

# Sprache – spräche – spriche – Sproch...

Die Konzeption einer Metalemmaliste zur Erschließung von  
Varietäten und Varianz

Luise Borek M.A.  
Dipl.-Math. Steffen Eger M.A.

9. September 2009



**Kompetenzzentrum**  
für elektronische Erschließungs-  
und Publikationsverfahren in  
den Geisteswissenschaften



**INSTITUT FÜR  
DEUTSCHE SPRACHE**

„Wechselwirkungen zwischen linguistischen und bioinformatischen  
Verfahren, Methoden und Algorithmen: Modellierung und  
Abbildung von Varianz in Sprache und Genomen“

# Förderprogramm

## „Wechselwirkungen zwischen Natur- und Geisteswissenschaften“

- Interdisziplinäre Forschungsverbünde geistes- und naturwissenschaftlicher Fächer
- Es sollen sowohl naturwissenschaftlich-technische, mathematische oder informationstechnologische Kompetenz einfließen
- Einsatz naturwissenschaftlicher Methoden in den Geisteswissenschaften und umgekehrt um nicht zuletzt neue Methoden entwickeln zu können

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

# Verbundpartner

- Prof. Dr. Claudine Moulin, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier
- Prof. Dr. Ludwig Eichinger, Institut für Deutsche Sprache, Mannheim
- Prof. Dr. Dietmar Seipel, Institut für Informatik, Lehrstuhl für Informatik I, Informationsstrukturen und wissensbasierte Systeme, Universität Würzburg
- Prof. Dr. Jörg Schultz, Dipl.-Biol., Theodor Boveri-Institut für Biowissenschaften/Bioinformatik, Universität Würzburg
- Prof. Dr. Werner Wegstein, Kompetenzzentrum EDV-Philologie, Universität Würzburg

# Verbundprojekt

## Teilvorhaben

- TV 1: Basislemmaliste der neuhochdeutschen Standardsprache
- TV 2: Klassifizierte Varietäten-Lemmalisten, semasiologische Vernetzung, Erkundungsmodul onomasiologische Vernetzung
- TV 3.1: Grammatik der Varianz in Genomen und Sprache, Quantitative vergleichende Analysen
- TV 3.2: Technik: Visualisierung, Gridifizierung, Modellierung der Vernetzung

# Einordnung der Metalemmaliste

## Datenbanken der Disziplinen



Ensembl ([www.ensembl.org](http://www.ensembl.org))



DWB ([www.dwb.uni-trier.de](http://www.dwb.uni-trier.de))

## Untersuchung von Varianz

→ benötigt ein Bezugssystem

- ist in der Biologie nicht realisierbar ✗
- entsteht (für die Sprachwissenschaft) in Form der Metalemmaliste ✓

# Was ist eine „Metalemmaliste“?

## Funktion

- Vernetzungstool zur Erschließung von Varianz
- Ein Metalemma ist ein „Gemeinsames Drittes“; Andockstelle für syn- und diachrone Varianten

## Merkmale

- Metalemma entspricht nhd. Standard, öffentlichem Sprachgebrauch
- Metalemma unterscheidet sich von einem Lemma ordnungsstrukturell
- (zunächst) rein semasiologisch angelegt

# Datengrundlage I

- Wörterbücher
  - Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm
  - Mittelhochdeutsches Wörterbuch von Benecke, Müller und Zarncke
  - Mittelhochdeutsches Handwörterbuch von Matthias Lexer
  - Findebuch zum mittelhochdeutschen Wortschatz
  - Pfälzisches Wörterbuch
  - Rheinisches Wörterbuch
  - Wörterbuch der elsässischen Mundarten
  - Wörterbuch der deutsch-lothringischen Mundarten
  - Goethe-Wörterbuch
- Korpusdaten



# Datengrundlage II

## DeReKo

- bildet mit über 3 Milliarden Wörtern die weltweit größte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus Gegenwart und neuerer Vergangenheit (50-60er Jahre)
- enthält belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten (Hauptbestandteil) sowie eine breite Palette weiterer Textarten
- wird im Hinblick auf Umfang, Variabilität und Qualität kontinuierlich weiterentwickelt
- Umfang und Zusammensetzung: [www.ids-mannheim.de/kl/projekte/korpora/archiv.html](http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html)

# Datengrundlage III

## DEReWo

- Lemmaliste der 30.000 „häufigsten“ deutschen Lemmata der Gegenwart und neueren Vergangenheit
- Basierend auf DEReKo, dem DEUTSCHEN REFERENZKORPUS
- Lemmaliste erzeugt durch Anwendung linguistischer Verarbeitungswerkzeuge (Lemmatisierer, Part-of-speech tagger) auf die Datensätze
- Quelle: [www.ids-mannheim.de/kl/projekte/methoden/derewo.html](http://www.ids-mannheim.de/kl/projekte/methoden/derewo.html)

# Datengrundlage IV

## Basislemmaliste

- Erweitert und ergänzt DEReWo
  - Quantität: Ziel ca. 100.000 Lemmata
  - Qualität: exaktere Analyse linguistischer Daten z. B. durch Zusammenarbeit verschiedener Analyse-Werkzeuge
  - Information: Wortklasse, Disambiguierung homographischer Formen
- Bereits (teilweise) realisierte Arbeitsschritte:  
Eigennamenerkennung, verbesserte Lemmatisierung,  
Disambiguierung, Präverbfügungen, erweiterte Lexikalische  
Abdeckung der Analysewerkzeuge, (Semantische Relationen)

# Beispiel „Sprache“:

spriche

sprâche

## Beispiel „Sprache“:

spriche

Sproch

sprâche

Sprach

# Beispiel „Sprache“:

spriche

Sproch

Sprache

sprache

sprâche

Sprach

# Struktur des Metalemmas

```
<metalemma id="1" value="Sprache" pos="nn" source="bll">
  <lemma corpus="bll" corpus_id="778"
    value="Sprache" lang="nhd">
    <lemma corpus="dwb" value="sprache">
    <lemma corpus="elswb" value="Sprach">
    <lemma corpus="lothrbw" value="Sproch">
    <lemma corpus="pfbw" value="Sprache">
    <lemma corpus="bmz" value="sprâche">
    <lemma corpus="bmz" value="spriche">
    <lemma corpus="lexer" value="sprâche">
    <lemma corpus="lexer" value="spriche">
  </metalemma>
```

# Komplexes Metalemma

```
<metalemma id="128" value="August" pos="nn"  
    source="bll">  
  <lemma corpus="bll" corpus_id="414"  
    value="August" lang="nhd"/>  
  <lemma corpus="gwb" value="August"/>  
  <lemma corpus="rhwb" value="August"/>  
  <lemma corpus="rhwb" value="August"/>  
  <lemma corpus="rhwb" value="August"/>  
  <lemma corpus="pfbw" value="August"/>  
  <lemma corpus="dwb" value="augst"/>  
  <lemma corpus="elwb" value="August"/>  
  <lemma corpus="lexer" value="ougest"/>  
  <lemma corpus="lexer" value="ougeste"/>  
  <lemma corpus="lexer" value="ougst"/>  
  <lemma corpus="lexer" value="ougste"/>  
  <lemma corpus="lexer" value="augustö"/>  
  <lemma corpus="lexer" value="ochste"/>  
  <lemma corpus="lexer" value="ogest"/>  
  <lemma corpus="lexer" value="ougest"/>
```

```
<lemma corpus="lexer" value="oust"/>  
<lemma corpus="lexer" value="ouste"/>  
<lemma corpus="lexer" value="ouwest"/>  
<lemma corpus="lexer" value="ouwestinne"/>  
<lemma corpus="lexer" value="owest"/>  
<lemma corpus="lexer" value="owist"/>  
<lemma corpus="lexer" value="öugestinne"/>  
<lemma corpus="nachtrlexer" value="ougest"/>  
<lemma corpus="findebuch" value="ougest"/>  
<lemma corpus="findebuch" value="augst"/>  
<lemma corpus="findebuch" value="oust"/>  
<lemma corpus="findebuch" value="ouwest"/>  
<lemma corpus="bmz" value="ougeste"/>  
<lemma corpus="bmz" value="ougste"/>  
<lemma corpus="bmz" value="oust"/>  
<lemma corpus="bmz" value="owest"/>  
<lemma corpus="bmz" value="ouwest"/>  
</metalemma>
```



# Zuordnungen

## Zuordnungsverfahren

- Aufgabe / Problem der Zuordnungsverfahren: Vernetzung diachroner und diatopischer Varianten des Deutschen
- Verankerungspunkt ist dabei der nhd. Standard bzw. das „Gegenwartsdeutsche“
- Generelle Möglichkeiten der Zuordnung:
  - linguistisches Expertenwissen
  - Verfahren des maschinellen Lernens / Statistik
  - Verweise in den Wörterbüchern

# String-Transformationen

## Problemstellung

- Problem: falls kein direkter Verweis im Wörterbuch, überführe Wortform der Sprachstufe  $x$  in Wortform der Sprachstufe  $y$
- Bsp: *flusz* (Grimm) nach *fluss* (Gegenwart)
- Vorteilhaft, zuerst benachbarte Sprachstufen zu betrachten
- Evtl. Kaskadierung, um von entfernteren Sprachstufen zur Gegenwart zu gelangen
- Verschiedene Möglichkeiten, um Transformation zu bewerkstelligen
  - Linguistisches Wissen (z. B. Lautgesetze)
  - selbständiges Lernen aus Daten

# Lautgesetze

## Beispiel

- Durch den Einsatz von Lautgesetzen auf ältere Sprachstufen lässt sich die neuhochdeutsche Form (re)konstruieren

- Beispiel

	Mhd.	Nhd.
Neuochhochdeutsche Diphthongierung	⟨î⟩	→ ⟨ei, ai⟩
	⟨û⟩	→ ⟨au⟩
	⟨iu⟩	→ ⟨eu, äu⟩
	bewîsen	→ bew <sup>ei</sup> sen
	tûchen	→ t <sup>au</sup> chen
	schîune	→ Sch <sup>eu</sup> ne

- Benötigt: linguistisches Expertenwissen

# Selbständiges Lernen: $n$ -gramme

## Beispiel

- Entdecke in Sprachstufe  $x$  überproportional repräsentierte  $n$ -gramme — Statistisch und automatisch

- Beispiel:

	Grimm	Gegenwart
Bigramme	sz, sc, th, tt	ss, sk
Trigramme	usz, osz, esz, asz, lic	ion, ung

- Finde zu substituierende Sprachkonstrukte in Sprachstufe  $y$
- Ersetze jeweilige Sprachkonstrukte und finde so Zuordnungen
- Beispiel:
  - ca. 300.000 Einträge im Grimm-Wörterbuch, davon ca. 70.000 nicht zuordenbar zu Gegenwartssprache.
  - Nach Änderung von  $sz \rightarrow ss$  und  $th \rightarrow t$ , 15.000 weitere zuordenbar.

# Selbständiges Lernen: Sequence Alignment

## Idee

- Motivation aus der Biologie / Bioinformatik
- Anordnung von Strings so, dass *score* maximiert wird
- Grundidee ist, dass ein String aus dem anderen durch Mutation hervorgegangen ist
- Beispiel:

v	e	r	p	f	i	c	h	t	i	g	e	n
v	e	r	p	f	i	c	h	t	-	-	e	n
- *Score* bewertet mittels Parameter: Alignierung übereinstimmender Buchstabensymbole, Alignierung unterschiedlicher Symbole, Lücken (-)
- Bestimmen der Parameter aus linguistischen Datensätzen

# Selbständiges Lernen: Bootstrapping

## Idee

- Generiere Anfangsdatensätze von Zuordnungen (händisch, semi-automatisch, ...)
- Lerne Übersetzungsregeln mittels Klassifizierer (Entscheidungsbaum, frequent episodes, etc.)
- Verbessere Performanz durch subsampling und resampling
- Verallgemeinere auf restliche Datensätze und erhalte Übersetzer zwischen Sprachstufen

# Kookkurrenzverhalten

## Idee

- Ermittle Kollokationen von Wörtern
- 1.) Varietäten: Aus Wörterbüchern
- 2.) Gegenwartssprache: Aus Kookkurrenzanalyse
- Letztere wird am IDS zur Identifikation von (Beinahe-)Synonymen verwendet
- Ordne Wörter aus verschiedenen Sprachstufen anhand der Ähnlichkeit des Kollokationsprofils zu
- Ergibt dann auch **onomasiologische** Zuordnung

# Kookkurrenzverhalten: Beispiel

## Beispiel

- `<form type="lemma">flusz</form>`  
...  
`<sense>...` da die bedeutung des rinnenden wassers  
mit der des flusses zusammentrifft, und wir einen  
flusz oder strom auch heute noch wasser oder  
gewässer nennen ....  
`</sense>`
- Kookkurrenzprofil von gegenwartssprachlich *fluss*: Seen, Ufer,  
Bäche, reißenden, **Wasser**, Stadt, ..., **Strom**, ..., **Wassers**, ...,  
**Gewässer**
- <http://corpora.ids-mannheim.de/ccdb/>



	In Basislemmaliste	Nicht in Basislemmaliste
In den Wörterbüchern	verknüpfbar	Pseudometalemma notwendig;  Erweiterung der Basislemmaliste erforderlich
Nicht in den Wörterbüchern	z. B. Entlehnungen jüngeren Ursprungs ( <i>Football</i> )	–

# Vorteile und Anwendungen

## Vorteile und Anwendungen

- Erleichterte Suche durch Erschließung von Varietätenlemmata; flexibler Einstieg über das Metalemma
- Erhalt der originalen Wörterbuch-Strukturen bei Mehrwert durch Vernetzung
- Ergänzung durch Onomasiologie
- Andockstelle für weitere Informationssysteme

# Praktisches Beispiel

## DAS HEINRICH-HEINE-PORTAL

Startseite | Aktuelles | Über Heinrich Heine | Werke | Briefwechsel | Suchen | Zeitgenössische Drucke | Über das Projekt | Kontakt | Hilfe

Titelseite | Inhaltsverzeichnis | Apparat

Klingt das Lied auch nicht ergötzlich,  
Hat's mich doch von Angst befreit.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H II.

H Ich weiß nicht, was soll es bedeuten,  
Daß ich so traurig bin;  
Ein Märchen aus alten Zeiten,  
Das kommt mir nicht aus dem Sinn.

5 Die Luft ist kühl und es dunkelt,  
Und ruhig fließt der Rhein;  
Der Gipfel des Berges funkelt  
Im Abendsonnenschein.

10 Die schönste Jungfrau sitzet  
Dort oben wunderbar,  
Ihr gold'nes Geschmeide blitzet,  
Sie kämmt ihr goldenes Haar.

DHA, Bd. 1/1, S. 207  
Aufschlagen

Titelseite | Inhaltsverzeichnis | Apparat

DHA, Bd. 1/1, S. 208  
Aufschlagen

### DIE HEIMKEHR

J<sup>1</sup> Sie kämmt es mit goldnem Kamme,  
Und singt ein Lied dabey;  
15 Das hat eine wundersame,  
Gewaltige Melodey.

D<sup>1</sup> R<sup>1</sup>H Den Schiffer, im kleinen Schiffe,  
Ergreift es mit wildem Weh;  
Er schaut nicht die Felsenriffe,  
20 Er schaut nur hinauf in die Höh'.

H Ich glaube, die Wellen verschlingen  
Am Ende Schiffer und Kahn;  
Und das hat mit ihrem Singen  
Die Lore-Ley gethan.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H IX.

Mein Heim, mein Heim, ich komme!

Gefördert durch die Kunststiftung NRW und die Deutsche Forschungsgemeinschaft

Heine *Lorelei* nach der Düsseldorfer Heineausgabe (<http://heine-portal.de/>)

# Praktisches Beispiel

## DAS HEINRICH-HEINE-PORTAL

Startseite | Aktuelles | Über Heinrich Heine | Werke | Briefwechsel | Suchen | Zeitgenössische Drucke | Über das Projekt | Kontakt | Hilfe

Titelseite | Inhaltsverzeichnis | Apparat

Klingt das Lied auch nicht ergötzlich,  
Hat's mich doch von Angst befreit.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H II.

H Ich weiß nicht, was soll es bedeuten,  
Daß ich so traurig bin;  
Ein Märchen aus alten Zeiten,  
Das kommt mir nicht aus dem Sinn.

5 Die Luft ist kühl und es dunkelt,  
Und ruhig fließt der Rhein;  
Der Gipfel des Berges funkelt  
Im Abendsonnenschein.

10 Die schönste Jungfrau sitzet  
Dort oben wunderbar,  
Ihr gold'nes Geschmeide blitzet,  
Sie kämmt ihr goldenes Haar.

DHA, Bd. 1/1, S. 207  
Aufschlagen

Titelseite | Inhaltsverzeichnis | Apparat

DHA, Bd. 1/1, S. 208  
Aufschlagen

### DIE HEIMKEHR

J<sup>1</sup> Sie kämmt es mit goldnem Kamme,  
Und singt ein Lied dabey;

15 Das hat eine wundersame,  
Gewaltige Melodey.

D<sup>1</sup> R<sup>1</sup>H Den Schiffer, im kleinen Schiffe,  
Ergreift es mit wildem Weh;  
Er schaut nicht die Felsenriffe,  
20 Er schaut nur hinauf in die Höh'.

H Ich glaube, die Wellen verschlingen  
Am Ende Schiffer und Kahn;  
Und das hat mit ihrem Singen  
Die Lore-Ley gethan.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H IX.

Mein Heim, mein Heim, ich komme!

Gefördert durch die Kunststiftung NRW und die Deutsche Forschungsgemeinschaft

Heine *Lorelei* nach der Düsseldorfer Heineausgabe (<http://heine-portal.de/>)

# Praktisches Beispiel

## DAS HEINRICH-HEINE-PORTAL

Startseite | Aktuelles | Über Heinrich Heine | Werke | Briefwechsel | Suchen | Zeitgenössische Drucke | Über das Projekt | Kontakt | Hilfe

Titelseite | Inhaltsverzeichnis | Apparat

Klingt das Lied auch nicht ergötzlich,  
Hat's mich doch von Angst befreit.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H II.

H Ich weiß nicht, was soll es bedeuten,  
Daß ich so traurig bin;  
Ein Märchen aus alten Zeiten,  
Das kommt mir nicht aus dem Sinn.

5 Die Luft ist kühl und es dunkelt,  
Und ruhig fließt der Rhein;  
Der Gipfel des Berges funkelt  
Im Abendsonnenschein.

Die schönste Jungfrau sitzet  
10 Dort oben wunderbar,  
Ihr gold'nes Geschmeide blitzet,  
Sie kämmt ihr goldenes Haar.

DHA, Bd. 1/1, S. 207  
Aufschlagen

Titelseite | Inhaltsverzeichnis | Apparat

DHA, Bd. 1/1, S. 208  
Aufschlagen

### DIE HEIMKEHR

J<sup>1</sup> Sie kämmt es mit goldnem Kamme,  
Und singt ein Lied dabey;  
15 Das hat eine wundersame,  
Gewaltige Melodey.

D<sup>1</sup> R<sup>1</sup>H Den Schiffer, im kleinen Schiffe,  
Ergreift es mit wildem Weh;  
Er schaut nicht die Felsenriffe,  
20 Er schaut nur hinauf in die Höh'.

H Ich glaube, die Wellen verschlingen  
Am Ende Schiffer und Kahn;  
Und das hat mit ihrem Singen  
Die Lore-Ley gethan.

J<sup>1</sup> D<sup>1</sup> R<sup>1</sup>H IX.

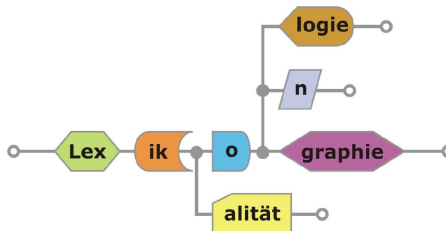
Märchen  
dabey  
Melodey  
gethan

gold'nes vs goldenes  
vs goldnem

Heine *Lorelei* nach der Düsseldorfer Heineausgabe (<http://heine-portal.de/>)

Gefördert durch die Kunststiftung NRW und die Deutsche Forschungsgemeinschaft

Herzlichen Dank für Ihre Aufmerksamkeit!



Luise Borek M.A.  
borek@uni-trier.de  
www.kompetenzzentrum.uni-trier.de

Dipl.-Math. Steffen Eger M.A.  
eger@ids-mannheim.de  
www.ids-mannheim.de