirkt, indem sie der Verdichtung von Links verschiedener Ressourcen dient, steht die Metalemmaliste auf gleichnamiger Website im Vordergrund. Sowohl die gesamte Liste als auch einzelne Metalemmata stehen sichtbar zur Verfügung und sind durchsuchbar. Wahlweise lässt sich eine Liste aller Metalemmata oder sämtlicher enthaltener Lemmata anzeigen. Über die Suchfunktion lassen sich zudem gezielt Anfragen an das Material stellen. Die Metalemmaliste selbst wird so zum Forschungsobjekt.

[10] Via http://metalemmaliste.de.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.



Meta Lemma

Value	Pferd
Source	bll
Pos	noun

Related Lemmata

Value	Corpus	Id	Actions
Pferd	BLL		Details
Perd	RhWB	RP01745	Details
Pferd	RhWB	RP02470	Details
Pferd	DWB	GP03285	Details
Pfërd	ElsWB	EP02749	Details
pfard	Lexer	LP00457	Details
phert	Lexer	LP00636	Details
phercht	Lexer	LP00610	Details
pherfrit	Lexer	LP00616	Details
pherift	Lexer	LP00619	Details
Perd	LothWB	CB00558	Details
phærît	BMZ	BP00248	Details
phert	BMZ	BP00346	Details
Pferd	PfWB	PB02803	Details
phert	FindeB	FP00235	Details

Abb. 4: Die Suchmaske des Trierer Wörterbuchnetzes mit der Visualisierung des unter dem Metalemma *Pferd* gebuchten Lemma-Aufkommens im Wörterbuchnetz auf www.metalemmaliste.de.

Abbildung 4 zeigt den Aufruf des Metalemmas „Pferd“ (*value*). Wie oben beschrieben, werden auch hier die obligatorischen Angaben zur Quelle (*source*) und der Wortklasse des Metalemmas beigegeben. Letztere gilt folglich für alle Lemmata desselben Clusters, die als *Related Lemmata* angeführt werden. Die Auflistung enthält verschiedene semasiologische Varianten, wie sie in den retrodigitalisierten Wörterbüchern des Trierer Wörterbuchs vorkommen. Ihre jeweilige Herkunft (*corpus*) wird in der zweiten Spalte der tabellarischen Ansicht angezeigt und ist im Design an das Trierer Angebot angelehnt, um einerseits einen vertrauten Umgang zu gewährleisten und um andererseits dem Benutzer zu signalisieren, dass ein Anklicken des jeweiligen Icons zum Lemma des jeweiligen Wörterbuchs im Wörterbuchnetz führt. Neben der Spalte zur ID findet sich zudem über die „Actions“-Spalte die Möglichkeit, weitere Informationen zu einem einzelnen Lemma aufzurufen.

3.2 Die Metalemmaliste als Verknüpfungswerkzeug

In ihrer Funktion als Verknüpfungswerkzeug kann die Metalemmaliste verschiedene elektronische Ressourcen miteinander assoziieren. Neben der Option, das reine Datenmaterial (z.B. via XML oder RDF) für Herstellen weiterer Verweise zu verwenden, soll im Folgenden eine Variante vorgestellt werden, die auf vorhandene Infrastrukturen zurückgreift.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Die von TextGrid entwickelte virtuelle Forschungsumgebung, das TextGridLab, schließt eine Reihe von Tools und Ressourcen ein und steht über die Projektwebsite zurzeit in der Version 2.1 zum freien Download zur Verfügung.[11] Die Wörterbücher des Trierer Wörterbuchnetzes sind hier als *Dictionary Tool* bereits integriert, weshalb sich das Operieren mit der Metalemmaliste anbietet. In seinem offenen, modularen Aufbau erlaubt das Lab die Integration weiterer Services, wie der prototypischen Version des *Dictionary Link Editors* (Leuk 2012). Die Abbildung 5 zeigt diesen kleinen Editor (unten), der auf die Ressourcen der Wörterbücher sowie die dazu implementierte Suchfunktion zurückgreift, in der Ansicht des TextGridLabs.

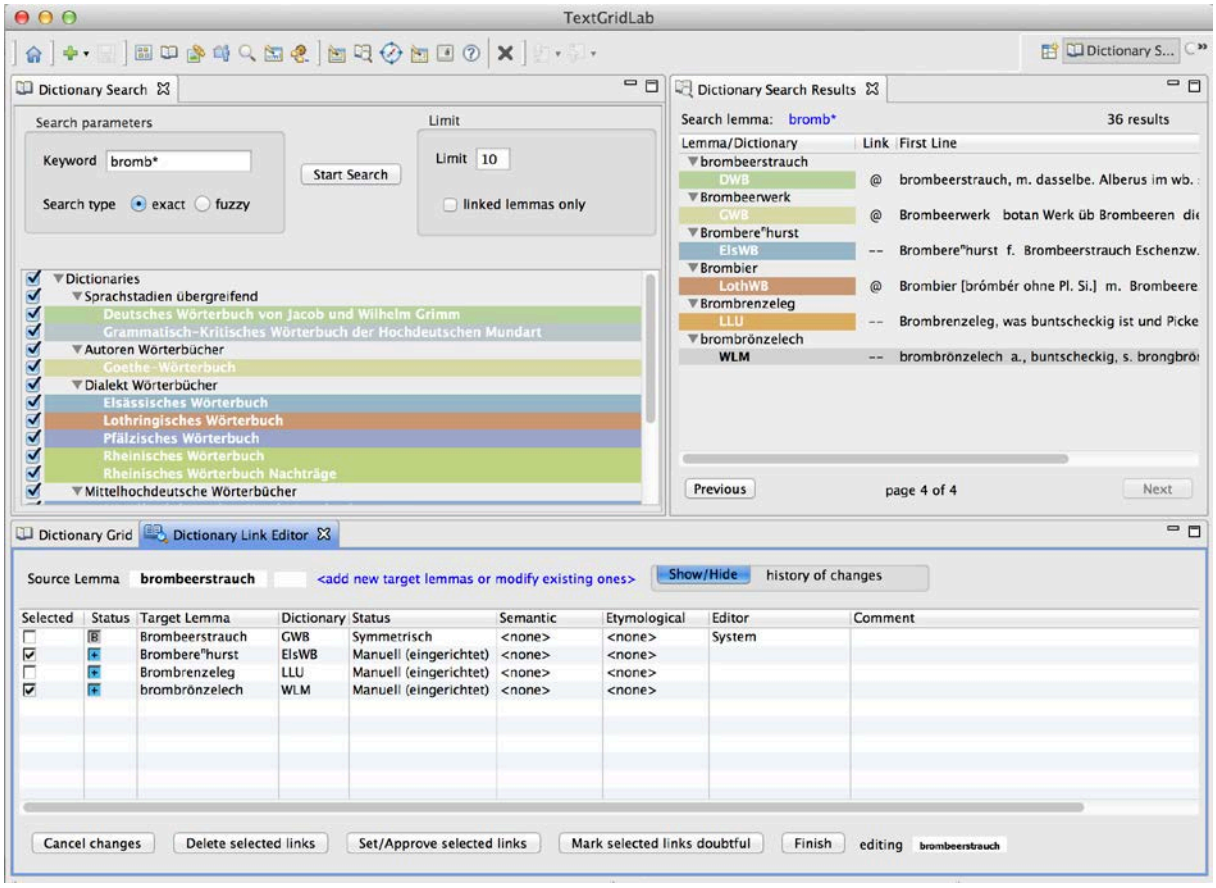


Abb. 5: Der Wörterbuch-Editor im TextGridLab. Die Ansicht zeigt einen Ausschnitt der hier zur Verfügung stehenden Wörterbücher mit eine exemplarischen Suchanfrage ‘bromb*’, deren Ergebnisse auf der rechten Seite dargestellt werden. Der untere Abschnitt bietet einen Blick auf das (nicht implementierte) Modul des ‘Dictionary Link Editor’, über das weitere Verknüpfungen händisch vorgenommen werden können.

Das Suchergebnis enthält in der zweiten Spalte den Hinweis auf bereits vorhandene Verweise. Ein Bearbeiter kann dank dieser Information zusätzliche Verweise eintragen und nach ihrer Art spezifizieren (in der gezeigten Ansicht stehen als Optionen *semantic* und *etymological* als Kategorien zur Verfügung, eine Individualisierung ist möglich). Auch ein kollaborativer Editionsprozess ist möglich, sodass mehrere Bearbeiter an einer Optimierung oder spezifischen Erweiterung der Verweisstruktur, z.B. einer Ausdehnung auf onomasiologische Verweise, arbeiten können. Weitere naheliegende Möglichkeiten wären Verknüpfungen von Wortfeldern oder die Zuordnung zu Wortfamilien.[12]

[11] Ein Download des TextGridLab wird auf der Projektseite von TextGrid angeboten: www.textgrid.de/registrierungdownload/download-und-installation/.

[12] Zur Identifizierung von Wortfamilien liegen aus dem Verbundprojekt Vorarbeiten vor, die aus dem dort entwickelten Annotationstool hervorgehen: siehe Borek/Rapp (2013, S. 24f.).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Das Tool ermöglicht jedoch nicht nur das Hinzufügen weiterer Informationen, es bietet ebenso die Möglichkeit, vorhandene Verweise zu bearbeiten, zu revidieren, zu klassifizieren oder zu kommentieren. Zusammenfassend lässt sich feststellen, dass über das Bearbeiten von Verweisen nicht nur eine Optimierung der Verweisstruktur erreicht wird, sondern auch die Einbeziehung von Nutzern möglich wird – ein Umstand, der bei retrodigitalisiertem Material dieser Art für gewöhnlich nicht ohne Weiteres gegeben ist.

4. Fazit und Ausblick

In diesem Beitrag wurde das Konzept der Metalemmaliste beschrieben, wie es im Projekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“ entwickelt und für das Erschließen sprachlicher Varianz verwendet wurde. Es wurde dargelegt, wie die Metalemmaliste dazu beiträgt, vorhandenes und vernetztes Wörterbuchmaterial weiter zu vernetzen. Doch müssen sich die entstehenden Verbindungen und Verweise keineswegs auf das Material des Trierer Wörterbuchnetzes beschränken. Eine Ausdehnung auf weitere Wörterbuchprojekte, z.B. aktuelle Online-Wörterbücher kann z.B. zu einer eleganten Verbindung von retrodigitalisiertem Wissen und gegenwartssprachlichem und modern aufbereitetem Sprachmaterial führen. Auch eine Verwendung außerhalb rein lexikografischer Interessen ist denkbar, z.B. der Einsatz in Editionsprojekten oder mittels einer Erweiterung auf Fremdsprachen. Auch zur Generierung von Beispielen für Lehrmaterial kann sich die Metalemmaliste als hilfreiche Quelle darstellen (z.B. für sprachliche Phänomene, Lautgesetze, Ausnahmen etc.). Zumindest für Verknüpfungen semasiologischer Art bieten die Metalemmata produktive Schnittstellen, über die effizient ganze Netze „mitverknüpft“ werden.

Aus den hier nur kurz dargelegten, projektspezifischen Gründen ist die Metalemmaliste rein semasiologisch angelegt.[13] Dies bedeutet aber nicht, dass sämtliche Verweise semasiologischer Natur sein *müssen*. Es ließen sich ebenso onomasiologische Bezüge herstellen oder Verweise auf andere Sprachen generieren. Das so entstehende Netz bedarf freilich eines kategorisierten Verweissystems, um die Art des Verweises sowie seine Entstehungsmethode kenntlich zu machen.

Die Schnittstellenfunktion, die die Metalemmaliste einnimmt, ermöglicht das Vernetzen verschiedener Informationssysteme im Sinne von Linked Open Data (LOD). Sie eröffnet damit das Bilden einer neuen Ressource und bietet vielfältige Abfragemöglichkeiten. Die Metalemmaliste führt zu einer neuen Dimension der Verknüpfung. Dabei stellen die Liste selbst sowie das aus ihr resultierende Wissensnetz neue Forschungsgegenstände dar und erweitern gleichzeitig das Spektrum zum Adressieren vorhandener Fragestellungen. Aus dem Entstehen dieser sprachwissenschaftlich fundierten LOD ergibt sich weiteres Potential: das Verwenden und Teilen von Ressourcen verknüpft nicht nur die jeweiligen Daten verschiedener Forschergruppen, es kann auch zur besseren Vernetzung von Forschungsprojekten und Fachwissenschaftlern beitragen. Durch die Integration in eine Infrastruktur, die z.B. über eine Nutzerverwaltung mit Rollensystem verfügt, lassen sich Forschungsdaten kollaborativ nutzen. Ein wichtiger Faktor ist dabei die Referenzierbarkeit der Daten, die über standardisierte Metadaten verfügen müssen und deren dauerhafte Erreichbarkeit gewährleistet ist.

Ein großer Nutzen ergibt sich zum Beispiel für Handschriften- und Editionsprojekte. Hier kann bei der Transkription jede Variante auf ein Metalemma gemappt und an ein Wörterbuch

[13] Für den Projektkontext siehe Borek/Rapp (2013).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

angeschlossen werden. Diese naheliegende Verknüpfungsart wird flankiert durch das Anschließen der Bildebene einerseits, indem der transkribierte Text mit seiner Image-Entsprechung auf der Handschriftenseite assoziiert wird, und dem Anschließen von weiteren Ressourcen (sog. „Weltwissen“). Auf diese Art entsteht eine multimodale Edition, die ihrerseits Teil einer Wissensbasis für andere Quellen und Ressourcen wird.

5. Elektronische Ressourcen

DEREKO: Das Deutsche Referenzkorpus. www.ids-mannheim.de/kl/projekte/korpora/archiv.html (Stand: 3.9.2014).

DEREWO: Korpusbasierte Grund-Wortformenlisten. http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html (Stand: 1.9.2014).

DWDS: Das Digitale Wörterbuch der deutschen Sprache. www.dwds.de (Stand: 1.9.2014).

P5: Guidelines for electronic text encoding and interchange. www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (Stand: 3.9.2014).

TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften. Seit 2009. www.textgrid.de (Stand 1.9.2014).

Trierer Wörterbuchnetz: Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften. Universität Trier. www.woerterbuchnetz.de (Stand: 1.9.2014).

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen. Gefördert vom Bundesministerium für Bildung und Forschung (BMBF) von 2008-2012. www.sprache-und-genome.de (Stand: 1.9.2014).

6. Literatur

Büdenbender, Stefan (2011): LexicoLux: EDV-philologische Perspektiven bei der Erstellung eines Wörterbuchnetzes der Großregion. In: Gilles, Peter/Wagner, Melanie (Hg.): Linguistische und soziolinguistische Bausteine der Luxemburgistik. (= Mikroglottika 4). Frankfurt a.M. u.a., S. 261-274.

Borek, Luise/Rapp, Andrea (2013): Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen. Darmstadt. http://dx.doi.org/10.2314/GBV:780785878 (Stand: 3.9.2014).

Burch, Thomas/Fournier, Johannes/Gärtner, Kurt (2000): Werkzeuge für Edition und Übersetzung. Mittelhochdeutsche Wörterbücher im elektronischen Verbund. Zur CD-ROM mit den wichtigsten Wörterbüchern zum Mittelhochdeutschen. In: editio 14, S. 117-129.

Burch, Thomas/Rapp, Andrea (2007): Das Wörterbuch-Netz. Verfahren – Methoden – Perspektiven. In: Burckhardt, Daniel/Hohls, Rüdiger/Prinz, Claudia (Hg.): Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006. (= Historisches Forum 10.1). Berlin, S. 607-627. http://edoc.hu-berlin.de/histfor/10_I/PHP/Woerterbuecher_2007-10-I.php#007001 (Stand: 1.9.2014).

Campe, Joachim Heinrich (1807-1811): Wörterbuch der deutschen Sprache. 5 Bde. Braunschweig.

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4., überarb. Aufl. Tübingen.

Flicek, Paul et al. (2012): Ensembl 2012. In: Nucleic Acids Research 40, Database issue: D84-D90 doi:10.1093/nar/gkr991. (Stand: 3.9.2014).

Hildenbrandt, Vera (2011): TEI-basierte Modellierung von Retrodigitalisaten (am Beispiel des Trierer Wörterbuchnetzes). In: Klosa, Annette/Müller-Spitzer, Carolin (Hg.): Datenmodellierung für Internetwörterbücher. 1. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. (= OPAL 2/2011). Mannheim, S. 21-35. http://pub.ids-mannheim.de/laufend/opal/pdf/opal2011-2.pdf (Stand: 8.9.2015).

Hockey, Susan (2004): The history of humanities computing. In: Schreibman, Susan/Siemens, Ray/Unsworth, John (Hg.): A companion to digital humanities. Oxford. http://digitalhumanities.org/companion (Stand: 3.9.2014).

Klappenbach, Ruth/Steinitz, Wolfgang (Hg.) (1961-1977): Wörterbuch der deutschen Gegenwartssprache. 6 Bde. Berlin.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich et al. (Hg.): Lexikographica. Berlin/New York, S. 79-93. www.dwds.de/static/website/publications/text/KleinGeykenDWDS.pdf (Stand: 1.9.2014).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Leuk, Michael (2012): Dictionary link editor (unveröffentlicher Prototyp).

Wooldridge, Russon (2004): Lexicography. In: Schreibman, Susan/Siemens, Ray/Unsworth, John (Hg.): A companion to digital humanities. Oxford. http://digitalhumanities.org/companion (Stand: 3.9.2014).

Seipel, Dietmar/Wegstein, Werner (2011): metaDictionary. Towards a generic e-Infrastructure for detecting variation in language by exploiting dictionaries. Proceedings of the International Symposium on Grids and Clouds (ISGC). http://www1.pub.informatik.uni-wuerzburg.de/databases/papers/isgc_2011.pdf (Stand: 17.4.2015).

Stadler, Heike/Wegstein, Werner (2012): Why and how to encode word structures and word-formation formula of a word family dictionary? A proposal based on <etym>. Annual Conference and Members Meeting of the TEI Consortium 2012, Texas AM, College Station.

Stadler, Heike (2013): Korpusbasierte Wortgrundformenliste. DEREWO v-ww-bll-320000g-2012-12-31-1.0. Benutzerdokumentation. Mannheim. www.ids-mannheim.de/derewo (Stand: 17.4.2015).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Vielfalt alignieren: ein halbautomatisches Werkzeug zum Erschließen varianter Lemmata in elektronischen Wörterbüchern

Dietmar Seipel (*Universität Würzburg*), *Luise Borek* (*Technische Universität Darmstadt*)

1. Einführung

Historische Wörterbücher bilden mit dem in ihnen verzeichneten synchronen und diachronen Wortschatz sprachliche Strukturen ab. Für das interdisziplinäre Forschungsprojekt „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen“,[1] gefördert vom BMBF von 2008 bis 2012, stellen sie daher die sprachwissenschaftliche Ressource der Analyse dar. Basierend auf verschiedenen elektronischen Wörterbüchern wurde ein MetaDictionary (Metalemmaliste) entwickelt, das der Erschließung und Analyse sprachlicher Vielfalt dient.

Das verwendete Wörterbuchkorpus umfasst retrodigitalisierte, historische Wörterbücher, die am Trier Center for Digital Humanities (TCDH) über das „Trierer Wörterbuchnetz“[2] öffentlich zur Verfügung stehen. Ergänzend konnte mit dem „Wörterbuch der Deutschen Sprache“ von Joachim Heinrich Campe und dem „Wörterbuch der deutschen Gegenwartssprache“ (WDG; Klappenbach/Steinitz 1977),[3] digitalisiert durch die Berlin-Brandenburgische Akademie der Wissenschaften, auf weitere einschlägige Wörterbücher des Deutschen zurückgegriffen werden. Damit liegt Material vor, das im Hinblick auf enthaltene sprachliche Vielfalt eine repräsentative Datenbank darstellt, auf deren Grundlage innerhalb des Projektes Methoden und Algorithmen zum Erschließen und Untersuchen von Varianz entwickelt und erprobt wurden.

Die sprachliche Vielfalt bezieht sich in diesem Kontext auf synchrone orthografische Varianten ebenso wie auf regionale und diachrone Phänomene. Aufgrund der nicht normierten Schreibung findet sich in historischen Wörterbüchern eine Vielzahl von Varianten belegt, deren Erschließung ein Hauptinteresse unserer Untersuchung darstellt. Die hier präsentierten Alignment-Verfahren bilden dabei einen wichtigen Teilaspekt zur Generierung des MetaDictionaries,[4] indem sie helfen, historische und dialektal bedingte Varianten automatisch aufzufinden und auszuzeichnen. Für diesen Zweck werden die vorhandenen Wörterbücher nicht vollständig miteinander aligniert, sondern zunächst stringbasiert auf der Lemmaebene verglichen. Über eine parallel stattfindende Auswertung der Wörterbuchartikel fließen enthaltene Informationen in die Analyse mit ein und dienen iterativ der Verbesserung des Ausgangsalignments.

In diesem Paper stellen wir unser Alignment auf Basis von mittelhochdeutschen Wörtern aus dem Mittelhochdeutschen Wörterbuch von Matthias Lexer (im Folgenden ‘Lexer’) vor, die mit diesem Verfahren ihren neuhochdeutschen Entsprechungen aus dem Deutschen Wörter-

[1] Das Projekt wurde im Rahmen der Förderrichtlinien des Bundesministeriums für Bildung und Forschung (BMBF) „Wechselwirkungen zwischen Natur- und Geisteswissenschaften“ gefördert. Siehe www.sprache-und-genome.de/.
[2] Vgl. www.woerterbuchnetz.de.
[3] Wir danken der BBAW für das Bereitstellen der ursprünglichen WDG-Daten, die die Grundlage für „Das Digitale Wörterbuch der deutschen Sprache“ (DWDS) bilden. Siehe www.dwds.de/.
[4] Vgl. hierzu den Beitrag zur „Metalemmaliste als Tool zum Erschließen von sprachlicher Varianz“ in diesem Band sowie Seipel/Wegstein (2011).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

buch der Brüder Grimm (im Folgenden ‘DWB’) zugeordnet werden. Wir weisen darauf hin, dass es sich dabei nicht um *Übersetzungen* im allgemeinsprachlichen Sinn handelt, da die Zuordnung nicht semantisch, sondern rein semasiologisch erfolgt. Für Hinweise auf einen möglichen Bedeutungswandel stehen die einzelnen Wörterbuchartikel zur Verfügung, deren Lemmata durch das Alignment der Stichwörter, bzw. deren Verlinkung, miteinander assoziiert werden.

Vor dem Hintergrund eines auf wechselseitigen Austausch zwischen verschiedenen Disziplinen bedachten Vorhabens standen bei der Entwicklung sowohl philologische als auch informatische Interessen im Vordergrund. Während aus sprachwissenschaftlicher Sicht eine möglichst geringe Fehlerquote sowie händische Eingriffsmöglichkeiten erforderlich sind, ist es ein zusätzlicher Anspruch der Informatik, diese Ziele möglichst effizient zu erreichen und die erforderliche Rechenzeit niedrig zu halten. Das hieraus resultierende Ineinandergreifen wird im Folgenden anhand unseres Werkzeugs zum Alignieren von Wörtern dargelegt.

In Abschnitt 2 wird zunächst die Datengrundlage kurz geschildert. Dabei gehen wir auf die Besonderheiten und Probleme beim Alignieren retrodigitalisierter Wörterbücher ein und erläutern die Auswahl unseres Beispiels sowie die Konzeption des Ausgangs-Alignments.

Anschließend beschreiben wir in Abschnitt 3 die Vorgehensweise des Alignment-Tools sowie integrierte und optionale Optimierungen. Weitere Ergänzungen in Form von extern generierten Vergleichslisten und Einschränkungsmöglichkeiten der Lemmaauswahl innerhalb des Alignierungsprozesses finden sich in Abschnitt 4, bevor im Fazit die Ergebnisse zusammengeführt und mit einem kurzen Ausblick resümiert werden.

2. Alignment und elektronische Wörterbücher

In unserem Paper geht es nicht um das vollständige Alignieren zweier Wörterbücher. Vielmehr greift das Alignment auf vergleichbare Einheiten der ausgezeichneten Artikelstrukturen zurück. Unser Hauptinteresse gilt dabei den Lemmata, die wir auf diese Art einander zuordnen wollen; Ziel ist somit das Alignment einzelner Wortpaare aus den beteiligten Wörterbüchern. Die dabei auftretenden Anforderungen sind jedoch deutlich komplexer, da es auch darum geht, möglichst schnell zu einem Wort des Ausgangswörterbuchs geeignete Kandidaten des Zielwörterbuchs zu finden, unter denen das entsprechend einer Ähnlichkeitsbewertung optimal passende Wort des Zielwörterbuchs sicher enthalten ist.

2.1 Zur Datengrundlage

Das Trierer Wörterbuchnetz, dessen Material einen Großteil unserer Untersuchung ausmacht und auf das wir mit dem Alignment aufbauen, weist bereits ein System von Verweisstrukturen zwischen den Wörterbüchern – meist derselben Sprachstufe – auf. Der wesentliche Schritt der Zuordnung bei den miteinander verlinkten Wörterbüchern ist zwischen dem Verbund der mittelhochdeutschen Wörterbücher und denen jüngerer Sprachstufen (Adelung, DWB, Dialektwörterbücher) zu leisten. Diesem Schritt galt daher unsere besondere Aufmerksamkeit und wir führen in diesem Paper stellvertretend für das weitere Datenmaterial Beispiele aus seinem Kontext an. Frühere Sprachstufen eignen sich auch deshalb besonders als Gegenstand für die Untersuchung von Varianz, weil sie aufgrund nicht vorhandener Normierung eine Vielzahl von Varianten aufweisen.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Für die Zuordnung durch die hier dargelegten Alignment-Verfahren dient das Wörterbuch von Matthias Lexer aufgrund seines – zu diesem Zweck günstigeren – Lemma-Ansatzes bei Verben als mittelhochdeutsche Grundlage. Stehen diese im BMZ in der ersten Person Singular, wird im Mittelhochdeutschen Wörterbuch von Lexer die Infinitivform der Verben herangezogen. Die elektronische Version des Wörterbuchs fungiert damit als Schnittstelle für alle vorliegenden mittelhochdeutschen Wörterbücher, da diese im Verbund untereinander vernetzt sind. Das DWB eignet sich aufgrund seiner frühen neuhochdeutschen Lemmaansätze besonders als Alignierungspartner, da der zu bewältigende Schritt etwas geringer ist als er es im Vergleich zur Gegenwartssprache wäre. Gleichzeitig bietet sein Umfang von ca. 300.000 Stichwörtern ausreichend Material, um auch archaisierende oder seltenere Wörter auffinden und zuordnen zu können. Für jedes Lemma aus dem Lexer suchen wir nach dem ähnlichsten Lemma im Grimm'schen Wörterbuch. Der Vergleich aller Paare (≈ 80.000 × 300.000) erfordert sehr viel Rechenzeit und verdeutlicht – besonders in Hinblick auf Verknüpfungs-, Visualisierungs- und Analysemöglichkeiten – die Dimensionen geisteswissenschaftlicher ‘Big Data’. Ergänzend finden sich im Etymologie-Teil der Artikel häufig Angaben zu mittelhochdeutschen Formen, die zusätzlich für die Analyse berücksichtigt werden (siehe Abschnitt 4.1).

2.2 Ausgangs-Alignment

Für das spätere Alignment ist eine Bewertung erforderlich, aufgrund derer die automatische Alignierung erfolgen kann. Diese Werte mussten zunächst anhand einer repräsentativen Auswahl voralignierter Lexeme ermittelt werden. Hierfür wurden Zufallsmengen aus der Lemmaliste des Lexer extrahiert und nach bestimmten Kriterien (in kurzer Form wird in Abschnitt 2.2 darauf eingegangen) auf ihr semasiologisch gleiches neuhochdeutsches Pendant manuell aligniert. Entsprechend konnten aus der Zufallsmenge nur solche Lemmata verwendet werden, die auch in (früh)neuhochdeutscher Form existieren und im DWB belegt sind. In Einzelfällen führte dies dazu, dass Komposita aus der Zufallsmenge als Einzelglieder aligniert wurden, um ein gültiges Pärchen zu bilden, dessen Alignment für die Berechnung von Wahrscheinlichkeiten verwendet werden kann. Auf die spätere Verlinkung hat dieses Vorgehen keinen Einfluss, da das für das manuelle Alignment verwendete Material noch nicht der Verlinkung, sondern zunächst der Generierung von Ausgangswerten für das automatische Alignment dient. Hierfür ist die korrekte Zuordnung auf Wortebene zunächst wichtiger als die Existenz korrespondierender Lemmata und Artikel der betrachteten Wörterbücher. Die so erstellte Liste umfasst ca. 2.000 alignierte Lemmata.

Ausgangs-Alignment zur Ermittlung der Ähnlichkeit von Buchstabenfolgen: Das manuelle Alignment wurde pragmatisch in einer fragmentarischen XML-Struktur vorgenommen. Unter einem alignment-Element gruppieren sich das mittelhochdeutsche und das neuhochdeutsche Lemma. Laute, die durch Vokal- oder Konsonantenverbindungen repräsentiert sind, werden als eigene Zeichen definiert, um ihre Bewertungen und die ihrer Einzelbestandteile nicht zu beeinträchtigen. Lautgruppierungen wurden mittels eckiger Klammern, Leerstellen mit ‘%’ realisiert:

```
<alignment>
       <lemma> ungel [au]pli[ch] </lemma>
       <lemma> ung%l[au]bli[ch] </lemma>
</alignment>
```

Aus der Testmenge der Wortpaare wurde sodann eine *Substitutionsmatrix* ermittelt, die angibt, wie häufig eine Buchstabenfolge in einem mittelhochdeutschen Wort durch eine Buchstabenfolge im entsprechenden neuhochdeutschen Wort ersetzt wird.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Die aus dem Ausgangs-Alignment resultierende *Substitutionsmatrix* enthält Ähnlichkeitswerte von Buchstabenfolgen:

Der Matrixeintrag $M(v, w) = p$ gibt die relative Häufigkeit der Substitution einer Buchstabenfolge *v* (zu sehen auf der y-Achse) durch eine Buchstabenfolge *w* (zu sehen auf der x-Achse) an. Der Eintrag $M(uo, u) = 0{,}8148$ in der Substitutionsmatrix in Abbildung 1 besagt beispielsweise, dass ein *uo* beim Übergang vom Mittelhochdeutschen zum Neuhochdeutschen in 81,48% aller Fälle durch ein *u* substituiert wird.

	ϵ	*a*	*b*	…	*u*	*uo*	…
ϵ	0	0,0104	0,0311	…	0	0	…
a	0,0025	0,8806	0	…	0,0025	0	…
b	0,0321	0	0,9231	…	0	0	…
uo	0	0	0	…	0,8148	0	…
…	…	…	…	…	…	…	…

Abb. 1: Substitutionsmatrix

Wir betrachten hier bei beiden Sprachstufen dieselben 87 Buchstabenfolgen. Die Matrix ist sehr dünn besetzt: von den 87 × 87 = 7.569 Einträgen sind nur 370 positiv. ϵ bezeichnet hier das leere Wort. Der Eintrag $M(a, \epsilon) = 0{,}0025$ besagt, dass 0,25 % der Vorkommen des Buchstabens *a* im Mittelhochdeutschen beim Übergang zum Neuhochdeutschen durch das leere Wort ersetzt werden – alternativ könnte man sagen, dass sie gelöscht werden. Meistens – in 88,06 % der Fälle – bleibt ein *a* aber ein *a*. Die Zeilen- und Spaltensummen sind jeweils 1. In etwa 11,44 % der Fälle verändert sich ein *a* also in eine hier nicht aufgelistete Buchstabenfolge.

3. Das Alignment-Tool

Die Größe und die Heterogenität der zu untersuchenden Datenmenge erfordern ein computergestütztes Verfahren, um ihre Erschließung zu ermöglichen. Lexikografische Regelmäßigkeiten, deren Auszeichnung in XML sowie die Regelbasiertheit des Forschungsgegenstandes *Sprachsystem* begünstigen dieses Vorgehen. Es ist jedoch zu beachten, dass all diese Faktoren fehleranfällig sind, da in einer natürlichen Sprache Ausnahmen und Vielfalt regelmäßig vorkommen. Mithilfe des Alignment-Tools lassen sich die wahrscheinlichsten Kandidaten für Varianten historischen, regionalen oder orthographiebedingten Ursprungs ermitteln. Das Werkzeug verknüpft lexikografische, linguistische und informatische Grundannahmen und ist doch generisch ausgelegt, indem sein Anwendungsbereich von dem vorgegebenen Ausgangs-Alignment abhängt und es weitere Einstellungen erlaubt. Auf das Vorgehen des Tools wird in diesem Abschnitt näher eingegangen.

3.1 Alignment von Wörtern

Alignment: Ein Alignment von zwei Buchstabenfolgen (Strings) y und x ist eine Segmentierung in (kleinere) Buchstabenfolgen y_i und x_i mit $y = y_1 y_2 \dots y_n$ und $x = x_1 x_2 \dots x_n$. Die Bedeutung des Alignments ist, dass durch die Ersetzung von y_i durch x_i der String y zum String x wird. Manche Buchstabenfolgen y_i oder x_i können leer sein; natürlich ergibt es aber keinen Sinn, Ersetzungen der Form $\in \mapsto \in$ zu betrachten, da diese redundant wären.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Ähnlichkeit: Die Ähnlichkeit *sim*(*y*,*x*) von *y* und *x* ergibt sich aus einem optimalen Alignment, bei dem das Produkt $M(y_1, x_1) \cdot M(y_2, x_2) \cdot \ldots \cdot M(y_n, x_n)$ maximal ist:

$$sim(y, x) = max_{y=y1\ y2\ldots y_n, x=x1x2\ldots x_n} M(y_1, x_1) \cdot M(y_2, x_2) \cdot \ldots \cdot M(y_n, x_n).$$

Wir gehen also von einer Unabhängigkeitsannahme aus, die besagt, dass die Ersetzungen $y_i \mapsto x_i$ unabhängig voneinander erfolgen – sonst könnte man *sim*(*y, x*) nicht als das angegebene Produkt berechnen.

Beispiel. Die Ähnlichkeit des mittelhochdeutschen Wortes *y* = *vluoch* und des neuhochdeutschen Wortes *x* = *fluch* ergibt sich als

$$sim\left(\overbrace{vluoch}^{mhd}, \overbrace{fluch}^{nhd}\right) = \overbrace{0{,}5875}^{M(v,f)} \cdot \overbrace{0{,}9599}^{M(l,l)} \cdot \overbrace{0{,}8148}^{M(uo,u)} \cdot \overbrace{0{,}9239}^{M(ch,ch)} = 0{,}4246$$

aus dem Alignment

$$y_1 = v, x_1 = f, y_2 = x_2 = l, y_3 = uo, x_3 = u, y_4 = x_4 = ch.$$

Für das alternative Alignment

$$y_1 = v, x_1 = f, y_2 = x_2 = l, y_3 = uo, x_3 = u, y_4 = x_4 = c, y_5 = x_5 = h,$$

bei dem *c* und *h* separat aligniert werden, ergibt sich die Ähnlichkeit 0, da *M*(*c, c*) = *0*.

Alignment und Wörterbücher: Das Ziel ist nun, zu jedem der etwa 80.000 Einträge *y* des Lexer den ähnlichsten der etwa 300.000 Einträge *x* im DWB zu finden; das ist der Eintrag *x* mit der maximalen Ähnlichkeit *sim*(*y, x*). Dabei wird die Ähnlichkeit – wie oben beschrieben – selbst wieder durch ein optimales Alignment berechnet.

Lösungsansatz: Es wurden bekannte Algorithmen aus der Informatik verwendet, die auch in der Bioinformatik bei der Genomanalyse zum Einsatz kommen. Die Problematik ist allerdings eine etwas andere, da wir die Lexempaare (*y, x*) möglichst nicht komplett alignieren wollen. Man kann die Suche abbrechen, sobald ein Teilalignment von *y* und *x* schlechter ist als der Ähnlichkeitswert des besten bisher gefundenen Alignments oder als eine vorgegebene Interessantheitsschranke M.

Zur Berechnung der Ähnlichkeit werden die Einträge $M(y_i, x_i)$ der Substitutionsmatrix negativ logarithmiert. Dann gilt

$$M'(y, x) = -\log\big(M(y_1, x_1) \cdot M(y_2, x_2) \cdot \ldots \cdot M(y_n, x_n)\big) = -\sum_{i=1}^{n} logM(y_i, x_i)$$
$$= \sum_{i=1}^{n} -logM(y_i, x_i) = \sum_{i=1}^{n} M'(y_i, x_i).$$

Also kann man das Ähnlichkeitsproblem auf ein Problem *kürzester Wege* in einem geeigneten kantengewichteten Graphen zurückführen und mit dem *Algorithmus von Dijkstra* lösen. Alternativ kann man eine Methode der dynamischen Programmierung aus der Mathematik verwenden, die in der Bioinformatik zur Alignierung von DNS-Sequenzen verwendet wird. In beiden Fällen muss versucht werden, ein Alignment der beiden Lexeme zu finden, sodass *M'*(*y, x*) minimal wird.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Wenn man naiv alle Lexempaare auf ihre Ähnlichkeit hin testet, dann liegt die Rechenzeit – je nach Leistungsfähigkeit des Computers – bei einigen Wochen. Wir konnten die Methoden geeignet verfeinern und verbessern, so dass es jetzt innerhalb von wenigen Tagen möglich ist, zu jedem *y* aus dem Lexer das ähnlichste Wort *x* aus dem DWB zu finden. Die wesentliche Beschleunigung des Verfahrens ergab sich durch den frühzeitigen Abbruch der Suche bei zu geringer Ähnlichkeit; diese ließ sich für den Dijkstra-Algorithmus besser umsetzen als für die dynamische Programmierung; sie machte den Dijkstra-Algorithmus zum schnelleren Verfahren.

Alignmentmethode: Im Folgenden erklären wir die Alignmentmethode anhand des mittelhochdeutschen Wortes *y = vluoch* und des neuhochdeutschen Wortes *x = fluch*. Es wird eine Alignmentmatrix, siehe Abbildung 2, durchlaufen, beginnend mit dem Kästchen links oben, das im Bild mit einem roten Pfeil markiert ist und das die Ähnlichkeit 1 enthält.

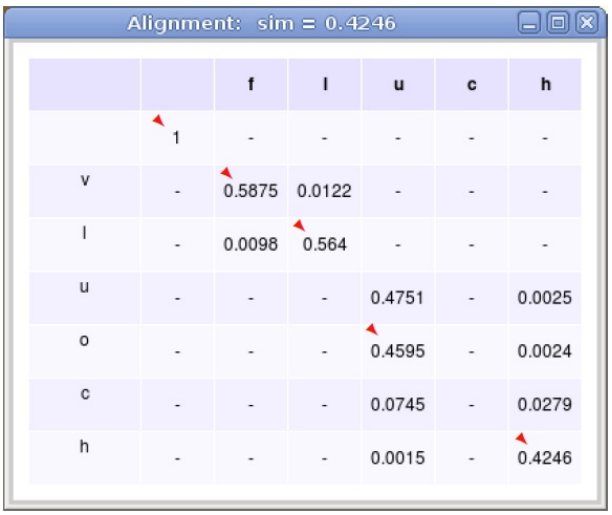


Abb. 2: Alignmentmatrix

In jedem Einzelschritt kann der Algorithmus einen oder mehrere Buchstaben aus der ersten Spalte – beginnend mit dem obersten Buchstaben *v* – und der ersten Zeile – links beginnend mit dem Buchstaben *f* – konsumieren. Wenn der erste Einzelschritt z.B. *v* und *f* konsumiert, dann zieht er zum Kästchen mit den Koordinaten (*3, 3*) weiter, das auch mit einem roten Pfeil markiert ist, und das die Ähnlichkeit $M(v, f) = 0{,}5875$ enthält.

- Nun kann der Algorithmus den Buchstaben *l* auf der y-Achse lesen, ohne etwas auf der x-Achse zu lesen. Dies würde bedeuten, dass das *l* aus dem Mittelhochdeutschen gelöscht wird – man könnte auch sagen, dass es durch das leere Wort $\in$ ersetzt wird – wenn man zum Neuhochdeutschen übergeht, was mit $M(l, \epsilon) = 0{,}01668$ passiert. Dann haben wir das Kästchen (*3, 4*) erreicht mit $sim(vl, f) = 0{,}0098$.
- Entsprechend kann der Algorithmus auch *l* auf der x-Achse konsumieren ohne etwas auf der y-Achse zu konsumieren. Dies würde bedeuten, dass beim Übergang vom Mittelhochdeutschen zum Neuhochdeutschen ein *l* eingefügt wird, was mit $M(\epsilon, l) = 0{,}02076$ passiert. Dann haben wir das Kästchen (*4, 3*) erreicht mit $sim(v, fl) = 0{,}0122$.
- Außerdem kann der Algorithmus *l* auch auf beiden Achsen konsumieren. Dies würde bedeuten, dass das *l* aus dem Mittelhochdeutschen mit dem *l* aus dem Neuhochdeutschen aligniert wird, was mit $M(l, l) = 0{,}96$ passiert. Dann haben wir das Kästchen (*4, 4*) mit der bisher besten Ähnlichkeit $sim(vl, fl) = 0{,}564$ ermittelt.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Ausgehend von (*4, 4*) gelangt man in einem nächsten Schritt zur besten Ähnlichkeit, indem man die mittelhochdeutschen Buchstaben *u* und *o* zusammen auf der x-Achse konsumiert und sie durch den einen neuhochdeutschen Buchstaben *u* auf der x-Achse ersetzt. Dann erreicht man (*5, 6*) mit $sim(vluo, flu) = 0{,}4595$. Die alternative Ersetzung des mittelhochdeutschen Buchstabens *u* durch den neuhochdeutschen Buchstaben *u* gefolgt von einer Löschung des nachfolgenden mittelhochdeutschen Buchstabens *o* ist unmöglich, da die Wahrscheinlichkeit einer Löschung von *o* gleich $M(o, \in) = 0$ ist. Schließlich werden die zwei mittelhochdeutschen Buchstaben *c* und *h* zusammen in einem Schritt durch dieselben neuhochdeutschen Buchstaben *c* und *h* ersetzt mit $M(ch, ch) = 0{,}9240$, um abschließend das Kästchen (*7, 8*) rechts unten mit $sim(vluoch, fluch) = 0{,}4246$ zu erreichen. Es erweist sich als günstiger, die beiden mittelhochdeutschen Buchstaben *c* und *h* gemeinsam durch die neuhochdeutschen Buchstaben *c* und *h* zu ersetzen. Die Wahrscheinlichkeit der Ersetzung von *c* durch c ist $M(c, c) = 0$. Die Ersetzung von *c* durch *c* und *h* mit $M(c, ch) = 0{,}0608$ gefolgt von der Löschung von *h* mit $M(h, \in) = 0{,}01976$ ergibt $M(c, ch) \cdot M(h, \in) = 0{,}0608 \cdot 0{,}01976 = 0{,}001201$, also einen kleineren Wert als $M(ch, ch) = 0{,}9240$.

Zugrundeliegende Theorie. Der Algorithmus von Dijkstra und die dynamische Programmierung unterscheiden sich beim Durchlauf der Alignmentmatrix. Bei der Verwendung des Algorithmus von Dijkstra fasst man die Alignmentmatrix als gerichteten, kantengewichteten Graphen auf, dessen Knoten die Kästchen der Alignmentmatrix sind und dessen Kanten sich aus den gelesenen Buchstabenfolgen ergeben. Die Dijkstra-Variante sucht dann einen kürzesten Weg von links oben nach rechts unten. In jedem Schritt werden die Nachfolger der Kästchen betrachtet, für die es bereits Eintragungen gibt. Es wird ein neuer Eintrag in den Nachfolger mit dem aktuell höchsten Wahrscheinlichkeitswert – dem niedrigsten negativen Logarithmus – gemacht. Man nennt diese Vorgehensweise ein Greedy-Verfahren. Sie garantiert, dass jede Eintragung optimal ist. Sobald das rechte untere Kästchen erreicht wird, ist die gesuchte Ähnlichkeit gefunden. Bei der dynamischen Programmierung werden dagegen die Kästchen in einer vorgegebenen Durchlaufordnung (z.B. zeilen-, spaltenweise oder diagonal links oben beginnend) systematisch angesehen. Im Gegensatz zur Dijkstra-Variante müssen hier alle Kästchen berücksichtigt werden.

3.2 Optimierungen

Es wurden verschiedene Maßnahmen getroffen, die die Qualität des Alignments verbessern und gewährleisten sollen. Idealerweise definieren sie nicht nur Grenzen, die dazu beitragen Fehler zu vermeiden, sondern führen zusätzlich zu einer größeren Effizienz, indem einige Berechnungen gar nicht erst durchgeführt werden müssen. Wenn man zu jedem der 80.000 Einträge des Lexer und jedem der 300.000 Einträge des Grimm die Ähnlichkeit berechnet, dann ergeben sich selbst für sehr schnelle Alignmentverfahren nicht tolerable Rechenzeiten.

Ähnlichkeitsschranke: Wenn man eine Schranke $M > 0$ vorgibt, dann gibt es nicht zu jedem Eintrag *y* des Lexer einen entsprechenden Eintrag *x* des Grimm, d.h. einen Eintrag mit $sim(y, x) \geq M$, und das, obwohl der Grimm etwa viermal so viele Einträge hat wie der Lexer.

Es hat sich aber bei der Entwicklung des Alignment-Tools herausgestellt, dass die Schranke *M* dazu benutzt werden kann, einen Berechnungsast bei der Alignierung frühzeitig abzubrechen, wenn die Teilalignierung der Anfangsstücke zu unähnlich ist. Anders als bei der DNS-Alignierung, ist dadurch der Algorithmus von Dijkstra besser geeignet als die dynamische Programmierung.

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Die Ergebnisse zeigen, dass die durchschnittliche Ähnlichkeit der gefundenen ähnlichsten Lexempaare mit steigender Wortlänge *N* etwa im Verhältnis *1/N* sinkt. Entsprechend sollte man die Interessantheitsschranke *M* geeignet in Abhängigkeit von der Wortlänge wählen.

Y- und Y X-Transitionen: Im Basisalgorithmus des Alignment-Tools wurden sogenannte Y-Transitionen betrachtet. Diese geben für eine Buchstabenfolge *v* an, in welche Buchstabenfolgen *w* sie sich entwickeln kann. Wenn man stattdessen zu den aktuellen Positionen in den beiden Wörtern eine passende Transition für die folgenden Buchstabenfolgen *v* und *w* sucht (Y X-Transitionen), dann kann man etwas Zeit sparen.

Dynamische Schrittweite: Eine Zeitersparnis wird auch erreicht, indem zunächst nur Transitionen für kürzere Buchstabenfolgen betrachtet werden und erst danach dynamisch die Schrittweite erhöht wird; dann kann man bereits die Ähnlichkeit des bisher gefundenen Alignments als höhere Schranke *M* benutzen, um die teurere Berechnung für größere Schrittweiten früher abbrechen zu können.

Verfeinerung des Basiskorpus: Das Basiskorpus wurde verfeinert, indem weitere Buchstabenfolgen identifiziert wurden, die sich häufig als Ganzes weiterentwickeln. Da $M(v_1 \cdot v_2, w_1 \cdot w_2) < M(v_1, w_1) \cdot M(v_2, w_2)$ sein kann, ergeben sich kleinere Ähnlichkeiten und es kommt vor, dass der ähnlichste Grimm-Eintrag zu einem Lexer-Eintrag ein anderer ist als vorher.

Die Erweiterung des Basiskorpus durch händische Auszeichnung weiterer Transitionen von Buchstabenfolgen brachte leider zunächst keine nennenswerte Verbesserung. Die automatisierte Suche nach solchen Transitionen führte zu einer erweiterten Substitutionstabelle, mit der 50% mehr Alignments mit im Schnitt 2,5 mal so großen Ähnlichkeiten gefunden werden konnten; allerdings hat sich dabei die Rechenzeit im Schnitt vervierfacht. Durch die Kombination der erweiterten Substitutionstabellen mit Y X-Transitionen, konnte die Rechenzeit dann aber wieder auf das 1,5-Fache des Ausgangswertes gedrückt werden. Die Verwendung der Y X-Transitionen brachte alleine keinen Gewinn.

Segmentierung in Morpheme und Annotation: Eine weitere Möglichkeit zum Erzielen besserer Resultate besteht, wenn einzelne Segmente für sich betrachtet werden (z.B. mithilfe von Morphemtrennstrichen der Wörterbuchausgaben)[5] – etwa ein Kompositum oder Derivat in $v = v_1 \cdot v_2$. Für diese werden dann separat die ähnlichsten Lexeme w_1 und w_2 aus dem DWB bestimmt. Dieses Verfahren hat den günstigen Nebeneffekt, dass es die Abhängigkeit von der Wortlänge erheblich minimiert. Man findet ein geeignetes Alignment, wenn $sim(v_1, w_1) \geq M$ und $sim(v_2, w_2) \geq M$ ist, selbst wenn es kein Alignment *v* für *w* gibt (z.B. da $w_1 \cdot w_2$ kein Eintrag im DWB ist) oder wenn die Ähnlichkeit die Schranke unterschreitet, d.h. wenn $sim(v_1 \cdot v_2, w_1 \cdot w_2) < M$.

Die Vorsegmentierung brachte eine Beschleunigung des Verfahrens. Dies ist insbesondere auch bei Komposita von Vorteil, wenn es zu einem Lexem *v* aus dem Lexer keine Entsprechung *w* im DWB gibt. Die Segmente können auf diese Art separat aligniert werden. Speziell beim mittelhochdeutschen Wörterbuch von Matthias Lexer stellt man fest, dass es in einigen Bereichen sehr viele Komposita aus einer Wortfamilie aufführt. Im Grimm'schen Wörterbuch gibt es dann oft keine Entsprechung. Durch die separate Alignierung konnte die Erfolgsquote hier deutlich erhöht werden.

[5] Die darüber hinaus gehenden Überlegungen richten sich nach Fleischer/Barz (2012).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Wenn die Lexeme annotiert sind, kann man durch die Betrachtung der Wortklasse ebenfalls die Liste der zu betrachtenden Lexempaare einschränken, um Rechenzeit zu sparen (siehe 4.2).

4. Lexikografische Einflüsse

4.1 Auswertung der mittelhochdeutschen Formen im DWB

In einer Vorauswahl wurde unter Anwendung des Tübinger Systems von Textverarbeitungs-Programmen (TUSTEP)[6] und PROLOG[7] eine Liste aller Artikel erstellt, in denen für das Lemma (oder Sub-Lemma) eine Angabe zu seiner mittelhochdeutschen Form vorliegt:

```
<hi rend ="italics"> mhd. </hi>
```

Im Anschluss wurde die mittelhochdeutsche Angabe isoliert und neben das Lemma gestellt. Die so entstandene Datei enthält ca. 15.500 Datensätze mit jeweils dem neuhochdeutschen Lemma und einem Kandidaten für seinen mittelhochdeutschen Ursprung. Der Etymologie-Teil ist beim Erfassen des Wörterbuchs als nicht regelmäßig enthaltener Abschnitt auch innerhalb des TEI-Markup nicht gesondert ausgezeichnet, so dass hier keine auswertbaren Elemente zur Verfügung stehen, was auf die Konzeption des Wörterbuchs zurückzuführen ist. Entsprechend führt das Extrahieren vorhandener, Layout-bezogener, Tags nicht immer zu einer eindeutigen Identifikation des mittelhochdeutschen Lemmas. Neben dem *regelmäßigen* Format

```
<entry xml:id ="GA01378" n="1.0116.5">
       <form type ="lemma"> abseite </form>
       <mhd> absîte </mhd>
</entry>
```

können auch mehrere mittelhochdeutsche Formen angeführt werden:

```
<entry xml:id ="GA01848" n="1.0167.38">
       <form type ="lemma"> achte </form>
       <mhd>ahte, ahtode</mhd>
</entry>
```

An einigen Stellen tritt auch ein Beleg an die Stelle der mittelhochdeutschen Form:

```
<entry xml:id ="GA01615" n="1.0143.11">
       <form type ="lemma"> abtreten </form>
       <mhd> eime sciltknehte wart lîhte ein spor hie ze hove
        abe getreten. </mhd>
</entry>
```

Hier war PROLOG besonders gut geeignet, um die Vielzahl der Alternativen abdecken zu können.[8]

Die Angabe unter *mhd.* stellt keinen verlässlichen Indikator für die Alignment-Analyse dar. Eine verfeinerte Anpassung der verwendeten Methoden und der Abgleich der ermittelten Kandidaten durch das bestehende Alignment führt jedoch zu einem qualitativ guten Ergebnis.

---

6 TUSTEP ist verfügbar via www.tustep.uni-tuebingen.de/.

7 PROLOG ist eine logische Programmiersprache. Die von uns verwendete Umgebung ist das frei verfügbare „SWI Prolog“: www.swi-prolog.org.

8 Für die Anwendung von PROLOG im Bereich NLP siehe auch Gazdar/Mellish (1989).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

4.2 Einschränkung der Lemmaauswahl

Part-of-Speech (*pos*): Die Wortklassenangabe der Wörterbücher wird in doppelter Hinsicht für das Alignment genutzt. Einerseits kann auf diese Art sichergestellt werden, dass ausschließlich Partner gleicher Wortklasse aligniert werden. Andererseits reduziert sich damit die Menge der Lemmata, für die wiederum Ähnlichkeiten errechnet werden müssen, was die Rechenzeit minimiert. Die verwendeten Kürzel für die Wortklassenangabe variieren nicht nur von Wörterbuch zu Wörterbuch, sondern auch innerhalb der betrachteten Wörterbücher Lexer und DWB. Beides wird aufgefangen, indem die *pos*-Angaben auf ein vereinfachtes Wortklassensystem gemappt werden.[9] Wir unterscheiden:

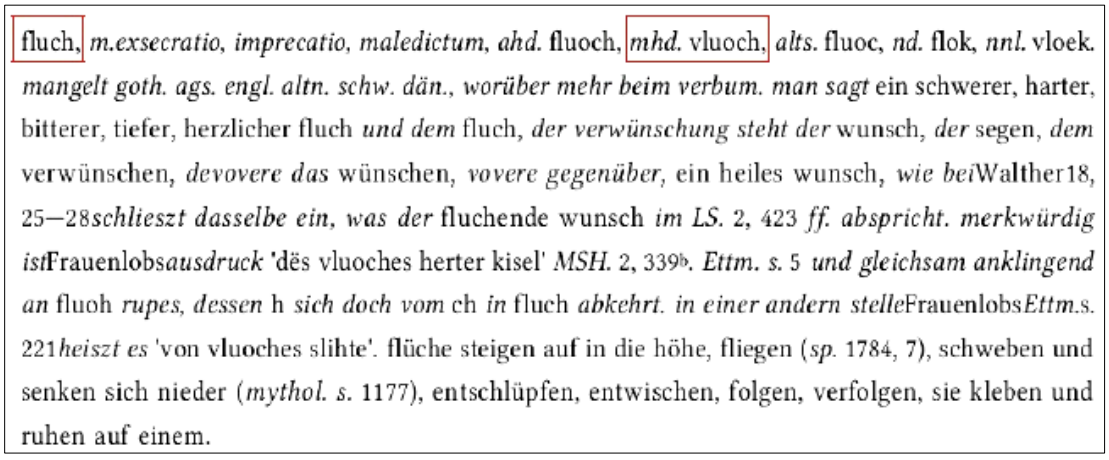
- Adjektiv, Adverb, Nomen, Partizip, Verb;
- Sonstige: Ableitungssilbe, Bildungssilbe, Fragewort, Interjektion, Konjunktion, Negationspartikel, Numerale, Präposition, Pronomen;
- keine Angabe.

Die unter „Sonstige“ zusammengefassten Angaben machen nur einen sehr kleinen Anteil verzeichneter Lemmata aus. Auch ist besonders in dieser Gruppe die Terminologie sehr heterogen, da sie nicht nur Wortklassen, sondern auch grammatische Einheiten etc. umfasst. Lemmata, denen eine *pos*-Angabe völlig fehlt, werden als einzige wortklassenübergreifend aligniert.

Anlaut: Ein weiteres Kriterium zur Reduktion der Abgleichsmenge bildet der Anlaut des betrachteten Wortes. Wenn die Angaben aus der Substitutionsmatrix auch für die Anlaute gelten, dann kann man bereits daraus ableiten, mit welchen neuhochdeutschen Wörtern ein mittelhochdeutsches Wort aligniert werden kann.

Zusammen mit der Part-of-Speech-Angabe dient auch dieses Vorgehen zur Beschleunigung des Alignment-Prozesses und zur Optimierung der Qualität, indem offensichtlich falsche Zuordnungen von Vornherein ausgeschlossen werden.

Es wurde ein Katalog erstellt, der für bestimmte Zeichen(folgen) im Anlaut des mittelhochdeutschen Wortes nur solche Strings im Neuhochdeutschen zulässt, die phonetisch und lautwandelbedingt möglich sind. Dies spezialisiert die Liste der Transitionen von Lauten, die man bereits aus der Substitutionsmatrix ableiten kann.



fluch, *m.exsecratio, imprecatio, maledictum, ahd.* fluoch, *mhd.* vluoch, *alts.* fluoc, *nd.* flok, *nnl.* vloek. *mangelt goth. ags. engl. altn. schw. dän., worüber mehr beim verbum. man sagt* ein schwerer, harter, bitterer, tiefer, herzlicher fluch *und dem* fluch, *der verwünschung steht der* wunsch, *der* segen, *dem* verwünschen, *devovere das* wünschen, *vovere gegenüber,* ein heiles wunsch, *wie bei*Walther18, 25—28*schlieszt dasselbe ein, was der* fluchende wunsch *im LS.* 2, 423 *ff. abspricht. merkwürdig ist*Frauenlobs*ausdruck* 'dës vluoches herter kisel' *MSH.* 2, 339ᵇ. *Ettm. s.* 5 *und gleichsam anklingend an* fluoh *rupes, dessen* h *sich doch vom* ch *in* fluch *abkehrt. in einer andern stelle*Frauenlobs*Ettm.*s. 221*heiszt es* 'von vluoches slihte'. flüche steigen auf in die höhe, fliegen (*sp.* 1784, 7), schweben und senken sich nieder (*mythol. s.* 1177), entschlüpfen, entwischen, folgen, verfolgen, sie kleben und ruhen auf einem.

Abb. 3: mhd. *vluoch* → frnhd. *fluch*

[9] Die im Alignment verwendeten Kürzel richten sich nach GOLD – General Ontology for Linguistic Description (2014).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

5. Fazit

In diesem Beitrag haben wir ein Alignment-Verfahren skizziert, das Wörterbücher verschiedener Sprachstufen (im beschriebenen Beispiel der mittelhochdeutschen und neuhochdeutschen) auf der Ebene ihrer Stichwörter vergleicht und den (bzw. die) ähnlichsten Kandidaten ermittelt. Die Ähnlichkeitsbetrachtung erfolgt stringbasiert und ist durch die integrierten linguistischen Regeln semasiologisch ausgerichtet. Das Verfahren bedarf einiger Vorarbeiten – in Form eines händisch erzeugten, idealen Alignments – aus denen die Ähnlichkeitswerte für das automatische Verfahren hervorgehen. Für den Übergang vom Mittelhochdeutschen zum Neuhochdeutschen lassen sich die bekannten Phänomene darin abbilden. Trotz der sprachlichen Vielfalt erweisen diese sich für die im Wörterbuch normiert aufgenommenen Wörter als produktiv. Zwar könnte anhand der Matrix aus den mittelhochdeutschen Wörtern auch ohne einen Vergleich das erwartete neuhochdeutsche Pendant erzeugt werden – dieses Vorgehen ginge jedoch von einer extremen Regelmäßigkeit aus, die für natürliche Sprache nicht gilt und daher eine höhere Fehlerquote zur Folge hätte. Faktoren, die das Alignment qualitativ begünstigen und gleichzeitig die Rechenzeit minimieren, sind das Einschränken auf vergleichbare Wortklassen und Anlaute sowie eine Segmentierung in Morpheme. Diese Informationen sind durch den Rückgriff auf Wörterbücher zum Teil bereits vorhanden. Ihre Nutzung als Datengrundlage hat den weiteren Vorteil, dass auch der semasiologisch angeleitete Stringvergleich semantisch nachvollziehbar wird, indem die zugehörigen Wörterbuchartikel der miteinander assoziierten Lemmata zur Verfügung stehen.

Das Alignment erlaubt uns nicht nur eine automatische Verknüpfung der als sicher zusammengehörig eingestuften Lemmata, es kann auch für die Überprüfung vorhandener Links herangezogen werden und anhand der Werte z.B. Hinweise darüber geben, ob eine semantische oder semasiologische Verknüpfung vorliegt. Idealerweise treffen sämtliche Annotationen (z.B. Informationen zu Wort- oder Morphemgrenzen) der Lemmata auch auf das jeweils andere zu und können übertragen werden.

Bei dem Annotationstool handelt es sich um ein flexibles Werkzeug, das mit einer entsprechend generierten Substitutionsmatrix leicht für ein Alignment anderer Sprachstufen, Dialekte oder Sprachen angewandt werden kann. Aus den beschriebenen Gründen empfiehlt sich dabei das Operieren mit elektronischen Wörterbüchern.[10]

Als Produkt einer projektspezifischen, interdisziplinären Fragestellung ist damit ein Werkzeug hervorgegangen, das auch in anderen Kontexten einen Beitrag zu Verknüpfungen oder zumindest einen Indikator zu ihrer Auswertung darstellen kann. Gleichzeitig stellt es ein Beispiel für die Kombination von händischen und automatischen Verfahren dar, das es AnwenderInnen erlaubt,[11] aufgrund von linguistischen Vorgaben und Einstellungen direkten Einfluss auf den Algorithmus zu nehmen. In einem iterativen Prozess führt dies zu einem gegenseitigen Prüfen der kombinierten Verfahren, das den bestmöglichen Ausgang zur Folge hat.

[10] Für die entsprechende Aufbereitung der Wörterbuchdaten siehe Schneiker et al. (2009).
[11] Zum Beispiel im Rahmen einer kollaborativen virtuellen Forschungsumgebung wie TextGrid (TextGrid 2009-).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

6. Elektronische Ressourcen

GOLD: General Ontology for Linguistic Description. www.isocat.org (Stand: 12.8.2014).

P5: Guidelines for electronic text encoding and interchange. www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (Stand: 12.8.2014).

SWI Prolog: www.swi-prolog.org/ (Stand: 12.8.2014).

TextGrid (2009-): Virtuelle Forschungsumgebung für die Geisteswissenschaften. www.textgrid.de (Stand: 12.08.2014)

The Protégé Ontology Editor. http://protege.stanford.edu (Stand: 12.8.2014).

TUSTEP: Tübinger System von Textverarbeitungs-Programmen. www.tustep.uni-tuebingen.de/ (Stand: 12.8.2014).

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen (2008-2012): Verbundprojekt gefördert durch das Bundesministerium für Bildung und Forschung (BMBF). 15.8.2014). www.sprache-und-genome.de (Stand 15.8.2014).

Wörterbuchnetz (2007-): Forschungsprojekt des Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (= Trier Center for Digital Humanities). www.woerterbuchnetz.de (Stand: 12.8.2014).

7. Literatur

Campe, Joachim Heinrich (1807-1811): Wörterbuch der deutschen Sprache. 5 Bde. Braunschweig

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4. Aufl. Tübingen.

Gazdar, Gerald/Mellish, Chris (1989): Natural language processing in PROLOG. An introduction to computational linguistics. Wokingham.

Klappenbach, Ruth/Steinitz, Wolfgang (1961-1977): Wörterbuch der deutschen Gegenwartssprache. 6 Bde. Berlin.

Schneiker, Christian et al. (2009): Declarative parsing and annotation of electronic dictionaries. Proceedings of the sixth International Workshop on Natural Language Processing and Cognitive Science, 6.-7. Mai 2009 (NLPCS). Mailand.

Seipel, Dietmar/Wegstein, Werner (2011): metaDictionary. Towards a generic e-infrastructure for detecting variance in language by exploiting dictionaries. Proceedings of the International Symposium on Grids and Clouds (ISGC). http://www1.pub.informatik.uni-wuerzburg.de/databases/papers/isgc_2011.pdf (Stand: 12.8.2014).

© 2016 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.