Schlussbericht

Zuwendungsempfänger:	Förderkennzeichen:
Julius-Maximilians-Universität Würzburg	01UA0815C

Vorhabenbezeichnung:

Verbundprojekt:

Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen, Teilprojekt: Bioinformatik und Informatik

Laufzeit des Vorhabens:	Berichtszeitraum
01.10.2008 – 31.12.2012	1.10.2008 – 31.12.2012

Verbundkoordinatorin:

Prof. Dr. Andrea Rapp, Institut für Sprach- und Literaturwissenschaft, Technische Universität Darmstadt (ab 1.5.2011, vorher Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier)

Verbundpartner:

Prof. Dr. Dr.h.c. mult. Ludwig Eichinger, Institut für Deutsche Sprache, Mannheim

Prof. Dr. Jörg Schultz, Dipl.-Biol.,Theodor Boveri-Institut für Biowissenschaften/Bioinformatik,

Prof. Dr. Dietmar Seipel, Institut für Informatik,

Prof. Dr. Werner Wegstein, Institut für deutsche Philologie, Sprachwissenschaft/Computer-philologie (Teilprojektverwaltung), alle Julius-Maximilians-Universität Würzburg

Teil I. Kurzdarstellungen

1. Aufgabenstellung

Das Verbundvorhaben wurde im Rahmen des BMBF-Förderschwerpunkts 'Wechselwirkungen zwischen Natur- und Geisteswissenschaften' gefördert. 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 festgehalten zu werden. (Zum weiteren Verlauf vgl. Dörries 2002, bes. Roberts 2002 und Suhr 2002).

Im Jahr 1863 veröffentlicht August Schleicher auf 30 Seiten einen Brief an seinen Kollegen Häckel unter dem Titel 'Die Darwinsche Theorie und die Sprachwissenschaft' und dem Untertitel 'Offenes Sendschreiben an Herrn Dr. Ernst Häckel, a. o. Professor der Zoologie und Director des zoologischen Museums an der Universität Jena', den er einleitend damit erläutert, dass Häckel ihm "nicht eher Ruhe gelassen, als bis ich 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" (S. 3). Dafür dankt er ganz ausdrücklich und bringt Darwins Darlegungen und Ansichten in Verbindung mit der Sprachwissenschaft und seiner 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 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." (S. 4)

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."

1869 wird Schleichers Schrift als 'Darwinism tested by the Science of Language' von Dr. Bikkers ins Englische übersetzt und im Januar 1870 im ersten Band der Zeitschrift 'Nature' von Max Müller rezensiert. 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 (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.).

1871 greift Darwin im ersten Band von 'The Descent of Man and Selection in Relation to Sex' Schleichers Schrift und Müllers Rezension wieder auf. Im zweiten Kapitel 'Comparison of the Mental Power of Man and the Lower Animals', Abschnitt 'Language' zitiert er 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.* 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.
>
> * See the very interesting parallelism between the development of species and languages, given by Sir C. Lyell in The Geological Evidences of the Antiquity of Man, 1863, chap. xxiii."

Diese frühe, inzwischen über 140 Jahre alte Wechselwirkung zwischen Natur- und Geisteswissenschaften kann als Ausgangspunkt unseres Projektes angesehen werden. Wenn es strukturelle Parallelitäten zwischen Biologie und Sprache gibt, können wir Methoden des einen Feldes nutzen, um Effekte des anderen Feldes zu erklären. Die erste Herausforderung eines solchen Projektes ist die Zusammenarbeit zwischen Natur- und Geisteswissenschaften, da nicht nur das Vokabular sondern auch die wissenschaftliche Vorgehensweise sehr unterschiedlich sind (Blythe/Croft 2010). Wir haben einen vergleichsweise formalen Weg gewählt, um eine gemeinsame Kommunikationsbasis zu erreichen. Der Fokus unseres Antrages waren Wörter und ihr Wandel in Zeit und Raum.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das kollaborative Projekt ist in drei Teilvorhaben gegliedert, die jeweils unterschiedliche Materialien, Verfahren und Fachkulturen in das Projekt einbringen, und die erst im Verbund Antworten auf die Ausgangsfragestellung von "Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen" erlauben.

Teilvorhaben 1

Institut für deutsche Sprache (Mannheim): Korpusbasierte Varietäten-Lemmalisten (TV 1)

Das IDS liefert mit der Basislemmaliste (BLL) eine Wortliste auf der Basis des Deutschen Referenzkorpus 'DeReKo', erstellt "in einer Kombination von automatischen, semi-automatischen und manuellen Verfahren, je nachdem, welche Vorgehensweise zur Lösung welcher partiellen Problemstellung sinnvoll bzw. erforderlich ist." (DeReWo-Wortliste 2013, S. 4). Die Basislemmaliste stellt damit das korpusbasierte sprachliche Referenzmaterial für den allgemeinen öffentlichen Gebrauch der neuhochdeutschen Standardsprache bereit. Sie dient auch als Schnittstelle für die Auswertung der wörterbuchbasierten Lemmalisten, die im Teilvorhaben 2 aus Daten des Trierer Wörterbuchnetzes erarbeitet und mit der Basislemmaliste zu einer Metalemmaliste verknüpft werden.

Institut für Deutsche Sprache: Allgemeine Anmerkungen zur Reihe DeReWo - Korpusbasierte Wortlisten. Technical Report IDS-KL-2013-01. Mannheim, Januar 2013
[http://www.ids-mannheim.de/kl/derewo/derewo-general-remarks.pdf]
[http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html]
Aktuelle DeReWo-Wortliste mit 320.000 Einheiten vom 18.1.2013:
http://www.ids-mannheim.de/kl/derewo/derewo-v-ww-bll-320000g-2012-12-31-1.0.zip
Institut für Deutsche Sprache, Mannheim, Januar 2013.

Teilvorhaben 2

Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft (anfangs Universität Trier, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften): Wörterbuchbasierte Varietäten-Lemmalisten/Metalemmaliste (TV 2)

Die Basis für die Erhebung und Dokumentation sprachlicher Varianz liefern die am Trierer Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften digitalisierten heterogenen Wörterbücher. Die Wörterbuchdaten sind über das Trierer Wörterbuchnetz zugänglich und wurden anfangs an der Universität Trier, später an der TU Darmstadt als wörterbuchbasierte Varietäten-Lemmalisten für die Metalemmaliste aufbereitet. Zur Aufbereitung und Auswertung tragen die in Teilvorhaben 3 entwickelten sprachwissenschaftlichen Kodierungskonzepte, die Werkzeuge und Verfahren aus der Informatik und die Analyse und Aufbereitung der Sprachdaten durch die Bioinformatik bei.

Teilvorhaben 3

Julius-Maximilians-Universität Würzburg, differenziert in die Arbeitsgruppe 'Bioinformatik' (TV 3.1) und die Arbeitsgruppe 'Informatik' und 'EDV-Philologie' (TV 3.2).

Arbeitsgruppe 'Bioinformatik' (TV 3.1)

Der interdisziplinäre Vergleich von Sprachdaten mit bioinformatischem Material kann aus methodologischer Sicht nur nach sorgfältiger Identifikation analoger Strukturen erfolgen. Hierzu hat die Arbeitsgruppe TV 3.1 'Bioinformatik' federführend mittels des Programms 'Protégé' eine Ontologie der für das Projekt relevanten sprachlichen Strukturen und Prozesse definiert, als Basis für den Vergleich mit existierenden biologischen Ontologien. Auf diese Weise sollen parallele Strukturen zwischen Genomen und Sprachen identifiziert und analysiert werden. Deren Resultate können ebenfalls wieder in 'Protégé' gespeichert werden.

Arbeitsgruppe 'Informatik' und 'EDV-Philologie' (TV 3.2)

Den sprachwissenschaftlichen Ausgangspunkt innerhalb des Projektes bilden die korpusbasierte Basislemmaliste (TV 1) und wörterbuchbasierte Lemmalisten aus heterogenen Wörterbüchern, die im Trierer Wörterbuchnetz in digitalisierter Form vorliegen (TV 2). Für die vergleichende Analyse ist es erforderlich, die Wörterbucheinträge morphologisch zu segmentieren und zu annotieren. Um dies in dem erforderlichen Umfang bewältigen zu können, hat die Informatik, basierend auf der deklarativen Programmiersprache Prolog ein Annotationstool entwickelt (Seipel/Wegstein 2011, Seipel/Borek 2012), das in Interaktion mit seinen Anwendern eine semi-automatisch annotierte Segmentierung liefert. Darüber hinaus erzielte die Informatik im 'Wechselwirkungsprojekt' weitere neue Forschungsergebnisse mit der Entwicklung eines Werkzeugs für die Alignierung von Worteinträgen aus Wörterbüchern zweier verschiedener Sprach-

stufen, z.B. dem Mittelhochdeutschen Wörterbuch von Matthias Lexer und dem (neuhochdeutschen) Wörterbuch von Jacob und Wilhelm Grimm. Ausgangspunkt für das Alignmenttool ist jeweils eine repräsentative Anzahl von in Vorarbeit alignierten Wortpaaren, aus denen Substitutionsmatrizen gewonnen werden, mit denen man dann auf die allgemeine Ähnlichkeit von Buchstaben- bzw. Lautgruppen der zu alignierenden Wörter schließen kann, d.h. das entwickelte Alignmenttool ist allgemein einsetzbar.

3. Planung und Ablauf des Vorhabens

3.1. Personalplanung

Das Vorhaben wurde unter schwierigen Voraussetzungen bei der Personalplanung durchgeführt. Für Informatik-Absolventen war und ist das Angebot auf dem Stellenmarkt derzeit ausgezeichnet. Deswegen war es sehr schwierig für die zur Verfügung stehende halbe Stelle geeignete Kandidaten zu finden, noch dazu möglichst mit sprachwissenschaftlichen Vorkenntnissen. Die Stelle konnte daher nicht kontinuierlich über die gesamte Projektlaufzeit besetzt werden. Teilweise ließ sich das kostenneutral durch eine Anpassung der Arbeitsstundenverteilung kompensieren. Dies gilt auch für die halbe Mitarbeiterstelle im Bereich Bioinformatik, die erst zum 1.9.2009 besetzt werden konnte. Der Fortschritt der Projektarbeiten in Würzburg und insgesamt wurde dadurch nicht verzögert, weil in der Bioinformatik vorrangig zunächst die für die vergleichenden Untersuchungen nötigen Sprachdatenstrukturen aufgebaut werden mussten und im Bereich Informatik der Projektleiter Prof. Seipel Entwicklungs- und Programmierarbeiten für das Annotations- und das Alignment-Tool übernommen hat.

3.2. Projektablauf

1. Einarbeitung in die projektrelevanten Strukturkonzepte von Sprache und Genomik, Einrichtung eines 'Protégé'-Servers für kollaboratives Arbeiten an einer Ontologie sowie Aufbau einer sprachwissenschaftlichen Ontologie als Basis für den Vergleich der Konzepte in Sprache und Genomen, die im Verlauf des Projekts in der sprachbezogenen Begrifflichkeit auf der Basis von Testsegmentierungen verfeinert und um Phänomene des Sprachwandels erweitert wurde.

2. Entwurf von XML-Kodierungsstrukturen (in Anlehnung an die TEI-Richtlinien) und Organisationskonzepten für die korpusbasierten DeReWo-Daten (Basislemmaliste) und die wörterbuchbasierten Lemmalisten aus heterogenen Wörterbüchern des Trierer Wörterbuchnetzes und Tests einer performanten Datenhaltung von XML-Daten.

3. Entwicklung von Verfahren, basierend auf der Programmiersprache 'Prolog', zur Ermittlung von sprachlicher Strukturinformation aus Wörterbüchern.

4. Erprobung und Bewertung von Verfahren zur Morphemzerlegung an Testlemmastrecken mit dem Programm 'glemm' (IDS Mannheim) und dem Open-Source-Programm 'Morfessor' an Daten aus den Wörterbüchern von Grimm, Campe, Adelung und dem 'Wörterbuch der deutschen Gegenwartssprache' (WDG) von Klappenbach/Steinitz.

5. Entwicklung und Erprobung von Verfahren zur Zuordnung von historischen Wortbelegen zum nhd. Basislemma (Alignment-Tool): Auf der Basis des Algorithmus von Dijkstra sowie dynamischer Programmierung (Wagner-Fisher-Algorithmus zur Analyse von Sequenzdistanzen) wurde

unter Verwendung von Substitutionsmatrizen (PAM-Matrizen von Magaret Dayhoff) ein Werkzeug entwickelt, implementiert und optimiert, das zum Aufbau der Metalemmaliste beitragen soll. Der statistische Ansatz wird in Zusammenarbeit mit der Sprachwissenschaft durch die Einbeziehung von Lautwandelgesetzen erweitert. Eine weitere Ergänzung durch Part-of-Speech-Markierungen wird erprobt.

6. Erarbeitung von Verfahren für netzwerkbasierte Vergleiche zwischen Wörterbüchern und erste Vergleichstests am Grimmschen 'Deutschen Wörterbuch' und dem 'Wörterbuch der deutschen Gegenwartssprache' (WDG), Untersuchungen zum Vernetzungsgrad und zu den Netzwerkeigenschaften (Grad, Clusterkoeffizient, Assortativität, Skaleninvarianz, Hierarchie) an quantitativen Testmengen von Sprachdaten aus dem Grimmschen Wörterbuch, dem Wörterbuch von Adelung und dem 'Wörterbuch der deutschen Gegenwartssprache' (WDG).

7. Tests mit quantitativen Vergleichen von Morphemen (Sprache) und Domänen (Genomik); Netzwerkvergleiche zwischen biologischen und linguistischen Daten (Grimmsches Wörterbuch – humane Proteindomänen) und zwischen Daten aus linguistischen Teilstrukturen (z.B. Basis- vs. Nicht-Basismorpheme, Morpheme mit höchstem Gradkoeffizienten im Grimmschen Wörterbuch und dem 'Wörterbuch der deutschen Gegenwartssprache' (WDG).

8. Erstellen eines Werkzeugs zur Morphemannotation von Wörterbucheinträgen aus synchronen und diachronen Wörterbüchern des Deutschen: Konzeption, Programmierung und Erprobung eines Informationssystems (mit graphischer Benutzerschnittstelle) auf der Basis von 'Prolog' in intensiver Zusammenarbeit zwischen Informatik und Sprachwissenschaft (Modellierung sprachwissenschaftlicher Regeln mit Mitteln der Logikprogrammierung und Erprobung in der Annotationspraxis) zum Aufbau einer relationalen Morphemdatenbank mit annotierten Morphemzerlegungen, die in die sprachwissenschaftliche Ontologie überführt werden können.

9. Quantitativer Vergleich ausgewählter Objekte und Prozesse unter Einsatz der getesteten und evaluierten Werkzeuge ('Morfessor') und Verfahren an synchronen, diachronen und regionalen Wörterbüchern des Deutschen, an Materialien aus verschiedene Zeitstufen der englischen Sprache sowie ein einem irischen Wörterbuch.

10. Evaluierung und Interpretation der quantitativen Unterschiede zwischen den irischen, englischen und den älteren deutschen Daten gegenüber den jüngeren deutschen Datenbefunden.

11. Vergleich von Analysen auf der Basis von Wörtern mit Analysen auf der Basis von Morphemen im Hinblick auf Signifikanz und Trennschärfe der angewandten Verfahren.

3.3 Beteiligung an Tagungen, Konferenzen und Workshops

2009

- Internationale Jahrestagung des Instituts für deutsche Sprache in Mannheim, 10. - 12. 3. 2009 (Wegstein, Posterpräsentation).
- Vielfalt in Sprachkorpora - Diversity in Language Corpora, Universität Basel, 22. - 23.4. 2009 (Wegstein).
- The 6th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2009), Mailand, 6.-7. 5. 2009 (Schneiker, Vortrag).

- Sprachressourcengipfel des BMBF-Projekts "D-Spin", Institut für deutsche Sprache, Mannheim, 14.5.2009 (Seipel, Wegstein).
- Third Linguistic Annotation Workshop (LAW III), 2009, held in conjunction with the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of NLP (ACL-IJCNLP 2009), Singapur, 2.-7.8.2009 (Schneiker, Vortrag).
- Wissenschaftssprache in Geschichte und Gegenwart, Universität Bamberg, 15.10.2009 (Wegstein).

2010

- Internationale Jahrestagung des Instituts für deutsche Sprache in Mannheim 'Sprachliches Wissen zwischen Lexikon und Grammatik', 9. - 11.03.2010 (Schultz und Wegstein, Posterpräsentation).
- Statusseminar des DLR für den Bereich der "Wechselwirkungen zwischen Natur- und Geisteswissenschaften", Bad Honnef, 15.-16.3.2010 (Schneiker, Schultz, Seipel - Präsentation).
- 24th Workshop on (Constraint) Logic Programming (WLP 2010), German University in Kairo, Ägypten, 14. - 16.9.2010 (Schneiker, Vortrag).

EU-Projekt COST Action A32: In our end are fresh beginnings. Perspectives for open scholarly communities in the web, Universität München, 29.9. - 2.10.2010, Vortrag: Organizing a networked access to German dictionary information based on corpus data, [http://www.uni-muenchen.de/aktuelles/pdf/cost.pdf](Wegstein und Borek).

- Tagung "Digitale Briefeditionen", Staatsbibliothek Berlin, 21.10.-22.10.2010. (Wegstein)

2011

- ISGC 2011 – 'International Symposium on Grids and Clouds' an der Academia Sinica in Taipei, Taiwan, 19.-25.3.2011 (Seipel, Vortrag)
- Jahrestagung der Gesellschaft für Wissenschaftsforschung an der Humboldt-Universität Berlin 1.4.2011, Vortrag: "Digitale Linguistik im Spannungsfeld von Wechselwirkungen zwischen Natur- und Geisteswissenschaften: ein Arbeitsbericht über 'Varianz'" (Wegstein).
- 'Wortbildung im elektronischen Wörterbuch', Institut für Deutsche Sprache (IDS), Mannheim. 26.-27.5.11 (Wegstein).
- Göttingen Centre for Digital Humanities (GCDH) und TextGrid Release 1.0, Digital Humanities Festakt, Universität Göttingen, 12.-13.7.2011 (Wegstein).
- Pre-Conference Workshop 'Analysing Electronic Dictionaries with TEI' zur korpusbasierten Analyse morphologischer Prozesse. Planung und Leitung: Dietmar Seipel, 11.10.2011. Teil der 2011 Annual Conference and Members' Meeting of the TEI Consortium, Universität Würzburg, 10.-16.10.2011. [http://www.tei-c.org/Vault/MembersMeetings/2011/tei_mm_2011/program/workshops/ index.html]
- Pre-Conference Workshop 'Tightening the representation of lexical data, a TEI perspective'. Planung und Leitung: Laurent Romary, 12.10.2011 Teil der 2011 Annual Conference and Members' Meeting of the TEI Consortium, Universität Würzburg, 10.-16.10.2011 (Präsentation: How to encode dictionaries of word families? Wegstein).

2012

- International Symposium on Grids and Clouds (ISGC 2012) an der Academia Sinica in Taipei, Taiwan, 26.2.- 2.3.2012.D. (Seipel, Vortrag)
- 11th International Conference on the Statistical Analysis of Textual Data, Liege (Belgium), 13.-15.6.2012. (Keller, Vortrag)
- Digital Humanities 2012, Universität Hamburg, 16.-22.7.2012 (Wegstein)
- 20.9.2012 Universität Göttingen: „Evolution der Informationsinfrastruktur: Forschung & Entwicklung als Kooperation von Bibliothek und Fachwissenschaft"; Beteiligung an der Podiumsdiskussion für den Bereich Geisteswissenschaften. (Wegstein)
- WLP 2012. The 26th Workshop on Logic Programming, Society of Logic Programming (GLP, Gesellschaft für Logische Programmierung e.V., Universität Bonn, 24.-25.9.2012 (Vortrag, Schneiker).
- TEI and the C(r|l)o(w|u)d. 2012 Annual Meeting of the TEI Consortium, Texas A & M University, College Station, Texas, USA, 7. -10.11.2012 (Vortrag Stadler und Wegstein).
- Varianz und Vielfalt interdisziplinär: Wörter und Strukturen. Arbeitstreffen des Projektes 'Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen', Universität Darmstadt, 29.11.-30.11.2012 (Vorträge Keller, Prätor, Seipel)

Die Beteiligung an einem Teil dieser Veranstaltungen wurden aus Eigenmitteln bzw. EU-Mitteln finanziert.

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

In den technischen Rahmenbedingungen und in der Verarbeitung der Sprachdaten befolgt das Projekt die Empfehlungen der DFG-Praxisregeln 'Digitalisierung' (2013), die durch den engen Bezug auf internationale Standards "einen Beitrag zur Nachhaltigkeit und Zukunftsfähigkeit der unterstützten Projekte" leisten wollen. Dem entspricht die geisteswissenschaftliche Aufbereitung von Daten in Übereinstimmung mit den vom TEI Consortium herausgegebenen Guidelines for Electronic Text Encoding and Interchange in der aktuellen Druckversion: TEI P5 Guidelines (2008) und der aktuellen Internetversion: Version 2.3.0. Last updated on 17th January 2013. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html> Darin wird ein hardware- und softwareunabhängiges Kodierungskonzept beschrieben, das auf dem Unicode-Standard als Zeichenformat und dem W3C-Standard XML (und damit verwandten Standards) als Auszeichnungssprache aufbaut.

4.1 Angewandte Normen:

ISO 12620:2009 Terminology and other language and content resources - Specification of data categories and management of a Data Category Registry for language resources.

ISO 24613:2008 Language resource management - Lexical Markup Framework (LMF).

4. 2 Fachliteratur sowie der benutzten Informations- und Dokumentationsdienste

Adams, Valerie: An Introduction to Modern English Word-Formation. London 1976.

Blythe, R.A./Croft, W.: Can a Science--Humanities Collaboration Be Successful? In: Adaptive Behavior February 2010, Vol. 18, No. 1, 12–20. Available: http://adb.sagepub.com/cgi/doi/10.1177/1059712309350969. [Accessed 8 October 2012].

Bühler, Karl: Sprachttheorie. Die Darstellungsfunktion der Sprache. Neudruck der Ausgabe von 1934. Stuttgart 1982. (Englische Übersetzung: Theory of Language. The representational function of language. Amsterdam/Philadelphia 2011.)

Darwin: Charles: The Descent of Man and Selection in Relation to Sex, in Two Volumes. Vol. I New York 1871.

Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Hrsg. v. Institut für deutsche Sprache. (Teil 1 – 5: Verb, Substantiv, Adjektiv, Substantivkomposita, Adjektivkomposita und Partizipialbildungen) 5 Bde. Düsseldorf und Berlin/New York 1973 – 1992.

DFG-Praxisregeln 'Digitalisierung' 2013. (Aktualisierte Überarbeitung der DFG-Praxisregeln "Digitalisierung" zu den Förderprogrammen der Wissenschaftlichen Literaturversorgungs- und Informationssysteme, April 2009) http://www.dfg.de/formulare/12_151/12_151_de.pdf

Dörries, Matthias (Ed.): Experimenting in Tongues. Studies in Science and Language. Stanford University Press. Stanford 2002.

Fleischer, Wolfgang/Barz, Irmhild: Wortbildung der deutschen Gegenwartssprache. 4. Aufl. Berlin/New York 2012.

Gallmann, Peter: Graphische Elemente der geschriebenen Sprache. Grundlagen für eine Reform der Orthographie. (Reihe Germanistische Linguistik Bd. 60). Tübingen 1985.

Henne, Helmut (Hrsg.): Deutsche Wörterbücher des 17. und 18. Jahrhunderts. Einführung und Bibliographie. Hildesheim/Zürich/New York 2001.

Johnson, Samuel: A Dictionary of the English Language in which the Words are deduced from their Originals, and Illustrated in their Different Significations by Examples from the best Writers. To which are prefixed, a History of the Language, and an English Grammar. In Two Volumes. London 1755.

Klappenbach, Ruth/ Steinitz, Wolfgang: Wörterbuch der Deutschen Gegenwartssprache. 6 Bde., Berlin 1961-1977.

Luna, Paul: The typographic design of Johnson's Dictionary. In: Anniversary Essays on Johnson's Dictionary, edited by Jack Lynch and Anne McDermott, 175–197. Cambridge 2005.

Müller, Max: The Science of Language. Darwinism tested by the Science of Language. Translated from the German of Professor August Schleicher, by Dr. Alex.V.W.Bickkers. (London: J.C.Hotten, 1869.) In: Nature Vol 1, Jan. 6, 1870, 256–259.

Osselton, Noel E: Hyphenated compounds in Johnson's Dictionary. In: Anniversary Essays on Johnson's Dictionary, edited by Jack Lynch and Anne McDermott, 160–174. Cambridge 2005.

Paul, Hermann: Prinzipien der Sprachgeschichte (1. Auflage 1880), 9.Auflage, Tübingen 1975.

Reddick, Allen: The Making of Johnson's Dictionary 1746 – 1773. Revised Edition. Cambridge 2006.

Richards, Robert J.: The Linguistic Creation of Man: Charles Darwin, August Schleicher, Ernst Haeckel, and the Missing Link in Nineteenth-Century Evolutionary Theory. In: Dörries, Matthias: Experimenting in Tongues (2002), S. 21 - 48

Schleicher, August: Die Deutsche Sprache. Stuttgart 1860.

Schleicher, August: 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 1863.

Splett, Jochen Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 18 Bde., Berlin/New York 2009.

Suhr, Stephanie: Is the Notion of Language Transferable to the Genes? In: Dörries, Matthias: Experimenting in Tongues (2002), S. 49 – 66.

4.3 Informations- und Dokumentationsdienste

DEREKO – Deutsches Referenzkorpus:
http://www.ids-mannheim.de/kl/projekte/korpora/

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

Das Digitale Wörterbuch der Deutschen Sprache (DWDS). Hg. von der Berlin-Brandenburgischen Akademie der Wissenschaften (für die elektronische Version des WDG)
http://www.dwds.de/panel/wdg_panel/

elexiko: http://www.owid.de/wb/elexiko/start.html

General Ontology for Linguistic Description (GOLD): http://www.isocat.org/

Morfessor 1.0 Software (Morfessor Baseline algorithm)
http://www.cis.hut.fi/projects/morpho/

Oxford English Dictionary Online
[http://public.oed.com/the-oed-today/guide-to-the-third-edition-of-the-oed/]

The Protégé Ontology Editor: http://protege.stanford.edu

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

5. Zusammenarbeit mit anderen Stellen.

Über die Tagung 'Wortbildung im elektronischen Wörterbuch' am Institut für Deutsche Sprache in Mannheim im Mai 2011 kam ein intensiver Kontakt zu Prof. Dr. Jochen Splett (Universität Münster) zustande, über sein Wortfamilienmaterial für das Althochdeutsche und das Neuhochdeutsche und Möglichkeiten einer Nachnutzung (Konvertierung der 18-bändigen Druckfassung). Mit dem Rechteinhaber, dem Verlag deGruyter in Berlin, wurde Kontakt aufgenommen. Für den projektinternen Gebrauch hat der Verlag einer Nutzung der Liste der Wortfamilienkerne zugestimmt. Die Beschäftigung mit den Daten führte zu ersten Überlegungen, wie Wortfamilien(wörterbücher) zu kodieren sind und wurde im Workshop 'Tightening the representation of lexical data, a TEI perspective' auf der 2011 Annual Conference and Members' Meeting of the TEI Consortium an der Universität Würzburg vorgestellt. Eine differenzierte Erweiterung des Kodierungskonzepts für Wortbildungen auf der Basis der TEI Guidelines wurde von Heike Stadler und Werner Wegstein auf der Jahrestagung der TEI 2012 in Texas präsentiert und soll im 'Journal of the Text Encoding Initiative' veröffentlicht werden.

Teil II. Eingehende Darstellung

1. Erzielte Ergebnisse

1.1 Teilvorhaben 3.1 Bioinformatik

1. Identifikation analoger Strukturen zwischen Sprache und Genomen

Der Fokus unseres Antrages waren Wörter und ihr Wandel in Zeit und Raum. Daher haben wir gemeinsam eine Ontologie erstellt, in der die relevanten Strukturen und Prozesse eindeutig definiert sind (Abbildung 1). Diese Ontologie bildet nicht nur die Basis unserer Arbeit. Da sie in Protégé, einem Open Source Ontologie Editor erstellt wurde, kann sie nachhaltig auch von weiteren Wissenschaftlern genutzt werden. Dazu haben wir die Ontologie IsoCat angepasst und stellen sie über http://www.sprache-und-genome.de zum download zur Verfügung.

Anhand der Ontologie konnten wir parallele Strukturen zwischen Genomen und Sprachen identifizieren. Im Antrag gingen wir davon aus, dass die Kombination von Wörtern in einem Satz vergleichbar der Kombination von Domänen in einem Protein ist. Anhand der entwickelten Ontologie wurde aber klar, dass dies nicht der korrekte Level ist. Sowohl Wörter als auch Proteine werden aus kleineren bedeutungstragenden / funktionstragenden Bestandteilen aufgebaut. Im Fall der Proteine sind dies Domänen, bei Wörtern Morpheme (Abbildung 2). Geht man von diesem Kern aus, kann die Analogie sinnvoll erweitert werden (Tabelle 1).

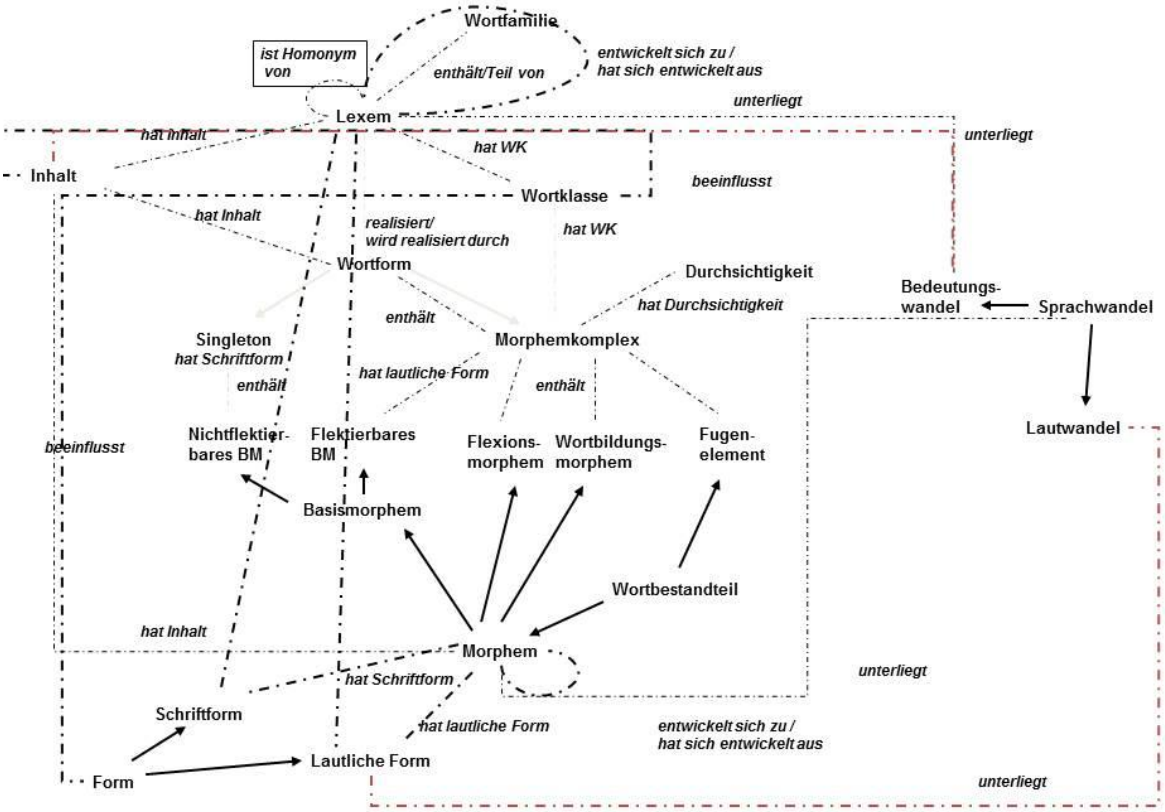

Abbildung1 : Struktur der Ontologie zur Wortbildung

Tabelle 1: Analoge Strukturen und Mechanismen zwischen Sprache und Genomen

Sprache	Genome
Buchstaben	Aminosäuren
Silben	Strukturbausteine
Morpheme	Domänen
Wörter	Proteine
Wortfamilien	Proteinfamilien
Wörter einer Sprache	Proteom
Lexem	Gen
Wörterbuch	Genom
Komposition	Domänenfusion

2. Konnektivität, nicht Gebrauch, bestimmt das Schicksal eines Morphems

Nachdem diese analogen Strukturen identifiziert wurden, konnten wir Methoden, die in der Bioinformatik zur Analyse bestimmter Entitäten entwickelt wurden, auf die entsprechenden Einheiten der Sprache anwenden. Wir haben uns besonders auf die Analogie Morphem – Domäne fokussiert, da dies bisher so noch nicht bearbeitet wurde. Dafür haben wir einen bioinformatischen Ansatz, der zur Analyse der Zusammensetzung von Domänen entwickelt wurde [4], angewendet und erweitert. Kern des Ansatzes ist es, die Morpheme und ihr Vorkommen in Wörtern in eine Netzwerkstruktur zu übersetzen. Netzwerke werden bereits

vielfach zur Repräsentation von linguistischen und kognitiven Strukturen verwendet [5], bisher aber nicht zur Analyse der morphologischen Struktur von Wörtern verwendet.

Klassischerweise wird Sprachwandel auf der Ebene von Wörtern untersucht. Aber es passiert nur relativ selten, dass einer bisher bedeutungslosen Zeichenfolge eine Bedeutung zugeordnet wird. Ein Beispiel aus dem Englischen ist das Wort ‚zilch', dass ‚nichts' bedeutet [6]. Häufiger wird ein neues Wort gebildet, indem schon in der Sprache vorhandene Wörter zusammengefügt werden. So entstand auch das Wort des Jahres 2010, ‚Wutbürger' [7]. Allein dieses Beispiel verdeutlicht, dass man Sprachwandel auch auf der Ebene der Morpheme verstehen muss. Aber Morpheme sind auch bei der mentalen Verarbeitung von Sprache relevant. Es wird kontrovers diskutiert, ob die morphologische Struktur von Wörtern auch im mentalen Lexikon (dem menschlichen Wortspeicher) reflektiert wird. Während die meisten Hypothesen heute davon ausgehen, dass dies zumindest teilweise der Fall ist [8,9], gibt es auch Modelle, in denen Morpheme nicht als diskrete Einheiten im mentalen Lexikon auftauchen [10,11]. Protagonisten dieser zweiten Variante würden ihre Hypothese verwerfen, wenn es gelänge Eigenschaften von Morphemen zu identifizieren, die sich nicht durch die Statistik der Wörter erklären lassen [12]. Mit der aus der Bioinformatik ausgeliehenen Methode konnten wir nach genau solchen Eigenschaften suchen.

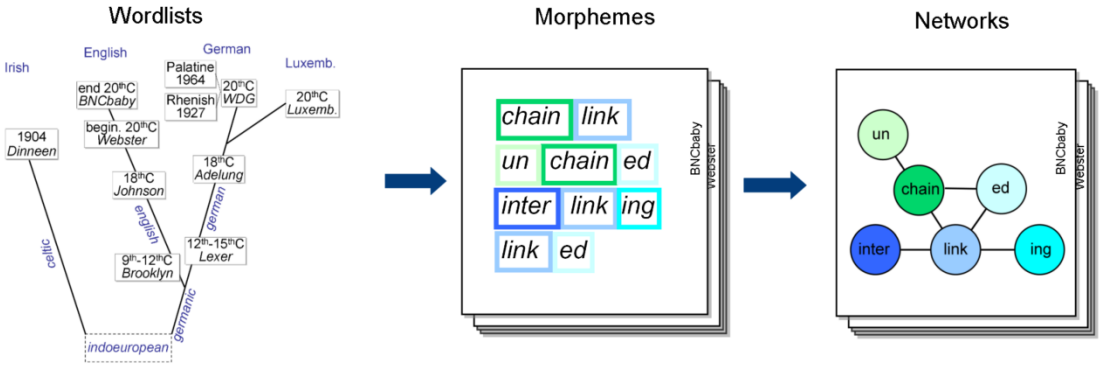

Abbildung 2: Verwendete Wortlisten, Zerlegung der Wörter in Morpheme und Darstellung der Morphemlisten als Netzwerk

Dafür haben wir Wortlisten aus dem Englischen und dem Deutschen analysiert, die mehr als 200 Jahre Sprachwandel abdecken. Diese lemmatisierten Corpora wurden automatisch in Morpheme zerlegt [13]. Anschließend wurde jeder Datensatz in ein Netzwerk transformiert, in dem Morpheme Knoten sind, zwischen denen eine Kante gezogen wurde, wenn diese Morpheme in einem Wort benachbart vorkommen (Abbildung 2). Diese Netzwerke hatten unabhängig von Sprache und Zeitstufe die gleichen globalen Eigenschaften, vermutlich begründet in der relativ engen Verwandtschaft der analysierten Sprachen. Ein wesentlicher Punkt ist, dass sie skalenfrei sind, das heißt, es existieren viele Morpheme, die nur mit wenigen anderen kombiniert sind und weniger, die mit sehr vielen anderen zusammen vorkommen (Hubs). Den Erwartungen entsprechend handelt es sich bei den stärksten Hubs um Affixe. In den Hubs, bei denen es sich um Basismorpheme handelt, konnten wir eine Reflexion kultureller Unterschiede identifizieren (Tabelle 2). Im nächsten Schritt haben wir die Netzwerke verschiedener Zeitstufen direkt aufeinander abgebildet. Dies ermöglichte uns, nicht nur das Entstehen und Sterben von

Morphemen zu verfolgen, sondern auch den zeitlichen Verlauf ihrer Kombination. Auch hier konnten wir kulturelle Einflüsse erkennen. So sind zum Beispiel die Morpheme ‚Auto', ‚Industrie' und ‚Film' im 18. Jahrhundert nicht vorhanden aber hochvernetzt im 20. Jahrhundert. Bisherige Versuche, kulturelle Entwicklungen aus Sprachwandel abzuleiten, basierten auf Häufigkeiten von Wörtern. Dies erschwert aber die Identifikation von Trends, da es deutlich mehr Wörter als Morpheme gibt und zur Identifizierung der Häufigkeiten auf Corpora gearbeitet werden muss. Durch die Fixierung auf Morpheme konnten wir beide Probleme umgehen und haben daher einen neuen Ansatz für ‚Culturomics'-Studien [14] gefunden.

Auf diesen Ergebnissen aufbauend haben wir uns die Frage gestellt, welche Faktoren hinter dem Wandel von Morphemen stehen könnten. Im Fall von Wörtern ist dies vor allem die Gebrauchshäufigkeit [15,16]. In unserer netzwerkbasierten Analyse konnten wir zeigen, dass für Morpheme die Konnektivität, d.h. die Anzahl unterschiedlicher benachbarter Morpheme, sowohl für das Überleben, als auch für die Anzahl neu entstehender Wörter wichtig ist. Nun sind aber Morpheme selbst auch Bestandteile von Wörtern und sollten daher von deren Gebrauchshäufigkeit beeinflusst werden. Anhand komplexer statistischer Verfahren (logistischer Regression und partieller Korrelation) konnten wir zeigen, dass der Einfluss der Gebrauchshäufigkeit deutlich unter dem der Konnektivität liegt. Interessanterweise kann man die Konnektivität auf ein bestehendes sprachwissenschaftliches Konzept abbilden. Hier wird die Anzahl der Wörter, in denen ein Morphem vorkommt, als dessen Typfrequenz bezeichnet. Im Gegensatz dazu versteht man unter Tokenfrequenz die Häufigkeit eines Morphems / Wortes in einem Korpus. Wir konnten also zeigen, dass für Morpheme die Typfrequenz einen stärkeren Einfluss hat als die Tokenfrequenz. Im Gegensatz dazu wird die Veränderung von Wörtern von ihrer Tokenfrequenz bestimmt. Wir haben damit also eine Eigenschaft von Morphemen gefunden, die nicht von den sie enthaltenden Wörtern gegeben ist. Die Identifikation einer solchen Eigenschaft wurde gefordert, um die Hypothese, dass Morpheme nicht im mentalen Lexikon vorkommen, zu verwerfen. Somit konnten wir durch die Analyse der diachronen Varianz von Sprache Eigenschaften des mentalen Lexikons deduzieren.

Interessanterweise sind unsere Ergebnisse konsistent mit Experimenten, die in der Psycholinguistik durchgeführt wurden. Hier hängt die Erkennungsgenauigkeit und Geschwindigkeit von Wörtern von ihrer Tokenfrequenz ab. Im Gegensatz dazu ist für die Erkennungsgeschwindigkeit von Morphemen ihre Familiengröße, die vergleichbar mit der Typfrequenz und damit der Konnektivität ist, bestimmend [17]. Somit haben wir eine erste Parallelität zwischen der individuellen Sprachverarbeitung (Psycholinguistik) und historischem Sprachwandel gefunden.

Tabelle 2: Ausgewählte Morpheme und ihr Rang nach der Konnektivität. Affixe haben sowohl in Englischen als auch im Deutschen die höchste Konnektivität. In beiden Sprachen und in den jeweils untersuchten Zeitstufen finden sich als Beispiele für hoch vernetzte Morpheme ‚Wasser', ‚Holz' und ‚Haus'. Die Konnektivität von Morphemen, die der Natur zugerechnet werden können, nimmt in beiden Sprachen über die Zeit ab, Im Gegensatz dazu finden steigt die Konnektivität von Morphemen, die Arbeit und Freizeit zugeordnet werden.

	German 18th century		German 20th century		English 18th century		English end 20th century	
	Morpheme	Rank	Morpheme	Rank	Morpheme	Rank	Morpheme	Rank
Affixes	En	1	en	2	Un	3	Un	10
	ver	4	ver	5	Ly	5	Ly	8
Common base morphemes	wasser (water)	15	wasser (water)	30	Water	199.5	water	148
	holz (wood)	18	holz (wood)	44	Wood	156.5	wood	102.5
	haus (house)	26	haus (house)	29	House	108.5	house	146
Terms from nature	baum (tree)	20	baum (tree)	82.5	Sea	88.5	sea	160
	kraut (herb)	23	kraut (herb)	549	Wort	95	wort	4987.5
Terms from work and leisure time	zeit (time)	89.5	zeit (time)	31	Time	298.5	time	152
	dienst (service)	159	dienst (service)	46	Life	1073	life	268
	spiel (game, play)	138.5	spiel (game, play)	50	Free	335	free	169

3. Wortbildung berücksichtigt die Größe der Morphemfamilie

Diese Ergebnisse zeigen, dass es möglich sein könnte, Reflexionen des mentalen Lexikons im Sprachwandel zu identifizieren. Dies haben wir in einem anschließenden Arbeitsschritt weiter verfolgt. Auch hier haben wir etablierte Verfahren aus der Biologie und im speziellen der Bioinformatik angewendet. Verfolgt man die Historie eines einzelnen Morphems können folgende Ereignisse eintreten. In einem Innovationsprozess kann ein neues Morphem in eine Sprache integriert werden. Wenn es produktiv ist, entstehen neue Wörter, die dieses Morphem enthalten. Genauso können Wörter dieses Morphems sterben. Dieses Sterben kann so weit gehen, dass kein Wort mit diesem Morphem mehr existiert; das Morphem selbst ist gestorben. Genau solche Prozesse werden in der vielen Bereichen der Biologie als ‚Birth-Death-Innovation'-Modelle angewendet [18]. Interessanterweise wurden sie auch schon zur Beschreibung der Evolution von Domänen in Genomen genutzt [19]. Analog haben wir ein Modell für die Evolution von Morphemen in einer Sprache entworfen. Anhand der Daten, die wir in dem

vorangegangenen Arbeitsschritt generiert hatten, konnten wir diese Modelle testen. Wie im Fall der Domänen reicht auch bei Morphemen das einfachste, neutrale Modell nicht aus, um die Daten gut zu fitten. Durch eine kleine Erweiterung konnte dies aber erreicht werden. Das heißt, dass die Verteilung von Morphemen über verschiedene Sprachen und Zeitstufen hinweg durch ein stochastisches Modell beschrieben werden kann. Die benötigten Erweiterungen verändern das neutrale Modell so, dass mit sehr häufigen Morphemen seltener neue Wörter gebildet werden.

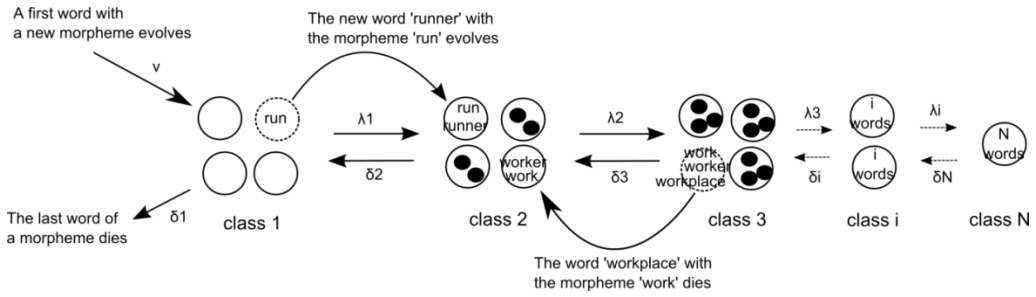

Abbildung 3: Birth-Death-Innovation-Modell für Morpheme – Kreise: Wortfamilien; Class: Anzahl der Wörter in der zugehörigen Wortfamilie (N ist die maximale Größe der Wortfamilie); δi: Birthrate für eine Wortfamilie der Größe i; λi: Death-rate für eine Wortfamilie der Größe i; v: Innovationsrate, mit der neue Morpheme entstehen.

Geht man von einem äußerungsbasierten Selektionsmodell des Sprachwandels aus [20], kann dies durch zwei möglicherweise kombinierte Effekte erklärt werden. Zum einen könnte es ein Effekt der Wortbildung sein, bei dem der Erfinder eines Wortes Morpheme vermeidet, die schon in zu vielen anderen Wörtern vorkommen, eventuell um eine Überladung der Bedeutung zu vermeiden. Zum anderen könnte es auf der Selektion von Wörtern liegen, bei der der Sprecher versucht neue Wörter, die zu häufig vorkommende Morpheme enthalten, zu vermeiden. In beiden Fällen muss aber ein Bewusstsein für die Typfrequenz eines Morphems im mentalen Lexikon vorhanden sein. Wie schon im vorhergehenden Arbeitsschritt wurden damit auch hier Reflexionen des mentalen Lexikons in historischem Sprachwandel identifiziert.

4. Morpheme und Domänen – Analoge Strukturen, unterschiedliche Mechanismen

Wir haben gezeigt, dass für Morpheme die Konnektivität wichtiger als der Gebrauch ist. Ausgangspunkt der Analysen war dabei eine von uns identifizierte Analogie zwischen Morphemen und Domänen. Daher haben wir in einem weiteren Arbeitsschritt untersucht, ob diese Befunde aus der Analyse von Sprachdaten auf genomische Daten übertragen werden können. Dazu haben wir die Gebrauchshäufigkeit eines Wortes mit der Expressionsrate eines Gens gleich gesetzt. Diese Analogie ist konsistent, da gezeigt wurde, dass so wie die Gebrauchshäufigkeit eines Wortes seine Veränderung bestimmt, die Evolutionsrate eines Gens durch die Expressionsrate vorhergesagt werden kann [21]. Im Fall der Gene ist dies durchaus unerwartet, da man eigentlich davon ausgegangen war, dass die Funktion (Krankheitsgen, Anzahl der Interaktionspartner…) die Rate bestimmt. Durch das Aufkommen verschiedener omics-Techniken ist inzwischen für Modellorganismen die Expressionsrate genomweit in unterschiedlichen Zelltypen bekannt. Analog zur Untersuchung der Morpheme konnten wir daher eine Expressionsrate für einzelne Domänen bestimmen. Der große Vorteil der Sprachdaten ist, dass explizite Daten aus der Vergangenheit zur Verfügung stehen. Dies ist für biologische Systeme nicht der

Fall. Daher mussten wir die Historie von Domänen unter Verwendung verwandter, komplett sequenzierter Spezies rekonstruieren. Dazu haben wir uns zum einen das Genus Drosophila, zum anderen das Genus Caenorhabditis ausgewählt, da hier sowohl Expressionsdaten als auch Genome von relativ eng verwandten Spezies vorliegen. Im Gegensatz zu den Sprachdaten, die einen Zeitraum von 200 Jahren umfassen, fand die Aufspaltung des Genus Drosophila vor ca. 40 Millionen Jahren statt. Trotz dieses vergleichsweise großen Zeitraums ist die Anzahl der neu entstandenen Domänen gering. Darüber hinaus kommen diese meist nur einzeln vor, d.h. nicht mit anderen Domänen zusammen in einem Protein. Im Fall des Genus Drosophila wären daher nur 1.7% der identifizierten Domänen für eine Analyse verwendbar gewesen. Belegbare statistische Aussagen können auf dieser Datenmenge nicht gemacht werden. Das wesentliche Ergebnis ist aber, dass der Mechanismus wie neue Gene entstehen, anders ist als bei der Entstehung neuer Wörter. Obwohl die Evolution neuer Domänenarchitekturen regelmäßig vorkommt, ist sie nicht der bestimmende Mechanismus. Hier ist die Genduplikation vermutlich deutlich stärker ausgeprägt.

Entgegen der ursprünglichen Hypothese konnten wir in diesem Arbeitsschritt zeigen, dass die Prozesse hinter der Evolution von Proteinen und Wörtern unterschiedlich sind. Damit wird Gesamtidee des Projektes gut zusammengefasst. Es existieren seltsame Parallelen zwischen Biologie und Sprache. Diese können ausgenutzt werden, um Analysemethoden zwischen den beiden Fächern auszutauschen. Dennoch sollte man sich nicht dazu verleiten lassen, aus der Analogie von Strukturen und vielleicht auch oberflächlichen Prozessen auf grundlegende Ähnlichkeiten in den zugrundeliegenden Prozessen zu schließen.

5. Literaturverzeichnis

4. Wuchty S (2001) Scale-free behavior in protein domain networks. Molecular biology and evolution 18: 1694–1702. Available:http://www.ncbi.nlm.nih.gov/pubmed/11504849. Accessed 13 February 2012.

5. Baronchelli A, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N, Christiansen MH (2013) Networks in Cognitive Science. Trends in cognitive sciences: 1–35. Available: http://www.ncbi.nlm.nih.gov/pubmed/23726319.

6. Harper D (2012) Online Etymological Dictionary. http://www.etymonline.com. Available:http://www.etymonline.com. Accessed 19 September 2012.

7. Gesellschaft für deutsche Sprache (2010) Pressemitteilung vom 17. Dezember 2010. http://www.gfds.de/presse/pressemitteilungen/171210-wort-des-jahres-2010/. Available:http://www.gfds.de/presse/pressemitteilungen/171210-wort-des-jahres-2010/. Accessed 17 December 2010.

8. Caramazza A, Laudanna A, Romani C (1988) Lexical access and inflectional morphology. Cognition 28: 297–332.

9. Baayen H, Schreuder R (1999) War and Peace: Morphemes and Full Forms in a Noninteractive Activation Parallel Dual-Route Model. Brain and Language 68: 27–32.

10. Rueckl JG, Mikolinski M, Raveh M, Miner CS (1997) Morphological Priming, Fragment Completion, and Connectionist Networks. Journal of Memory and Language 405: 382–405.

11. Plaut DC, Booth JR (2000) Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. Psychological review 107: 786–823. Available:http://www.ncbi.nlm.nih.gov/pubmed/11089407. Accessed 8 June 2012.

12. Seidenberg M, Gonnerman L (2000) Explaining derivational morphology as the convergence of codes. Trends in cognitive sciences 4: 353–361. Available:http://www.ncbi.nlm.nih.gov/pubmed/10962617.

13. Creutz S, Lagus K (2005) Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Publications in Computer and Information Science. Helsinki: University of Technology.

14. Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. (2011) Quantitative analysis of culture using millions of digitized books. Science (New York, NY) 331: 176–182. Available:http://www.ncbi.nlm.nih.gov/pubmed/21163965. Accessed 19 July 2011.

15. Lieberman E, Michel J-B, Jackson J, Tang T, Nowak M (2007) Quantifying the evolutionary dynamics of language. Nature 449: 713–716. Available:http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2460562&tool=pmcentrez&rendertype=abstract. Accessed 18 July 2011.

16. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449: 717–720. Available:http://www.ncbi.nlm.nih.gov/pubmed/17928860. Accessed 18 July 2011.

17. Bertram R, Baayen RH, Schreuder R (2000) Effects of Family Size for Complex Words. Journal of Memory and Language 42: 390–405. Available:http://linkinghub.elsevier.com/retrieve/pii/S0749596X99926815. Accessed 30 March 2012.

18. Novozhilov AS, Karev GP, Koonin E V. (2006) Biological applications of the theory of birth-and-death processes. Briefings in Bioinformatics 7: 70–85. Available:http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbk006. Accessed 21 July 2012.

19. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin E V (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. BMC Evolutionary Biology 26.

20. Croft W (2001) Explaining language change: an evolutionary approach. Pearson Education ES. p.

21. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Molecular biology and evolution 23: 327–337. Available:http://mbe.oxfordjournals.org/cgi/content/abstract/23/2/327. Access 11 June 2011.

1.2 Teilvorhaben 3.2 'Informatik' und 'EDV-Philologie'

Der interdisziplinäre Vergleich von Sprachdaten mit entsprechendem bioinformatischen Material kann aus methodologischer Sicht nur nach sorgfältiger Identifikation analoger Strukturen erfolgen.

Den sprachwissenschaftlichen Ausgangspunkt bildet innerhalb des Projektes eine Reihe von Lemmalisten der heterogenen Wörterbücher.

Die kleinsten bedeutungstragenden Einheiten der Sprache bilden jedoch die Morpheme, in die die Lemmalisten zu diesem Zweck zerlegt und auf diese Art angereichert werden mussten.

Um dies in dem erforderlichen Umfang bewältigen zu können, wurde basierend auf der deklarativen Programmiersprache Prolog ein Annotationstool entwickelt (Seipel/Wegstein, 2011, Seipel/Borek 2012), das in Interaktion mit seinen Anwendern eine semi-automatisch annotierte Segmentierung liefert.

1. Das Annotationstool

Die Herausforderung war, ein Tool zu entwickeln, das mit den vorhandenen sehr großen digitalen Datenmengen vernünftig umgehen kann - ein Unternehmen, das vorher einen riesigen Aufwand für Sprachwissenschaftler darstellte.

Um die große Anzahl von Wörterbucheinträgen, die leicht 100.000 Einheiten übersteigen kann, annotieren zu können, benötigt man linguistisches Wissen und geeignete Tools und Konzepte aus der Informatik: einen verlässlichen Morphemzerleger, eine geeignete kompakte Wissensrepräsentation, Inferenzmethoden für Annotationsregeln und eine graphische Benutzeroberfäche.

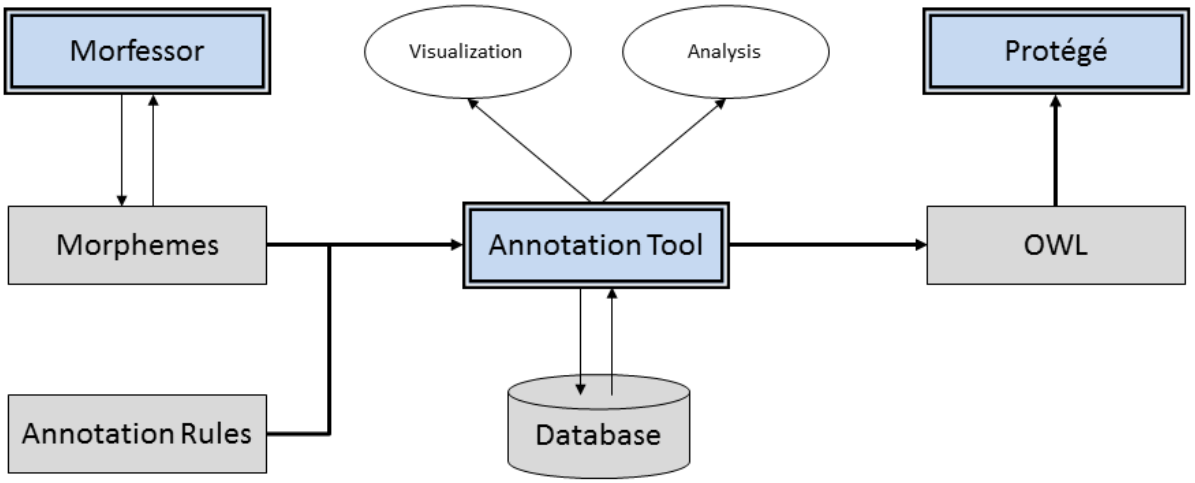

Abbildung 4: Systemarchitektur.

Die Architektur unseres Annotationstools, das wir in der deklarativen Programmiersprache Prolog implementiert haben, wird in Abbildung 4 gezeigt.

Die initialen Morphemzerlegungen werden durch etablierte Tools wie Morfessor gewonnen. Sie werden dann basierend auf den fein-granularen Informationen elektronischer Wörterbücher vorannotiert. Dazu benutzen wir z.B. Informationen zu den Wortklassen und zur Nutzung der Wörter.

Die grundlegenden morphologischen Einheiten können später mit unserem Tool korrigiert, verfeinert und interaktiv annotiert werden.

Wir versuchen mittels unserer Parsing-Techniken, möglichst viele Annotationen aus den zugrundeliegenden elektronischen Wörterbüchern zu extrahieren.

Annotierte Morphemterme

Wir benutzen eine kompakte Wissensrepräsentation für die annotierten Morphemterme in Form einer intuitiven Termnotation

(t1 + t2 + ... + tn) * a1 * a2 * ... * am,

in Prolog, die auch für Sprachwissenschaftler gut verständlich ist. Dabei sind die ti die Komponenten und die aj die Annotationen. Derartige Morphemterme kann man in Prolog auch effizient und übersichtlich verwalten.

Mittels eindeutiger Abkürzungen wie 'lm' für lexikalisches Morphem, 'da' für derivational affix und 'le' für linking element kann der Morphemterm ((craft+s)+man)+ship als

(((craft*lm*noun+s*le+man*lm*noun)*noun*comp + ship*da*suffix)*noun*derv

annotiert werden. Dies wird in Abbildung 5 veranschaulicht.

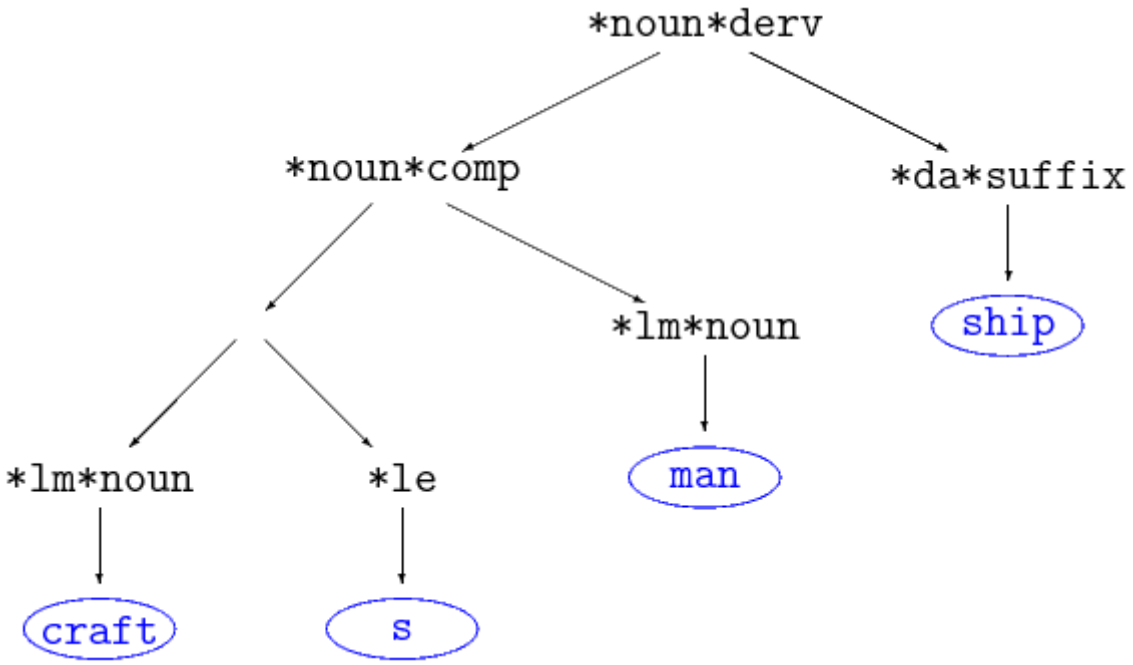

Abbildung 5: Annotierte Morphemzerlegung.

Zur persistenten Speicherung der Morphemterme und zum effizienten Zugriff auf der Basis der Textform wurde eine relationale Datenbank benutzt. Diese speichert für jeden Morphemterm weitere Information wie die Textform, den Identifikator aus dem Wörterbuch und den Namen des Benutzers, der den Term annotiert hat, zusammen mit einem Zeitstempel.

Annotationsregeln

Die Benutzung spezifischer, systematischer sprachlicher Regeln zur automatischen Annotation war die eine Säule, die Behandlung von Ausnahmen und das Ermöglichen von Benutzerinteraktion die andere. Es sollten auch dynamisch Annotationsregeln modifiziert und hinzugefügt werden können.

Wir haben in Prolog ein System von etwa 50 allgemeingültigen Wortbildungsregeln implementiert, die zusammen mit Ausschlussregeln die automatisierte Ableitung von Wortklassen und Wortbildungsprozessen ermöglichen.

Prolog ist hervorragend geeignet, Morphemstrukturen abzubilden, da Datenstrukturen in Prolog als Terme realisiert sind. Prolog ermöglicht eine sehr flexible Kodierung und Verarbeitung.

Es wird eine Menge von Annotationsregeln verwaltet, die automatisch aus den von den Benutzern bereits eingegebenen Annotationen und Segmentierungen neue Annotationen ableiten kann. Mittels der folgenden logischen Annotationsregel wird z.B. der Term ((craft+s)+man)*noun+ship als Nomen (noun) erkannt und weiter zu (((craft+s)+man)*noun+ship)*noun annotiert:

```
has_word_class(X, noun) :-
    mc(X, A, B), has_word_class(A, noun),
    text_form(B, [ship, ...]).
```

Außerdem haben wir eine Hierarchie von Regeln etabliert: die meisten werden automatisch angewendet, andere führen nur zu Vorschlägen, die der Benutzer bestätigen oder ablehnen kann. Bestätigte Regeln steigen in der Hierarchie und werden daraufhin automatisch angewendet.

Wir haben dabei einen Regelsatz erarbeitet, um bereits existierende Annotationen aus der Datenbank des Annotationstools bzw. aus anderen Datenquellen mit den neu gewonnenen Annotationen kombinieren zu können. Dabei müssen über Ausschlussverfahren geeignete Annotationen ermittelt werden und eine Auswahlfunktion für alternative Annotationen eingerichtet werden.

Der Annotationsworkflow

Zur Untersuchung geeigneter Reihenfolgen der Annotierung wurden die Gebrauchshäufigkeit und die Konnektivität der Morpheme als Kriterium herangezogen.

Es wurde auch untersucht, wieviele Wörter durch das Annotieren einer geeigneten Liste von Morphen bereits vollständig erschlossen werden können. Eine Untersuchung ergab z.B., dass durch die Annotierung der 18 häufigsten Morphe - meist Affixe - bereits etwa 10.000 Lexeme des Wörterbuchs betroffen sind. Nach der Bearbeitung der 1.000 häufigsten Morphe konnten für etwa 15.000 Lexeme des Wörterbuchs vollständige Annotationsvorschläge gemacht werden.

Zur Reduzierung der Bearbeitungszeit durch die Sprachwissenschaftler hat sich die konsequente Tastaturbedienung der Oberfläche als sehr wichtig herausgestellt.

Es wurden Konzepte zum Umgang mit Über-, Unter- und Fehlsegmentierung entwickelt. Durch Verschmelzen von unterschiedlichen Annotationen eines Wortes kann man vermeiden, den sprachwissenschaftlichen Benutzer mit einer Vielzahl von ähnlichen Vorschlägen zu überhäufen. Durch weitere Ausschlussregeln werden auch ungültige Positionen für Annotationen verhindert; so kann z.B. ein Suffix nicht am Wortanfang stehen. Außerdem kann eine ungültige Anzahl von Komponenten eines Wortes erkannt werden.

Zukünftige Entwicklungen

Sowohl das Annotationstool als interaktive Forschungsumgebung als auch das in ihm etablierte linguistische Regelwerk bieten Anwendungsmöglichkeiten, die über die ursprüngliche Zielsetzung des Projektes hinausgehen und zukünftig der Allgemeinheit zur Verfügung gestellt werden können.

Das Annotationstool wurde zusätzlich zur Integration in TextGrid als Webversion umgesetzt. Gleichzeitig wird das enthaltene Regelwerk veröffentlicht und ein Export in die Web Ontology Language (OWL) bzw. deren Regelerweiterung, die Semantic Web Rule Language (SWRL), ermöglicht.

Aufgrund der bereits aufbereiteten Daten lassen sich zudem weitere Teilmengen generieren, die eine Ergänzung des vorhandenen Wörterbuchs und des Korpusmaterials bilden. Beispiele hierfür wären das Generieren von Ausgangsmaterial für eine Liste der Basismorpheme des Deutschen sowie von Wortfamilienkernen.

2. Das Alignmenttool

Außerdem wurde an einem Alignmenttool für die Lexeme zweier Wörterbücher verschiedener Sprachstufen gearbeitet. Wir haben als Beispiel das Mittelhochdeutsche Wörterbuch von Matthias Lexer und das Deutsche Wörterbuch von Jacob und Wilhelm Grimm verwendet. Das entwickelte Alignmenttool ist aber allgemein einsetzbar.

Dabei werden aus einer repräsentativen Anzahl von in Vorarbeit alignierten Wortpaaren Substitutionsmatrizen gewonnen, mit denen man dann auf die allgemeine Ähnlichkeit von Buchstaben- bzw. Lautgruppen schließt. Dabei kommt es sehr darauf an, möglichst viele Buchstaben- und Lautgruppen zu identifizieren und als Ganzes zu betrachten.

Durch Extraktion aus den elektronischen Datenquellen konnten weitere ähnliche Wortpaare – wie unten das mittelhochdeutsche *vluoch* und das neuhochdeutsche *fluch* – gefunden werden.

Das manuelle Alignment wurde in einer fragmentarischen XML-Struktur vorgenommen. Unter einem alignment–Element gruppieren sich das mittelhochdeutsche und das neuhochdeutsche Lemma. Lautgruppierungen wurden mittels eckiger Klammern, Leerstellen mit % realisiert.

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.

<lemma>ungel[au]pli[ch]</lemma>

<lemma>ung%l[au]bli[ch]</lemma>

Aus der Testmenge wurde sodann eine Substitutionsmatrix ermittelt, deren Einträge M(a,b) angeben, wie häufig eine Buchstabengruppe a in einem mittelhochdeutschen Wort durch eine Buchstabengruppe b im entsprechenden neuhochdeutschen Wort ersetzt wird.

Ein Alignment von zwei Strings x und y ist eine Segmentierung in Buchstabengruppen x1 , . . . , xn und y1 , . . . , yn mit y = y1 y2 . . . yn und x = x1 x2 . . . xn . Die Bedeutung des Alignments ist, dass durch die Ersetzung von yi durch xi der String y zum String x wird. Manche Buchstabengruppen, z.B. yi oder xi können leer sein. Natürlich ist es auch nicht sinnvoll, Ersetzungen der Form ε → ε zu betrachten, da diese redundant wären.

Gesucht ist ein optimales Alignment; bei diesem soll das Produkt M(y1 ,x1) · M(y2,x2) · . . . · M(yn,xn) der Ähnlichkeiten der Teilstrings minimal sein. Wir gehen von einer Unabhängigkeitsannahme aus, die besagt, dass die Ersetzungen yi → xi unabhängig voneinander erfolgen. Beispielsweise ergibt sich für y = vluoch und x = fluch aus M(v,f) = 0.5875, M(l,l) = 0.9599, M(uo,u) = 0.8148 und M(ch,ch) = 0.9239 die Ähnlichkeit sim(vluoch, f luch) = 0.5875 · 0.9599 · 0.8148 · 0.9239 = 0.4246.

Alignmentverfahren

Auf der Basis der Substitutionsmatrizen suchen wir dann für jedes Lexem x aus Wörterbuch A das ähnlichste Lexem y aus Wörterbuch B. Die zugehörige Alignierung von x und y soll das Produkt der Ähnlichkeiten der einzelnen alignierten Buchstaben- bzw. Lautgruppen minimieren. Dazu 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 Paare möglichst nicht komplett alignieren wollen. Sobald ein Teilalignment eines Lexems y aus B schlechter ist als der Ähnlichkeitswert des besten bisher gefundenen Alignments oder als eine vorgegebene *Interessantheitsschranke*, kann man die Suche abbrechen.

Die verwendeten Alignmentverfahren sind der *Dijkstra-Algorithmus* und die *dynamische Programmierung*. 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 v aus A das ähnlichste Wort w aus B 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, und sie machte den Dijkstra-Algorithmus zum schnelleren Verfahren.

Ergebnisse

Die Ergebnisse zeigen, dass die durchschnittliche Ähnlichkeit der gefundenen ähnlichsten Wortpaare mit steigender Wortlänge N etwa im Verhältnis 1/N sinkt. Deswegen sollte man die Interessantheitsschranke geeignet in Abhängigkeit von der Wortlänge wählen.

Eine weitere Beschleunigung des Verfahrens ergab sich, wenn man die Lexeme vorsegmentiert. Dies ist insbesondere auch bei Komposita von Vorteil, wenn es zu einem Lexem v aus A keine Entsprechung w in B gibt. Dann kann man die Teile separat alignieren.

Speziell beim Mittelhochdeutschen Wörterbuch von Matthias Lexer stellt man fest, dass es in einigen Bereichen sehr viele Komposita aus einer Wortfamilie aufführt. Im Wörterbuch von

Jacob und Wilhelm Grimm gibt es dann oft keine Entsprechung. Durch die separate Alignierung konnte die Erfolgsquote hier deutlich erhöht werden.

3. Synergieeffekte

Mit dem Alignmenttool können Wörter unterschiedlicher Sprachstufen aligniert werden, um damit Segmentierungen und Annotationen aus dem Annotationstool in andere Sprachstufen übertragen zu können.

2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Art der Ausgabe	**Zeitraum**	**Summe**
Personalausgaben		
Pos. 0812 Beschäftigte (E13)	1.10.2008 - 31.12.2012	191.438,12 €
Pos. 0822 Beschäftigungsentgelte (Hilfskräfte)	1.10.2008 – 31.12.2012	27.585,07 €
Sächliche Verwaltungsausgaben		
Pos. 0843 Sonstige allge,meine Verwaltungsausgaben	1.10.2008 – 31.12.2012	1.122,34 €
Pos. 0846 Dienstreisen	19.10.2009 - 29.12.2011	13.114.34 €

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Das Forschungsprojekt 'Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen' konnte nur im Zusammenwirken der Einrichtung durchgeführt werden, die sich zu diesem Forschungsverbund zusammengeschlossen haben: der Bereich 'Sprache' erforderte nicht sprachwissenschaftlich-philologische Fachkompetenz, sondern auch Erfahrung im Umgang mit und Zugriff auf entsprechende Sprachdaten: Korpusdaten einerseits und Wörterbuchdaten andererseits. Dafür waren die Teilprojekte 1 und 2 die idealen Projektpartner, mit denen in Teilprojekt 3 entsprechende Partner mit Forschungskompetenz für Bioinformatik und Informatik kooperierten.

Wie die Einleitung zeigt, sind standen solche Wechselwirkungen zwischen Natur- und Geisteswissenschaften bereits am Anfang der Entwicklung in Biologie (Darwin) wie in der Sprachwissenschaft (Schleicher). Durch die technische Entwicklung war unser Verbundprojekt jetzt in der Lage, auf einem kleinen Sektor der Phänomene, große Datenmengen aus Biologie bzw. Sprachdaten zu untersuchen und für die Interpretation der Wechselwirkungen neue Akzente zu setzen. Die Konzentration auf den Bereich Wort und im Verlauf des Projekts die Konzentration auf Morphemstrukturen erwies sich dabei als notwendig im Blick auf Sorgfalt im Umgang mit den Daten udn auf ihre Verarbeitbarkeit und angemessen im unter dem Blickwinkel der Kooperation im Verbund.

Eine schöne Bestätigung dafür finden wir in dem Erfahrungsbericht ' Can a Science—Humanities Collaboration Be Successful?' von Blythe (School of Physics and Astronomy, University of Edinburgh, UK) und Croft (Department of Linguistics, University of New Mexico, Albuquerque, USA) aus dem Jahre 2010. Sie fassen ihre Erfahrungen folgendermaßen zusammen:

> "Using our interdisciplinary research collaboration as a case study, we discuss the question of whether formal modeling and empirical approaches can be successfully integrated into a single line of research. We argue that to avoid an undesirable disconnect between the two, one needs considerable time and patience for a science—humanities collaboration to bear fruit. In our collaboration and, we believe, in science—humanities collaborations in general, certain shared goals are required for success, including: starting with simple models before moving to more complex models; the importance of continually comparing models with empirical data where possible; and a focus on explaining statistical patterns rather than accounting for single data points individually."
>
> [http://adb.sagepub.com/content/18/1/12.short]

4. Voraussichtlicher Nutzen

Für die Anforderungen des Projekts mussten diese Daten in geeignete Analysestrukturen aufbereitet werden. Im Hinblick auf Nachnutzung und Nachhaltigkeit für virtuelle Forschungs-umgebungen haben wir dabei auf Normen im Bereich von Sprachdaten und Kodierung (gerade im Bereich der Lemmalisten) geachtet, die über die Webseite des Projekts zugänglich gemacht werden können. Auch bioinformatisch motivierte Arbeiten, z.B. die Ontologie, sind über die Webseite des Projekts nachnutzbar. Die Forschungsarbeiten der Informatik werden auf der Webseite der 'Research Group on Databases and Logic Programming' von Prof. Dietmar Seipel zugänglich gemacht.Unmittelbar sichtbar wird der Nutzen des Projekts an der Aufbereitung der Sprachdaten am Institut für deutsche Sprache in Mannheim. Bei Projektbeginn stand dort eine Liste der 30.000 häufigsten Wörter des DeReWo-Korpus zur Verfügung. Heute weist die DeReWo-Liste über 300.000 der häufigsten Wörter des DeReWo-Korpus nach, ein Volumen, das in etwa dem Umfang des historischen 'Deutschen Wörterbuchs' von Jacob und Wilhelm Grimm entspricht.

5. Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Fortschritte auf dem Gebiet des Vorhabens bei anderen Stellen sind nicht bekannt geworden. Lediglich im Bereich der Thematik 'Varianz' deuten sich in anderen Forschungsbereichen neue Aktivitäten an (z.B. in einer Pilotstudie der Mellon-Foundation, auf die wir aufmerksam gemacht wurden (Pilot Study to explore the relationship between linguistic variation in a large corpus of early modern English texts and the development of genres of printed discourse). Schließlich wurde auf der Digital Humanities Konferenz in Hamburg ein Visualisierungsprojekt vorgestellt (und mit einem Preis ausgezeichnet) zum Thema: 'Patchworks and Field-Boundaries: Visualizing the History of English'. Das Projekt arbeitet mit Daten des 'Historical Thesaurus of English', also völlig anderem Datenmaterial, wäre aber aus dem Blickwinkel 'Visualisierung' von großem Interesse für eine Fortsetzung der Forschungsarbeiten des Wechselwirkungsprojekts.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses

- Christian Schneiker, Dietmar Seipel, Werner Wegstein: Schema and Variation: Retro-Digitizing Printed Dictionaries. Teilnahme mit Vortrag (Christian Schneiker) und Veröffentlichung an der Tagung Third Linguistic Annotation Workshop (LAW III), 2009, held in conjunction with the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of NLP (ACL-IJCNLP 2009), Singapur.

- Christian Schneiker, Dietmar Seipel, Werner Wegstein, Klaus Prätor: Declarative Parsing and Annotation of Electronic Dictionaries. Teilnahme mit Vortrag (Christian Schneiker) und Veröffentlichung an der Tagung 6th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2009), Mailand, Italien.

- Christian Schneiker, Mohamed M. Khamis, Dietmar Seipel: Prolog Server Faces - A Declarative Framework for Dynamic Web Pages; Paper zu Web-Benutzerschnittstellen. Teilnahme mit Vortrag (Christian Schneiker) und Veröffentlichung an der Tagung 24th Workshop on (Constraint) Logic Programming (WLP 2010), an der German University in Kairo, Ägypten. http://met.guc.edu.eg/events/ wlp2010/proceedings/ Declarative%20Framework.pdf

- Dietmar Seipel, Werner Wegstein: MetaDictionary - Towards a Generic e-Infrastructure for Detecting Variance in Language by Exploiting Dictionary Information. Teilnahme mit Vortrag (Dietmar Seipel) und Veröffentlichung an der Tagung ISGC 2011 – 'International Symposium on Grids and Clouds' an der Academia Sinica in Taipei, Taiwan.
http://event.twgrid.org/isgc2011/slides/ HumanitiesandSocialSciencesApplications/1/Seipel.pdf

- Keller DBK, Schultz J (2012) Morpheme networks reveal language dynamics. In: Actes des 11es Journées Internationales d'Analyse Statistique des Données Textuelles / Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data.

- D. Seipel, L. Borek: A Tool for Collaborative Rule-Based Morpheme Annotation. Teilnahme mit Vortrag (Dietmar Seipel) und Veröffentlichung an der Tagung International Symposium on Grids and Clouds (ISGC 2012) an der Academia Sinica in Taipei, Taiwan.

- C. Schneiker, D. Seipel: Declarative Web Programming with Prolog and XUL. Teilnahme mit Vortrag (Christian Schneiker) und Veröffentlichung an der Tagung 26th Workshop on Logic Programming (WLP 2012).

- Laurent Romary, Werner Wegstein: Consistent Modeling of Heterogeneous Lexical Structures. In: Journal of the Text Encoding Initiative, Issue 3, November 2012, TEI and Linguistics. [http://jtei.revues.org/540]

- Keller DBK, Schultz J (2013) Connectivity, not frequency, determines the fate of a morpheme. In: PLoS One (in press).

- Keller DBK, Schultz J (2013) Word formation is aware of morpheme family sizes. (Manuscript in preparation).

- Heike Stadler, Werner Wegstein: Why and how to encode word structures and word-formation formula ? A proposal based on <etym> [Arbeitstitel]. Veröffentlichung im 'Journal of the Text Encoding Initiative'.

- Andrea Rapp, Werner Wegstein, Luise Borek (Hrsg.): Varianz und Vielfalt interdisziplinär. Wörter und Strukturen. Interdisziplinärer Workshop im Rahmen des Projektes „Wechsel-wirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen.“ in Darmstadt, 29. und 30. November 2012. Mannheim, Institut für Deutsche Sprache, 2013. [Veröffentlichung voraussichtlich im 4. Quartal 2013, mit Beiträgen von Daniela Keller und Dietmar Seipel]

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel Schlussbericht des Verbundprojekts: Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen, Teilprojekt: Bioinformatik und Informatik	
4. Autor(en) [Name(n), Vorname(n)] Schultz, Jörg Seipel, Dietmar Wegstein, Werner	5. Abschlussdatum des Vorhabens 31.12.2012 6. Veröffentlichungsdatum 7. Form der Publikation
8. Durchführende Institution(en) (Name, Adresse) Julius-Maximilians-Universität Würzburg Sanderring 2 97070 Würzburg	9. Ber. Nr. Durchführende Institution 10. Förderkennzeichen 01UA0815C 11. Seitenzahl 27
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben 40 14. Tabellen 2 15. Abbildungen 5
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung Der interdisziplinäre Vergleich von Sprachdaten mit bioinformatischem Material kann aus methodologischer Sicht nur nach sorgfältiger Identifikation analoger Strukturen erfolgen. Hierzu definierte die Arbeitsgruppe 'Bioinformatik' federführend mittels des Programms 'Protégé' eine Ontologie der für das Projekt relevanten sprachlichen Strukturen und Prozesse, als Basis für den Vergleich mit existierenden biologischen Ontologien und führte mit bioinformatischen Methoden und Verfahren Analysen an Sprachdaten aus dem Projekt durch. Für die vergleichende Analyse erarbeitete die Arbeitsgruppe ''Informatik' und 'EDV-Philologie' ein Annotationstool für die morphologische Kennzeichnung von in Morphemstrukturen segmentierten Wörterbucheinträgen und Korpusdaten, basierend auf der deklarativen Programmiersprache Prolog sowie ein Alignmenttool für die Alignierung von Worteinträgen aus Wörterbüchern zweier verschiedener Sprachstufen.	
19. Schlagwörter Ontologie,Kodierungsverfahren, Prolog, Morphemsegmentierung, Annotation, Alignment, Systemlinguistik, Wortbildung, XML, TEI P5 Guidelines	
20. Verlag	21. Preis