Historische formelhafte Sprache als "harte Nuss" der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt

Natalia Filatkina (Trier)

Abstract

The following article tackles not so much the corpus and computer linguistic questions in a narrow sense. It rather focuses on such linguistic phenomena as formulaic patterns in the history of German language and describes them from a corpus and computer linguistic perspective. Since July 2007, historical formulaic language has been a subject of investigation in the research group "Historical Formulaic Language and Traditions of Communication" at the University of Trier. Corpus and computer linguistic methods are not in the middle of the research interest in this project but they constitute its important methodological part. After a short introduction (Chapter 1), Chapter 2 gives a brief outline about the state of the art in the field of formulaic language within the framework of corpus and computer linguistics. Chapter 3 analyzes some problems in this area with regard to modern languages. The issues tackled here turn to be even more problematic from an historical point of view, as shown in the following Chapter 4. Finally, Chapter 5 suggests a possible solution that was developed and implemented in the HiFoS-group.

1 Einführung

Im Mittelpunkt des vorliegenden Beitrags stehen nicht primär die Fragen des Aufbaus eines elektronischen Korpus oder seiner Struktur, sondern eher eine sprachliche Ebene – die der formelhaften Wendungen –, die aus corpus- und computerlinguistischer sowie sprachhistorischer Perspektive problematisiert wird. Historische deutsche formelhafte Sprache bildet den Gegenstand der seit Juli 2007 an der Universität Trier etablierten Nachwuchsforschergruppe "Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)".[1] Die Forschungsziele des Projekts gehen dabei weit über die korpus- und computerlinguistischen Fragestellungen hinaus; diese machen allerdings einen wichtigen methodischen Teil des Projekts aus. Nach einigen Überlegungen zu der noch ausbaubedürftigen Verknüpfung der gegenwartsbezogenen Phraseologieforschung und Korpus- und Computerlinguistik (Abschnitt 1) sollen einige theoretische und methodische Schwierigkeiten erläutert werden, die bis jetzt für die Bezeichnung der formelhaften Sprache als "harte Nuss" der Korpus- und Computerlinguistik gesorgt haben (Abschnitt 2). Diese Problembereiche werden in Abschnitt 3 aus historischer Perspektive nochmals verschärft. Abschnitt 4 enthält abschließend einen Vorschlag für ihre mögliche Lösung aus der Sicht des HiFoS-Projekts.

[1] Die Gruppe besteht aus 2 wissenschaftlichen Mitarbeitern, einem technischen Mitarbeiter und einigen studentischen Hilfskräften. Die Nachwuchsforschergruppe finanziert sich aus dem Sofja Kovalevskaja-Preis-Programm der Alexander von Humboldt-Stiftung (Stifter: Bundesministerium für Bildung und Forschung).

2 Formelhafte Wendungen in der Korpus- und Computerlinguistik

Formelhafte Wendungen (oder enger Phraseologismen[2]) sind seit den 40er Jahren des 20. Jahrhunderts Gegenstand der Phraseologieforschung. Der Begriff Phraseologismus bezieht sich auf solche Wendungen, die aus mehr als einem Wort bestehen (Merkmal Polylexikalität), syntaktisch mehr oder weniger stabil sind (Merkmal Festigkeit) und semantisch irregulär oder anders nicht kompositionell (Merkmal Idiomatizität) sein können (Burger 2007: 11–15). Über diese Merkmale verfügen unterschiedliche Phraseologismus-Typen (Idiome, Kollokationen, Routineformeln, Paarformeln, Sprichwörter usw.) in unterschiedlichem Maße. Als usualisierte Wortverbindungen werden Phraseologismen in einer bestimmten, mehr oder weniger festen Form und Bedeutung wiederholt von Sprechern einer Gemeinschaft reproduziert (Merkmal: Reproduzierbarkeit). Sie sind nicht nur Elemente des Zeichensystems Sprache, des mentalen Lexikons, sondern auch Träger der Kultur, denn einige Phraseologismen tradieren in ihrer bildlichen Grundlage (oder inneren Form, *image component*) kulturell geprägte oder gar kulturspezifische Gegebenheiten. Seit den 60er Jahren gehört die Phraseologieforschung zu den international etablierten linguistischen Teildisziplinen.

Seit dieser Zeit werden formelhafte Wendungen auch im Bereich der Korpus- und Computerlinguistik berücksichtigt, in dem sie im weiten Sinn als *multiword expressions* (MWEs) bezeichnet werden. Zurückblickend auf die nun mehr fast 50jährige Tradition, werden formelhafte Wendungen seitens der Korpus- und Computerlinguisten immer noch als "harte Nuss" oder "pain in the neck" bezeichnet. Sie stellen sowohl bei der Korpuserstellung als auch bei der Korpusannotation und erst recht bei der Korpusbenutzung eine Herausforderung dar und haben sich bis jetzt einer zufrieden stellenden Erfassung mit modernen Technologien entzogen.

Korpus- und Computerlinguistik gehört zu den wissenschaftlichen Teildisziplinen, die in den 90er Jahren empirisch anhand elektronischer Textkorpora nachweisen konnte, in welch hohem Maße die meisten (auf dem Territorium Europas verbreiteten) Sprachen formelhaft (syntaktisch und semantisch) geprägt sind.[3] Der Hintergrundgedanke ist dabei, dass Wörter einer Sprache nicht isoliert voneinander funktionieren, sondern syntagmatisch und eben phraseologisch miteinander verbunden sind und nur in dieser Verbundenheit in verschiedenen Kommunikationssituationen Sinn ergeben. Zu nennen wären hier exemplarisch J. Sinclairs *idiom principle* (1987), M. Hoyes dymanisches Lexikonmodell *lexical priming* (2005) oder auch A. Wrays Begriff *formulaic language* (2005, 2008). Sag et al. (2001) nehmen anhand Korpusrecherchen an, dass die Anzahl an formelhaften Wendungen, die in ein adäquat funktionierendes Natural Language Processing-System aufgenommen werden müssten, mindestens genauso hoch sein muss wie die Zahl der Einzellexeme. R. Jackendoff behauptet für das Englische: "The lexicon of a language available to speakers in everyday situations contains at least as many multi-word expressions as single words" (1995: 156; cf. auch 1997). Mit Hilfe des *Oxford Hector Pilot Corpus* ermittelt R. Moon (1998) eine auffällige Frequenz der verbalen Phraseologismen sowohl in *types* als auch in *tokens*. A. Cowie (1992: 1) stellt nach der Auswertung eines Korpus der englischen Nachrichten fest, dass dort Verb-Substantiv-Kollokationen des Typs *to make a proposal, give figures, try the case* usw. einen

[2] Die Begriffe haben einige Gemeinsamkeiten, sind jedoch nicht deckungsgleich. Ausführlicher zum Begriff *formelhafte Sprache* aus historischer Perspektive cf. Filatkina/Gottwald/Hanauska/Rößger (2009) und Filatkina (2009).

[3] Zur ähnlichen Schlussfolgerung kommen auch die Untersuchungen aus den Bereichen der Formulierungstheorie, Textsortenlinguistik und Ritualforschung: Viele der in der Kommunikation auszuführenden Sprachhandlungen sind konventionalisierte, ritualisierte Kommunikationsformen, die das Formulieren ökonomischer gestalten. Die Konventionalisierung, Routine in der Mündlichkeit (Coulmas 1981) wie in der Schriftlichkeit (Gülich 1997) oder – mit Feilke (1994: 366) idiomatische Prägung entsteht zu einem wesentlichen Teil durch die Verwendung von formelhaften Wendungen unterschiedlichster Typen.

erheblichen Teil des Wortschatzes ausmachen. Die Frequenz der verbalen idiomatischen wie nicht idiomatischen Phraseologismen im Korpus "Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS)" hat Ch. Fellbaum (2007: 2) erlaubt, den Untersuchungsgegenstand des Projekts "Kollokationen im Wörterbuch" auf diesen Typ der formelhaften Wendungen zu beschränken.

Den korpusbasierten Analyseverfahren, die seit dem Ende der 90er Jahre auch innerhalb der traditionellen Phraseologieforschung ein methodologischer Standard sind, ist ein Wechsel der Blickrichtung sowie der Materialbasis in dieser Disziplin zu verdanken. Während zuvor vor allem Wörterbücher den Gegenstand phraseologischer Untersuchungen bildeten, stellen neuere Studien Primärtexte verstärkt in den Mittelpunkt. Dieser Blickrichtungswechsel hat neue Einblicke in das tatsächliche Funktionieren der Phraseologismen in Texten ermöglicht: Korpusrecherchen haben für die meisten modernen Sprachen erwiesen, dass Phraseologismen nicht nur lange "vorgeformte" Einheiten sind. Obwohl sie von Sprechern einer Sprache als lexikalische Ganzheiten erkannt und als solche gespeichert werden, weisen Phraseologismen vielfältige Möglichkeiten der morphosyntaktischen und lexikalischen Variation auf, die durch mehrere linguistische Ebenen, inklusive Diskurs, verlaufen (Fellbaum 2006; Heid 2008: 343–350). Für das Englische stellt R. Moon (1998: 120–177) fest, dass etwa 40% aller im Korpus vorkommenden formelhaften Wendungen lexikalische Varianten aufweisen, wobei das prinzipiell mögliche Variationspotential weder vom Typ der Wendung[4] noch von ihrer *token*-Frequenz abhängt. In dem korpusbasierten *Collins Cobuild Dictionary of Idioms* weist ein Drittel aller 4000 verzeichneten formelhaften Wendungen lexikalische Variation auf (Moon 2007: 1055). Diese lexikalische Einheit, die fehlenden Grenzen zwischen Grammatik und Lexik bzw. Semantik einerseits sowie kompliziertes grammatisches Verhalten in Texten und starke Betonung der Syntagmatik statt Paradigmatik andererseits stellten bis jetzt eine Herausforderung für die konventionellen (zumeist paradigmatisch organisierten) grammatischen Theorien sowie für die Korpus- und Computerlinguistik dar.

3 Formelhafte Wendungen als Herausforderung für Korpus- und Computerlinguistik

Trotz der Tatsache, dass Korpus- und Computerlinguistik (sowie die maschinelle Übersetzung) sich mit Phraseologismen beschäftigt haben, lange bevor es in der Germanistischen Linguistik bzw. Anglistik eine etablierte Phraseologieforschung gab, hatten die beiden Disziplinen bis jetzt wenig theoretische Berührungspunkte. Vor allem folgende drei Bereiche erweisen sich heute noch als ungenügend geklärt bzw. besonders problematisch:

1) die Größe der Korpora für Untersuchungen der Formelhaftigkeit und ihre Komposition,
2) die Annotation oder Mark up auf Ebene der Formelhaftigkeit und
3) die Korpustools für das Identifizieren und Retrieval der formelhaften Wendungen.

In Bezug auf die Größe der Korpora für Untersuchungen mit formelhaften Fragestellungen wurde bereits mehrmals hervorgehoben, dass sie höher sein muss als bei den Korpora, die für Untersuchungen anderer Phänomene benutzt werden, da davon auszugehen ist, dass eine formelhafte Wendung in einem Text weniger frequent vorkommt als etwa ein bestimmter Artikel. Laut A. Geyken et al. (2004) ist sogar ein Korpus von 100 Mio. Wörtern nicht ausreichend für phraseologische Untersuchungen.[5] Aus diesem Grund können auch nur Volltexte und keine Textausschnitte von bestimmter Seitenzahl herangezogen werden, da die Distribution der formelhaften Wendungen im ausgewählten Ausschnitt nie im Voraus erkannt

[4] So hat sich z. B. die Annahme, dass vor allem Idiome und Sprichwörter verstärkt lexikalischer Variation unterliegen, nicht bestätigt.

[5] Die Existenz großer Korpora hält z. B. Sailer (2007: 1065) für wichtiger als z. B. ihre tiefe Annotation.

werden kann. Über die Komposition der Korpora ist in der Korpuslinguistik noch weniger bekannt. R. Moon (2007: 1052f.) behauptet vage für das moderne Englisch: "genres may be marked by their phraseology, and at the same time phraseologies may be determinants of those genres". Auch die traditionelle Phraseologieforschung kann hier den Korpus- und Computerlinguisten kaum behilflich sein, weil sie bis vor kurzem vor allem Wörterbücherein-träge und nicht Primärtexte als Untersuchungsmaterial hatte.

Der Frage nach der Tiefe der Annotation wird im Vergleich zu der Frage nach der Größe des Korpus etwas weniger Beachtung geschenkt. Allerdings erweist sie sich auf Grund der unter-schiedlichen Blickrichtungen in der Korpuslinguistik und Phraseologieforschung als nicht trivial: Während sich die Annotationsverfahren bis jetzt auf die Morphologie und Syntax konzentriert haben, ist die Phraseologieforschung überwiegend semantisch orientiert. Denn sie hat u. a. Einheiten zu analysieren, die lexikalisch und morphologisch nach den gängigen grammatischen Regeln aufgebaut sein können (wie etwa *ins Wasser fallen*), semantisch aber irregulär sind. Man denke hier an die figurative Bedeutung 'scheitern' beim Idiom *ins Wasser fallen*. Die Berücksichtigung der semantischen Irregularität ist wichtig, würde aber die Annotation auf mehreren Ebenen erfordern, wie dies im XML-Format repräsentiert und bis jetzt selten durchgeführt ist.[6] Auf diese Weise wird bei der Annotation dem Zusammenspiel der Syntax und Semantik in der Struktur der formelhaften Wendungen Rechnung getragen. Hilfreicher wäre es, wenn solche Annotationen nicht wortbasiert (im Sinne von heute eher veralteten *words-with-spaces*-Verfahren) wären, sondern die Wendung als Ganzheit erfassen würden. Die Entwicklung der Annotationsstandards im Textkorpus bzw. einer formalisierten Sprache zur Repräsentation von formelhaften Wendungen in einem elektronischen Wörter-buch oder einer Datenbank stellt sich als ein weiteres Problem in diesem Zusammenhang dar. Zu einem großen Teil fehlen sie, weil der Forschung bis jetzt nicht genug Informationen über das tatsächliche Funktionieren der Wendungen in Texten und das zu erwartende Variations-spektrum, das sich gleichzeitig auf mehrere und unterschiedliche sprachliche Ebenen erstreckt (Morphosyntax, Semantik, Pragmatik, Diskurs), vorliegen (Tschichold 2008: 367). Die ersten Versuche auf diesem Gebiet kommen aus dem Bereich des *Language Engineering*[7] und konzentrieren sich auf die Repräsentation eines bestimmten Typs von formelhaften Wendun-gen.[8]

Im Gegensatz zu den zwei letzten Fragen ist die Retrieval-Frage nicht neu. Die bis jetzt in Erwägung gezogenen Verfahren lassen sich wie folgt zusammenfassen:[9]

a) Extrahieren mit Hilfe der *association measures* (statistische Verfahren)
b) Extrahieren auf Grund formaler Preferenzen (symbolische Verfahren)
c) Extrahieren mit Hilfe der Distributionssemantik.

Die zwei ersten Verfahren sind bekanntlich die ältesten und suchen nach statistisch signifikanten Kookkurrenzen in unterschiedlich und nur sehr flach annotierten Korpora. Während die Verfahren der ersten Gruppe lexikalisch feste Ausdrücke gut auffinden können, ermitteln die Tools der zweiten Gruppe formelhafte Wendungen auf Grund der morphosyn-taktischen Restriktionen und Idiosynkrasien. Bei allen Vorteilen dieser Methoden gehen sie

---

[6] Cf. etwa die Annotation mit Hilfe der *source* und *taget frames* im SALSA-Projekt (www.coli.uni-saarland.de/projects/salsa/page.php?id=index).

[7] Cf. das ISLE-Projekt (www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm).

[8] Am meisten wurden bis jetzt Kollokationen bzw. Funktionsverbgefügen berücksichtigt, so etwa im Xmellt-Projekt (www.cs.vassar.edu/~ide/XMELLT.html). Kollokationen sind auch im ISO Standard Lexical Markup Framework berücksichtigt. Einen Vorschlag zur standardisierten Repräsentation von Idiomen äußert Odijk (2004). Im Gegensatz dazu sind verschiedene Typen der formelhaften Wendungen in der Datenbank Keil (1997) und im PhraseManager (Tschichold 2008: 368–375) berücksichtigt.

[9] In Anlehnung an Heid (2008: 350) und Heid (2007: 1041).

von Einzellexemen aus und sind vor allem eben für die Suche nach lexikalisch und/oder morphosyntaktisch festen Wendungen geeignet, und auch nur nach solchen, die nicht aus mehr als zwei Wörtern bestehen. Absolut feste Wendungen finden sich aber auch im modernen Sprachgebrauch selten. Als Beispiel könnten hier Paarformeln oder Wendungen mit unikalen Konstituenten (wie etwa *Maulaffen feilhalten*) dienen. Verfahren der dritten Gruppe scheinen deshalb bei der Suche nach formelhaften Wendungen Erfolg versprechend zu sein, ebenso wie die neusten Verfahren, die die Funktionsspezialisierung (also ein pragmatisches Kriterium) in den Vordergrund stellen.

4 Historische formelhafte Sprache in der Korpus- und Computerlinguistik

Historische formelhafte Sprache stellt für die Korpus- und Computerlinguistik eine noch größere Herausforderung dar. Das größte Problem besteht darin, dass heutzutage für die ältesten Sprachstufen des Deutschen seit Beginn seiner Überlieferung im 8. Jahrhundert bis in die Frühe Neuzeit (ca. 1650) kein umfangreiches philologisch verlässliches elektronisches Referenzkorpus vorhanden ist, das die Untersuchung der Dynamik der formelhaften Wendungen epochenübergreifend und empirisch abgesichert erlauben würde. Während für das moderne Deutsch die Größe des Korpus, das der Beantwortung formelhafter Fragestellungen dienen soll, bei mehreren Millionen Wörtern angesetzt wird, ist es für das historische Material alleine schon auf Grund der spärlichen Überlieferung (so etwa für das Alt- und Frühmittelhochdeutsche) problematisch, diesem Anspruch gerecht zu werden.

Die bereits existierenden Korpora, die auf handschriftengetreuen Transkriptionen der Originalquellen basieren, enthalten meistens keine Volltexte, sondern beschränken sich auf ausgewählte Textausschnitte, auf Texte einzelner Autoren bzw. verschiedene Überlieferungsträger eines Textes.[10] Da die Distribution von formelhaften Wendungen in einem Text nicht im Voraus einzuschätzen ist, eignet sich die Auswahl einzelner Ausschnitte nicht für die Beantwortung phraseologischer Fragestellungen. Auch die Fokussierung auf ausgewählte Autoren bzw. einzelne Texte bringt die Forschung m. E. nicht wesentlich voran, sondern würde ihren bisherigen punktuell greifenden Charakter nur stärken.[11] Andere Korpora setzen sich zwar aus Volltexten zusammen, bestehen aber nicht aus originalen Überlieferungsträgern, sondern Editionen, die zum Teil älteren Datums und folglich stark normiert sind.[12] Den Editionskriterien des 19. Jahrhunderts entsprechend stellen solche Editionen aus verschiedenen Handschriften von Sprachhistorikern kompilierte "Urtexte" dar, die in dieser Form in historischen Sprachstufen aber nie existierten. Während solche Korpusdaten für die Repräsentation der Semantik in Sprachstufenwörterbüchern ausreichend sein mögen, sind sie wenig für die Untersuchung der morphosyntaktischen und lexikalischen Variation geeignet.

Ferner liegen der Forschung nur sporadische Informationen über die text(sorten)spezifische Distribution von formelhaften Wendungen vor. Ältere Texte, in denen die Verwendung formelhafter Wendungen augenfällig ist, bildeten bereits Gegenstand zahlreicher Untersuchungen. Systematische Informationen in Bezug auf verschiene Textsorten fehlen zum großen Teil komplett. D. h., dass hier der Korpuserstellung entsprechende sprachhistorische Untersuchungen vorangehen müssten.[13] H. Burger hat bereits 1983 die Wichtigkeit des Althochdeutschen für die Untersuchung der Formelhaftigkeit hervorgehoben. In der HiFoS-

[10] Cf. den Überblick in Hoffmann (1998: 875).

[11] Zum Forschungsstand im Bereich der historischen formelhaften Sprache vgl. Filatkina (2007) und (im Druck).

[12] Beispiele hierzu bringt ebenfalls Hoffmann (1998: 875–887).

[13] Dazu sind in der HiFoS-Nachwuchsforschergruppe zwei Dissertationen entstanden, die die Verwendung formelhafter Wendungen in Kölner Stadtchroniken (cf. Hanauska 2009: 45–65) und Nürnberger Fastnachtspielen (cf. Gottwald 2009: 11–43) untersuchen.

Datenbank sind derzeit 4613 althochdeutsche Belege verzeichnet,[14] wobei 2673 den Bibelübersetzungen und Kommentaren, 641 den gelehrten Texten für den Schulunterricht, 217 der pastoralen Literatur, den Predigten und Gebetstexten entstammen; 170 Belege kommen in den Ordensregeln und Ordensliteratur, 51 in der Heldenepik vor.[15]

Eine weitere Schwierigkeit, die bereits bei der Identifikation von formelhaften Wendungen in Texten in Erscheinung tritt, ist theoretischer bzw. terminologischer Art. Sobald man versucht, in älteren Texten nach Phraseologismen zu suchen und die für Phraseologismen geltenden Merkmale anzuwenden, bemerkt man, dass man relativ bald an Grenzen stößt. Denn diese Merkmale sind nicht eins zu eins aus der bis jetzt vorwiegend gegenwartssprachlich orientierten Phraseologieforschung auf das ältere Material übertragbar. So ist z. B. das Problem der Wortgrenzen für historische Sprachstufen so alt wie die Erforschung derselbigen. Das zweite Kriterium der relativen syntaktischen Festigkeit steht dem Variantenreichtum nicht kodifizierter älterer Sprachstufen entgegen. Das Festigkeitskriterium bezieht sich immer auch auf die Gebräuchlichkeit der Einheit, und auch da hat man es aufgrund der Zufälligkeit der schriftlichen Überlieferung mit großen Problemen zu tun. Das hermeneutische Zugangsproblem zur Bedeutung in historischen Texten relativiert ferner die Gültigkeit des dritten Merkmals (Idiomatizität).

Wenn soeben von einer starken Variation der historischen formelhaften Wendungen die Rede war, so muss diese Behauptung auf der heutigen Untersuchungsetappe den Charakter einer Annahme haben. Während die Untersuchung der gegenwartssprachlichen Variation im Bereich der formelhaften Wendungen bereits seit einiger Zeit verstärkt in Angriff genommen wurde und wird,[16] ist die Erforschung der Dynamik der formelhaften Wendungen sowohl in der Diachronie als auch synchron in den älteren Sprachstufen des Deutschen nach wie vor ein Forschungsdesiderat. Die Annahme, dass die Verfestigungs- bzw. Konventionalisierungsprozesse bei formelhaften Wendungen über die Reduzierung des Variantenspektrums verlaufen (Burger/Linke 1998), lässt die Vermutung zu, dass für das historische Deutsch mit einem stärkeren Grad an morphosyntaktischer, lexikalischer und semantischer Variation zu rechnen ist, der mit der Variation auf orthographischer Ebene einhergeht, und sich mit dem teilweise anderen Status der formelhaften Wendungen im älteren Deutsch überschneidet. Ein Beispiel soll dies veranschaulichen:

(1) *nach jemandes Pfeife tanzen* 'alles tun, was jmd. von einem verlangt; jmdm. gehorchen'

Die Recherchen im Archiv für geschriebene Sprache des COSMAS II_{win}-Systems am Institut für deutsche Sprache in Mannheim gibt für das Idiom (1) ca. 160 Treffer aus der Zeitspanne 1960 bis 2007. Die in den idiomatischen Wörterbüchern des Deutschen verzeichnete lexikalische Variante *nach jemandes Geige tanzen* kommt hingegen nur dreimal vor. Die Leerstelle *jemandes* wird in überwiegender Mehrzahl der Fälle mit Possessivpronomina bzw. Personennamen besetzt, in 8 Fällen kommen Adjektive in Possessivbedeutung vor (z.B. *nach amerikanischer/orientalischer/schwarzer* (im Sinne der Parteizugehörigkeit) *Pfeife tanzen*). Für Österreich und die Schweiz kann ferner mit drei Belegen die Ersetzung der Leerstelle durch den Artikel attestiert werden *(nach der/einer Pfeife tanzen)*. Die substantivische Konstituente *Pfeife* erweist sich als konstant und stabil; die verbale Konstituente *tanzen* wird (eingeschränkt auf Präsens) flektiert. Korpusanalysen veranschaulichen außerdem die Mög-

[14] Stand: März 2009.

[15] Der frappierende Unterschied zwischen dem bisherigen Forschungsstand und den Ergebnissen des HiFoS-Projekts in Bezug auf nur einen Autor – Notker den Deutschen von St. Gallen – wird an einer anderen Stelle thematisiert, cf. Filatkina/Gottwald/Hanauska/Rößger (2009).

[16] Cf. exemplarisch die Ergebnisse des Projekts "Kollokationen im Wörterbuch" (Fellbaum 2007) oder das Projekt "Usuelle Wortverbindungen" am Institut für deutsche Sprache in Mannheim (Steyer 2003; 2004).

lichkeit des Aktionsartwechsels: *jmdn. nach jmds. Pfeife tanzen lassen*. Dem gegenwartssprachlichen Befund stehen folgende historische Belege entgegen:[17]

(2a) *Wol ûf, swer tanzen welle nâch der gîgen* (Walther v. d.Vogelweide, 19, 37, vor 1230)
(2b) *Und must am lesten tantzen als die von zwürch pfiffen* (Justinger, Berner Chron. 339, S. 207, 20; 1420–30)
(2c) *Vor dich, vor dich, vordencke mich nicht, Noch deyner pheyffe tantze ich nicht* (Prov. Frid., 131; 2. H. 15. Jh.)
(2d) *Went gy moten na myner pypen springen* (Totentanz, 20;1463)
(2e) *Went se hebben ... noch myt leve noch myth pranghe Nummende konen darto brynghen, De de wil na erer pipen springhen* (Redentiner Ostersp., 1464; 1214)
(2f) *He moste al na syner pypen dantzen* (Hagen, Helmst. Chron. 162, S. 119; 1491)
(2g) *He heft leeff den ... de so dantzet, alze he vore synget* (Reinke Vos, 3893; 1498)
(2h) *Der gemein man můß im nach tantzen, wie er pfeiffet* (Geiler, Brösamlin, I, 59d; 1517)
(2i) *De na der horen pypen danset, de is der schmede vry. Depudet, obsequitur scorto quicunque bilingui* (Tunnicius, 645; 1513)
(2j) *Tantzen nach irer alten geigen* (Sachs VI, 381, 33; 1523)
(2k) *Du dantzt nach deyner alten geygen* (Sachs IV, 46, 27; 1534)
(2l) *Alls ir mir vor tanczt, also sol ich nach springen* (Füetrer, Lanzelot, 152; um 1467)

Vom ersten deutschen Beleg bei Walther von der Vogelweide bis in die 2. Hälfte des 16. Jahrhunderts hinein ist das heute eher umgangssprachliche Idiom vielfältig in der Chronistik, in einem Totentanz, in einem geistlichen Spiel, in der didaktischen Literatur, in Predigten und satirischen Werken überliefert. Es findet Eingang in vier parömiologische Sammlungen mit Autoritätsstatus, was möglicherweise seine Verbreitung begünstigt hat und auf die Zugehörigkeit der Wendung zu einem gehobenen Stilregister schließen lässt. Bei der Konstanz der aktuellen Bedeutung 'alles tun, was jmd. von einem verlangt; jmdm. gehorchen' und dem seit Beginn der Überlieferung hohen Idiomatizitätsgrad bleiben die morphosyntaktische Struktur und lexikalische Besetzung bis hin zu Hans Sachs und über ihn hinaus beweglich. Die Beispiele (2) veranschaulichen die lexikalische Variation im Bereich der substantivischen und verbalen Konstituenten (z. B. die Substitution *pfeife* durch *geige* bzw. durch den Nebensatz *wie er ihm vorsingt, wie er pfeift; tanzen* durch *springen* oder die Erweiterung der Struktur durch Adjektiveinschub) sowie die morphosyntaktische Variation (etwa die Negation oder die heute zumindest für den binnendeutschen Sprachraum untypische Null-Besetzung der Leerstelle *jemandes*).

Der erstmaligen systematischen quellenkundlichen Dokumentation und Untersuchung dieser historischen Gebrauchsdynamik ist die Arbeit in der Nachwuchsforschergruppe "Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)" gewidmet.

5 Die Nachwuchsforschergruppe "Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)"

5.1 Zielsetzung und Fragestellungen

Wenn man heute allgemein von einer gewissen "Phraseologie-Zentrierung" der sprachwissenschaftlichen Forschung (Dobrovol'skij 1992: 29) sprechen kann, so gilt das erst seit sehr kurzer Zeit für die historische Sprachwissenschaft. Wenige bereits vorliegende Untersuchungen knüpfen an die so genannte Literaritäts-Forschung an und erklären die Verwendung der

[17] Bei dieser Liste handelt es sich um eine exemplarische Auswahl; diese und weitere Belege, u. a. auch aus anderen mittelalterlichen Sprachen, sind im TPMA (1996; Bd. 11: 268–270) verzeichnet.

formelhaften Ausdrücke durch die mnemotechnische Notwendigkeit.[18] In der letzten Zeit ist ein verstärktes Interesse an Formelhaftigkeit seitens der Literaturgeschichte zu vermerken: Es liegen bereits einige umfangreiche Monographien zur Texttheorie und Formelhaftigkeit und zur herausragenden Rolle der Sentenzen in mhd. Artus-Romanen vor.[19] Im Gegensatz dazu wendet sich eine Vielzahl von Studien der Geschichte der Sprichwörter (Parömiologie) zu, wobei es sich vor allem um volkskundliche Untersuchungen zu ihrer Herkunft handelt. Beschränkt ist ferner das Repertoire wissenschaftlicher Hilfsmittel (Wörterbücher), in denen formelhafte Wendungen nicht als "versteckte Informationen" mit zufälligem "Beispielsatz"-Charakter betrachtet werden würden.

Aus diesem Grund sind in der gegenwärtigen Untersuchungsetappe eher keine kleineren Studien zu einzelnen Autoren oder Textsorten nötig, sondern eine erste Bestandsaufnahme anhand der möglichst breiten Textsortenpalette aus unterschiedlichen Epochen, die die textuelle Distribution der formelhaften Wendungen quellenbasiert dokumentiert sowie Einblicke in ihre Verfestigungsprozesse und das funktionale Spektrum gewährt. Diese Forschungslücke versucht die Nachwuchsforschergruppe "Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)" teilweise zu schließen, indem es eine epochenübergreifende Dokumentation und Kommentierung der historischen Variation und Gebrauchsdynamik der Formelhaftigkeit in älteren deutschen Texten unterschiedlichster Textsorten (literarische Werke, Gebrauchstexte (Kochrezepte usw.), Urkunden und Rechtstexte, Tagebücher und Reiseberichte, politische Texte der Reformationszeit (Flugblätter) und ansatzweise ältere phraseologische Sammlungen, Grammatiken, Stillehren sowie Formular- und Sprachlehrbücher) aus der Zeitspanne zwischen ca. 750 bis ca. 1700 anstrebt. Dabei handelt es sich hier in mehrfacher Hinsicht um eine erste Bestandsaufnahme bzw. Grundlagenforschung, die die Beantwortung weiterer Fragen ermöglichen soll[20] und in ihren Mittelpunkt zunächst die Dokumentation und die linguistische Annotation des Variantenspektrums der älteren Wendungen stellt. Der Begriff *Variation* bezieht sich vor allem auf semantische, morphosyntaktische und lexikalische Prozesse der Konventionalisierung der älteren Wendungen, ohne aber die Konventionalisierung teleologisch als einen notwendigen Endpunkt in der Verwendung einer Formel zu implizieren. Die Forschungsziele der HiFoS-Nachwuchsforschergruppe gehen somit einerseits weit über rein korpus- und computerlinguistische Fragestellungen hinaus, andererseits liefert das Projekt genaue Angaben zur Verwendung der formelhaften Wendungen in Texten und ihrer Variation, auf deren Grundlage in weiteren Untersuchungen u. a. formalisierte Annotationsstandards für die Ebene der Formelhaftigkeit entwickelt werden könnten. Derzeit enthält die Datenbank über 9000 Belege (Stand: März 2009), die primär den ahd. sowie einigen (im Moment noch stichprobenweise ausgewerteten) mhd. und frnhd. Texten entstammen.

---

[18] Angesichts des Fehlens der Schriftlichkeit diktiert sie sowohl die ausdrucksseitige Gestaltung durch Reim und Rhythmus, syntaktische Symmetrien, phonetische Assonanzen und Alliterationen als auch die inhaltsseitige Prägung der Figuren.

[19] Cf. exemplarisch Tomasek (2009); Tomasek (2005); Tomasek/Eikelmann (2002); Eikelmann (1998; 2002).

[20] So z. B.: Welche formelhaften Wendungen sind in welchen Texten überliefert? Inwiefern hängt Formelhaftigkeit von der Textsorte ab? In welchen Kommunikationssituationen und mit welchen Funktionen werden formelhafte Wendungen verwendet? Wie ist ihr "Sitz im Leben" zu beschreiben? Welche Bereiche (Ausgangskonzepte) der realen bzw. imaginären Wirklichkeit konstituieren die bildlichen Grundlagen der historischen formelhaften Sprache? Welche thematischen Bereiche (Zielkonzepte) werden mit ihrer Hilfe versprachlicht? Begleitend zur Forschung wird im Projekt außerdem noch eine online-Bibliographie mit internationaler Sekundärliteratur zur historischen deutschen Formelhaftigkeit erstellt. Diese bislang einzige Bibliographie mit dem genannten Schwerpunkt ist bereits im Netz frei zugänglich (www.hifos.uni-trier.de/Bibliographie.php); sie ist dynamisch aufgebaut und wird regelmäßig aktualisiert. Derzeit enthält sie 801 Einträge (Stand: März 2009) und bietet eine gute Grundlage für eigenständige Recherchen.

5.2 Annotation und Analyse formelhafter Sprache im HiFoS-Projekt

Da momentan umfangreiche philologisch verlässliche Korpora für die historischen Sprachstufen des Deutschen fehlen und das HiFoS-Projekt die Erstellung eines solchen elektronischen Korpus nicht anstrebt, werden Texte in Originalhandschriften und Drucken zunächst manuell ausgewertet. Wenn für die Zeit vor 1050 auf Grund der relativ überschaubaren Überlieferung alle ahd. Texte (mit Ausnahme der Glossen) zur Analyse herangezogen werden konnten, stellt sich ab der spätmhd. Zeit verschärft die Frage nach den Kriterien der Textauswahl. Dabei ist jetzt schon davon auszugehen, dass das in der Korpuslinguistik gängige dreidimensionale Raster *Raum – Zeit – Textsorte* für die Fragestellung des Projekts wenig ergiebig sein wird. Die in den Texten identifizierten Belege werden in eine Datenbank eingetragen. Für die Eingabe der Daten wurde eine webbasierte Anwendung implementiert. Die grafische Weboberfläche ist in HTML realisiert, die Anwendungslogik und die Kommunikation mit der Datenbank regeln mehrere PHP-Skripte. Das technische Konzept sieht die Nutzung einer relationalen MySQL-Datenbank vor, die über standardisierte Schnittstellen erreichbar ist. Die Bestandsdaten sind somit von der Anwendungs- und Präsentationslogik getrennt und stehen als eigenständige Ressource zur Verfügung. Diese Architektur ermöglicht eine dezentrale Datenpflege und eine simultane Nutzung der Arbeitsplattform.[21] Die historischen Belege werden in der Datenbank (und nicht in den Texten) nach dem im Projekt entwickelten Kriterienkatalog kommentiert. Die Benutzerschnittstelle besteht aus:

– einem Formular, in dem bibliographische Angaben zu den ausgewerteten Quellen erfasst werden,
– einem Belegkorpus,
– einer Maske zur Belegsuche und
– einem Tool zur Belegcorpusverwaltung.

Den Kern der Benutzerschnittstelle machen fünf Erfassungsmasken aus, die die Analyse eines jeden Belegs nach drei semiotischen Dimensionen – Lexik/Semantik, Syntaktik und Pragmatik – sowie nach seinem kulturhistorischen Hintergrund ermöglichen. Die Erfassungsmasken sollen im Folgenden vorgestellt werden.

[21] Für die theoretische und praktische Unterstützung bei der Konzipierung der Datenbank bedanke ich mich herzlich bei Prof. Dr. Christiane Fellbaum und ihren Mitarbeiterinnen und Mitarbeitern im Projekt "Kollokationen im Wörterbuch", die mir während meines Besuchs an der Berlin-Brandenburgischen Akademie der Wissenschaften im Februar 2007 die Datenbank und die Arbeitsweise des Projekts ausführlich vorgestellt und ein kompetentes Forum für die Diskussion einiger methodischer Probleme angeboten haben.

Abbildung 1: Erfassungsmaske BELEG

In der ersten Erfassungsmaske wird der Kontext notiert, in dem der Beleg im originalen Textzeugen (Handschrift, Druck oder philologisch verlässliche diplomatische Edition rezenteren Datums) vorkommt. Beim Speichern wird ihm automatisch eine ID-Nummer zugeteilt. Der Datensatz, der als Beleg bezeichnet wird, ist meistens auch ein Leitbeleg; er ist von "Abhängigen Textzeugenbelegen" zu unterscheiden.[22] Verglichen mit einem gegenwartssprachlichen Wörterbuch wären das die Beispielsätze, die traditionell z. B. in einem idiomatischen Nachschlagewerk einer Nennform zugeordnet werden. Da die Annahme einer Nennform im historischen Kontext anders als im Neuhochdeutschen willkürlich wäre, werden im HiFoS-Projekt die Belege nicht als Beispielsätze betrachtet, sondern primär als Untersuchungsgegenstand verstanden. Die Eingabe erfolgt handschriftengetreu und unicode-konform, inklusive Sonderzeichen und Zeilenumbrüche. Die Länge des Kontextes ist nicht festgelegt und muss jedes Mal vom Bearbeiter neu bestimmt werden. Als Kriterium gilt dabei die Relevanz für das Verständnis des jeweiligen Belegs. Der Beleg selbst ist im Kontext farbig unterlegt; auf ihn beziehen sich alle weiteren Kommentierungen in der Datenbank. Entsprechend der in Abschnitt 3 angeführten Definition der formelhaften Sprache werden in der aktuellen Untersuchungsetappe auch Belege im Status der "Beleg-Kandidaten" aufgenommen: Ihr formelhafter Charakter muss durch das Vorhandensein ähnlicher Belege in den späteren Sprachstufen bzw. in anderen Sprachen erst bewiesen werden.[23]

In dieser Erfassungsmaske werden außerdem Angaben zur Belegstelle und -quelle gemacht, die die Verbindung zum jeweiligen Quellen-Datensatz gewährleisten. Der Beleg wird ferner ins Neuhochdeutsche übersetzt. Die Übersetzung stellt keinen literarischen Anspruch und soll im Gegenteil möglichst textnah sein. Außerdem wird hier der Typ des Belegs kommentiert, wozu zunächst die Mischklassifikation H. Burgers (2007: 33–58) übernommen wurde, die sich im Großen und Ganzen bewährt hat. Allerdings sind auch in diesem Bereich Abweichungen von der Gegenwartssprache festzustellen. Begrifflichkeiten wie *Sprichwort* oder *geflügeltes Wort* sind sprachhistorisch problematisch, weil der Grad ihrer allgemeinen Geläufigkeit auf Grund singulärer Belege und fehlender metasprachlicher Äußerungen nicht immer nachgewiesen werden kann. Ebenfalls schwierig ist die Klassifizierung *phraseolo-*

[22] Cf. dazu unten die Darstellung der Erfassungsmaske 6.

[23] Cf. ausführlicher zu dieser Problematik Filatkina (2009).

gischer Terminus, besonders bezogen auf die ältesten Sprachstufen, in denen von einer etablierten Fachsprache oft noch nicht die Rede sein kann. Es konnten ferner Typen von formelhaften Wendungen aufgedeckt werden, für die das in der Phraseologieforschung etablierte typologische Raster nicht genügt. Die Klassifikation versteht sich deshalb als ein offenes ausbaufähiges Raster, das auch die Möglichkeit der Mehrfachzuordnung vorsieht.

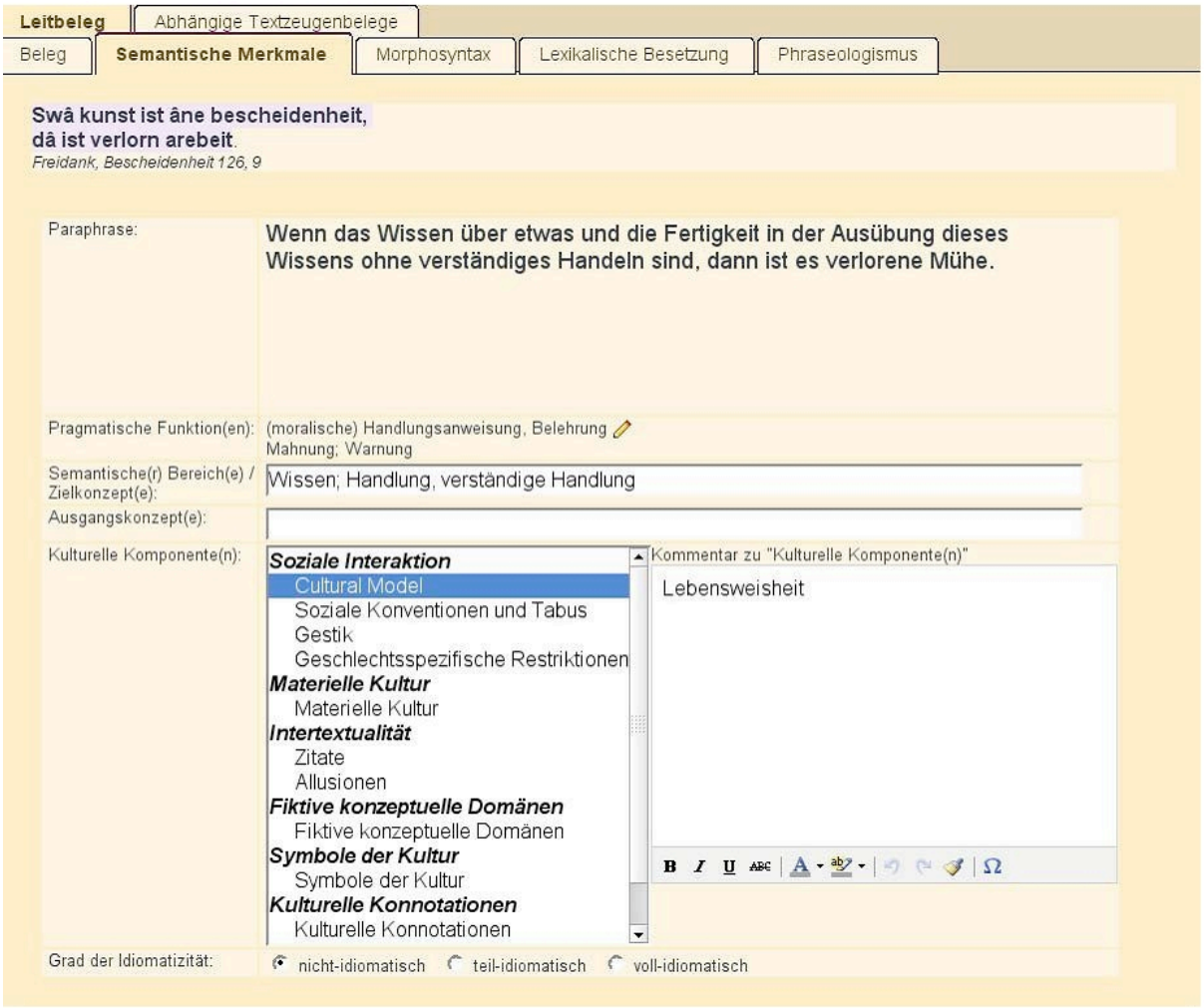

Abbildung 2: Erfassungsmaske SEMANTISCHE MERKMALE

Der nächste Bereich bezieht sich auf die Semantik des Belegs. Im Feld "Paraphrase" wird die aktuelle Bedeutung des Belegs kommentiert, die er im gegebenen Kontext trägt. Im Gegensatz zu einigen existierenden linguistischen Datenbanken, die die standardisierte Metasprache der lexikalischen Semantik verwenden, wird die Paraphrase in der HiFoS-Datenbank frei formuliert. Dies ermöglicht die ausführliche und kontextbezogene Beschreibung des Belegbrauchs, in die auch stilistische und teilweise pragmatische Angaben einfließen, um auf diese Weise eine exakt dokumentierte Belegbasis zu erstellen, die Vielfalt des Gebrauchs einer formelhaften Wendung zu dokumentieren, bzw. ihre feinere Analyse in einem semantischen Feld (in der Synchronie) durchzuführen. Dieses Verfahren wird auch bei der Ermittlung der semantischen Unterschiede gegenüber dem Neuhochdeutschen (in der Diachronie) ergiebig sein.

Das nächste Feld dient der Erfassung der pragmatischen Funktion(en) des Belegs im Kontext. Mit dem Begriff "Pragmatik" ist vor allem das illokutive Potenzial einer formelhaften Wendung (und nicht etwa ihre stilistische Markierung) gemeint. Die Angaben in diesem Feld erlauben im Gegensatz zur konventionellen lexikographischen Praxis, das illokutive Potenzial

als festen Bestandteil der Semantik der formelhaften Wendungen zu beschreiben. Sie bilden die Grundlage für die Beantwortung der Frage nach der Rolle der Formelhaftigkeit in der Sprachgeschichte. Die pragmatische Funktion wird aus einer Vorschlagsliste gewählt. Diese Liste ist offen und wird sukzessive erweitert bzw. spezifiziert. Die bislang benutzten Kategorien sind den Untersuchungen zur formelhaften Sprache am Gegenwartsdeutschen entnommen und werden am historischen Material überprüft. Da es oft schwierig ist, die pragmatische Funktion eines Belegs mit wenigen Begriffen zu erfassen, ist das Feld "Pragmatische Funktion(en)" zusätzlich um das Feld "Funktionsspektrum" ergänzt. Es bietet die Möglichkeit, die erfolgte Einordnung nun durch Wahl zusätzlicher Begriffe aus einer mit einem Begriffscluster benannten Kategorie zu spezifizieren oder aber ebenfalls im Beleg vorhandene, aber nicht dominierende pragmatische Funktionen anzuführen, etwa durch Nennung der entsprechenden Begriffe aus oben erwähnter Liste.

An das Feld "Paraphrase" ist das Feld "Semantischer Bereich, Zielkonzept" gekoppelt. Es handelt sich dabei um einen relativ abstrakten Begriff, der mittels einer gegebenen Wendung versprachlicht wird und der sich aus der Paraphrase ergibt. Angaben zum semantischen Bereich werden bei allen Belegen mit einer referentiellen (nominativen wie propositionalen) Funktion gemacht. Dabei spielt weder der Grad der Idiomatizität noch der Typ der Bildlichkeit eine Rolle. Bei diesem Feld wurde bewusst auf die Übernahme der fertigen Begriffsraster verzichtet, um sich ausschließlich vom eigenen Textmaterial leiten zu lassen. Im Zuge der ersten Auswertungen wurde allerdings eine Liste mit solchen Begriffen erarbeitet. Sie ist nicht geschlossen und wird sukzessive ergänzt.

Das Feld "Ausgangskonzept" bezieht sich vor allem auf die enge Klasse der idiomatischen Belege, die zugleich bildlich (metaphorisch und/oder symbolisch motiviert) sind. Der Begriff wird in Anlehnung an die *Figurative Language Theory* (Dobrovol'skij/Piirainen 2005) als das Konzept, das einen Bereich des realen und/oder imaginären Lebens darstellt, der als Grundlage für metaphorische bzw. symbolische Übertragung benutzt wird, definiert. Dieses Feld ist an das Feld "Etymologie" gekoppelt[24] und unterscheidet sich deshalb grundlegend von dem in der Linguistik ebenfalls verbreiteten Terminus *Sachfeld* oder *Sachbereich*. Die Einträge in diesem Feld ermöglichen die Beantwortung der aus kulturhistorischer Perspektive wichtigen Frage nach den Bereichen, die von einer historischen Sprache als Grundlagen für metaphorische Umdeutungen benutzt werden.

Diesem Ziel dient auch das nächste Feld. Die so genannte "Kulturelle Komponente" gibt Auskunft über die kulturellen Rahmenbedingungen, unter denen eine formelhafte Wendung entstanden ist bzw. über den kulturellen Aspekt, der bei ihrer Herausbildung eine Rolle gespielt hat. Die Klassifikation dieses Bereichs wurde zunächst der *Figurative Language Theory* entnommen, obwohl die Notwendigkeit der Modifikation aus sprachhistorischer Perspektive sich schnell gezeigt hat. Sie betrifft vor allem die Kategorie "Cultural Model", die durch 10 Unterkategorien präzisiert werden konnte. Abweichend vom gegenwartssprachlichen Befund konstituiert sich für historisches Material ferner die Kategorie "Fiktive konzeptuelle Domänen", denn das Wissen, das heute als fiktiv charakterisiert werden kann (z. B. Temperamentenlehre, das Wissen über Tiere oder menschliche Körperorgane), kann für das Mittelalter nur bedingt als fiktiv gelten. Im Feld "Kommentar" kann sich der Bearbeiter zu den vorgenommen Zuordnungen äußern.

In dieser Erfassungsmaske wird außerdem durch die Wahl einer vorgegebenen Option (nicht-idiomatisch, teil-idiomatisch, voll-idiomatisch) der Grad der Idiomatizität kommentiert.

[24] Cf. dazu die Ausführungen zu Erfassungsmaske 5.

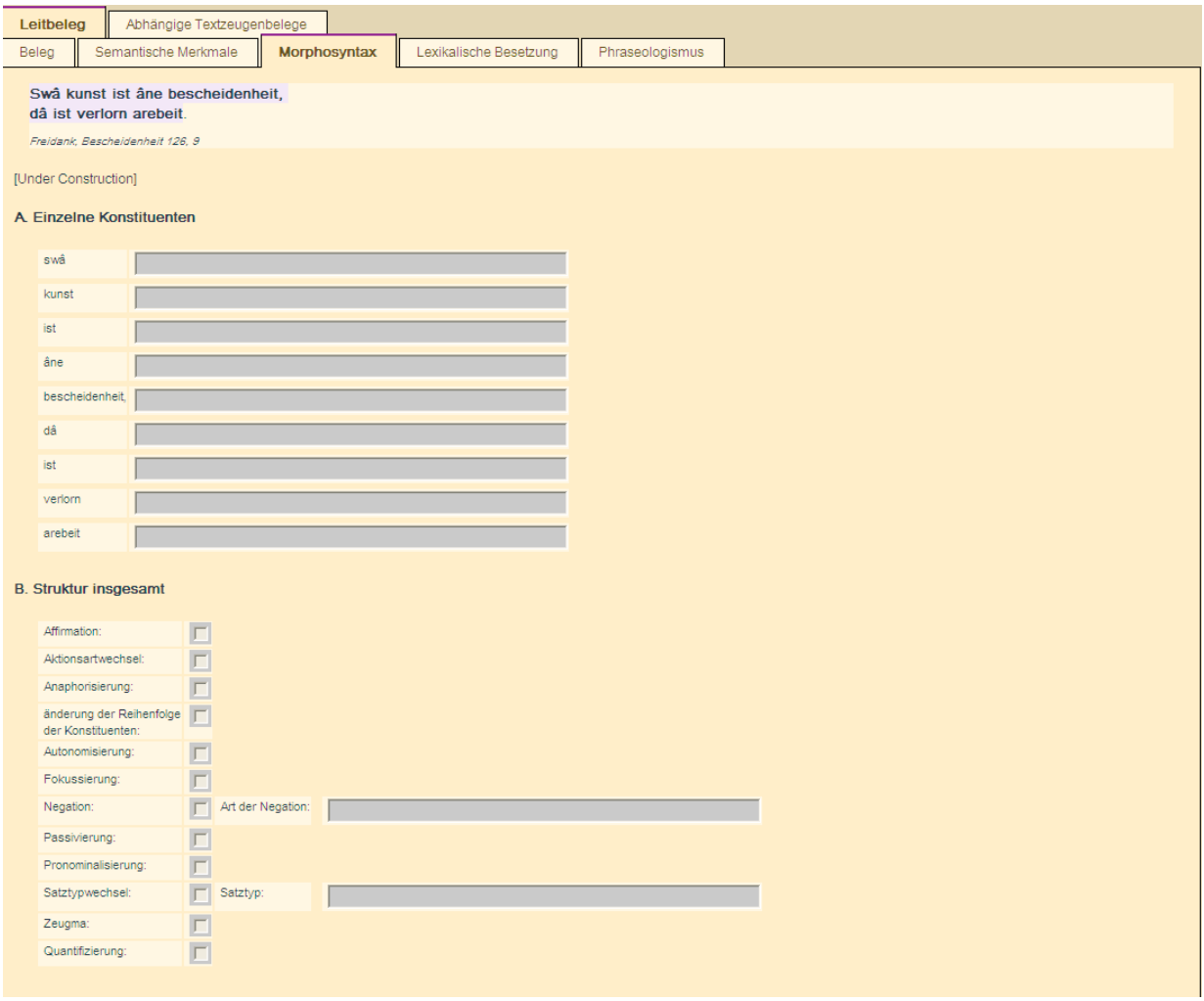

Abbildung 3: Erfassungsmaske MORPHOSYNTAX

Der nächste Bereich widmet sich der Morphosyntax der Belege und befindet sich noch in der Testphase. Möglich wäre die Verlagerung dieser Art von Kommentierungen auf die annotierten elektronischen Volltexte, um so die Verteilung der Informationsstrukturen zwischen dem Textkorpus und der Datenbank zu optimieren. Konzipiert sind hier aber einerseits Kommentierungen, die sich auf grammatische Formen einzelner Konstituenten beziehen und in diese freien Felder eingetragen werden. Andererseits sollen Kommentierungen vorgenommen werden, die sich auf mögliche syntaktische Variationen in der gesamten Struktur der Belege beziehen. Die bisherige Forschung stellt kein einheitliches Raster dieser Variationsmöglichkeiten zur Verfügung. Im HiFoS-Projekt werden deshalb die syntaktischen Eigenschaften mit Hilfe des Rasters erfasst, das in Kooperation mit dem Projekt "Kollokationen im Wörterbuch" an der Berlin-Brandenburgischen Akademie der Wissenschaften (Prof. Dr. Christiane Fellbaum) erstellt wurde. Allerdings musste dieses Raster für sprachhistorische Fragestellungen modifiziert werden. In Planung ist ferner seine Ergänzung um weitere sprachstufenspezifische Kriterien. Die Felder sind momentan einem Beleg zugeordnet, obwohl ihre Füllung erst in einer Gruppe der grammatischen Varianten eines Belegs möglich sein wird. Dies bedeutet, dass als Vergleichsbasis nicht wie im Neuhochdeutschen die Nennform gilt, sondern andere Belege in dieser Gruppe.

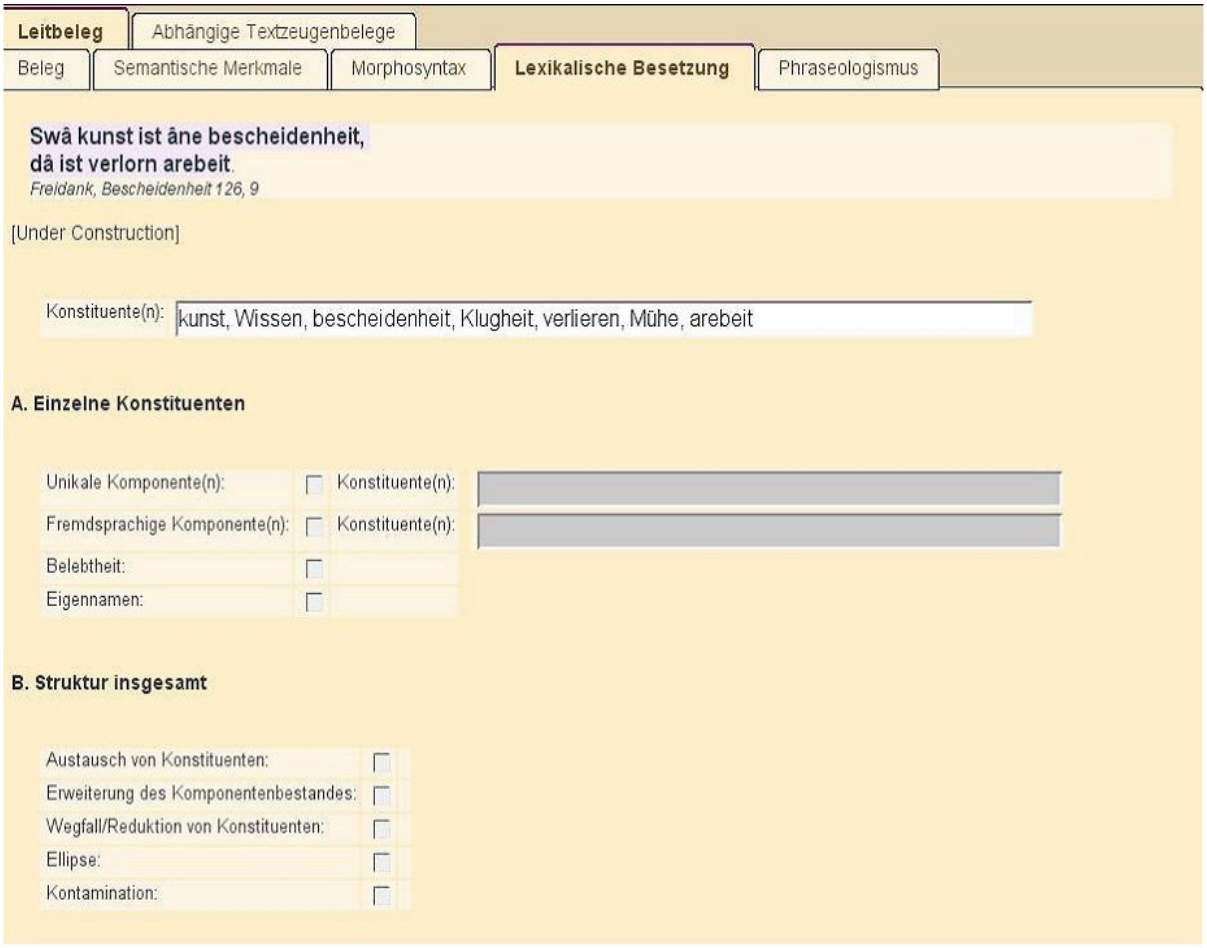

Abbildung 4: Erfassungsmaske LEXIKALISCHE BESETZUNG

Eine ähnliche Vorgehensweise ist für die Kommentierung der lexikalischen Variation im nächsten Bereich "Lexikalische Besetzung" vorgesehen. Als Erstes werden hier die Konstituenten des jeweiligen Belegs notiert, die für die Konstituierung seiner aktuellen Bedeutung ausschlaggebend sind. Die Konstituenten werden in unflektierten, meistens neuhochdeutschen Formen angeführt, gegebenenfalls auch in Althochdeutsch, Mittelhochdeutsch und Frühneuhochdeutsch, wenn sie ausgestorben sind oder starke Bedeutungsveränderungen aufweisen. Dieses Feld ermöglicht dem späteren Benutzer die diachrone Suche über Sprachstadien hinweg. Kommentiert werden hier ferner die lexikalischen Besonderheiten einzelner Konstituenten wie Unikalia, Fremdwörter, Eigennamen bzw. Belebtheit oder Unbelebtheit.

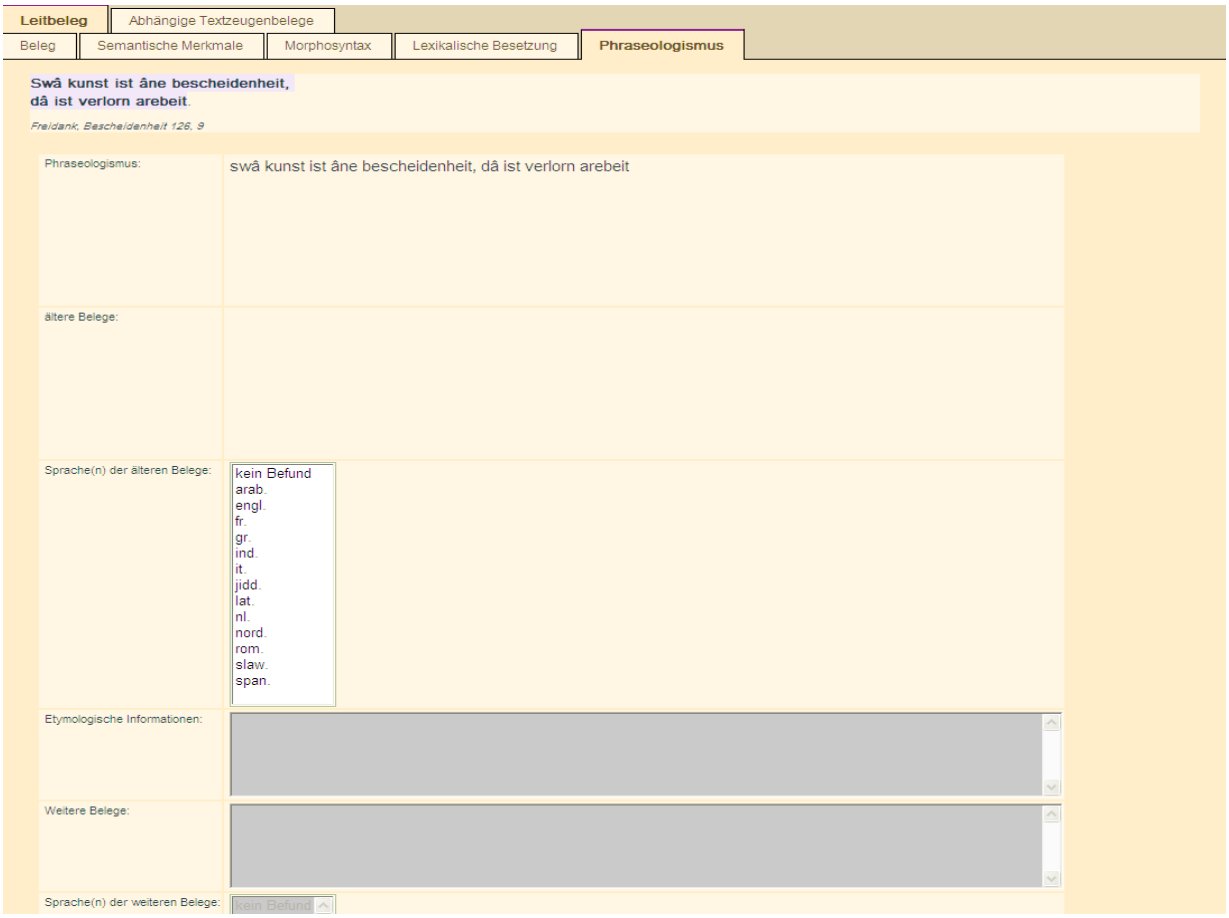

Abbildung 5: Erfassungsmaske PHRASEOLOGISMUS

Zum Schluss wird jeder Beleg in der letzten Erfassungsmaske dem so genannten "Phraseologismus" zugeordnet. Da im Projekt bis jetzt primärquellenbasiertes Sammeln und Dokumentieren von Belegen im Vordergrund stand, sind einzelne Belege noch relativ lose und isoliert voneinander in der Reihenfolge ihrer Bearbeitung im Belegkorpus aufgelistet. Um aber die Analyse der Variationstypen möglich zu machen, sollen die Varianten einer potentiellen formelhaften Wendung über einen gemeinsamen Nenner zusammengeführt werden. Als dieser gemeinsame Nenner ist das Feld "Phraseologismus" vorgesehen.

Es handelt sich dabei um eine vom Bearbeiter abstrahierte Phraseologismus-Form, die rein dienende Funktionen der Zuordnung der grammatischen bzw. lexikalischen Varianten zu einem Phraseologismus hat. In der neuhochdeutschen Lexikographie ist sie ungefähr mit dem Begriff "Nennform" vergleichbar. Sprachhistorisch hat sie allerdings keinen Aussagewert, erhebt keinen Anspruch darauf, die sprachhistorisch geläufigste, "korrekte" oder älteste Form zu sein, und hat rein technische Funktionen der Belegbündelung. Die adäquate Formulierung dieser Phraseologismus-Form verlangt eine gewisse Belegzahl, die die Abstraktion zulässt. Im Moment wird der Phraseologismus in der Form aufgenommen, in der er im Beleg vorkommt. Abstraktionsmöglichkeiten bestehen nur bei der Markierung von Leerstellen und bei der Rückführung von Verben auf den Infinitiv.

In den weiteren Feldern werden Angaben zu den früheren Belegen sowie ihrer Sprache gemacht. Sie werden dokumentiert und dienen nicht als Grundlage für die Schlussfolgerung, dass ein Lehnprozess vorliegt. Obwohl es im Projekt nur ein Bereich ist, der nicht unmittelbar im Vordergrund steht, werden diese Angaben als ein Beitrag zur Geschichte der europäischen Phraseologie betrachtet. Die Etymologie wird ebenfalls kommentiert, obwohl ihre Ermittlung nicht im Vordergrund des Projekts steht. Informationen dieser Art werden der Sekundär-

literatur entnommen und bei der Formulierung der kulturhistorisch relevanten Aussagen über die Ausgangskonzepte berücksichtigt.

Alle Erfassungsmasken enden mit den Feldern "Notizen" (freie Kommentierungen der Bearbeiter), "Laufzettel" (mit den Siglen der Hauptnachschlagewerke), "Bearbeitungsstand" und "Bearbeitungskommentar". Darauf kann von allen Registern der Erfassungsmaske in gleichem Maße zurückgegriffen werden. Sie dienen der Dokumentation der Bearbeitungsschritte und somit der Arbeitsorganisation im Projektteam.

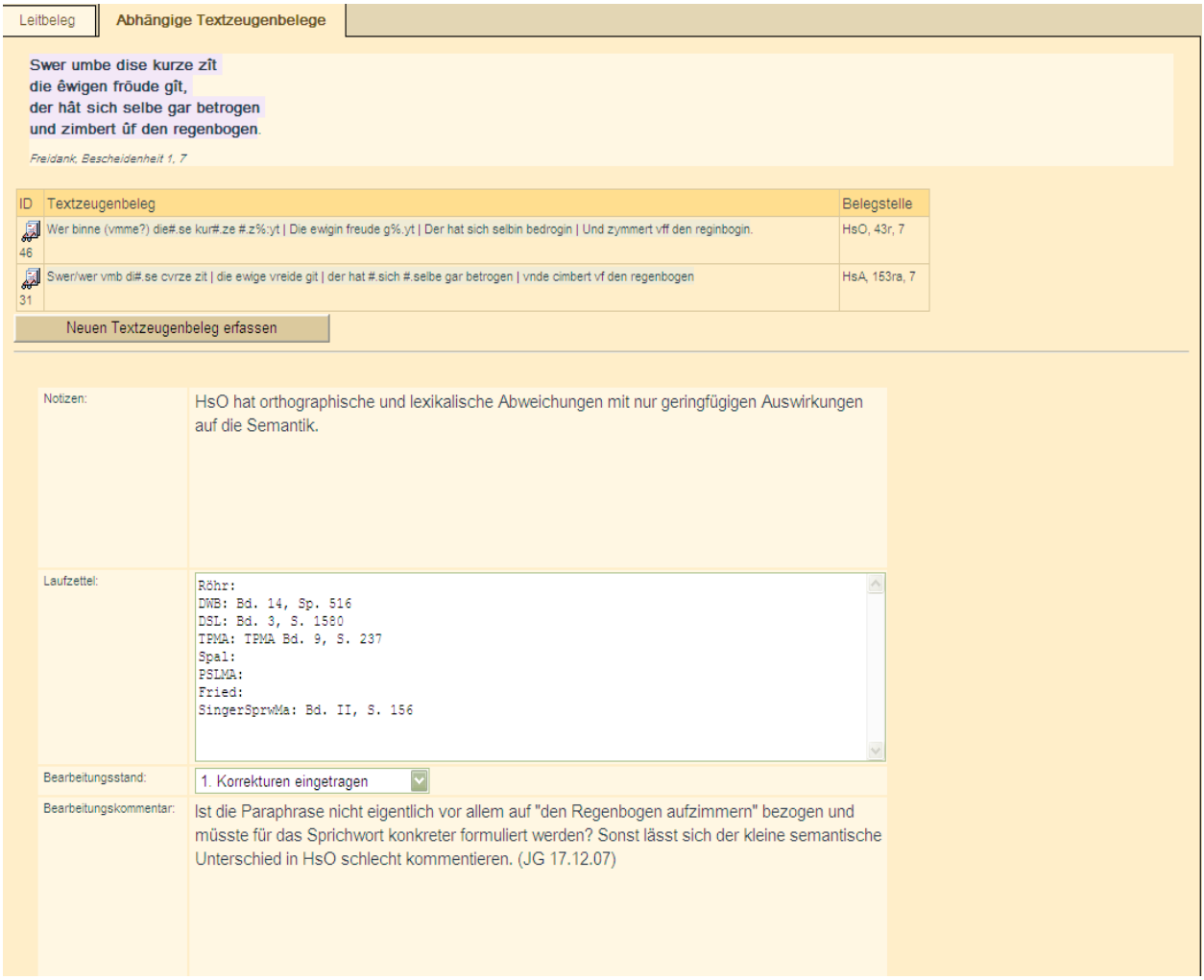

Abbildung 6: Erfassungsmaske ABHÄNGIGE TEXTZEUGENBELEGE

Mit der Erfassungsmaske 6 ist die Möglichkeit vorgesehen, bei einem Text, der z.B. in mehreren Handschriften und/oder Drucken überliefert ist, Belege aus mehreren Textzeugen einzutragen und dabei ihre Zugehörigkeit zu dem Beleg zu visualisieren, der als Leitbeleg betrachtet wird. In solchen Fällen, in denen Textzeugen diatopisch und zeitlich zerstreut sind, ermöglicht die Zusammenführung der Belege die Untersuchung der Dynamik innerhalb eines Textes. Abhängige Textzeugenbelege verfügen ihrerseits über die oben erwähnten Erfassungsmasken (mit Ausnahme der Erfassungsmaske "Phraseologismus") und können in Bezug auf dieselben Kriterien kommentiert werden. Dabei werden stets nur Abweichungen gegenüber dem Leitbeleg manuell erfasst, die übrigen Kommentierungen werden automatisch vom Leitbeleg ererbt.

5.3 Beleg(corpus)verwaltung

Die Belegverwaltung sieht die Möglichkeit vor, isoliert stehende Belege nach unterschiedlichen Kriterien für sprachhistorische Analysen zu bündeln. Diese Kriterien können von den Mitarbeitern abhängig von der Fragestellung beliebig definiert werden. Abfragbar sind alle Ebenen (Belege sowie Quellen) und alle Felder der Datenbank, so dass die Belege zu ganz unterschiedlichen Gruppen und zu mehreren Gruppen gleichzeitig zugeordnet werden können. Es ist z. B. vorgesehen, semantisch ähnliche Belege, Belege aus einer Quelle, einer Textsorte, eines Autors bzw. Belege, die zu einem thematischen Bereich oder einem Ausgangskonzept gehören, zusammenführen und diese Gruppen auch auf Dauer speichern zu können. Grammatische und lexikalische Varianten zu einem "Phraseologismus" können über das Kriterium "Phraseologismus" gruppiert werden. Die Gruppierung erfolgt im Moment manuell: Nachdem das Kriterium für die Gruppenbildung definiert wurde (also "Phraseologismus"), kann sich der Bearbeiter über die Suchfunktion die entsprechenden Belege anzeigen lassen. Interessiert man sich z. B. für die grammatischen und lexikalischen Varianten der Paarformel *gut und übel*, so sollen in der Suchmaske die Konstituenten *gut* und *übel* in Nhd. und zur Einschränkung der Suche noch der Typ "Paarformel" angeben werden. Im Output erscheint ein Referenzkorpus bzw. ein Subkorpus mit derzeit 25 Belegen, die nun zu engerer Betrachtung zu einer Gruppe zusammengefasst und als Gruppe für einen unbegrenzten Zeitraum gespeichert werden können.

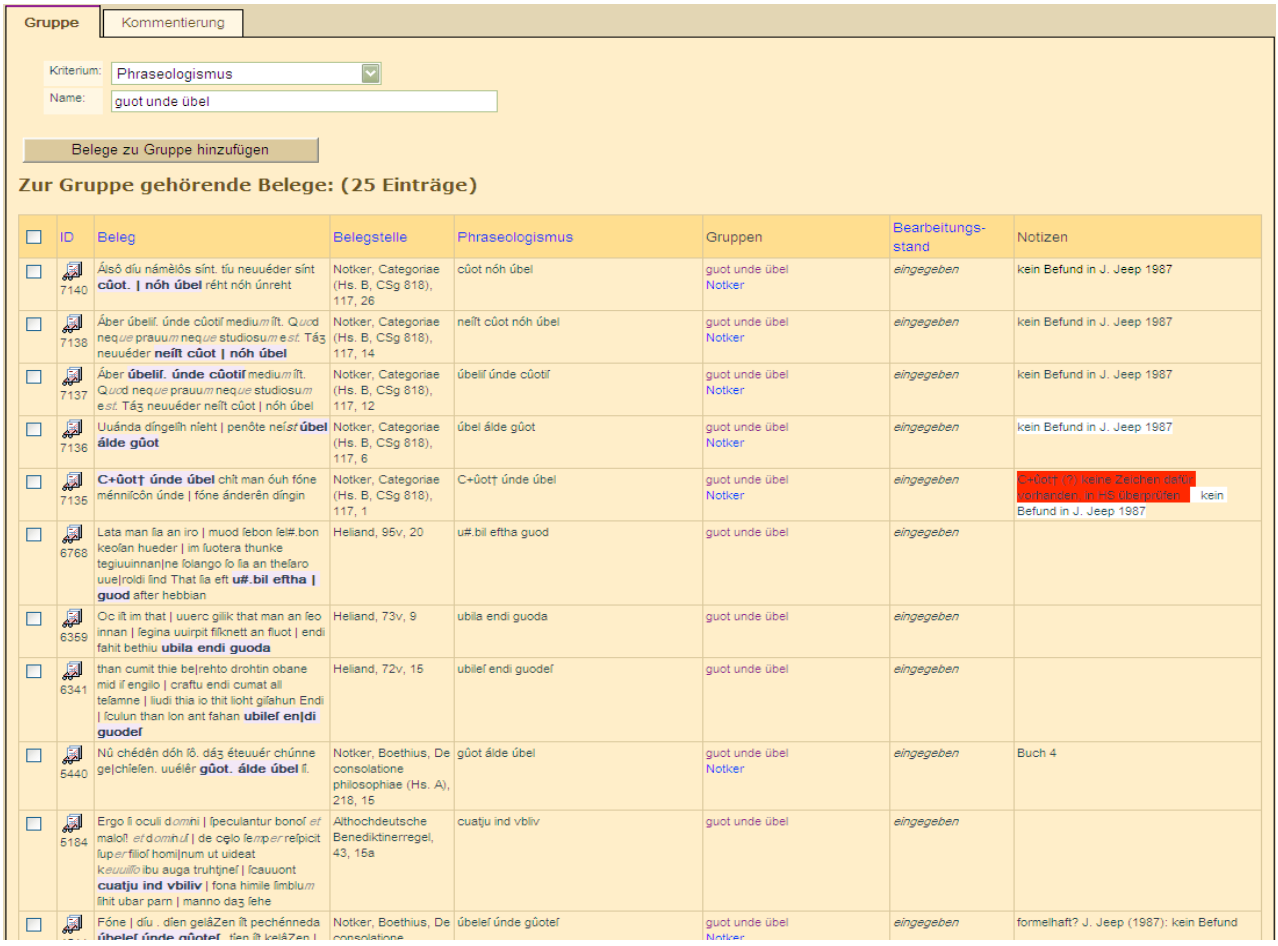

Abbildung 7: BELEGGRUPPE "GUOT UNDE ÜBEL"

Das Analyseergebnis wird in den Kommentarfeldern notiert. So ist in Bezug auf diese Paarformel festzuhalten, dass sowohl die Morphosyntax als auch die Semantik in Bewegung sind: Die Konstituenten können in Plural versetzt, mit dem bestimmten Artikel versehen und

substantiviert werden; ihre Reihenfolge kann vertauscht werden. Es liegt Varianz in der Konjunktion vor (Substitution von *und* durch *oder* oder *wider*). Die Paarformel kann durch eine Präposition (z.B. *ze* oder *an*) erweitert werden.

Auch wenn die Suchfunktion das Formulieren komplexer Anfragen ermöglicht, ist dieses Werkzeug für die Belegbündelung nicht optimal. Insbesondere das Fehlen einer einheitlichen Rechtschreibung erschwert das Auffinden aller relevanten Belege. Daher wird zur Zeit ein Programm entwickelt, das zu dem aktuell bearbeiteten Beleg selbständig in der Datenbank nach ähnlichen Belegen sucht und dem Bearbeiter eine Liste mit den besten Treffern liefert.[25] Hierzu werden Verfahren verwendet, welche Ähnlichkeiten zwischen Zeichenketten messen und somit ähnliche Wörter erkennen können. Das Ähnlichkeitsmaß kann der jeweiligen Fragestellung angepasst werden, indem die Gewichtung der einzelnen Felder festgelegt wird.

6 Desiderata und Perspektiven

Im vorliegenden Beitrag wurde der Versuch unternommen, die korpus- und computerlinguistischen Ansätze für die Untersuchung der historischen formelhaften Sprache fruchtbar zu machen. Dabei hat sich herausgestellt, dass dies u. a. angesichts der fehlenden philologisch verlässlichen elektronischen Korpora für ältere Sprachstufen des Deutschen sowie der vielfältigen Variation der historischen Belege bei der (oft) gleichzeitigen Singularität ihrer Überlieferung nur bedingt möglich ist. Mit dem Bewusstsein für diese Umstände, aber auch mit dem Blick auf den heutigen Stand der Untersuchungen im Bereich der historischen formelhaften Sprache wurde in der HiFoS-Nachwuchsforschergruppe eine Vorgehensweise entwickelt, die einerseits dieser Forschungslage Rechnung trägt, aber andererseits empirisch abgesicherte Angaben zur Verwendung und Dynamik der historischen formelhaften Wendungen im Sinne der Grundlagenforschung liefert.

Durch die Anbindung der HiFoS-Datenbank an ein Textkorpus sowohl im Rahmen eines Konsultationsparadigmas (Steyer 2004: 93) als auch im Rahmen eines Analyseparadigmas könnten diese Angaben für den breiteren Benutzerkreis nachvollziehbarer gemacht werden. Beim ersten würden sich die Überprüfung der Geläufigkeit der Belege, die Visualisierung des größeren Kontextes und somit auch die Verifizierung der Kommentierungen lohnen. Im Rahmen des Analyseparadigmas gewährleistet die Korpusbenutzung die Annäherung an die Frage nach dem Extrahieren von formelhaften Wendungen aus historischen Texten.

Nicht trivial wird ferner die geplante Verlinkung der HiFoS-Datenbank mit anderen an der Universität Trier im Entstehen begriffenen Datenbanken zu jiddischen Phraseologismen und zu den so genannten Sprichwortbildern sein.[26] Sie bringt eine Erhöhung der Komplexität mit sich, denn es werden Belege in einem anderen Schriftsystem und zusätzlich zu verbalen Informationen visuelle Informationen in die Datenstruktur aufgenommen. Damit öffnen sich durch formelhafte Fragestellungen Türen für zukünftige Forschung, die über Probleme der sprachlichen Repräsentation oder technischen Organisation hinausgehen und den schlechten Ruf der formelhaften Wendungen als "harte Nuss der Korpus- und Computerlinguistik" hoffentlich bald widerlegen werden.

---

[25] Cf. dazu ausführlicher H. Dostert (2009).

[26] Ausführlicher zu den Projekten "Form und Formung sprachkonzeptueller Wissensräume: Jiddische Phraseologie im Kontext europäischer Sprachen (JPhras)" (Leitung: Dr. Ane Kleine) und "GnoViS – Gnomik Visuell. Gnomisches Wissen im Raum der Bilder. Die Visualisierung von Sprichwörtern in der Kunst des Mittelalters und der Frühen Neuzeit" (Leitung: Dr. Birgit Ulrike Münch) vgl. www.hkfz.uni-trier.de (Arbeitsgruppe "Wissensraum Kommunikation: Kulturelle Praktiken, Tradition und Wandel").

Literatur

Burger, Harald (1983): "Neue Aspekte der Semantik und Pragmatik phraseologischer Wortverbindungen". In: Matešič, Josip (ed.): *Phraseologie und ihre Aufgaben. Beiträge zum 1. Internationalen Phraseologie-Symposium vom 12. bis 14. Oktober 1981 in Mannheim*. Band 3. Heidelberg, Groos: 24–34.

Burger, Harald (2007): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 3. Auflage. Berlin: Erich Schmidt Verlag.

Burger, Harald/Linke, Angelika (1998): "Historische Phraseologie". In: Besch, Werner et al. (eds.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. Band 1. 2., vollständig neu bearbeitete und erweiterte Auflage. Berlin/New York, de Gruyter: 743–755.

Coulmas, Florian (1981): *Routine im Gespräch. Zur pragmatischen Fundierung der Idiomatik*. Wiesbaden: Athenaion.

Cowie, Antony (1992): "Multiword lexical units and communicative language teaching". In: Arnaud, Pierre J. L./Béjoint, Henri (eds.): *Vocabulary and Applied Linguistics*. London, Macmillan: 1–12.

Dobrovol'skij, Dmitrij (1992): "Angewandte Phraseologie: Zu einigen aktuellen Problemen". In: Große, Rudolf et al. (eds.): *Beiträge zur Phraseologie, Wortbildung und Lexikologie. Festschrift für Wolfgang Fleischer zum 70. Geburtstag*. Frankfurt a. M., Suhrkamp: 29–36.

Dobrovol'skij, Dmitrij/Piirainen, Elisabeth (2005): *Figurative Language. Cross-cultural and cross-linguistic perspectives*. Amsterdam etc: Elsevier.

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.

Eikelmann, Manfred (1998): "Autorität und ethischer Diskurs. Zur Verwendung von Sprichwort und Sentenz in Hartmanns von Aue 'Iwein'". In: Anderson, Elizabeth et al. (eds.): *Autor und Autorschaft im Mittelalter. Kolloquium Meißen 1995*. Tübingen, Niemeyer: 73–100.

Eikelmann, Manfred (2002): "Zur historischen Pragmatik des Sprichworts im Mittelalter". In: Hartmann, Dietrich/Wirrer, Jan (eds.): *Wer A sägt, muss auch B sägen. Beiträge zur Phraseologie und Sprichwortforschung aus dem Westfälischen Arbeitskreis*. Baltmannsweiler, Schneider: 95–105.

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

Fellbaum, Christiane (ed.) (2006): *Corpus-based studies of German idioms and light verbs. International Journal of Lexicography* 19/4.

Fellbaum, Christiane (ed.) (2007): *Idioms and collocations. Corpus-based linguistic and lexicographic studies*. London/New York: continuum.

Filatkina, Natalia (2007): "Formelhafte Sprache und Traditionen des Formulierens (HiFoS): Vorstellung eines Projekts zur historischen formelhaften Sprache". *Sprachwissenschaft* 32–2: 217–242.

Filatkina, Natalia (2009): "Historical Phraseology of German: regional and global". Erscheint in: Korhonen, Jarmo (ed.): *Europhras 2008. Akten der internationalen Tagung am 13.08.–16.08.2008 an der Universität Helsinki*.

Filatkina, Natalia/Gottwald, Johannes/Hanauska, Monika/Rößger, Carolin (2009): "Formel–hafte Sprache im schulischen Unterricht im Frühen Mittelalter: Am Beispiel der so genannten "Sprichwörter" in den Schriften Notkers des Deutschen von St. Gallen". Erscheint in *Sprachwissenschaft*.

Filatkina, Natalia (im Druck): "Und es duencket einem noch/wann man euch ansiehet / daß ihr Sand in den Augen habt. Phraseologismen in ausgewählten historischen Grammatiken

des Deutschen". Erscheint in: Földes, Csaba (ed.): *Europhras 2006: Phraseologie disziplinär und interdisziplinär*.

Geyken, Alexander (2004): "What is the optimal corpus size for the study of idioms?". Lecture presented at the annual meeting of the DGfS, Mainz 2004.

Gottwald, Johannes (2009): "Formelhaftigkeit im städtischen Schrifttum. Nürnberger Fastnachtspiele des 15. und 16. Jahrhunderts". In: Libuše Spácilová et al. (eds.): *Akten der 26. Tagung des Internationalen Arbeitskreises Historische Stadtspachenforschung. Olomouc, 09.–11. Oktober 2008*.

Gülich, Elisabeth (1997): "Routineformeln und Formulierungsroutinen. Ein Beitrag zur Beschreibung 'formelhafter Texte'". In: Wimmer, Rainer/Berens, Franz-Josef (eds.): *Wortbildung und Phraseologie*. Tübingen, Niemeyer: 130–175.

Hanauska, Monika (2009): "Formelhafte Sprache im städtischen Schrifttum: Die Kölner Stadtchroniken im Spätmittelalter". In: Libuše Spácilová et al (eds.): *Akten der 26. Tagung des Internationalen Arbeitskreises Historische Stadtsprachenforschung. Olomouc, 09.–11. Oktober 2008*.

Heid, Ulrich (2007): "Computational linguistic aspects of phraseology II". In: Burger, Harald et al. (eds.): *Phraseologie/Phraseology. An International Handbook of Contemporary Research/Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York, de Gruyter: 1036–1044.

Heid, Ulrich (2008): "Computational phraseology. An overview". In: Granger, Sylviane/ Meunier, Fanny (eds.): *Phraseology. An interdisciplinary perspective*. Amsterdam, Benjamins: 337–360.

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

Hoffmann, Walther (1998): "Probleme der Korpusbildung in der Sprachgeschichtsschreibung und Dokumentation vorhandener Korpora". In: Besch, Werner et al. (eds.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. 2., vollständig neu bearbeitete und erweiterte Auflage. Berlin/New York, de Gruyter: 875–889.

Keil, Martina (1997): *Wort für Wort. Repräsentation und Verarbeitung verbaler Phraseologismen (Phraseo-Lex)*. Tübingen: Niemeyer.

Jackendoff, Ray (1995): "The boundaries of the lexicon". In: Everaert, Michael et al. (eds.): *Idioms: Structural and psychological perspectives*. New York, Hillsdale: 133–165.

Jackendoff, Ray (1997): "Twistin' the Night away". *Language* 73: 534–539.

Moon, Rosamund (1998): *Fixed Expressions and Idioms in English. A corpus-based approach*. Oxford: Clarendon Press.

Moon, Rosamund (2007): "Corpus linguistic approaches with English corpora". In: Burger, Harald et al. (eds.): *Phraseologie/Phraseology. An International Handbook of Contemporary Research/Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York, de Gruyter: 1045–1059.

Odijk, Jan (2004): "A proposed standard for the lexical representation of idioms". In: Williams, Geoffrey/Vessier, Sandra (eds.): *Proceedings of the 11th Euralex International Congress, EURALEX 2004, Lorient, France, July 6–10*. Vol. 1. Lorient, Université de Bretagne-Sud: 153–164.

Sag, Ivan A. et al. (2002): "Multiword expressions: A pain in the neck for NLP". lingo.stanford.edu/pubs/WP-2001-03.pdf, Stand Januar 2009.

Sailer, Manfred (2007): "Corpus linguistic approaches with German corpora". In: Burger, Harald et al (eds.): *Phraseologie/Phraseology. An International Handbook of Contemporary Research/Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York, de Gruyter: 1060–1071.

Sinclair, John McHardy (1987): "Collocation: A progress report". In: Steele, Ross/Threadgold, Terry (eds.): *Language Topics: Essays in Honour of Michael Halliday. II*. Amsterdam, Benjamins: 319–331.

Steyer, Kathrin (2003): "Korpus, Statistik, Kookkurrenz. Lässt sich Idiomatisches 'berechnen'?". In: Burger, Harald et al. (eds): *Flut von Texten – Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie*. Baltmannsweiler, Schneider: 33–46.

Steyer, Kathrin (2004): "Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven". In: Steyer, Kathrin (ed.): *Wortverbindungen – mehr oder weniger fest*. Berlin/New York, de Gruyter: 87–116. (= *Jahrbuch des Instituts für Deutsche Sprache* 2003).

Tomasek, Tomas (2005): "Sentenzverwendung im höfischen Roman des 12./13. Jahrhunderts. Vom Diskurs zur Konvention". In: Lutz, Eckhard Conrad et al (eds.): *Literatur und Wandmalerei 2. Freiburger Colloquium vom 29. August bis 1. September*. Tübingen, Niemeyer: 47–63.

Tomasek, Thomas (2009): *Handbuch der Sentenzen und Sprichwörter im höfischen Roman des 12. und 13. Jahrhunderts*. Band 2: *Artusromane nach 1230, Gralromane, Tristanromane*. Berlin/New York: de Gruyter.

Tomasek, Tomas/Eikelmann, Manfred (2002): "Sentenzverwendung in mittelhochdeutschen Artusromanen. Ein Zwischenbericht mit einem Beispiel aus dem späten Artusroman". In: Meier, Christel et al (eds): *Pragmatische Dimensionen mittelalterlicher Schriftkultur. Akten des Internationalen Kolloquiums 26.–29. Mai 1999*. München, Fink: 135–160.

TPMA. *Thesaurus proverbiorum medii aevi. Lexikon der Sprichwörter des germanisch-romanischen Mittelalters*. Begründet von Samuel Singer. Hg. vom Kuratorium Singer der Schweizerischen Akademie der Geistes- und Sozialwissenschaften. Berlin/New York 1996–2002.

Tschichold, Cornelia (2008): "A computational lexicography approach to phraseologisms". In: Granger, Sylviane/Meunier, Fanny (eds.): Phraseology. An interdisciplinary perspective. Amsterdam/Philadelphia, Benjamins: 361–376.

Wray, Alison (2005): *Formulaic language and the lexicon*. Cambrigde etc.: Cambridge University Press.

Wray, Alison (2008): *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.