INSTITUT FÜR INFORMATIK
Lehr- und Forschungseinheit für
Programmier- und Modellierungssprachen
Oettingenstraße 67, D–80538 München

Ludwig
Maximilians–
Universität
München

# Eine linguistische Wissensbank

**Heribert Schütz und Dietmar Zaefferer**

http://www.pms.informatik.uni-muenchen.de/publikationen
Forschungsbericht/Research Report PMS-FB-1996-17, November 1996

# Eine linguistische Wissensbank

Heribert Schütz[†] Dietmar Zaefferer[‡]

[†]Institut für Informatik
heribert.schuetz@informatik.uni-muenchen.de
[‡]Institut für Deutsche Philologie
ue303bh@sunmail.lrz-muenchen.de

Ludwig-Maximilians-Universität München

**Zusammenfassung**

Wir stellen in diesem Beitrag die Arbeit an einer Datenbank vor, in der Wissen über sprachliche Phänomene verschiedenster Einzelsprachen verwaltet wird. Zweck dieser Datenbank ist die Unterstützung der linguistischen Forschung.

Das zentrale Problem bei der Konzeption einer solchen Datenbank ist das der Vergleichbarkeit der Daten: Damit keine Birnen in der Kategorie der Äpfel aufgeführt werden, müssen alle Sprachbeschreibungen mithilfe einer einheitlichen Terminologie durchgeführt und gegliedert werden, die auf einheitlichen theoretischen Grundannahmen beruhen. Wir beschreiben einen logikbasierten Ansatz, der eine solche terminologische Standardisierung unterstützen soll.

## 1 Einführung

Die Linguistik befasst sich mit Sprache im Allgemeinen und Sprachen im Besonderen, d.h. einerseits mit der spezifisch menschlichen Fähigkeit, höchst komplexe Zeichensysteme zu erlernen, zu modifizieren und weiterzugeben, und andererseits mit den verschiedenen Einzelsystemen, die sich aufgrund dieser Fähigkeit in den verschiedenen Sprachgemeinschaften herausgebildet haben, je nach Zählung etwa fünf- bis sechstausend. Einen wichtigen Zugang zum Verständnis des menschlichen Sprachvermögens bietet die vergleichende Untersuchung der unterschiedlichen Einzelsprachen. Viele Sprachen sind erst höchst mangelhaft beschrieben, aber selbst das vorhandene Wissen über die besser dokumentierten Sprachen ist für vergleichende Untersuchungen zum Teil viel zu schwer zugänglich. Hier bietet die moderne Informationstechnologie einen Ausweg an.

Um die linguistische Forschung zu erleichtern, ist es wünschenswert, vorhandenes Wissen über möglichst viele Eigenschaften möglichst vieler Sprachen in einer Datenbank zu verwalten. In diesem Papier werden die besonderen Anforderungen an eine solche Datenbank erläutert (Abschnitt 2) und der von uns verfolgte Lösungsansatz dargestellt (Abschnitt 3). Wir stellen einige der noch offenen Probleme vor (Abschnitt 4) und berichten schließlich über den Stand der Arbeit und geben einen Ausblick auf mögliche zukünftige Entwicklungen (Abschnitt 5).

## 2 Szenario

Informationen, die in der Datenbank verwaltet werden sollen, können beispielsweise sein:

- Das Vokalinventar des Italienischen
- Die Regeln der Vokalharmonie im Türkischen
- Die Flexionsformen der Verben im Finnischen
- Die Regeln zur Bildung von Relativsätzen im Tagalog
- Der Bestand von Witterungsprädikaten im Deutschen
- Die Markierung von Fragesätzen im Bulgarischen

Wir erwarten, dass typischerweise Spezialisten für bestimmte Sprachen Informationen über die jeweiligen Sprachen in die Datenbank eintragen, während Spezialisten für Teilgebiete der Linguistik Anfragen über bestimmte sprachliche Phänomene in allen Sprachen stellen. Vorstellbare Anfragen sind beispielsweise:

- Welchen Umfang haben typischerweise Vokalinventare?
- In welchen Sprachen gibt es Vokalharmonie?
- Gibt es Sprachen, in denen Labiodentale mit der Oberlippe und den unteren Schneidezähnen gebildet werden?
- Gibt es Sprachen ohne Verben?
- Gibt es Sprachen ohne Verbflexion (unterschiedliche Verbformen)?
- Welche Sprachen unterscheiden mehrere Stufen von Vergangenheitsformen?
- Auf welche Weisen können Relativsätze gebildet werden?
- Gibt es Sprachen ohne Markierung von Fragesätzen?

Abbildung 1 verdeutlicht die zwei Dimensionen, nach denen die linguistischen Informationen im Wesentlichen eingeteilt werden.

## 3 Lösungsansatz

Es ist sicher nicht mit vertretbarem Arbeitsaufwand möglich, die (Einträge in die) Datenbank so weit zu formalisieren, dass alle Anfragen dieser Art vollautomatisch beantwortet werden können. Viele der verwalteten Informationen werden vielmehr informell, also als natürlichsprachige Texte verwaltet werden. Dies kann insbesondere deshalb in Kauf genommen werden, da unser vorrangiges Ziel nicht die automatische Sprachverarbeitung ist, sondern – wie bereits gesagt – die Unterstützung von Forschern aus verschiedensten Teilbereichen der Linguistik.

Dennoch soll es bis zu einem gewissen Grad möglich sein, gezielt auf Wissen über bestimmte Phänomene in verschiedenen Sprachen zuzugreifen. Dies

| | Kisuaheli | Deutsch | Tagalog | Japanisch | Italienisch | ... | | |
|---|---|---|---|---|---|---|---|---|
| **Vokalinventar** | | | | | | | | |
| **Konsonanteninventar** | | | | | | | | |
| **Assimilationen** | | | | | | | | |
| **Dissimilationen** | | | | | | | | |
| **...** | | | | | | | | |
| **Bildung der Vergangenheitsformen** | | | | | | | | |
| **...** | | | | | | | | |
| **Bildung von Relativsätzen** | | | | | | | | |
| **...** | | | | | | | | |
| | | | | | | | | |

Abbildung 1: Wissen wird Sprache für Sprache eingetragen und nach sprachlichen Phänomenen abgefragt.

ist kein Problem, wenn Phänomene eindeutige Namen haben. Leider werden jedoch in den Beschreibungen verschiedener Sprachen vielfach sehr unterschiedliche Terminologien verwendet, d.h. gleiche oder ähnliche Phänomene werden verschieden bezeichnet und umgekehrt wird der gleiche Ausdruck für verschiedene Phänomene verwendet. Dadurch werden einerseits echte Gemeinsamkeiten und andererseits echte Unterschiede zwischen Sprachen verdeckt. Anfragen an die Datenbank liefern dann falsche Antworten.

Für die Lösung dieses Problems bieten sich Techniken des Information Retrieval an, etwa eine Thesaurus-unterstützte Suche in einer weitgehend unstrukturierten Volltext-Datenbank. Wir erhoffen uns jedoch von einer weitergehenden Formalisierung die Möglichkeit eines gezielteren Zugriffs auf Informationen. Der von uns gewählte Ansatz basiert auf einer standardisierten Terminologie. Bezeichnungen für Teilbereiche der Linguistik, Phänomen-Klassen und Phänomene sollen den in die Datenbank eintragenden Sprach-Spezialisten bis zu einem gewissen Detaillierungsgrad vorgegeben werden.

Eine ähnliche Strategie wurde auch bei einer Reihe von Sprachbeschreibungen in Buchform, den Grammatiken der Descriptive Grammars Reihe bei Routledge, (vgl. [CS77]) verfolgt. Jedes Buch dieser Reihe beschreibt eine andere Sprache, aber alle Bücher haben die gleiche Gliederung. Jeder Punkt dieser Gliederung entspricht einem Teilgebiet der Linguistik, einer Gruppe von Phänomenen oder einem einzelnen Phänomen.

Traditionell haben Klassifikationen die Form eines Baums oder eines zyklenfreien Graphen, etwa als Klassenhierarchie einer Objekt-orientierten Datenbank oder Programmiersprache. Abbildung 2 enthält eine solche (verein-

```
cooccurrence restriction
  assimilation
    total assimilation
      total assimilation of vowels
        phonological total assimilation of vowels
        grammatical total assimilation of vowels
          outward total assimilation of vowels
          inward total assimilation of vowels
      total assimilation of consonants
        ...
    partial assimilation
      partial assimilation of vowels
        ...
      partial assimilation of consonants
        ...
  dissimilation
    dissimilation of vowels
      ...
    dissimilation of consonants
      ...
```

Abbildung 2: Eine Klassenhierarchie für Kookkurrenz-Beschränkungen. Drei Punkte stehen jeweils für Subklassen analog zu den Subklassen von `total assimilation of vowels`.

fachte) Hierarchie für Kookkurrenz-Beschränkungen, d.h. für Regeln über das gemeinsame Auftreten von Lauten in Silben, Wörtern oder Wortgruppen.[1] Es fällt auf, dass verschiedene Klassen auf die gleiche Weise in Unterklassen aufgeteilt werden. Eine derartige Gruppe von Klassen mit einheitlicher Subklassen-Struktur besteht aus `total assimilation of vowels`, `total assimilation of consonants`, `partial assimilation of vowels`, etc. Dies hat ja in der Abbildung die abkürzende Verwendung von drei Punkten erlaubt. Eine weitere Gruppe einheitlich strukturierter Klassen besteht aus `total assimilation`, `partial assimilation` und `dissimilation`.

Diese Regelmässigkeit kann in der traditionellen Klassenhierarchie nicht repräsentiert werden. Dies hat wesentliche Nachteile: Falls wir bei den grammatikalisch bedingten Kookkurrenzbeschränkungen etwa zusätzlich zu den auswärts (vom Stamm zum Affix) und einwärts (vom Affix zum Stamm) wirkenden Beschränkungen noch stamminterne berücksichtigen wollen, müssen wir an mehreren Stellen der Klassenhierarchie neue Klassen einfügen. Falls wir mehrere Arten von Dissimilation unterscheiden wollen, müssen wir einen Teil der Klassenhierarchie noch weitergehend umstrukturieren. Ähnliche Probleme treten auf, wenn wir eine Unterscheidung aufgeben wollen, etwa die zwischen Vokalen und Konsonanten.

Ein weiterer Nachteil dieser Klassenhierarchie ist, dass beispielsweise eine Klasse für alle Kookkurrenz-Beschränkungen auf Vokalen fehlt. Klassen dieser

[1] Für das Verständnis dieses Papiers ist die Kenntnis der genauen Bedeutung der einzelnen Klassen nicht erforderlich.

```
cooccurrence restriction
  assimilation
    total
    partial
  dissimilation

cooccurrence restriction
  on vowels
  on consonants

cooccurrence restriction
  phonological
  grammatical
    outward
    inward
```

Abbildung 3: Drei orthogonale Klassenhierarchien für Kookkurrenz-Beschränkungen.

Art können zwar eingeführt werden, wenn wir einen beliebigen zyklenfreien Graph und nicht nur einen Baum als Struktur der Klassenhierarchie zulassen, aber die zuvor beschriebenen Probleme verschlimmern sich noch: Mit welchen alten Klassen soll die neu eingeführte Klasse eine „Durchschnittsklasse" bilden?

Die Klassenhierarchie aus Abbildung 2 kann gewissermaßen als kartesisches Produkt der drei zueinander orthogonalen Hierarchien aus Abbildung 3 betrachtet werden. Mit dieser Darstellung sind die erwähnten Änderungen der Klassenhierarchie trivial.

Ein Kookkurrenzbeschränkungs-Phänomen muss nun in jede der drei Hierarchien eingeordnet werden. Zu den Subklassen, zu denen die spezielle Vokalharmonie des Türkischen gehört, können wir auf folgenden Pfaden navigieren:

- `cooccurrence restriction` – `assimilation` – `partial`
- `cooccurrence restriction` – `on vowels`
- `cooccurrence restriction` – `grammatical` – `outward`

Diese drei Pfade können wir in einem Baum zusammenfassen, den wir als Term der Prädikatenlogik erster Stufe ausdrücken:

```
cooccurrence_restriction(assimilation(partial),
                         vowels,
                         grammatical(outward))
```

Intuitiv gilt: Superklassen und Subklassen entsprechen Funktionssymbolen und ihren Argumenten, während orthogonale Klassifikationen durch mehrere Argumente eines Funktionssymbols ausgedrückt werden. Es ist möglich, die Klassenhierarchien aus Abbildung 3 so in eine Typisierung der Logiksprache zu transformieren, dass die sinnvollen Terme gerade die korrekt typisierten Terme sind.

Die Termdarstellung bietet weitere Vorteile:

- Superklassen können als Terme mit Variablen dargestellt werden. Beispielsweise werden alle Vokalharmoniesysteme durch

  ```
  cooccurrence_restriction(assimilation(X),
                           vowels,
                           grammatical(Y))
  ```

  dargestellt. Die Termdarstellung einer Unterklasse erhält man hieraus durch Einsetzen von Termen an Variablenpositionen.

  Ein variablenhaltiger Term wird typischerweise für Anfragen an die Wissensbasis verwendet um Informationen über alle den Instanzen des Terms entsprechenden Subklassen zu erhalten.

- Häufig will man bei Assimilationen nicht nur danach unterscheiden, ob sie total oder partiell sind, sondern auch danach, ob der Öffnungsgrad, die Zungenstellung oder die Lippenhaltung betroffen ist. Dies kann durch Hinzufügen weiterer Argumente zum Funktionssymbol `assimilation` geschehen. Bei der türkischen Vokalharmonie sind Zunge und Lippen betroffen:

  ```
  cooccurrence_restriction(assimilation(partial,
                                        no, yes, yes),
                           vowels,
                           grammatical(outward))
  ```

- In der bisherigen Termdarstellung entsprechen zueinander orthogonale Hierarchien verschiedenen Argument*positionen* eines Funktionssymbols. Diese Zuordnung ist nicht intuitiv, wie gerade das letzte Beispiel zeigt. Daher gehen wir zu „Feature-Termen" (z.B. [BS95]) über, in denen Argumente nicht durch ihre Position, sondern durch einen *Namen* identifiziert werden. Die Verwendung von Feature-Termen bietet sich auch deshalb an, weil sie in der Linguistik weit verbreitet und damit vielen potentiellen Benutzern bereits bekannt sind. Ein weiterer Vorteil von Feature-Termen ist, dass Superklassen noch einfacher dargestellt werden können, nämlich als Terme mit fehlenden Features (Argumenten).

Die Datenbank besteht nun im Wesentlichen aus einer einzigen Relation mit dem Schema

(Sprache, Phänomen-Klasse, Phänomen-Beschreibung),

wobei die Phänomen-Klasse ein (Feature-)Term ist. Die Phänomen-Beschreibung kann, wie bereits gesagt, formal oder ein natürlichsprachlicher Text sein.

Benutzer, die in die Datenbank neues Wissen eintragen, können bei der Phänomen-Klassifikation anhand des Typsystems (d.h. anhand der codierten Klassenhierarchie(n)) maschinell geführt werden. Sie werden dabei gezwungen, die durch das Typsystem repräsentierte Terminologie zu verwenden. An jeder Position in einem Term müssen sie dabei nur zwischen wenigen Alternativen unterscheiden.

# 4 Offene Probleme

Probleme, für die noch keine endgültige Lösung gefunden wurde, sind unter anderem:

**Flexibilität:** Da nicht alle möglichen sprachlichen Phänomene von Anfang an bedacht werden können, ist eine gewisse Flexibilität in der Klassenhierarchie nötig. Andererseits muss verhindert werden, dass diese Flexibilität zu einem „Wildwuchs“ bei der Pflege der Phänomen-Klassifikation führt. Bisher sehen wir keine technische Lösung für dieses Problem, sondern nur eine admistrative: Die Klassenhierarchie darf nur von einem besonders privilegierten Benutzer, dem Datenbank-Verwalter, verändert werden. Wenn ein Sprachbeschreiber meint, in einer Sprache eine neuartige Klasse von Phänomenen gefunden zu haben, dann muss er den Datenbankverwalter davon überzeugen und ihn um eine entsprechende Erweiterung der Klassenhierarchie bitten.

Vermutlich müssen die Klassenhierarchien bisweilen auch auf Grund neuer linguistischer Erkenntnisse oder Sichtweisen umgestellt werden. Dabei müssen auch alte Klassifikationsterme in neue transformiert werden. Dafür dürften sich Logikprogramme unmittelbar anbieten. Eine Methodologie für die Erstellung dieser Programme fehlt jedoch.

**Negatives Wissen:** Soll das Fehlen eines Eintrags zu einem Paar Sprache–Phänomen bedeuten, dass das jeweilige Phänomen in der Sprache nicht auftritt („closed-world assumption“), oder dürfen wir das nicht annehmen, da die Datenbank unvollständig sein kann („open-world assumption“)? Da wir realistischerweise nicht von einer vollständigen Datenbank ausgehen können, also mit der open-world assumption arbeiten müssen, müssen wir explizit ausdrücken, dass ein Phänomen in einer Sprache *nicht* auftritt. Dies führt jedoch bei einer naiven Vorgehensweise zu einer immensen Vergrößerung der Datenbank.

Hier hilft wiederum die hierarchische Struktur unserer Klassifikation: Wenn es etwa in einer Sprache keine grammatisch induzierten Kookkurrenzbeschränkungen gibt, dann braucht auch nicht explizit vermerkt zu werden, dass sie keine Vokalharmonie aufweist, denn letztere ist als Spezialfall von ersterer definiert.

# 5 Schlussbemerkungen

Die in das vorliegende Papier eingeflossenen linguistischen Überlegungen stammen aus einem DFG-Projekt mit der Bezeichnung „Allgemein-vergleichende Grammatik 2.0“, das unter der Leitung des zweiten Autors am Lehrstuhl für Theoretische Linguistik des Instituts für Deutsche Philologie der LMU läuft und das die Entwicklung eines universellen Rahmens für die Beschreibung von Sprachen in einer flexiblen Datenbank zum Ziel hat. Nachdem in einer ersten Phase mit Hypertext-Applikationen experimentiert worden war, werden jetzt in Zusammenarbeit mit Informatikern die Möglichkeiten der Datenbanktechnologie im engeren Sinn exploriert.

Derzeit läuft am Institut für Informatik eine Diplomarbeit zu dem in diesem Papier beschriebenen Thema. Sie soll Mitte November 1996 abgeschlossen sein.

In dieser Arbeit werden unter anderem einige der oben angegebenen Techniken untersucht. Einige Algorithmen, die mit Klassifikationstermen arbeiten, wurden prototypisch implementiert.

Wir denken, dass die Technik der Klassifikation durch Terme auch in anderen Gebieten eingesetzt werden kann, und hoffen, beim Tag der Informatik dafür mögliche weitere Anwender zu finden. In der anderen Richtung hoffen wir auch auf Anregungen für das Projekt der linguistischen Wissensbank. Weitere Arbeitsgebiete in diesem Projekt sind:

**Benutzer-Schnittstelle:** Die zu entwickelnde Datenbank soll im Endausbau von der gesamten linguistischen Fachwelt verwendet werden. Dies führt zu den folgenden beiden Anforderungen an die Benutzer-Schnittstelle:

- Der Zugriff auf die Datenbank muß über das Internet erfolgen können, idealerweise über das World-Wide Web.
- Die Benutzung muß so einfach wie möglich sein, da eine intensive Einarbeitung vieler Benutzer nicht geleistet werden kann.

**Information-Retrieval:** Da das in der Datenbank gespeicherte Wissen nicht vollständig formalisiert ist, ist zu überlegen, ob zusätzlich zu der oben beschriebenen Anfrage-Methode über die Phänomen-Klassen auch Information-Retrieval-Methoden zur Verfügung gestellt werden sollen.

**Formale Beschreibung der Phänomene:** In diesem Papier wurde beschrieben, wie sprachliche Phänomene *formal klassifiziert* werden können. Wir sind jedoch andererseits davon ausgegangen, dass die Phänomene *informell beschrieben* werden. Natürlich ist es für die Linguistik auch interessant, die Beschreibung der Phänomene zu formalisieren. Dabei kann vermutlich vielfach auf Erkenntnisse der Computerlinguistik zurückgegriffen werden.

Ermutigend erscheint uns in diesem Zusammenhang, daß viele klassifikatorische linguistische Termini Definitionen in Form von Regelschemata nahelegen. So läßt sich z.B. Vokalharmonie definieren als Gültigkeit einer Regel, die die Übereinstimmung bestimmter Merkmale von Stammvokal und Affixvokal verlangt. Auf diese Weise ließen sich aus den Phänomenbezeichnungen Schemata für Phänomenbeschreibungen generieren, in die dann nur noch die jeweiligen Spezifika eingetragen werden müssten.

# Literatur

[BS95] Rolf Backofen and Gert Smolka. A complete and recursive feature theory. *Theoretical Computer Science*, 146:243–268, 1995.

[CS77] Bernard Comrie and Norval Smith. Lingua descriptive studies: Questionnaire. *Lingua*, 42:1–72, 1977.