

Wortbedeutungsdisambiguierung im Kontext des
Sanskrit

- Magisterarbeit -

vorgelegt von

Jonas Soiné

Seminar für Indologie

der Eberhard Karls Universität Tübingen

Lehrstuhlinhaber : Prof. Dr. Klaus Butzenberger

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Betreuer : Dr. Oliver Hellwig

Priv. Doz. an der
Ruprecht-Karls-Universität Heidelberg

August 2011

Inhaltsverzeichnis

1	Einleitung	1
2	Die Entstehung formaler Grammatiken	2
2.1	Sprache als System	2
2.2	Generative Grammatik	3
3	Ambiguitäten	6
3.1	Exkurs: Logischer Fehlschluss durch Ambiguität	7
3.2	Lexikalische Ambiguität	8
3.2.1	Der Wahrheitsgehalt von Sätzen	8
3.2.2	Semantische Implikation und Hyponymie	9
3.2.3	Homonymie und Polysemie	10
3.3	Strukturelle Ambiguität	12
3.4	Skopus-Ambiguität	13
4	Bedeutungswandel im Sanskrit	14
4.1	Ursachen des Bedeutungswandels	14
4.1.1	Extra-linguistische Ursachen	15
4.1.2	Linguistische Ursachen	16
4.2	Tendenzen und Klassifizierungen	17
4.2.1	Metapher	19
4.2.2	Metonymie	20
4.2.3	Volksetymologie	21
4.2.4	Ellipse	22
4.3	Computational Sanskrit	23
5	Wortbedeutungsdisambiguierung	25
5.1	Wissensbasierte Methoden	26
5.1.1	Der Lesk-Algorithmus	26
5.1.2	Semantische Similarität	27
5.1.3	Lokale und globale Kontexte	28
5.1.4	Selektionale Präferenzen und Heuristik-Methoden	29
5.2	Überwachte Methoden	30
5.2.1	Probabilistische Methoden	32
5.2.2	Methoden, die auf der Similarität der Beispiele beruhen	33
5.2.3	Methoden, die auf Regel-Kombinationen beruhen	33
5.2.4	Kernel-basierte Verfahren	34
5.2.5	Empirische Auswertung von NB,kNN,DI,AB,SVM auf dem DSO-Corpus	35
5.3	Un- und semi-überwachte Methoden	36
5.4	Evaluation von WSD-Systemen	37

6	Der SanskritSemAnnotator	40
6.1	Tabellen-Schemata	40
6.2	Die Benutzer-Schnittstelle	43
7	Auswertung der Annotationen mit SanSemAn	44
7.1	Inter-Annotator-Übereinstimmung	45
7.2	Konsistenz der Annotationen	49
7.2.1	Bedeutungen von śārdūla	49
7.2.2	Bedeutungen von jana	51
7.3	Probleme aufgrund der Granularität	57
7.4	Cohens κ	59
8	Die WSD durch DL	61
8.1	Das finale Trainings-Corpus	61
8.2	Der DL-Algorithmus nach Yarowsky	62
8.3	Evaluierung der WSD durch DL	71
8.4	Ausblick	72
	Glossar	73
	Sanskrit Abkürzungen	77
	Stellenverzeichnis	78
	Abbildungsverzeichnis	79
	Code-Schnipsel	79
	Tabellenverzeichnis	79
	Literatur	80

1 Einleitung

Schon die alten Gelehrten Indiens haben das Phänomen der Mehrdeutigkeit und die damit verbundenen Probleme, sowie die daraus entstehenden stilistischen Möglichkeiten für die Poetik gekannt. Auch in der modernen Wissenschaft spielt Mehrdeutigkeit eine entscheidende Rolle. In Zeiten der maschinellen Übersetzung oder der Kommunikation zwischen Mensch und Maschine mit Hilfe von allen zur Verfügung stehenden technischen Errungenschaften wie visuelle Bildverarbeitung und Mustererkennung¹, rückt die damit verbundene Problematik entschieden in den Vordergrund und beschäftigt eine Vielzahl von Wissenschaftlern und akademischen, sowie wirtschaftlichen Projekten.

In dieser Arbeit wird ein kleiner Teil der Künstlichen Intelligenz (KI), die automatische *Wordbedeutungsdisambiguierung* bzw. WSD, behandelt und in Zusammenhang mit der Verarbeitung von digitalen Sanskrit-Texten gebracht².

In diesem Zuge wird zuerst die notwendige linguistische Begrifflichkeit geklärt und etwas linguistisches Hintergrundwissen eingebracht – z. B. der Ansatz, Sprache als System aufzufassen oder die Entwicklung erster formaler Grammatiken, sowie die unterschiedlichen Erscheinungsformen der Ambiguität.

Der nächste Teil widmet sich dem Problem des *semantischen Wandels*³ im Sanskrit und ein kleiner Überblick über die bisher geleistete (doch unpopuläre) Arbeit im Bereich der maschinellen Verarbeitung des Sanskrit wird skizziert.

Der letzte Teil wird der technologische Aspekt der *Disambiguierung*, genauer der *Wordbedeutungsdisambiguierung* sein. Es werden die Methoden der Computer Linguistik (CL) und KI erläutert, mit welchen die Aufgabe der Wortbedeutungsdisambiguierung⁴ vollbracht wird. Die Darstellung der Web-Anwendung *SanSemAn*, eines Multi-Annotatoren-Systems zur Schaffung eines bedeutungsannotierten Test-Corpus, leitet den experimentellen bzw. forschenden Teil dieser Arbeit ein, in welchem ein System zur Disambiguierung einer Menge von zwei Sanskrit-Begriffen (*śārdūla* und *jana*) entwickelt und evaluiert

¹Stichworte hier seien z. B. computergestützte Mimik- und Gestikererkennung oder Lippenlesen.

²Die in der KI unlängst hervorgegangenen und in Fußnote 1 erwähnten Gebiete sind im Feld des Sanskrit jedoch nur von beschränktem Nutzen, weil deren Anwendung auf das Sanskrit noch in sehr weiter Ferne liegt, nicht zuletzt, da eine größere, statistisch verwertbare Menge an Sanskrit-Sprechern schwer (wenn überhaupt) an einem Ort zu vereinen ist.

³Auch Bedeutungswandel; engl. *semantic change*.

In der *Onomasiologie* bezeichnet der Bedeutungswandel alle Veränderungen, bei denen sich mit ein und demselben Wortkörper eine neue Bedeutung verbindet oder bei denen Wortbedeutungen ihren Anwendungsbereich verändern. Vgl. LEWANDOWSKI [41, S. 158–159].

⁴Im Folgenden werden Wortbedeutungsdisambiguierung und WSD synonym verwendet.

wird. Die Textstellen die hierzu verwendet werden, stammen aus dem DCS, im Speziellen aus dem Teil des DCS, welcher durch das Epos des Rāmāyaṇa konstituiert wird.

2 Die Entstehung formaler Grammatiken

2.1 Sprache als System

Heute wird von nahezu allen linguistischen Schulen, wenngleich mit unterschiedlichen Akzentuierungen, die Auffassung vertreten, Sprache nicht bloß als amorphe Akkumulation von Äußerungen anzusehen, sondern als zusammenhängendes System. Diese Auffassung ist im Rahmen einer deutlichen Gegenstandsbestimmung der Linguistik von F. de Saussure in seinem Cours formuliert worden⁵. Im Anschluss daran wurde der System-Begriff der Sprache näher an die in der Mathematik und Kybernetik verwendeten Termini System und Struktur gerückt. Zweifellos erhoffte sich Saussure eine den Naturwissenschaften vergleichbare Wissenschaftlichkeit in der Sprachwissenschaft⁶. Sein Ausgangspunkt ist die Kritik an der junggrammatischen Schule. So sind die wohl bekanntesten begrifflichen Differenzierungen die Paare *langue/parole* oder *Synchronie/Diachronie*. In der bekannten Ausgangsfrage im Cours über den Gegenstand der Sprachwissenschaft, wird der „Gesichtspunkt, der das Objekt erschafft“⁷ thematisiert. Der Erfassung eines auf der Basis der Evidenz zustande gekommenen Gegenstandes, bedarf es entgegen dem „sensualistischen Positivismus der junggrammatischen Schule“⁸, welche auf induktivem Wege zur Erkenntnis des Gegenstandes zu gelangen versuchte, einer leitenden theoretischen Überlegung⁹.

Saussure hat in seinem Streben, den Gegenstand der Sprachwissenschaft zu differenzieren, einige miteinander schwer vereinbare Parameter ins Spiel gebracht, die sich jedoch späterhin als gleichermaßen fruchtbar erwiesen. Selbst Begriffe wie *langue* und *parole* wurden nicht deutlich eingeführt. Offenbar konkurrieren im Cours die Bestimmungen der *langue* als „soziales Ergebnis“ mit der im Individuum verankerten psychischen Gegebenheit und dem System von Oppositionen. Hier liegen scheinbar im Falle der *langue* mehrere Bedeutungen vor.

Dem System, welches zwar im sprechenden Individuum verinnerlicht ist, sich seinem

⁵Wenn im Nachfolgenden von Saussure die Rede ist, so ist dies eine *façon de parler*: gemeint ist der *Premier cours de linguistique générale* (1907). Zitate Vgl. SAUSSURE [59].

⁶Vgl. WOLSKI [65, S. 7].

⁷Vgl. SAUSSURE [59].

⁸Vgl. JÄGER [34, S. 20].

⁹Vgl. WOLSKI [65, S. 7].

direkten Zugriff aber entzieht, liegt offenbar ein wechselseitiges Voraussetzungsverhältnis zwischen individuellem und interindividuellem Existenzmodus der Sprache zugrunde. Die Vorstellungen

- des Sprachsystems als Zeichenvorrat, dessen Elemente in Opposition zueinander stehen,
- der Trennung des synchronischen und diachronischen Aspektes,
- des sozial verbindlichen Charakters der *langue*

haben zu einer statischen Rezeption des Systems geführt. Im Rahmen einer sprachzeichentheoretisch fundierten Semantik, sind Unstimmigkeiten des Cours, in der für den Strukturalismus bestimmenden Unterscheidung von Sprachsystem (*langue*) und dessen Realisierung (*parole*) bereinigt worden. So gab der Cours Anlass für eine über ihn weit hinausgehende Reflexion über die Bereichsgliederung der Sprache, welche dem logischen Positivismus die Möglichkeit gab, sich auf die Linguistik zu entfalten.

2.2 Generative Grammatik

Die Entwicklung der Automatentheorie, insbesondere Turing-Maschinen und [Semi-Thue-Systeme](#) lieferten Noam Chomsky die Voraussetzungen, den Grammatik-Begriff in Form der *generativen Grammatik*¹⁰ neu zu definieren. Die Grammatik versteht sich nach Chomsky¹¹ nicht mehr als eine Theorie zur Beschreibung beobachtbarer sprachlicher Daten, sondern vielmehr als Beschreibung von unbeobachtbaren grammatischen Fakten. Als prominentes Beispiel gibt er eine generative Grammatik für den englischen Satz 'the man hit the man' mit den nachfolgenden Ersetzungsregeln an:

$S \rightarrow NP + VP$	Erzeugbare Sätze:
$NP \rightarrow T + N$	
$VP \rightarrow \text{Verb} + NP$	the man hit the man
$T \rightarrow \text{the}$	the ball hit the man
$N \rightarrow \text{man, ball, etc.}$	the man hit the ball
$V \rightarrow \text{hit, etc.}$	the ball hit the ball

In seiner ersten Grammatiktheorie hat Chomsky¹² an die Beschreibungen der [PSG](#)

¹⁰Vgl. CHOMSKY [15, S. 19]. Siehe auch [gTG](#) und [PSG](#).

¹¹Vgl. CHOMSKY [13].

¹²Vgl. CHOMSKY [13].

angeschlossen und sie in einen rekursiven Regelapparat umgedeutet¹³. Inhaltliche Beziehungen zwischen Sätzen, Satzteilen und Wörtern wurden jedoch zunächst außer Acht gelassen.

„What we are suggesting is that the notion of 'understanding a sentence' be explained in part in terms of the notion of 'linguistic level'. To understand a sentence, then, it is first necessary to reconstruct its analysis on each linguistic level.“¹⁴

Technisch kann die gTG als aus einem **Semi-Thue-System** hergeleitet angesehen werden. Sie erzeugt bzw. deduziert eine rekursiv aufzählbare Menge von Ausdrücken und repräsentiert die den Sprechern zugeschriebene „kreative“ Fähigkeit, von „endlichen Mitteln unendlichen Gebrauch“¹⁵ machen zu können. Die gTG ist somit ein algorithmisches Entscheidungsverfahren, um zu ermitteln, welche Sätze Teilmengen der Kalkülsprache sind. Die Menge der zugrunde gelegten Sätze (Konkationen von Worten aus Σ^*) wird in zwei disjunkte Teilmengen zerlegt, eben genau die Teilmenge, welche sich im Kalkül ableiten lässt, also die grammatischen Sätze und genau die, welche sich mit der jeweiligen Grammatik nicht ableiten lässt, also die ungrammatischen Sätze.

¹⁶Der Begriff der *Sprachebene*¹⁷ unterscheidet mindestens zwei verschiedene Ebenen. Während Sequenzen wie (1a) und (1b) die Notwendigkeit einer 'Morphem-Ebene' zeigen,

(1a) the sun's rays meet

(1b) the sons raise meet

zeigen Sequenzen wie (2) die Notwendigkeit der Ebene der 'hierarchischen bzw. Phrasenstruktur':

(2) the old man and women

(3) the shooting of the hunters

Die Mehrdeutigkeit einer Sequenz kommt nach WELLS [64] nur dann zu tragen, wenn die **IC-Analyse** der entsprechenden Sequenzen problematisch ist. Nach diesem Prinzip kann die Mehrdeutigkeit von (3) oder wie sie in zweimorphemigen Sequenzen vorkommen

¹³Vgl. WOLSKI [65, S. 13].

¹⁴Vgl. CHOMSKY [13, S. 87].

¹⁵Vgl. CHOMSKY [15, S. 19].

¹⁶Die nachfolgenden Beispiele sind im Wesentli-

chen FRIES [20] entnommen.

¹⁷Man vergleiche mit dem Zitat von Chomsky auf Seite 4, wo er von „linguistic level“ spricht.

kann, nicht dargestellt werden. Sequenzen wie (3) haben grundsätzlich zwei unterschiedliche Lesearten, die sich formal in Sequenzen wie (4) und (5) wiederfinden und sollten deshalb in die syntaktische Analyse miteinbezogen werden.

(4) the growling of lions

(5) the raising of flowers

Weil sich die unterschiedlichen Strukturen von (4) und (5) beide in (3) auffinden lassen, sollte die Mehrdeutigkeit von (3) als 'syntaktische' **Ambiguität** beschrieben werden, somit rechtfertigen Sequenzen wie (3),(4),(5) die abstrakte Ebene der Transformationen. (3) liegen also zwei unterschiedliche Kernsätze zugrunde, die mit entsprechenden Transformationen in (3) überführt werden können. Diese beiden Kernsätze liegen auch formal-syntaktisch Sequenzen wie (4) und (5) zugrunde.

Die Standard-Theorie ist die Erweiterung der frühen Modelle der **gTG**, den *Syntactic Structures*, um eine semantische Komponente. Im Kontext von Mehrdeutigkeiten heißt das, dass zur 'syntaktischen **Ambiguität**' der Begriff der 'semantischen Ambiguität' hinzukommt, welcher versucht, die betreffende Mehrdeutigkeit allein in der semantischen Komponente der Grammatik zu beschreiben. Wird die Bedeutung eines Satzes nur von seinen lexikalischen Einheiten und deren syntaktischen Beziehungen bestimmt, bedeutet 'semantische Ambiguität' also, dass die jeweilige Mehrdeutigkeit im Lexikon zu repräsentieren ist.

Es ergibt sich nun eine Gliederung von Mehrdeutigkeiten in drei Klassen¹⁸: Solche, die von der syntaktischen Komponente erfasst werden, solche die in der semantischen Komponente, dem Lexikon, beschrieben werden und Mehrdeutigkeiten, die außerhalb davon liegen, weil sie auf Faktoren beruhen, die der Performanz zuzuschreiben sind, das sind 'Vagheiten' in der betreffenden Bedeutungshinsicht. Der damit verbundene Anspruch an gegenseitiger Abgrenzbarkeit der genannten Klassen, führt jedoch zu weiteren Problemen, wie der Tatsache, dass nicht über ein absolutes Unterscheidungsmerkmal zwischen außer- und innergrammatischen Phänomenen verfügt wird. Gemeint sind z. B. Fälle, deren syntaktische Wohlgeformtheit nicht eindeutig bestimmbar ist oder Mehrdeutigkeiten lexikalischer Einheiten und deren Ketten. Diese Zuordnungsschwierigkeiten haben unter Anderem zu den Weiterentwicklungen der Standard-Theorie (Erweiterte Standard-Theorien¹⁹) und den verschiedenen Ausprägungen der **GS** geführt. In dieser

¹⁸Vgl. FRIES [20, S. 24].

und **EST III**.

¹⁹Erweiterte Standard-Theorien: **EST I**, **EST II**

Hinsicht besonders problematisch sind Sätze mit Quantifikatoren, manche Sätze mit Negationen oder pronominalisierte Sätze:

- (6) Hans sagt, er habe mit der Frau geschlafen, die aus Schweden kommt.

In (6) kann sich die Sequenz *die aus Schweden kommt* sowohl auf die Aussage von Hans beziehen, als auch auf die des Sprechers. Zur grammatikalischen Erfassung dieser Mehrdeutigkeit müsste ein 'Sprecher' in die grammatische Analyse miteinbezogen werden oder abstrakte, zugrunde liegende 'logische' Strukturen müssten postuliert werden. Wird jedoch die **Ambiguität** von (6) aus der grammatischen Beschreibung ausgeschieden, wird damit die Forderung der **Standard-Theorie** verletzt, die Kompetenz des idealen Sprecher/Hörers abzubilden und ambigen Sätzen verschiedene Interpretationen zuzuordnen. Ein weiteres Beispiel, wo eine Mehrdeutigkeit auf tiefenstruktureller Ebene disambiguiert werden müsste ist (7):

- (7) Zwei Hähne befruchteten zwanzig Hühner.

In (7) ist nicht klar zu erkennen, wieviele Hühner tatsächlich befruchtet wurden. Mögliche Interpretationen sind, daß zwanzig oder vierzig Hühner befruchtet wurden²⁰. Da kaum denkbar ist, dass diese semantischen Verhältnisse in syntaktischen Tiefenstrukturen ohne logische Quantoren, repräsentiert werden können, erklären sich Entwicklungen der Standard-Theorie zur **GS**.

3 Ambiguitäten

Der Begriff der Ambiguität bezeichnet eine Form der Unbestimmtheit der Bedeutung von sprachlichen Zeichen, bzw. von Sequenzen²¹ von sprachlichen Zeichen. Das Phänomen, dass manche Sätze außerhalb des Gebrauchskontextes mehrdeutig sind, kann in allen natürlichen Sprachen beobachtet werden.

²⁰In der zweiten Leseart wäre auch möglich, dass weit mehr als vierzig Hühner befruchtet wurden, nimmt man an, die Hähne haben verschiedene Hühner mehrfach befruchtet. Vgl. hierzu auch die Fußnoten 19–20 in FRIES [20, S. 27].

²¹Beispielsweise Phrasen oder Sätze.

3.1 Exkurs: Logischer Fehlschluss durch Ambiguität

In der philosophischen Logik kann eine sprachliche Ambiguität zu einem logischen Fehlschluss führen, wenn eine Form der linguistischen Ambiguität die logische Form eines Argumentes als gültig erscheinen lässt, obwohl sie es nicht ist. Die folgende Argumentation könnte man im Kalkül des natürlichen Schließens nach Gentzen²² mit einem Beweisbaum²³ darstellen, dessen Hypothesenmenge $Hyp(D) = \{\varphi \rightarrow \psi, \psi \rightarrow \perp\}$ ist:

President Clinton should have been impeached only if he had sexual relations with Monica Lewinsky.

He did not have sexual relations with Lewinsky.

Therefore, he should not have been impeached.

$$\frac{\frac{[\varphi]^{(1)} \quad \varphi \rightarrow \psi}{\psi} \rightarrow E \quad \psi \rightarrow \perp}{\frac{\perp}{\varphi \rightarrow \perp} (\rightarrow I: 1)} \rightarrow E$$

- (8) sexual relations
- (8a) eine sexuelle Beziehung
- (8b) Geschlechtsverkehr

Es gilt also $Hyp(D) \vdash \neg\varphi$, jedoch nur dann, wenn eine eindeutige Verwendung von (8) vorausgesetzt wird. Aufgrund der sprachlichen Ambiguität von (8) mit den beiden Interpretationen (8a) und (8b), müssen (8a) und (8b) auch zwei verschiedene Variablen zugewiesen werden. Daraus ergibt sich $Hyp(D) = \{\varphi \rightarrow \psi, \sigma \rightarrow \perp\}$ und eine Anwendung der Beweisfigur *modus tollendo tollens* ist nicht möglich.

Beweis:

Man wähle eine Belegung v mit $\llbracket\varphi\rrbracket_v = 1$, $\llbracket\psi\rrbracket_v = 1$ und $\llbracket\sigma\rrbracket_v = 0$ daraus folgt $\varphi \rightarrow \psi$, $\sigma \rightarrow \perp \not\vdash \varphi \rightarrow \perp$, denn $\llbracket\varphi \rightarrow \perp\rrbracket_v = 0$. Der Korrektheitssatz²⁴ besagt aber, dass wenn $\Gamma \vdash \varphi$ gilt, dann gilt auch $\Gamma \vDash \varphi$. Und damit $\{\varphi \rightarrow \psi, \sigma \rightarrow \perp\} \not\vdash \varphi \rightarrow \perp$.

²²Vgl. GENTZEN [23] und GENTZEN [24].

²³Die Schlussfigur des hier gezeigten Baumes ist natürlich eine Version von *modus tollendo tollens*. Die Schlussregel $\rightarrow E$ ist die Schlussfigur *modus ponendo ponens*.

²⁴Für den Beweis des Korrektheitssatzes sei auf das Theorem 7.2 im Skript zur Mathematischen Logik von Herrn Prof. Dr. Schroeder-Heister der Abteilung Logik und Sprachtheorie des Wilhelm-Schickard-Instituts für Informatik der Universität Tübingen verwiesen.

3.2 Lexikalische Ambiguität

Worte wie 'Note', 'Schloss', oder 'Bank' sind semantisch mehrdeutig. Mit dem englische Wort 'note' werden wie im Deutschen verschiedene Bedeutungen assoziiert, zusätzlich ist sogar seine lexikalische Klasse ambig, es kann sowohl Verb, als auch Substantiv sein. Diese auch **lexikalische Ambiguität** genannte Form der Mehrdeutigkeit, erfordert im Kontext der Sprachwissenschaft eine genauere Analyse, denn sie umfasst mindestens zwei weitere in der Sprachwissenschaft und Philosophie oft diskutierte Begriffe: **Homonymie** und **Polysemie**. Zu deren Erfassung sind jedoch einige Bemerkungen zur Semantik notwendig.

3.2.1 Der Wahrheitsgehalt von Sätzen

Mit den Begriffen Objekt- und Metasprache erfordert die modelltheoretischen Semantik²⁵ eine Trennung der Sprache, die zur Wahrheitsbeschreibung eines Satzes dienlich sein soll: ein objektsprachlicher und deshalb diakritisch gekennzeichnete Satz 'p' unterscheidet sich von seiner metasprachlichen Übersetzung p. Tarski beginnt mit einem konkreten Beispiel:

The sentence 'snow is white' is true if, and only if, snow is white.

Aus diesem Satz leitet er eine verallgemeinerte, formale Prozedur ab, Äquivalenzen dieser Art darzustellen. Ein willkürlich gewählter Satz wird durch den Buchstaben 'p' ersetzt. Der Name des Satzes wird gebildet und durch den Buchstaben 'X' ersetzt. Somit beschreibt er dann das logische Verhältnis der Sätze 'X is true' und 'p' als eine Äquivalenz der Form (T):

(T) X is true if, and only if, p.

Hiermit betrachtet er die Verwendung und Definition von 'true' bzw. 'wahr' als adäquat aus materieller Sicht, wenn 'true' bzw. 'wahr' in solcher Weise benutzt wird, dass alle Äquivalenzen der Form (T) gelten und er spricht von einer *adäquaten Definition von Wahrheit*²⁶, wenn all diese Äquivalenzen aus ihr (der Definition) folgen.

²⁵Vgl. TARSKI [61].

²⁶In TARSKI [61, Kap. I.4] wird der Ausdruck „*material adequacy of the definition*“ (*of truth*) genannt.

3.2.2 Semantische Implikation und Hyponymie

In diesem Abschnitt bezieht sich *wahr* immer auf die Definitionen im Abschnitt 3.2.1.

(9a) Max managed to finish *Infinite Jest*.

(9b) Max finished *Infinite Jest*.

Angenommen (9a) ist *wahr*. Dann ist (9b) auch *wahr*. Es gibt keinen möglichen Zustand in dem (9a) *wahr* und (9b) *falsch* ist. In diesem Fall sagt man (9a) impliziert semantisch (9b) und es ergibt sich eine generelle Definition von semantischer Implikation²⁷:

(10) Ein Satz (S_1) impliziert einen Satz (S_2) genau dann,
wenn gilt: wann immer S_1 *wahr* ist, dann ist auch S_2 *wahr*.

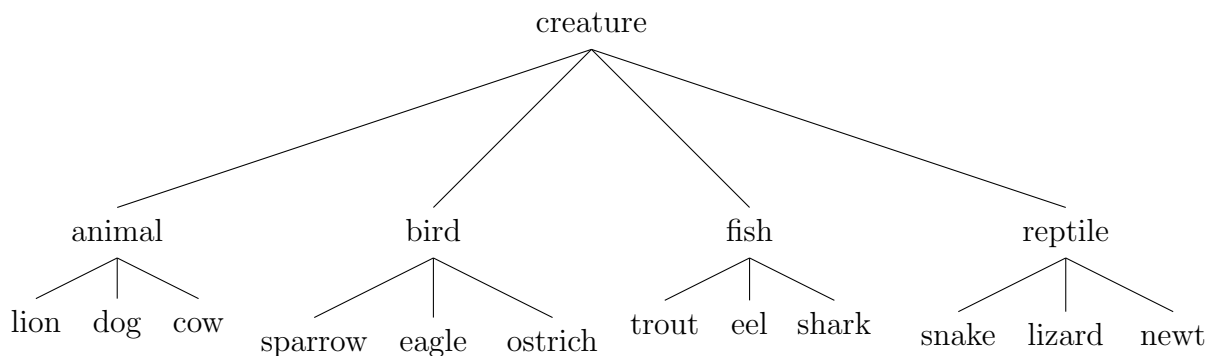
Mit Hilfe der semantischen Implikation zwischen (11a) und (11b), sowie zwischen²⁸ (11b) und (11c) und des generellen Schemas (12) kann man nun das Gerüst der **Hyponymie** konstruieren:

(11a) The thing in the cage is a lion.

(11b) The Thing in the Cage is an animal.

(12) 'X is a lion' impliziert semantisch 'X is an animal'.

Im nachfolgenden Baum kann man für diesen Ausschnitt des Englischen die mehrschichtige Taxonomie erkennen, welche durch die semantische Relation der Hyponymie in Form von Hyponymen, Ko-Hyponymen und Hyperonymen definiert ist:



²⁷Vgl. RADFORD [52, S. 171].

²⁸Und aufgrund der Transitivität natürlich auch zwischen (11a) und (11c).

3.2.3 Homonymie und Polysemie

In der griechischen Antike beschreibt Aristoteles den Begriff der 'Homonymie' in den *Kategorien*²⁹ mit dem 'Wesen der Dinge', mit ihrer Unterteilung in Gattungen, ihren spezifischen Unterschieden usw., diese Verwendung von 'Homonym' zielt darauf ab, dass das, worauf dieses Wort in der außersprachlichen Welt verweist, 'wesensmäßig' differenziert werden kann. Aus linguistischer Sicht ist ein Wort nicht deshalb als lexikalisch ambig anzusehen, weil seine Denotate in Ober- und Unterklassen aufgeteilt werden können³⁰. Das Wort 'Mensch' ist nicht homonym, weil es weiße, braune oder gelbe Menschen gibt, obwohl die Gattung 'Mensch' in verschiedene 'Arten' aufgeteilt werden kann. Die 'Gattungsbedeutung'³¹ von Wörtern kann zwar die Ursache für Homonymie sein, bestimmt aber deren Klassifikation als Homonyme nicht zwangsläufig; vielmehr ist die Unterteilung von Gattungsbegriffen in Artbegriffe generell „unspezifiziert“³². Die 'Gattungsbedeutung' von Wörtern erfordert eine Fallunterscheidung: Einerseits sind z. B. Wörter wie 'Person' für Bedeutungsmerkmale wie <weiblich>, <männlich> nicht spezifiziert. 'Homonym' bedeutet demgegenüber, dass ein Wort bei identischer Wortform im Lexikon gleichzeitig für unterschiedliche Bedeutungsmerkmale spezifiziert wird. Andererseits meint 'Gattungsbedeutung' semantische Hierarchien zwischen lexikalischen Einträgen, **Hyponymien**. Wegen diesen verschiedenen Auffassungen von 'Gattungsbedeutung' sollte hier eine Abgrenzung zwischen 'Referenz-' und 'Inhaltssemantik' getroffen werden, auf welcher viele Verwechslungen zwischen Linguisten und Philosophen beruhen.

Im Rahmen der Referenzsemantik entscheidet darüber, ob ein Wort Γ homonym ist, die Tatsache ob sich die Denotate im außersprachlichen Bereich in zwei Klassen $\{\alpha, \beta\}$ mit den Eigenschaften $\{A, B\}$ aufspalten lässt. Auch der Vagheits-Begriff wird dadurch definiert, dass Γ dann 'vage' ist, wenn Unklarheit herrscht, ob Γ auf ein bestimmtes Objekt in der außersprachlichen Welt anwendbar ist. Im Gegensatz dazu, ist der Vagheits-Begriff der Inhaltssemantik absolut, d. h. Γ ist bezüglich eines Objektes genauso vage wie bezüglich eines anderen Objektes; die Unbestimmtheit von Γ hängt nicht von der außersprachlichen Objektwelt ab.

Diese Vorgehensweise der lexikalischen Semantik ist natürlich auch mit diversen Schwierigkeiten verbunden: es werden Bedeutungen sprachlicher Zeichen auf Bedeutungen an-

²⁹Vgl. ROLFES [58, Kap. I].

wendeten 'Generality'.

³⁰Beachte die Darstellungen in Übertschrift 3.2.2.

³²Vgl. FRIES [20, S. 46].

³¹FRIES [20] identifiziert diesen Begriff mit dem von Max Black 1949 (und später Anderen) ver-

derer sprachlicher Zeichen zurückgeführt. Oft wird behauptet³³, hierbei gehe es darum, eine metasprachliche Beschreibung zu leisten. Der Status von Beschreibungseinheiten wird dann durch spezielle Kennzeichnung zum Ausdruck gebracht. Ein solcher Anspruch wird u. a. deshalb nicht einlösbar, als entsprechende Beschreibungseinheiten ihren Zeichencharakter nicht dadurch verlieren, dass sie in Umgebung von Klammern auftreten; sie sind nicht getrennt von entsprechenden klammerlosen Ausdrücken interpretierbar.

Damit soll eine Bezugnahme auf Beschreibungseinheiten nicht als unbrauchbar oder unnütz verworfen werden. Die unbestreitbar nützliche Verwendung in der modernen CL und KI steht außer Frage.

Ein für die Spezifikation bzw. Abgrenzung von Homonymie und Polysemie in der traditionellen Sprachbeschreibung häufig angebrachtes Kriterium ist der diachrone Aspekt dieser Phänomene. Wörter, die zufällig den selben Sprachkörper haben, sich möglicherweise etymologisch aus zwei verschiedenen Wurzeln herleiten lassen, werden in historischer Sichtweise als 'Homonyme' bezeichnet, z. B. *kosten*₁ (mhd. *kosten* aus afrz. *coster*, vlat. *costare*, klass.-lat. *cō-stāre*) 'eine Preis haben' und *kosten*₂ (mhd. *kosten*, ahd. *kostōn*, lat. *gustare*) 'abschmecken, versuchen'. Polyseme besitzen zwar unterschiedliche Bedeutungen, die sich aus einer gemeinsamen sprachlichen Wurzel in Form von Metaphern entwickelt haben, z. B. *Pferd*₁ (Tier), *Pferd*₂ (Turngerät), *Pferd*₃ (Spielzeug), *Pferd*₄ (Schachfigur), die Wörter sind aber mit verschiedenen Bedeutungen in den Wortschatz aufgenommen worden. Unter einer synchron ausgerichteten Sprachanalyse spielen die getroffenen Entscheidungen keine so große Rolle³⁴. *Schloss*₁ (Gebäude) und *Schloss*₂ (Verschluss) werden z. B. trotz gemeinsamen Ursprungs als getrennte, eigenständige Wörter empfunden. So könnte man synchronisch 'Homonymie' dadurch definieren, dass wenn Wortformen, die sich mit zwei oder mehreren verschiedenen Inhalten verbinden, zwischen denen keine Beziehungen bestehen, diese als Homonyme betrachtet werden können, sind Beziehungen zwischen den Wortformen jedoch existent, kann man von Polysemen sprechen.

FRIES [20, S. 61] bemerkt abschließend zur Abgrenzung von Homonymie und Polysemie bezüglich Diachronie und Synchronie, dass „die Beschreibung lexikalischer Mehrdeutigkeit, die traditionelle Trennung zwischen Homonymie und Polysemie insofern nutzbar

³³Vgl. GREIMAS [26] und WOLSKI [65, S. 46].

³⁴Vgl. hierzu auch RICHTER [57], welche keine Polysemie kennt, sondern nur „echte Homonyme“, Lautreihen, die denselben akustischen Eindruck hinterlassen und „unechte Homony-

me“, in erweiterter Bedeutung, die „einzeln gesprochen“ den selben akustischen Eindruck vermitteln. Ob ein Wort homonym ist oder nicht, bestimmt sich demzufolge nach seinem Stellenwert in der jeweiligen Artikulation.

machen kann, als sie die Klassifizierung der im traditionellen Sinne Homonyme als bloße Unbestimmtheiten (Vagheiten) mit großer Wahrscheinlichkeit ausschalten kann und demgegenüber ein Großteil der im traditionellen Sinne Polyseme eher vage als ambig erscheint.“

3.3 Strukturelle Ambiguität

Eine weitere Gruppe von Mehrdeutigkeit ist die **strukturelle** oder **syntaktische Ambiguität**, welche das Phänomen beschreibt, dass manche Sätze, obwohl deren Konstituenten allein nicht ambig sind, mehrere Lesearten ermöglichen. So hat (9) die beiden möglichen (u.v.m.) Interpretationen (9a) und (9b):

- (9) Er entdeckte den Mann mit einem Fernglas.
- (9a) Er entdeckte den Mann durch ein Fernglas.
- (9b) Er entdeckte den Mann, welcher ein Fernglas hatte.

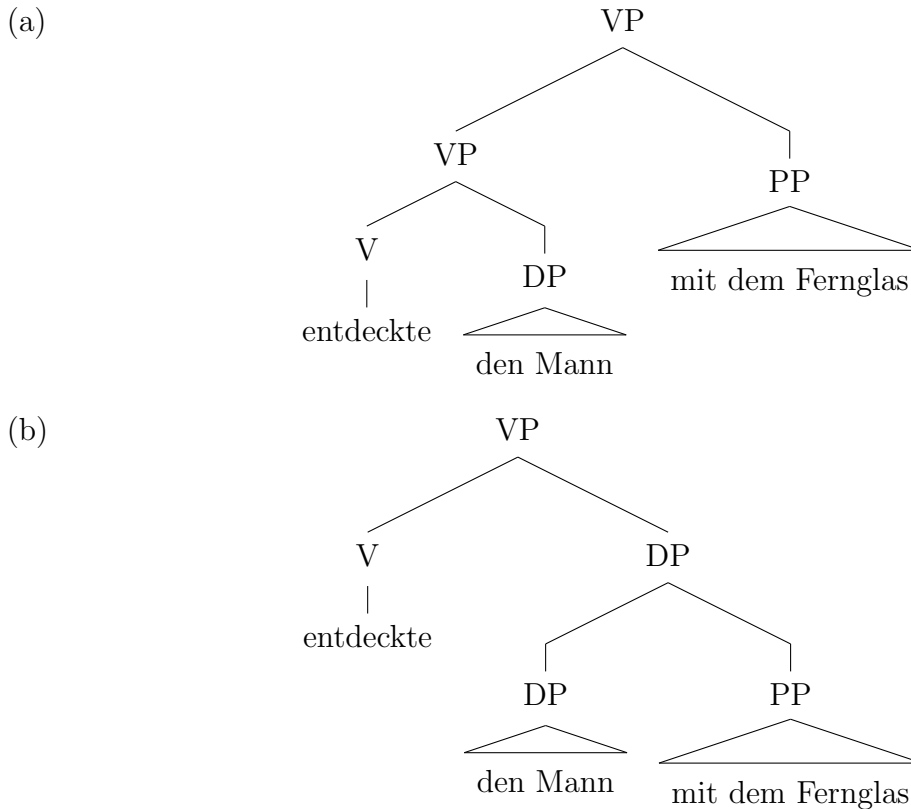
Daraus lässt sich das für das Verständnis von struktureller Ambiguität essentielle Kompositionalitätsprinzip³⁵ ableiten:

Die Interpretation eines Satzes wird durch die Interpretation der Wörter, die in dem Satz vorkommen und durch seine syntaktische Struktur bestimmt.

Man kann nun davon ausgehen³⁶, dass der *mit*-Satz in (9) ein Adjunkt ist und somit die Ambiguität von (9) darin liegt, ob der *PP mit einem Fernglas* Adjunkt für den *DP den Mann* oder den *VP entdeckte den Mann* ist. Wenn man weiterhin davon ausgeht, dass ein Adjunkt die syntaktische Eigenschaft hat, sich mit einer Phrase zu verbinden, um sie zu einer noch größeren Phrase des selben Typs zu erweitern, erzeugt das Hinzufügen des *PP mit einem Fernglas* zu dem *VP entdeckte den Mann* den größeren *VP entdeckte den Mann mit einem Fernglas*, Das Hinzufügen des *PP mit einem Fernglas* zu dem *DP den Mann*, wird hingegen den größeren *DP den Mann mit einem Fernglas* bilden. Daraus folgt dann, dass die beiden unterschiedlichen Interpretationen des *VP entdeckte den Mann mit einem Fernglas* zwei syntaktische Strukturen (a) und (b) haben, die sich aus verschiedenen Anordnungen der einzelnen Elemente ergeben:

³⁵Auch **Frege-Prinzip**; benannt nach dem deutschen Mathematiker, Logiker und Philosophen Gottlob Frege (1848-1925).

³⁶Vgl. dazu Abschnitt 18 und die Seiten 330f. aus RADFORD [52].



3.4 Skopus-Ambiguität

Ein dritter Typ sprachlicher Mehrdeutigkeit ist die **Skopus-Ambiguität**³⁷, ein klassisches Beispiel wäre (10), mit mindestens den Lesearten (10a) und (10b):

Auch in der mathematischen Quantoren-Logik spielt der Bereich des Skopus in Verbindung mit gebundenen und freien Variablen eine Rolle. So liegt eine Variable im Skopus eines Quantors $\in \{\forall, \exists\}$, wenn sie in der Menge der gebundenen Variablen $BV(\phi)$ vorkommt. In der Formel $\phi \simeq \forall x: x + y = y + x$ ist: $BV(\phi) = \{x\}$ und $FV(\phi) = \{y\}$.

- (10) Einen Computer benutzen alle Indologiestudenten.
 (10a) $\exists x: \text{computer}(x) \quad \forall y: \text{indolgiestudent benutzt}(y, x)$
 (10b) $\forall y: \text{indologiestudent} \quad \exists x: \text{computer}(x) \text{ benutzt}(y, x)$

Als Phänomen der Linguistik, wäre über die Skopus-Ambiguität noch weit mehr anzumerken, insbesondere da ihre automatische Disambiguierung sehr aufwändig ist. Da sie

³⁷In der Regel betrifft der Skopus, also der Geltungsbereich, quantifizierende Wörter wie 'alle' oder 'keine' und Zahlwörter. An dieser Stelle halte ich es für sinnvoll, sich in der Interpretati-

on von quantifizierten Sätzen, mathematischer Hilfsmittel der Quantoren-Logik, wie \forall oder \exists zu bedienen.

jedoch für die WSD keine Rolle spielt, werden hier weitere Ausführungen außer Acht gelassen.

4 Bedeutungswandel im Sanskrit

Da der Bedeutungswandel eine der Hauptquellen³⁸ für ambige Begriffe des Sanskrit in Form von Homonymen³⁹ und Polysemen ist, sollen in diesem Kapitel einige Anmerkungen bezüglich des Bedeutungswandels im Sanskrit⁴⁰ generell und an speziell ausgesuchten Beispielen gemacht werden. Zuerst werden einige Ursachen für Bedeutungswandel vorgestellt, die als extra-linguistisch bzw. als linguistisch bezeichnet werden können. Anschließend möchte ich noch ein wenig auf verschiedene Tendenzen und Klassifizierungsmöglichkeiten dieses Phänomens eingehen.

4.1 Ursachen des Bedeutungswandels

Anmerkung: die nachfolgenden Zitate aus dem ṚV werden akzentuiert angegeben. Bei allen anderen Zitaten aus dem Bereich der Objektsprache wird auf eine Akzentuierung verzichtet. Wird bei der Klärung der Bedeutung eines Lexemes dieses in die Metasprache eingebunden, so geschieht dies ohne Akzentuierung.

Als Ursachen für Bedeutungswandel kommen in allen natürlichen Sprachen viele Gründe in Betracht. Unbestritten haben religiöse Faktoren einen starken Einfluss auf das tägliche Leben und somit auf die Sprache. In besonderem Maße trifft dies auf die Menschen des indischen Subkontinents zu – wo sich doch hier Religion in einer speziellen Weise in der heiligen Sprache des Sanskrit manifestiert.

³⁸Die intendierte Ambiguität von künstlichen Begriffen in der Kunstdichtung soll hier außer Acht gelassen werden.

³⁹H. aufgrund von Lautverschmelzung sind im Sanskrit eher selten, die meisten H. entstanden aufgrund des semantischen Auseinanderlaufens eines Begriffes. Siehe KĀMBOJA [35, S. 302]. Zum Beispiel das Wort *abhyāsa*₁ (Wiederholung) und *abhyāsa*₂ (nahe, zur Hand), welches von *abhyāśā* abstammt und seine phonetische Form als *abhyāsa* entwickelte. Siehe z. B. in

Mbht. VII.61.10. Oder auch durch Vermischung mit dravidischen Wörtern: *arka*₁ (von \sqrt{arc} strahlen, loben) mit den Bedeutungen 'Strahl, Sonne, Feuer, Lied, Hymne' und *arka*₂ aus dem Dravidischen als Bezeichnung der Pflanze *calotropis gigantea*.

⁴⁰Die hier verwendeten Daten reichen vom frühen Ṛg Veda bis hinein in die epische Periode mit Sammlungen wie dem Mahābhārata und die klassische Periode mit dem Pañcatantra.

4.1.1 Extra-linguistische Ursachen

Ein Beispiel für einen extra-linguistischen Bedeutungswandel ist das Wort *brahman* (Formung, Gestaltung), welches ursprünglich von $\sqrt{bṛh}$ (kräftigen, groß machen⁴¹) stammt und einen ganz enormen Bedeutungswandel erfahren hat. Im **RV** wird es in der Bedeutung von 'Zauberwort, Gedicht' verwendet, wenn es auf der ersten Silbe akzentuiert ist:

tát tvā yāmi bráhmaṇā vādamānah

Das erflehe ich von Dir,

mit diesem Gedicht freundlich empfangend ...

RV 1.24.11

Liegt der Akzent auf der zweiten Silbe, so bezeichnet es denjenigen, der mit *bráhman* versehen ist, den *brahmán*⁴²:

yó brahmāṇe prathamó gá avindat

..., welcher zuerst für den, der mit bráhman versehen ist, die Kühe fand, ...

RV 1.101.5

Später erhielt dieses Wort auch die Bedeutung als 'Sammlung von Hymnen':

brahma brahmacāribhirudakrāmat

Der Veda erhob sich mit den Schülern⁴³.

AV 19.19.8

Mitunter die wichtigste Bedeutung von *brahman* im weiteren Verlauf der Entwicklung der vedischen Religion ist 'Weltgeist, Absolutum'⁴⁴. Diese Bedeutung ist weit verbreitet in philosophischen Werken, den Upaniṣaden und der Bhagavad Gītā:

ya evaṃ veda ahaṃ brahmāsmīti sa idaṃ sarvaṃ bhavati

„Ich bin das Brahman“, wer das weiß, der wird zu diesem ganzen.

Br. Up. 1.4.10

An diesem Beispiel lässt sich ein religiös motivierter Bedeutungswandel des Wortes *brahman* erkennen.

⁴¹Zur Übersetzung von $\sqrt{bṛh}$ siehe **BARH** in MAYRHOFER [45].

⁴²GELDNER [22] übersetzt *brahmán* mit 'der, der heiligen Rede Kundige'. Außerdem bezeichnet *brahmán* den Priester im vedischen Opfer, der

die RV-Verse rezitiert.

⁴³Vgl. hierzu KĀMBOJA [35, S. 23].

⁴⁴Zu diesen Bedeutungen vgl. MONIER-WILLIAMS [50].

Auch das Sprachtabu spielt eine wichtige Rolle für den Bedeutungswandel. Der vedische Gott Rudra wurde mithin sehr gefürchtet, in ṚV 4.3.6 erhält er z. B. das Epithet *nṛhán* (Männer tödend). Da es für die Opfernden zu gefährlich war diesen Namen auszusprechen, wurde er später mit dem euphemistischen Epithet *śiva* (vielversprechend) bekannt. Er wird dann auch *śānkara* (wohltätig) oder *mahādeva* (der große Gott) genannt.

*dve tanū tasya devasya vedajñā brāhmaṇā viduḥ
ghorām anyāṃ śivāṃ anyāṃ*

Zwei Körper des Gottes kennen die vedakundigen Brāhmaṇen:

Der eine ist fürchterlich, der andere vielversprechend.

Mbht. 8.161.3

Doch können noch viele andere Faktoren auf die Bedeutung eines Begriffes einwirken und als Ursache für einen Bedeutungswandel fungieren. Es ließen sich noch soziale, kulturelle und wirtschaftliche oder geographische Veränderungen, sowie Fremdspracheinwirkung und psychologische Faktoren wie z. B. die Ironie aufzeigen und mit Beispielen belegen, dies soll hier nicht unternommen werden. Ich verweise für weitere Ausführungen auf KĀMBOJA [35, S. 19–34].

4.1.2 Linguistische Ursachen

Eine Vielzahl linguistischer Ursachen für Bedeutungswandel sind zusätzlich zu den oben genannten Gründen auch für das Sanskrit bekannt. In frühen indischen Schriften lässt sich zum Beispiel ein freier Austausch von [r] und [l] erkennen⁴⁵. Das Wort *śukrá* von $\sqrt{śuc}$ (leuchten, glühen, brennen) taucht im ṚV mit der Bedeutung 'klar, licht, hell, weiß' auf:

śukrá vāsānāḥ sváravo na águḥ
weißgekleidet sind die Pfosten zu uns gekommen⁴⁶.

ṚV 3.8.9

Und im ŚB zeigt sich ein *śukla* mit gleicher Bedeutung:

yacchuklaṃ tadagneyaṃ yat kṛṣṇaṃ tat saumyaṃ
Was weiß ist, ist agnisch, was schwarz ist, ist somisch⁴⁷.

ŚB 1.6.3.41

⁴⁵Die eckigen Klammer zeigen an, dass es sich um Phone handelt.

⁴⁶Diese Übersetzung ist aus GELDNER [22].

⁴⁷Siehe dazu auch EGGELING [19]: 'that which is white is related to Agni, and that which is black is related to Soma'.

Später findet man das Wort *śukra* mit verschiedenen Bedeutungen⁴⁸ vor, überwiegend jedoch mit der Bedeutung 'Samen':

svapne siktva brahmacārī dvijaḥ śukraṃ akāmataḥ

A twice-born student, who has involuntarily wasted his manly strength during sleep, ...⁴⁹

Manu. 2.181

Das Wort *śukla* tritt hauptsächlich mit der Bedeutung von 'weiß' auf: in der Entwicklung von [r] und [l] zu zwei getrennten Phonemen /r/ und /l/ liegt also ein Bedeutungswandel begründet. Ähnliches ließe sich z. B. bei Worten wie *aśrīra/aślīla* oder \sqrt{car}/\sqrt{cal} erkennen⁵⁰.

Auch Entlehnung und Rückentlehnung tragen zum Bedeutungswandel von Sanskrit-Begriffen (natürlich auch nicht-Sanskrit-Begriffen) bei. Einerseits in Form des Substrat-Einflusses und andererseits durch Sprachkontakte mit beispielsweise den Arabern oder Griechen. Nach KĀMBOJA [35] entstammt z. B. das Wort *paṅgu* (lahm) dem Munda und taucht als erstes im *Atharva Veda-Pariśiṣṭa* auf, gefolgt von *paṅguka*, *paṅgula* im Mahābhābrata. Andere Varianten sind *vaṅku* (gekrümmt gehen), *bhaṅga* (Lamheit, Kurve) oder *bhaṅgura* (gebogen). Ursprünglich scheint das Wort eine Bedeutung wie 'krumm, gebogen' gehabt zu haben und als Lehnwort im Sanskrit zu 'lahm' modifiziert worden zu sein⁵¹.

4.2 Tendenzen und Klassifizierungen

Da ein Bedeutungswandel nicht zwingendermaßen eine vollständige Veränderung eines Begriffes mit sich bringen muss, möchte ich einige Tendenzen oder Richtungen in die sich der Bedeutungswandel entwickeln kann, aufgreifen. Allgemein bekannt sind die Begriffe Bedeutungsverengung und Bedeutungserweiterung. Der Ausdruck *śṛṣṭīriṣṭakāḥ*, der mit *śṛṣṭi* (Erschaffung, Emission) als Abstraktnomen von \sqrt{srj} (loslassen, ergießen, erschaffen) in der Form einer Apposition für *iṣṭakā* (Backstein, Ziegelstein) erscheint, nimmt später in gewissem Kontext elliptisch die Bedeutung von *iṣṭakā* an⁵² und somit verengt sich in diesem Kontext seine Bedeutung. Im Gegensatz dazu liegt bei dem Wort *varṣa*

⁴⁸Unter anderem als Name eines Sommermonats oder eines Marut. Vgl. MYLIUS [51].

⁴⁹Vgl. BÜHLER [12].

⁵⁰Auch hier verweise ich auf KĀMBOJA [35, S. 49].

⁵¹Vgl. KĀMBOJA [35, S. 51].

⁵²Vgl. hierzu *śṛṣṭīrupadadhāti*, 'er legt die *śṛṣṭi*-Ziegelsteine'. *Mīmāṃsā* 1.4.23.

(regnend, Regen⁵³), welches sich von $\sqrt{vr̥ṣ}$ (regnen) ableitet, eine Bedeutungserweiterung vor, wenn es später mit 'Jahr' übersetzt wird⁵⁴.

Die Liste der Nuancen des Bedeutungswandels könnte noch ganz beachtlich weitergeführt werden. Ähnlich den oben angesprochenen Veränderungen des Umfangs der Bedeutung eines Begriffes, wären hier die verschiedenen Ausprägungen der 'moralischen' Bewertung, also pejorative und meliorative Entwicklungen aufzuführen. Doch auch diese Aufführungen würden hier den Rahmen sprengen, es sei also erneut auf KĀMBOJA [35] verwiesen.

An den angebrachten Beispielen kann man gut erkennen, dass wohin sich die Bedeutung eines Begriffes auch entwickelt, so bleibt eine gewisse Verbindung zwischen neuer und alter Bedeutung bestehen. BUCK [6] sagt in seinem Vorwort, „The associations underlying semantic changes are so complex that no rigid classification of the latter is possible (...). Nevertheless, there are certain types which it is convenient to recognise.“ Wenngleich der Bedeutungswandel an den verwendeten Beispielen gezeigt wurde, ist bezüglich dieser „komplexen Assoziation“ anzumerken, dass man heute der Auffassung ist, dass sich Bedeutungswandel nicht am einzelnen Wort, sondern an Wörtern als Elementen eines Feldes (Wortfeld, lexikalisches Feld, assoziatives Feld) vollzieht, nicht zuletzt weil die Bedeutung selbst dem Wandel unterliegt, da viele Bedeutungen vage bzw. unscharf sind⁵⁵. Somit kann eine kleine Veränderung eines Elementes eines Feldes zu einer Anpassung aller Elemente dieses Feldes führen. Nach ULLMANN [62, S. 211] war es Léonce Roudet, welcher als erster Linguist, den Strukturalismus Saussures mit der Philosophie Bergsons verband und eine Klassifikation, welche auf Assoziation beruhte zu postulieren. Er verbindet sein Schema mit der analytischen Definition, welche besagt, dass Bedeutung eine reziproke und umkehrbare Beziehung zwischen Name und Bedeutung sei. Dementsprechend ist Bedeutung eine Beziehung oder Assoziation. Nimmt man diese Definition als Arbeitshypothese, so erhält man zwei Kategorien von Bedeutungswandel: entweder auf einer Assoziation zwischen Bedeutungen oder auf einer Assoziation zwischen Namen beruhend. Da auch Assoziation aus zweierlei Arten besteht, Similarität und Kontiguität, welche alle Arten assoziativer Relationen umfasst, die nicht der Similarität zuzuordnen sind, erhält man vier Hauptklassifikationsmerkmale:

⁵³Zu dieser Bedeutung siehe ŚB 1.5.2.19; EGGE-LING [19] übersetzt *sa yadi vr̥ṣṭikāmaḥ syāt* mit 'Should he (the sacrificer) be desirous of rain'.

⁵⁴In z. B. AV 12.1.36 übersetzt BLOOMFIELD [5] *varṣāṇi* mit 'rainy season'. Der Übergang zum

'Jahr' ist nicht mehr weit: von der Regenzeit als Zeiteinheit, zum Zählen der Jahre durch das Zählen der Regenzeiten. Vgl. KĀMBOJA [35, S. 90 Fn. 1].

⁵⁵Vgl. LEWANDOWSKI [41, S. 159].

1. Similarität von Bedeutungen, d. h. Metapher.
2. Kontiguität von Bedeutungen, d. h. Metonymie.
3. Similarität von Namen, d. h. Volksetymologie.
4. Kontiguität von Namen, d. h. Ellipse.

Zu den Punkten 1. bis 4. soll nun jeweils ein Beispiel aus dem Sanskrit mit Belegstellen angeführt werden. Auch wenn die einzelnen Punkte noch weitere Unterpunkte beinhalten, halte ich eine prägnantes Beispiel für ausreichend um einen gewissen Überblick der mannigfaltigen Schichten des Bedeutungswandels zu schaffen.

4.2.1 Metapher

Die Metapher ist zweifelsohne eine treibende Kraft für Synonymie und Polysemie⁵⁶. Wenn Similarität zwischen zwei oder mehreren Bedeutungen vorliegt, kann das denotative Wort der einen Bedeutung auch für die andere oder anderen verwendet werden.

Das Wort *phala* bedeutet im RV noch sehr konkret 'Frucht, Baumfrucht'⁵⁷:

vrkṣām pakvām phálamañkīva dhūnuhīndra sampāraṇam vāsu
Schüttle, Indra, aus der Not helfendes Gut herab wie einer mit
dem Haken die Reife Frucht vom Baum (schüttelt)⁵⁸.
RV 3.45.4

Im AV entwickelt sich eine abstraktere Bedeutung von *phala* zu 'Gewinn, Profit':

śunām no astu prapaṇó vikráyaś ca pratipaṇāḥ phalīnam mā
kṛṇotu
Handel und Verkauf soll für uns erfolgreich sein. Tauschhandel
soll mich zu einem Mann mit Früchten⁵⁹ machen.
AV 3.15.4

Die abstrakteste Stufe von *phala* lässt sich in z. B. philosophischen Schriften wie der Bg. feststellen, wo die metaphorische Bedeutung von 'Frucht' in Form von 'Konsequenz, Effekt, Ergebnis' zum Ausdruck kommt:

karmaṇyevādhikāraṣṭe mā phaleṣu kadācana

⁵⁶Vgl. ULLMANN [62, S. 211].

⁵⁷Zu dieser Bedeutung siehe MAYRHOFER [45].

⁵⁸Zu dieser Übersetzung siehe GELDNER [22].

⁵⁹Das in-Suffix wird unter Anderem verwendet,

um Adjektive zu bilden, die den Besitz anzeigen. Vgl. MACDONELL [44] und WACKERNAGEL [63].

mā karmaphalahetur bhūr mā te saṅgo 'stvakarmanī

Nur in der Tat ist Deine Herrschaft, niemals in dem Ergebnis⁶⁰.

Sei nicht einer dessen Motiv das Ergebnis der Tat ist – Du sollst
nicht Affektion zur Untätigkeit haben!

Bg. 2.47

4.2.2 Metonymie

Auch die Metonymie spielt eine große Rolle im Bedeutungswandel. Die Metonymie, im Gegensatz zur Metapher, bezeichnet eine reale, d. h. kausale, räumliche oder zeitliche Beziehung zwischen zwei sprachlichen Zeichen. Die Metapher wird durch Überschneidung von Bedeutungsmerkmalen (Semen) konstituiert, die Metonymie beruht dagegen auf ihrer Nichtüberschneidung. Sie wird also durch den gemeinsamen Einschluss von Bedeutungsmerkmalen zu einer Bedeutungsmerkmal-Einheit konstituiert⁶¹.

Das Wort *go* bedeutet im ṚV und späteren vedischen Schriften generell 'Kuh'. Man findet jedoch an vielen Stellen *go* auch für alle möglichen Rindsprodukte wie Milch, Joghurt, Ghee, Haut, Leder etc. als metonyme Begriffe. Mādhava⁶², der solche Bedeutungen erklärt, rechtfertigt seine Stellung indem er sagt:

vikāre prakṛtiśabdah

Ein Effekt wird durch seine Ursache bezeichnet.

Mādh. zu ṚV 1.137.1

Auch Yāska teilt diese Ansicht und sieht *go* in diesen Bedeutungen als Sekundärform ohne Sekundärsuffix an:

athāpyasyāṃ tādhitena kṛtsnavannigamā bhavanti

Und in Bezug auf diese (Bezugswort?) sind Vedastellen mit abgeleitetem (Ergänzung?) wie vollständig (ursprünglich?)

Nir. 2.5

Hieran lässt sich schon die metonymische Verwendung von *go* erkennen: alles was von einem Ochsen oder einer Kuh stammt wird mit *go* bezeichnet. Einige Vorkommen des Wortes in diesen Bedeutungen werden nun aufgeführt:

ná sá rájā vyathate yásminn índras tīvrāṃ sómaṃ pībati gó-

⁶⁰Zu dieser Übersetzung siehe auch BUITENEN [9].

⁶¹Vgl LEWANDOWSKI [42, S. 686].

⁶²Nach W. Slaje in ZDMG 2010, Heft 2, ist der

Name des berühmten Kommentators des ṚV nicht Sāyaṇa sondern Mādhava.

sakhāyam

Der König kommt nicht ins Schwanken, bei welchem Indra den scharfen Soma trinkt, den mit Milch Verbundenen⁶³.

RV 5.37.4

té somādo hārī indrasya niṃsate ṃśúṃ duhánto ádhy āsate gávi
Die Somaesser suchen Indra's Falbenpaar auf, die Somapflanze melkend sitzen sie auf der Stierhaut⁶⁴.

RV 10.85.9

Andere Stellen im RV, wo *go* in ähnlichen Bedeutungen vorkommt sind z. B. 8.13.14; 9.46.4; 10.16.7; 10.94.9 oder 6.75.11.

Das Wort *kāṇḍa* heißt in *Manu.* noch 'Glied':

udbhijjāḥ sthāvarāḥ sarve bījakāṇḍaprarohinaḥ

Alle Pflanzen⁶⁵ aus Samen oder Glied wachsend, keimen.

Manu. 1.46

Durch Metonymie erlangte es z. B. im *Mbht.* die Bedeutung 'Stamm, Stiel':

paṭṭiśaṃ ca tribhir vānaiś ciccheda tilakāṇḍavat

Und den Paṭṭiśa⁶⁶ mit drei Hölzern schnitt er wie einen Sesam-Stamm.

Mbht. 6.113.41

4.2.3 Volksetymologie

Nach ULLMANN [62, S. 101f] ist die treibende Kraft der Volksetymologie das Verlangen zu beleben, was undurchsichtig bzw. opak wurde. Dieses Verlangen ist mehr psychologisch als historisch und basiert auf der Beziehung zwischen Klang und Bedeutung.

In der späteren Sanskrit-Literatur wurde das Wort *dampatī* als Dvandva-Kompositum mit der Bedeutung 'Ehemann und Ehefrau' interpretiert. Es wurde als korrumpierte Form von *jāyāpatī* angesehen, die sich über *jampatī* entwickelt haben soll. Im Amara wird es mit den folgenden Synonymen erklärt:

dampatī jampatī jāyāpatī bhāryāpatī ca tau

⁶³Nach GRASSMANN [25, S. 414] bedeutet *gó-sakhi* 'mit Milch verbunden'.

⁶⁵Nach BÖHTLINGK [11] heißt *sthāvara* 'pflanzlich, zur Pflanzenwelt gehörig'.

⁶⁴Zu 'Stierhaut' siehe auch GELDNER [22], der 'Stier(haut)' übersetzt.

⁶⁶Nach BÖHTLINGK [11] ein Speer oder eine Waffe mit drei Spitzen.

Amara 2.6.38

Dem gegenüber ist *dampati* ganz klar ein Tatpuruṣa-Kompositum mit dem Vorderglied *dam*⁶⁷ (Haus) und dem Hinterglied *pati* (Gatte, Herr). Im ṚV findet *dam* und auch seine thematische Erweiterung *dama* häufig⁶⁸ Verwendung:

*tám tvā suśipra dampate stómair vardhanty átrayo gīrbhīḥ śumbhanty
átrayah*

Als den stärken dich, oh schönlippiger Hausherr, die Atri-s, schmücken
(dich) die Atri-s mit Lobliedern.

ṚV 5.22.4⁶⁹

Diese rege Verwendung von *dam* bzw. *dama* hat sich in der späteren Sanskrit-Literatur nicht fortgesetzt. Doch hat in dem Kompositum *dampatī* (N. d.) das Wort *dam* als vermeintliches Allomorph von *jam* bzw. *jāyā* weiter gelebt.

4.2.4 Ellipse

Die Ellipse ist im Wesentlichen ein morphologischer Prozess, in welchem signifikante Elemente aus einem Ausdruck weggelassen werden und der verbleibende Teil die gesamte Bedeutung des ursprünglichen Ausdrucks übernimmt.

Das Wort *candra* leitet sich wohl von $\sqrt{\text{cand}}$ (glänzend, schimmern)⁷⁰ ab und bedeutet 'schimmernd, licht'. Im ṚV und später wird es häufig als Adjektiv für z. B. Licht⁷¹, Tropfen⁷² oder Feuer⁷³, Wasser oder Wagen verwendet:

yás cāpāḥ candrā bṛhatīr jajāna ...

... und welcher die weiten, schimmernden Wasser erschaffen hat.

ṚV 10.121.9

vāyav ā candréṇa ráthena yāhí sutásya pītáye

Oh Vāyu, komm mit dem schimmernden Wagen zum Trank des
Ausgepressten.

ṚV 4.48.1

In Verbindung mit *mās* (Mond, Monat) bildet sich das Kompositum *candramās* (glänzender

⁶⁷Siehe auch *dám* in MAYRHOFER [45].

unter **cand**.

⁶⁸Noch: z. B. ṚV 1.1.8; 75.5; 120.6; 149.1; 8.69.16;
84.7; 10.61.20; 99.6.

⁷¹ṚV 1.48.9.

⁷²ṚV 3.40.4.

⁶⁹Vgl. KUPFER [38, S. 196].

⁷³ṚV 5.10.4.

⁷⁰Nur in ṚV 5.43.4 belegt. Siehe MAYRHOFER [45]

Mond, Mond) und durch den Prozess der Ellipse erscheint als erstes im AV *candra* mit der Bedeutung von 'Mond':

yathā sūryaś ca candraś ca na bibhīto na ṛisyataḥ

Wie Sonne und Mond weder fürchten noch Schaden nehmen, ...

AV 2.15.3

Ursprünglich bedeutet *vinaśana* 'Untergang, Verderben'. Im Laufe der Zeit nahm der Ausdruck eine Bedeutung wie 'Name des Ortes, wo der Fluß Sarasvatī versiegt' an:

himavadvindhyaḥ madhyaḥ yat prāg vinaśanād api

Was als Mitte von Himalaya und Vindhya östlich von Vinaśana

...

Manu. 2.21

Das ist das Resultat der Kürzung des Ausdrucks *sarasvatyāḥ vinaśanaḥ*, der in früherer Literatur auftaucht:

etad vinaśanaḥ nāma sarasvatyā viśāṃpate

dvāraḥ niṣādarāṣṭrasya yeṣāḥ doṣāt sarasvatī

praviṣṭā pṛthivīm vīra mā niṣādā hi māḥ viduḥ

Dies ist nämlich das Verschwinden der Sarasvatī, Oh Herr der

Leute, das Tor zum Reich der Niṣāda-s,

aufgrund deren Verbrechen Sarasvatī in die Erde eintrat, Oh

Held, denn nicht sollen die Niṣāda-s sie⁷⁴ kennen!⁷⁵

Mbht. 3.130.3-4

4.3 Computational Sanskrit

In diesem Kapitel wird ein kurzer Einblick in den Bereich der maschinellen, d. h. computergestützten Verarbeitung des Sanskrit geschaffen. Da dieses Feld in Forschung und Wissenschaft noch bei weitem nicht so große Verwendung gefunden hat, wie für andere Sprachen, sei hier im Wesentlichen eine Zustandsbeschreibung nach HELLWIG [29] gegeben.

Das erste öffentlich zugängliche Analyse-System des Sanskrit stellte Gérard Huet mit einem Internet-Dienst zur Analyse von Roh-Sanskrit-Texten vor: die *Sanskrit-Heritage Site*⁷⁶. Dieser Ansatz kombiniert endliche Automaten mit syntaktischen Regeln, um

⁷⁴*māḥ* wurde durch *tāḥ* ersetzt, da hier nicht Sarasvatī, sondern Lomaśa spricht.

⁷⁵Vgl. auch BUITENEN [8].

⁷⁶Siehe <http://sanskrit.inria.fr>.

eine Menge linguistischer Analysen für einen Satz zu liefern⁷⁷. HELLOWIG [30] stellt einen SanskritTagger vor, welcher Phoneme lokalisiert, die durch *sam̐dhi* entstanden sein könnten. Ein Teil der Daten, welche durch SanskritTagger generiert wurden, sind im WWW als *Digital Corpus of Sanskrit*⁷⁸ erreichbar. Das System sieht auch eine semantische Repräsentation in Form eines semantischen Baumes vor, dessen Einträge jedoch händisch zugewiesen bzw. mit den entsprechenden lexikalischen Einheiten verknüpft werden müssen. Um diesen Prozess zu automatisieren, könnten computergestützte WSD-Systeme von großem Nutzen sein. Eine weitere Schicht linguistischer Annotation könnte syntaktische Informationen enthalten, in diesem Bereich ist die Forschung jedoch noch nicht weit vorangekommen⁷⁹. So sieht HELLOWIG [29] im Wesentlichen zwei Bereiche für zukünftige Forschung als notwendig: i) Eine Reihe von Anwendungen zur elektronischen Bearbeitung von Sanskrit-Ressourcen muss geschaffen werden:

1. Ein Austauschformat für linguistisch annotierte Sanskrit-Texte, welches in XML realisiert werden könnte.
2. Ordentlich strukturierte elektronische Wörterbücher und Corpora müssen geschaffen werden, idealerweise sollten Wörterbücher geschaffen werden, welche direkt mit den entsprechenden Quelltexten verknüpft sind.
3. Vorhanden *Tokenizer* und *Lemmatizer* sollten regelbasierte Analyseverfahren mit statistischen verbinden.
4. Syntaktische und semantische Analyseverfahren sollten aus dem Zustand der 'Pseudo-Code-Formulierungen' in reale Implementationen gehoben und der wissenschaftlichen Gemeinschaft zur Verfügung gestellt werden.

Als Anwendungsgebiet mit den somit geschaffenen technischen Grundvoraussetzungen, sollte im Feld der indologischen Arbeit und generell in den Geisteswissenschaften eine Verbindung moderner statistischer Methoden mit den gewohnten textkritischen Aufgaben, die sich seit dem 19. Jarh. kaum verändert haben, angestrebt werden.

Ein wichtiger Beitrag hierzu ist auch die vorliegende Arbeit, welche versucht moderne WSD-Verfahren auf das Sanskrit anzuwenden und zu überprüfen, ob man im Kontext des Sanskrit überhaupt zu guten Ergebnissen in der automatischen Bedeutungsunterscheidung gelangen kann und in wie weit sich Probleme der Granulärität von z. B. einem Wort wie *jana*, dessen Bedeutungen sich oft überschneiden und nahe bei einander liegen,

⁷⁷Siehe HUET [31].

⁷⁸Siehe DCS.

⁷⁹Siehe HELLOWIG [29].

darauf auswirken.

5 Wortbedeutungsdisambiguierung

Die Wortbedeutungsdisambiguierung, die im Folgenden vereinfacht mit der englischen Abkürzung **WSD** bezeichnet wird, dient der Feststellung der jeweiligen Bedeutung eines sprachlichen Zeichens innerhalb eines gegebenen Kontextes. Der natürlichsprachliche Prozess der Disambiguierung scheint im Wesentlichen unbewusst von statten zu gehen⁸⁰. Als Problem der Berechenbarkeit wird er häufig als 'KI-vollständig' referenziert, was bedeutet, dass er ein Problem ist, dessen Lösung eine vollständige Lösung des Verständnisses von natürlicher Sprache voraussetzt.

Für die WSD sind nur lexikalische Ambiguitäten von Interesse, also Polyseme und Homonyme, welche in einer Hierarchie von grob- bis feinkörnig gegliedert⁸¹ sind. Auf einer grobkörnigen Ebene hat ein Wort oft eine kleine Menge von Bedeutungen, die klar verschieden und sehr wahrscheinlich vollständig unverwandt sind – gewöhnlich sind diese Mehrdeutigkeiten Spezialfälle von Homonymie, Homographen⁸². Je mehr man sich in die feinkörnige Ebene begibt, desto komplexer wird die Struktur wechselseitiger Beziehungen, mit Phänomenen wie Polysemie oder metaphorischer Bedeutungserweiterung. In der modernen WSD ist ein Trend zu erkennen, das 'Bedeutungsinventar' nicht zu feinkörnig zu gestalten und somit die grobkörnige Disambiguierung den Hauptstellenwert der WSD darstellt⁸³. Methoden der WSD werden meist entsprechend der für sie verwendeten Quellen klassifiziert. Wissens- oder Wörterbuchbasierende⁸⁴ Methoden, welche sich hauptsächlich auf Wörterbücher, Thesauri oder Lexika berufen, unterscheiden sich von den unüberwachten⁸⁵ Methoden, welche direkt auf rohen, unannotierten Corpora arbeiten und den überwachten bzw. semi-überwachten⁸⁶ Methoden, welche annotierte Corpora bzw. eine sehr kleine handverlesene Menge an Trainingsdaten benutzen, um von diesen zu lernen. Zusätzlich gibt es noch verschiedene Kombinationen der einzelnen Ansätze. Auswertungen wie Senseval-2 oder Sensval-3 haben gezeigt, dass überwachte Methoden klar über wörter- buchbasierte Methoden dominieren und bezüglich der englischen Sprache zwischen 71,8% und 72,9% korrekte Disambiguierungen vornehmen können.

⁸⁰Vgl. AGIRRE [1, S. 1].

⁸¹Vgl. EDMONDS [18], der diese Hierarchie wie folgt angibt: Wortklasse, Homographie, Polysemie, reguläre Polysemie, Wortgebrauch, fester Ausdruck.

⁸²Beispielsweise 'modérn/móder'n' mit den Bedeu-

tungen 'fortschrittlich/verwesen'.

⁸³ Vgl. AGIRRE [1, S. 9 u. Kap. 3].

⁸⁴Der englische Fachbegriff dafür ist 'knowledge-based' oder 'dictionary-based'.

⁸⁵Engl.: 'unsupervised'.

⁸⁶Engl.: 'supervised' und 'semi-supervised'.

Im Folgenden sollen einige Verfahren und Methoden der WSD dargestellt werden, besonderer Augenmerk richtet sich dabei auf wissensbasierte und überwachte Methoden, da diese später kombiniert auf die *WSD im Kontext des Sanskrit* angewendet und ausgewertet werden.

5.1 Wissensbasierte Methoden

5.1.1 Der Lesk-Algorithmus

Als einer der ersten Algorithmen zur WSD benutzt der Lesk-Algorithmus⁸⁷ nur eine Menge von Einträgen eines Wörterbuches, mit einem Eintrag für jede mögliche Wortbedeutung und Kenntnis über den direkten Kontext in welchem die Disambiguierung erfolgt⁸⁸. Die Idee dahinter ist, dass für ein gegebenes Wortpaar W_1 W_2 die größte Schnittmenge von Begriffen aus den Bedeutungsdefinitionen des Wörterbuchs berechnet wird und den beiden Wörtern dann die Bedeutungen der für diese Schnittmenge verantwortlichen Bedeutungen zugeordnet werden.

Nachfolgend das Verfahren des Lesk-Algorithmus in Pseudo-Code:

```

1 foreach sense  $i$  of  $W_1$ 
2   foreach sense  $j$  of  $W_2$ 
3     compute  $Overlap(i, j)$ , the numbers of words in
4     common between the definitions of sense  $i$  and sense  $j$ 
5 find  $i$  and  $j$  where  $Overlap(i, j)$  is maximized
6 assign sense  $i$  to  $W_1$  and sense  $j$  to  $W_2$ 

```

Pseudo-Code-Bsp. 1: Lesk-Algorithmus

Dieser Algorithmus soll am prominenten Beispiel aus dem Englischen, dem Wort Paar *pine cone*, 'Kiefernzapfen' kurz veranschaulicht werden, indem man die Definitionen von *pine* und *cone* aus dem *Oxford Advanced Learner's Dictionary* vergleicht:

pine

1. seven kinds of evergreen tree with needle-shaped leaves

⁸⁷Vgl. LESK [40].

⁸⁸Vgl. MIHALCEA in AGIRRE [1, S. 108].

2. pine
3. waste away through sorrow or illness
4. pine for something, pine to do something

cone

1. solid body which narrows to a point
2. something of this shape, wheter solid or hollow
3. fruit of certain evergreen trees

Betrachtet man nun die einzelnen Worte in den Bedeutungsdefinitionen als Elemente der Mengen W_1^i bzw. W_2^j , so ergibt sich, dass $W_1^1 \cap W_2^3 = \{evergreen, tree\}$ maximal ist. Somit werden die Wortbedeutungen W_1^1 und W_2^3 gewählt. Dieser Algorithmus erreicht⁸⁹ eine Präzision von 50-70%.

Seit der ursprünglichen Definition des Algorithmus durch LESK [40] wurden diverse Variationen davon entwickelt: **i**) eine Version⁹⁰ des Algorithmus, die versucht die kombinatorische Explosion durch mögliche Wortdefinitionen bei Analyse von mehr als zwei Worten zu umgehen, **ii**) eine Variation⁹¹, bei welcher nur ein Wort disambiguiert wird, indem der *Overlap* der Lexikon-Einträge und des umgebenden Satz-Kontextes berechnet wird und **iii**) eine Alternative⁹², bei der das semantische Feld, welchem die jeweilige Wortbedeutung entstammt in die Evaluierung miteinbezogen wird, um einen vergrößerten Kontext der entsprechenden Wortbedeutung zu erhalten. Abschließend lässt sich bemerken, dass Variation **ii**) , der *vereinfachte Lesk-Algorithmus* die Variation des ursprünglichen Algorithmus ist, die die beste Effizienz und Präzision aufweist⁹³.

5.1.2 Semantische Similarität

Eine der wichtigsten Bedingungen der Disambiguierung beruht auf der Hypothese nach HALLIDAY UND HASAN [27], dass die Wörter in einer Rede für deren Verständnis semantisch verwandt sein müssen. Somit kann man die passenden Bedeutungen finden,

⁸⁹Vgl. LESK [40].

⁹⁰*Simulierte Abkühlung* bzw. *simulated annealing* nach Cowie et al. 1992 in *Proceedings of the International Conference on Computational Linguistics (COLING)*, Nantes, France, S. 157–161.

⁹¹Der *Vereinfachte Lesk-Algorithmus* bzw. *simplified lesk algorithm*.

⁹²Der *angepasste Lesk-Algorithmus* bzw. *adapted lesk algorithm* nach Banerjee & Pederson 2002 in *Proceedings of the Conference on Computational Linguistics and intelligent Text Processing (CICLING)*, Mexico City, Mexico, S.136 – 145.

⁹³Vgl. Mihalcea, R. in AGIRRE [1, S. 113].

indem man die Bedeutungen der kürzesten semantischen Entfernung wählt. Diese Methode beschränkt sich jedoch auf den *lokalen Kontext* eines gegebenen Wortes. Weitere Methoden⁹⁴ ziehen einen *globalen Kontext* mit in Betracht und erweitern ihren Skopus damit über das schmale Fenster, das wenige Zielwörter umgibt. Eine ganze Reihe an Größen zur Bestimmung der semantischen Similarität zweier Wörter wurde entwickelt. Ein⁹⁵ Formalismus, welcher von MIHALCEA UND MOLDOVAN [48] gegeben wurde und auf der [WordNet](#)-Hierarchie gut arbeitet, wird nun gegeben:

$$\text{Similarity}(C_1, C_2) = \frac{\sum_{k=1}^{|CD_{12}|} W_k}{\log(\text{descendants}(C_2))} \quad (5.1)$$

Formel (5.1) gibt die semantische Similarität zwischen unabhängigen Hierarchien, inklusive Hierarchien verschiedener Wortarten an. MIHALCEA UND MOLDOVAN [48] erzeugen virtuelle Pfade zwischen unterschiedlichen Hierarchien durch die Glossen in [WordNet](#). In (5.1) ist $|CD_{12}|$ die Anzahl gemeinsamer Wörter der Definitionen in den Hierarchien C_1 und C_2 , $\text{descendants}(C_2)$ repräsentiert die Anzahl der Konzepte in der Hierarchie C_2 und W_k ist das Gewicht (*weight*), welches mit jedem Konzept assoziiert und durch die Tiefe des Konzepts in der semantischen Hierarchie determiniert wird. Eine andere Möglichkeit, wenn die Similarität in einem Baumdiagramm der Taxonomie gemessen wird, ist die minimale Länge des Pfades zwischen *synsets*⁹⁶, welche die Eingabe-Wörter enthalten, zu ermitteln. In (5.2) nach LEACOCK U. A. [39] steht $\text{Path}(C_1, C_2)$ für die Länge des Pfades zwischen zwei Konzepten und D ist die Gesamttiefe der Taxonomie:

$$\text{Similarity}(C_1, C_2) = -\log\left(\frac{\text{Path}(C_1, C_2)}{2D}\right) \quad (5.2)$$

5.1.3 Lokale und globale Kontexte

Da ein Text üblicherweise mehr als nur ein ambiges Wort beinhaltet, hat die Disambiguierung meist Mengen von ambigen Worten zum Gegenstand, in welchen die Entfernung eines Wortes zu allen anderen Worten im Kontext die Bedeutung des Wortes beeinflusst. *Lokale Kontexte* als weitere Bedingung, um die Anzahl der Wörter in diesen ambigen Mengen zu begrenzen, sowie *syntaktische Abhängigkeiten* in Verbindung mit einem sehr einfachen Maß semantischer Similarität, haben auch recht gute Ergebnisse erzielt.

⁹⁴*Lexikalische Ketten* oder *lexical chains* sind ein Beispiel semantischer Relationen, die sich über mehrere Wörter innerhalb eines Textes ziehen.

⁹⁵Für einen Überblick, der gängigsten Maße verweise ich auf MIHALCEA in AGIRRE [1, Kap. 5.3.1].

⁹⁶Siehe [WordNet](#).

Innerhalb eines *globalen Kontextes* finden *lexikalische Ketten*⁹⁷ Verwendung. Durch einen generischen Verkettungs-Algorithmus werden für die Kandidaten-Wörter über den ganzen Kontext bzw. über große Teile davon, Ketten gebildet um Wortbedeutungen zu bestimmen. Alle Ketten, die ein bestimmtes Maß überschreiten, werden gewählt.

5.1.4 Selektionale Präferenzen und Heuristik-Methoden

Die ersten Algorithmen zur WSD beruhen auf *selektionalen Präferenzen*⁹⁸. *Selektionale Präferenzen* speichern Information von möglichen Relationen zwischen Wortarten und repräsentieren den gesunden Menschenverstand bezüglich Klassen von Konzepten: z. B. **EAT-FOOD**, **DRINK-LIQUID**. Obwohl für Menschen *selektionale Präferenzen* intuitiv erscheinen, ist deren Verwendung für die WSD schwierig.

In AGIRRE UND MARTÍNEZ [2] wurden vergleichende Auswertungen der verschiedenen Ansätze zur Feststellung *selektionaler Präferenzen* gemacht. Der klassische Ansatz stammt von RESNIK [54], welcher *selektionale Assoziationen* als Maßstab der semantischen Nähe zwischen einem Wort und einer semantischen Klasse vorschlägt⁹⁹. Bei *selektionalen Assoziationen* wird der Beitrag einer semantischen Klasse in einer gegebenen Relation dadurch quantifiziert, dass der Beitrag aller Konzepte, die durch diese Klasse subsumiert werden, quantifiziert wird. Bei gegebenem Wort W und semantischer Klasse C , verbunden durch die Relation R , wird die *selektionale Assoziation* wie in Gleichung (5.3) bewertet, basierend auf den Gleichungen (5.4) und (5.5):

$$A(W, C, R) = \frac{P(C|W, R) \log(P(C|W, R)/P(C))}{\sum_C P(C|W, R) \log(P(C|W, R)/P(C))} \quad (5.3)$$

$$P(C|W, R) = \frac{\text{Count}(W, C, R)}{\text{Count}(W, R)} \quad (5.4)$$

$$\text{Count}(W, C, R) = \sum_{W' \in C} \frac{\text{Count}(W, W', R)}{\text{Count}(W')} \quad (5.5)$$

Eine einfache und doch präzise Methode, Wortbedeutungen zu bestimmen, ist die Verwendung von schlichten Heuristiken. Gute Ergebnisse erzielen z. B. Wortbedeutungsfestlegungen aufgrund der häufigsten Verwendung¹⁰⁰ (einer Bedeutung), aufgrund der

⁹⁷Eine *lexikalische Kette* ist eine Sequenz semantisch verwandter Wörter: z. B. Rome → capital → city → inhabitant.

⁹⁸Engl. *selectional preferences*.

⁹⁹Im Speziellen dreht sich RESNIK [54] um Verben und die semantische Klasse deren Hauptwort-Argumente.

¹⁰⁰Engl. *most-frequent-sense heuristic*.

Tendenz, dass ein Wort innerhalb eines Diskurses¹⁰¹ bzw. einer Kollokation¹⁰² bzw. einer Domäne die selbe Bedeutung hat. Diese Algorithmen werden häufig als 'baseline', als Richtlinie im Vergleich der Auswertungen mehrerer WSD-Systeme verwendet.

5.2 Überwachte Methoden

Der *überwachte*¹⁰³ Ansatz zur WSD ist die automatische Herleitung von Regeln oder Modellen zur Klassifikation aus einem semantisch annotierten Corpus. Das Ziel von überwachtem (maschinell) Lernen zur Klassifizierung besteht darin, aus einer Trainingsmenge S eine Hypothese h einer unbekannt Funktion f herzuleiten, die dem Eingaberaum X einen diskreten ungeordneten Ausgaberaum $Y = \{1, \dots, K\}$ zuordnet.

Die Trainingsmenge besteht aus m Paaren¹⁰⁴ $S = \{(x^1, y^1), \dots, (x^m, y^m)\}$ mit $x \in X$ und $y = f(x)$. Die x -Komponente von jedem Tupel ist üblicherweise ein Vektor $x = (x_1, \dots, x_n)$, dessen Komponenten Attribute oder *features* heißen, diskret oder reelwertig sind und die relevanten Informationen des Beispiels beschreiben. Die Werte des Ausgaberaumes Y , die mit jedem Trainingsbeispiel assoziiert werden, heißen Klassen bzw. *classes* oder Kategorien. Das folgende Beispiel von MÀRQUEZ U.A. in AGIRRE [1, Kap. 7.1] verdeutlicht das Vorgehen am Beispiel der Disambiguierung des Verbs 'to know' in dem Satz 'there is nothing in the whole range of human experience more widely known and universally felt than spirit':

Die Bedeutungen von 'know' sind die Klassen des Klassifizierungsproblems, die den Ausgaberaum Y definieren. Jedes Vorkommen des Wortes in einem Corpus wird in ein Trainings-Beispiel (x^i) , das mit den korrekten Bedeutungen annotiert ist, kodiert. (Die nachfolgenden Bedeutungen sind aus WordNet, welches in Version 3.0 elf Bedeutungen auflistet, entnommen und sind die Bedeutungen, welche positive log-likelihood-Werte in Tabelle 1 erhalten haben.)

know

1. know, cognize, cognise (be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about) 'I know that the President lied to the people'; 'I want to know who is winning the game!'; 'I know it's time'
2. know (be familiar or acquainted with a person or an object) 'She doesn't

¹⁰¹Engl. *one-sense-per-discourse heuristic*.

¹⁰²Engl. *one-sense-per-collocation heuristic*.

¹⁰³Engl. *supervised methods*.

¹⁰⁴Trainings-Beispielen.

know this composer'; 'Do you know my sister?'; 'We know this movie'; 'I know him under a different name'; 'This flower is known as a Peruvian Lily'

Eine Entscheidungsliste oder *decision list* (DL) ist ein einfacher Lern-Algorithmus, der zur WSD angewendet werden kann. Er akquiriert eine Liste von Klassifizierungsmerkmalen der Form¹⁰⁵:

if (feature = value) then class

Im Wesentlichen kann eine DL als gewichtete Liste (*Bedingung, Klasse, Gewicht*) dieser *if-then-else Regeln* angesehen werden. Wenn nun ein neues Beispiel x klassifiziert wird, wird die Liste von Regeln der Reihe nach überprüft und der erste Treffer angewendet. Angenommen, dass solch eine Liste von Klassifizierungsregeln aus den Trainings Beispielen erzeugt wurde, enthält Tabelle 1 alle Regeln, die zu dem Beispielsatz passen.

(Nur Einträge mit positivem log-likelihood sind aufgelistet):

Feature	Value	Sense	Log-likelihood
+3-word-window	'widely'	2	2,99
word-bigram	'known widely'	2	2,99
word-bigram	'known and'	2	1,09
sentence-window	'whole'	1	0,91
sentence-window	'widely'	2	0,69
sentence-window	'known'	2	0,43

Tabelle 1: Klassifizierungs-Beispiel von 'know' durch *decision list*.

Wird das Beispiel durch die ersten beiden Regeln, welche durch die direkte Nachbarschaft von 'know' und 'widely' aktiviert werden, klassifiziert, so wird die Bedeutung $know_2$ zugewiesen.

¹⁰⁵Vgl. MÀRQUEZ et al. in AGIRRE [1, S. 170].

5.2.1 Probabilistische Methoden

Statistische Methoden schätzen üblicherweise eine Menge probabilistischer Parameter, welche die Wahrscheinlichkeitsverteilungen von Kategorien und Kontexten durch *features* ausdrücken. Diese Parameter können dann benutzt werden, um jedem neuen Beispiel die Kategorie zuzuweisen, die die bedingte Wahrscheinlichkeit einer Kategorie bezüglich des beobachteten Kontext-*features* maximiert. Der Naïve Bayes (NB) Algorithmus ist der einfachste dieser Art¹⁰⁶. In diesem Modell wird ein Beispiel dadurch 'generiert', dass zuerst die Bedeutung s stochastisch gewählt wird und dann jedes der *features* unabhängig, entsprechend der jeweiligen Verteilung $P(x_i|s)$.

Die Klassifikationsregel eines neuen Beispiels $x = (x_1, \dots, x_m)$ besteht daraus, die Bedeutung s zuzuweisen, welche die bedingte Wahrscheinlichkeit der Bedeutung bezüglich der beobachteten *features* maximiert. Dies drückt sich in nachfolgender Ungleichung aus:

$$\arg \max_s P(s|x_1, \dots, x_m) = \arg \max_s \frac{P(x_1, \dots, x_m|s)P(s)}{P(x_1, \dots, x_m)} \quad (5.6)$$

$$\geq \arg \max_s P(s) \prod_{i=1}^m P(x_i|s) \quad (5.7)$$

Gleichung (5.6) ist der Satz von Bayes, während die Faktorisierung auf der Annahme der Unabhängigkeit beruht: $P(x_i|s, x_{j \neq i}) = P(x_i|s)$. Die probabilistischen Parameter $P(s)$ und $P(x_1|s)$ können durch relative Häufigkeitszählung¹⁰⁷ aus der Trainingsmenge geschätzt werden. Die Anfangswahrscheinlichkeit der Bedeutung s , $P(s)$ wird als Verhältnis der Anzahl von Beispielen mit der Bedeutung s und der Anzahl aller Beispiele geschätzt. $P(x_i|s)$ ist die Wahrscheinlichkeit, das *feature* x_1 zu beobachten, vorausgesetzt die beobachtete Bedeutung ist s . In diesem Fall ist die MLE die Anzahl der Beispiele mit Bedeutung s mit aktivem feature x_i , geteilt durch die Anzahl aller Beispiele mit Bedeutung s .

¹⁰⁶Siehe AGIRRE [1, S. 185].

¹⁰⁷Z. B. *maximum likelihood estimation*, MLE.

5.2.2 Methoden, die auf der Similarität der Beispiele beruhen

Der am weitesten verbreitete¹⁰⁸ Repräsentant dieser Methode ist der k-Nearest Neighbor Algorithmus (kNN). Die Klassifikation eines neuen Beispiels wird dadurch gewonnen, dass die Menge der k ähnlichsten Beispiele (*nearest neighbors*) aus einer gespeicherten Menge semantisch annotierter Beispiele gefunden und ein 'Durchschnitt' ihrer Bedeutungen berechnet wird, um die Vorhersage der entsprechenden Bedeutung zu treffen.

Um die Menge der *nearest neighbors* zu erhalten, wird das zu klassifizierende Beispiel $x = (x_1, \dots, x_m)$ mit jedem gespeicherten Beispiel $x^i = (x_1^i, \dots, x_m^i)$ verglichen und deren Entfernung wird berechnet. Das gängigste Entfernungsmaß ist das *Überlappungsmaß*¹⁰⁹ in Gleichung (5.9), wo w_j das Gewicht des j -ten *features* und $\delta(x_j, x_j^i)$ die Entfernung zwischen zwei Werten ist, mit der Fallunterscheidung in Gleichung (5.8).

$$\delta(x_j, x_j^i) = \begin{cases} 0, & \text{wenn } x_j = x_j^i, \\ 1, & \text{sonst.} \end{cases} \quad (5.8)$$

$$\Delta(x, x^i) = \sum_{j=1}^m w_j \delta(x_j, x_j^i) \quad (5.9)$$

5.2.3 Methoden, die auf Regel-Kombinationen beruhen

Ein sehr erfolgreicher Algorithmus dieser Methoden ist AdaBoost (AB), der viele einfache und möglicherweise nicht sehr akkurate Klassifizierungsregeln (*weak rules*) linear zu einem starken Klassifizierer kombiniert, mit einer auf der Trainingsmenge geringen Fehlerquote. Die *weak rules* werden sequentiell, eine pro Zeitpunkt, gelernt. Konzeptionell wird die *weak rule* bei jedem Durchlauf darauf ausgerichtet, die Beispiele zu klassifizieren, welche durch das Zusammenspiel vorhergehender *weak rules* am schwierigsten zu klassifizieren waren. AdaBoost unterhält einen Vektor von Gewichten als Verteilung D_1 über Beispielen. In Runde t ist die Aufgabe des Lernalgorithmus, eine *weak rule*, $h_1: X \rightarrow \mathfrak{A}$ zu finden, mit verhältnismäßig kleiner Fehlerquote bezüglich der Gewichtsverteilung D_1 . In dieser Abstimmung machen *weak rules* reelwertige Voraussagen. Anfänglich ist D_1 einheitlich, aber nach jeder Iteration vergrößert (verkleinert) AB exponentiell die Ge-

¹⁰⁸Vgl. AGIRRE [1, S. 176].

¹⁰⁹Auch *Hamming Distance* genannt.

wichte $D_t(i)$, für welche $h_t(x^i)$ ein gute (schlechte) Voraussage trifft, mit einer Variation, die proportional zum 'Vertrauen' $|h_t(x^i)|$ ist. Die letzte kombinierte *weak rule* $h_t: X \rightarrow \mathfrak{R}$ berechnet Ihre Voraussagen durch eine gewichtete Wahl der *weak rules*:

$$f(x) = \sum_{t=1}^T \alpha_t \cdot h_t(x) \quad (5.10)$$

Für jedes Beispiel x wird das Vorzeichen von $f(x)$ als die vorhergesagte Klasse interpretiert¹¹⁰. Der Betrag $|f(x)|$ wird als Maß des Vertrauens in die Vorhersage interpretiert.

5.2.4 Kernel-basierte Verfahren

Das Gebiet der linearen Klassifizierer, welches im Bereich der WSD nicht zu überzeugen vermochte¹¹¹, kann in seiner Ausdrucksstärke dadurch erweitert werden, dass man das Lernen von nicht linearen Funktionen erlaubt. Diese nicht-linearen Funktionen sind ein nicht lineares *mapping* der Eingabe-*features* auf einen n -dimensionalen *feature*-Raum, in welchem neue *features* dadurch ausgedrückt werden, dass grundlegende *features* kombiniert werden und dann lineares Lernen durchgeführt wird. Wenn Beispiel-Vektoren nur innerhalb von Skalarprodukt-Operationen im Lern-Algorithmus und der Klassifizierungsregel vorkommen, dann kann das nicht-lineare Lernen mit sogenannten *kernel functions* durchgeführt werden. Diese *kernel functions* bieten eine sehr bequeme Möglichkeit, anwendungsspezifische *kernels* zu definieren, um die Eigenheiten der Daten auszuschöpfen und Hintergrundinformationen einzuführen.

Support Vector Machines (*SVM*) ist die bekannteste *kernel*-Methode. Der Lern-Algorithmus besteht darin, die Hyperebene zu wählen, welche die positiven Beispiele von den negativen Beispielen mit maximalem Abstand (*margin*) trennt. In der einfachsten Form der SVM, wird der lineare Klassifizierer durch zwei Elemente definiert: a) der Gewichtsvektor w , der pro *feature* eine Komponente hat und b) ein Ausrichtung b , welche für die Entfernung der Hyperebene vom Ursprung steht. Die Klassifizierungs-Regel weist einem neuen Beispiel x die Werte $+1$ oder -1 zu, wie folgt:

$$h(x) = \begin{cases} +1, & \text{wenn } (w \cdot x) + b \geq 0, \\ -1, & \text{sonst.} \end{cases} \quad (5.11)$$

¹¹⁰Der einfache AB arbeitet nur mit binären Ausgaben (-1 und +1). ¹¹¹Siehe MÀRQUEZ in AGIRRE [1, S. 180].

Die positiven und negativen Beispiele, die der (w, b) Hyperebene am nächsten sind, heißen *support vectors*. Das Lernen der maximalen Abstand (*margin*) Hyperebene (w, b) , ist letztlich ein konvexes quadratisches Optimierungsproblem mit einer eindeutigen Lösung:

Minimiere $\|w\|$ bezüglich der Bedingungen¹¹² $y_i[(w \cdot x_i) + b] \geq 1$, die anzeigen, dass alle Trainings-Beispiele klassifiziert wurden, mit einem Abstand (*margin*) größer als 1.

5.2.5 Empirische Auswertung von NB,kNN,DI,AB,SVM auf dem DSO-Corpus

In diesem Kapitel werden kurz die Ergebnisse der Tests, die MÀRQUEZ U.A. in AGIRRE [1] vorstellen, besprochen. Sie testeten die oben erwähnten Algorithmen auf dem DSO-Corpus. 13 Nomina (*age, art, body, car, child, cost, head, interest, line, point, state, thing, work*) und 8 Verben (*become, fall, grow, lose, set, speak, strike, tell*) wurden getestet und als unabhängige Klassifikations-Probleme behandelt. Für die Disambiguierung wurde *local* und *topical context* verwendet. Fünfzehn *feature*-Muster, welche teilweise¹¹³ für AB und SVM in eine binäre Form kodiert wurden, kamen zum Einsatz: Es sei $[w_{-3}, w_{-2}, w_{-1}, w, w_{+1}, w_{+2}, w_{+3}]$ der Kontext aufeinanderfolgender Wörter um das Zielwort w und p_i , $-3 \leq i \leq 3$, der POS-Tag des Wortes w_i . Dann gibt es die 15 *features* bezüglich des *local context*: $p_{-3}, p_{-2}, p_1, p_{+1}, p_{+2}, p_{+3}, w_{-1}, w_{+1}, (w_{-2}, w_{-1}), (w_{-1}, w_{+1}), (w_{+1}, w_{+2}), (w_{-3}, w_{-2}, w_{-1}), (w_{-2}, w_{-1}, w_{+1}), (w_{-1}, w_{+1}, w_{+2})$ und (w_{+1}, w_{+2}, w_{+3}) . Der *topical context* wird durch die Menge der *bag of words* $\{c_1, \dots, c_m\}$ geleistet, welche eine ungeordnete Menge von m Wörtern offener Klasse ist, die in dem Satz vorkommen. Durch die Kodierung in eine binäre Form entstanden mehrere Tausend *features*.

Die 10-fache Kreuzvalidierung zeigte klar, dass alle Methoden die Most-Frequent-Sense-Classifer-Kennlinie (ca. 46,5%) übertrafen. Die besten Ergebnisse lieferten SVM und AB mit einer Genauigkeit von ca. 66,5%. Im Mittelfeld war kNN mit ca. 63,5% und die schlechtesten Ergebnisse erzielten NB und DL mit ca. 61,4%. Die ungewohnt schlechten Ergebnisse von DL führen MÀRQUEZ U.A. auch auf die verwendete DL-Variante *simple smoothing* zurück. Generell sind die Gesamtergebnisse (61%–67%) dieser Versuchsreihe nicht besonders beeindruckend, allerdings sind die hier Verwendeten DSO-Wörter unter den am meisten polysemen Wörtern des Englischen und die WordNet-Bedeutungen sind sehr feinkörnig.

¹¹²Eine für jedes Trainings-Beispiel.

¹¹³Die *local context features*. Die *topical context features* blieben binäre Tests auf An- bzw. Abwesenheit von bestimmten Worten im Satz.

5.3 Un- und semi-überwachte Methoden

Aufgrund des Flaschenhalses der Wissensakquise in Form manueller Bedeutungsannotation, erfreuen sich die unüberwachten bzw. semi-überwachten Methoden in der CL immer größerer Beliebtheit. Weil für die WSD der Sanskrit-Begriffe, um welche es in dieser Arbeit geht, nur überwachte Methoden verwendet werden, sollen diese Methoden hier jedoch nur kurz erwähnt werden.

Zwei Hauptansätze zu den überwachten Methoden der WSD finden (auch unter dem Begriff der *Word Sense Discrimination*) Verwendung: i) *distributional methods* und ii) *translational-equivalence methods*. Version i) identifiziert Wörter in ähnlichen Kontexten ohne Rücksicht auf vorhandene Bedeutungs-Inventare. In diesem Sinne möglich¹¹⁴ ist eine Betrachtung dieser Methode der WSD als zweiteiligen Prozess: Im ersten Schritt wird unter den verschiedenen Bedeutungen eines Zielwortes unterschieden, indem die Kontexte in welchen es auftritt, in Cluster aufgeteilt werden, die verteilte Charakteristika gemeinsam haben. Dann wird jeder Cluster mit einer Glosse annotiert, welche die zugrundeliegende Bedeutung des Zielwortes in diesen Kontexten beschreibt. Diese Methode ist in der Tat ein gänzlich anderer Ansatz zu sonstigen Methoden der WSD, in welchen immer eine Art der Annotation als vorausgesetzt angesehen wird. Aufgrund verschiedener Schwierigkeiten¹¹⁵ des zweiten Schrittes dieser Interpretation der WSD, werden meist Kompromisse eingegangen und die Kontext-Cluster, welche von den *distributional methods* gefunden werden, müssen durch Verwendung von Informationen aus existenten wissensreichen Ressourcen annotiert werden. Methoden der Version ii), die *translational-equivalence methods*, haben Potential Wortbedeutungsunterscheidungen mit Fokus auf maschinelle Übersetzung zu machen. Hier werden im Wesentlichen die Wortbedeutungen dadurch hergeleitet, dass jedes Vorkommen eines Zielwortes mit der passenden Übersetzung aus dem parallelen Corpus der zweiten Sprache annotiert wird. Am Ende erhält man einen 'bedeutungsannotierten' Corpus, dessen Tags die Übersetzungsäquivalenzen der Zielwörter in diesem Kontext sind. Bekannte Algorithmen sind hier HAL, LSA und CBC.

Einer der ersten Ansätze zur semi-überwachten WSD ist der *Bootstrapping* Algorithmus von Yarowsky. Dieser iterative und inkrementelle Algorithmus arbeitet auf einer kleinen *seed*-Menge annotierter Beispiele, welche repräsentativ für die jeweiligen Bedeutungen sind, auf einer großen Menge von Beispielen, die klassifiziert werden müssen

¹¹⁴Siehe PEDERSEN in AGIRRE [1, S. 137].

¹¹⁵Siehe PEDERSEN in AGIRRE [1, Kap. 6].

und auf einem überwachten Lern-Algorithmus¹¹⁶. Initial wird der Lern-Algorithmus mit der *seed*-Menge trainiert und klassifiziert dann die große Menge von unannotierten Beispielen. Nur die Beispiele, die innerhalb eines gewissen Konfidenzintervalls klassifiziert werden, bleiben als Zusatz-Beispiele für den nächsten Durchlauf erhalten. Der Algorithmus läuft solange, bis von einem zum nächsten Durchlauf keine Veränderungen mehr feststellbar sind.

5.4 Evaluation von WSD-Systemen

Die Aufgabe, verschiedene WSD-Systeme miteinander zu vergleichen, drängt sich aufgrund der Fülle an Methoden und Ansätzen fast auf. Die Fragestellung liegt auf der Hand: Welche Methode eignet sich am besten um die automatische Unterscheidung der möglichen Bedeutungen eines Wortes in einem gegebenen Kontext zu vollbringen? Aufgrund der Verschiedenheit der Ansätze und einzelnen Algorithmen, lohnt es sich, ein wenig Licht auf die Terminologie im Zuge der Auswertung von WSD-System zu werfen und gängige Vergleichsverfahren aufzuführen.

Das **Bedeutungsinventar**¹¹⁷ nimmt eine entscheidende Rolle im Design von WSD-Systemen ein. Als ein *berechenbares Lexikon* bzw. *maschinen-lesbares* Wörterbuch, sind hier idealerweise für jedes Wort eindeutige und konsistente Bedeutungen hinterlegt. Bei der Wahl des Bedeutungsinventars sind Konsistenz, Granularität und Aufbau des Inventars miteinzubeziehen. Insbesondere im Vergleich zweier oder mehrerer Systeme sollten diese Faktoren bedacht werden.

Hinsichtlich der **Anwendungsabhängigkeit** von WSD-System werden zwei Typen der Evaluation unterschieden. Bei *in-vitro*-Auswertungen, wird das WSD-System unabhängig von einem Anwendungsgebiet, quasi in künstlicher Umgebung getestet. In der *in-vivo*-Auswertung wird der Nutzen des WSD-Systems innerhalb oder für eine Anwendung gemessen. Da das Ziel dieser Arbeit eher genereller Natur ist, wird hier nicht näher auf die Bindung an bestimmte CL-Aufgaben eingegangen. Ein WSD-System kann entweder auf eine *kleine Menge von Beispiel-Wörtern*¹¹⁸ oder auf *alle Wörter*¹¹⁹ in einem spezifischen Kontext angesetzt werden. Die *all-words-task* stellt natürlich besondere Anforderungen an das zugrundeliegende Bedeutungsinventar, welches alle vorkommenden Wörter auch beinhalten sollte und an die Art der Annotation. In der *lexical-sample-task*

¹¹⁶In diesem Fall meist mit DLs.

¹¹⁷Engl. *sense inventory*.

¹¹⁸Engl. *lexical sample*.

¹¹⁹Engl. *all-words*.

wird eine Beispiel-Menge zu unterscheidender Wörter aus einem Lexikon gewählt, zusammen mit einer Anzahl von Corpus-Instanzen für jedes Wort. Diese Beispiel-Wörter müssen vom WSD-System in kurzen Text-Ausschnitten mit Bedeutungen versehen werden.

Das **Corpus** für die Beispiel-Wörter ist üblicherweise eine größere Menge von natürlichen Sätzen, in welchen die zu disambiguierenden Beispiel-Wörter enthalten und mit einem Zeiger auf eine Bedeutung im *sense inventory* annotiert sind. Ein Teil der annotierten Daten kann für **überwachte Lern-Methoden** als Trainings-Menge zurückbehalten werden, während ein anderer Teil als Vergleichsmenge zum Testen verwendet wird, welche als 'Goldstandard' erachtet werden kann. Um beste Resultate zu erzielen, werden große Verhältnisse von 5 : 1 oder 10 : 1 zwischen Trainings-Menge und Ziel-Menge verwendet; nach PALMER U.A. in AGIRRE [1, S. 76] wäre jedoch ein Verhältnis von 2 : 1 realistischer.

Ein einfaches Kriterium, ein WSD-System auf ein bestimmtest Test-Beispiel auszuwerten (**Scoring**), ist das Kriterium der exakten Übereinstimmung¹²⁰. In diesem Fall erhält das System für die Zuweisung einer Bedeutung den Wert 1, genau dann wenn die Bedeutung exakt mit der korrekten Bedeutung übereinstimmt, sonst den Wert 0. Wenn ein System einer Instanz w mehrere Bedeutungen (mit den jeweils verbundenen Wahrscheinlichkeiten) zuweist, kann der *Score* nach PALMER U.A. in AGIRRE [1, S. 78] als die Wahrscheinlichkeit berechnet werden, die es der korrekten Bedeutung c zuweist, unter Voraussetzung von w und seinem Kontext $context(w)$:

$$Score = P(c|w, context(w)) \quad (5.12)$$

Wenn ein Beispiel mehr als eine korrekte Bedeutung hat, ist der Score schlicht als Summe der Wahrscheinlichkeiten der einzelnen korrekt zugewiesenen Bedeutungen zu berechnen:

$$Score = \sum_{t=1}^C P(c_t|w, context(w)) \quad (5.13)$$

Bei hierarchischen Bedeutungsinventaren wird zwischen drei Ebenen des *Scoring* unterschieden: fein-, grob- und gemischtkörnig. Bei feiner Körnung treffen nur identische Bedeutungen, bei grober Körnung werden alle Bedeutungen des 'Goldstandards' und der System-Zuweisung auf die jeweilige *top-level*-Bedeutung abgebildet und das System

¹²⁰Engl. *exact-match criterion*.

erhält einen *Score* von 1, wenn die Zuweisung die selbe *top-level*-Bedeutung wie die korrekte Bedeutung hat. Gemischtkörniges *Scoring* findet Verwendung bei hierarchischen Inventaren in Verbindung mit Systemen die irgend eine Bedeutung der selben Hierarchie zuweisen dürfen¹²¹. Bedeutungen sind in einem Baum strukturiert und jede zugewiesene Bedeutung produziert dann einen *Score* von 1, wenn sie ein Nachkomme (*child*) der korrekten Bedeutung ist. Wenn eine Bedeutung zugewiesen wurde, die Vorfahre (*parent*) der korrekten Bedeutung ist, wird der *Score* durch die Annahme berechnet, dass die Wahrscheinlichkeit eines Vorfahren einheitlich über seine Nachkommen verteilt ist: der $Score_{mixed_anc}$ für den zugewiesenen Vorfahren c_{anc} ist dann einfach die bedingte Wahrscheinlichkeit des korrekten Nachkommens c_{desc} mit vorausgesetztem Vorfahren:

$$Score_{mixed_anc} = P(c_{desc}|c_{anc}) \quad (5.14)$$

Die **Abdeckung**¹²² eines Systems beschreibt den Prozentsatz von Beispielen aus der Auswertungs-Menge, für welche das System eine Bedeutung zuweist. Die **Präzision**¹²³ wird berechnet, wenn man die *Scores* aller Beispiele summiert, für die das System eine Bedeutung zuweist und durch die Anzahl der annotierten Beispiele teilt. Die **Trefferquote**¹²⁴ wird berechnet, indem die Summe aller *Scores* durch die Anzahl der Beispiele geteilt wird¹²⁵. Im Falle der Aufgabe der Bedeutungs-zuweisung wird die Trefferquote mit **Exaktheit**¹²⁶ referenziert, da hier für jedes Beispiel ein Bedeutung zugewiesen werden soll und somit *precision = recall* gilt.

Zwei weitere wichtige Begriffe sind **untere**¹²⁷ und **obere Schranke**¹²⁸. Die untere Schranke stellt einen unteren Grenzwert dar, der mit geringen Kosten berechnet werden kann¹²⁹ und nützlich ist, um zu erkennen ob es von statistischem Wert ist, den Mehraufwand eines komplexeren Systems überhaupt einzugehen. Eine obere Schranke die meist verwendet wird¹³⁰, ist z. B. die menschliche *ITA*, die berechnet wird, indem man zählt wie oft die selben Bedeutungen durch zwei oder mehrere Menschen zugewiesen werden, denen die selben Richtlinien zur Bedeutungs-Annotierung gegeben wurden.

¹²¹Bzw., wenn Menschen irgend eine Bedeutung der selben Hierarchie zuweisen dürfen. Vgl. MELAMED UND RESNIK [46].

¹²²Engl. *coverage*.

¹²³Engl. *precision*.

¹²⁴Engl. *recall*.

¹²⁵Wenn für ein Beispiel keine Bedeutung zugewiesen wurde, wird hier einfach ein *Score* von 0

verwendet.

¹²⁶Engl. *accuracy*.

¹²⁷Engl. *lower bound*.

¹²⁸Engl. *upper bound*.

¹²⁹Z. B. *most-frequent-sense* oder *LESK*.

¹³⁰Zu *kappa coefficient* und *replicability* siehe PALMER U.A. in AGIRRE [1, Kap. 4].

6 Der SanskritSemAnnotator

Um eine statistisch verwertbare Menge an semantisch annotierten Sanskrit-Lexemen zu erhalten, wurde ein Multi-Annotatoren-System entwickelt, welches beliebig vielen Annotatoren ermöglicht, über das Internet die zur Annotation zur Verfügung stehenden Lexeme in einem bestimmten Kontext mit Bedeutungen zu versehen, mit genau einer Bedeutung pro Lexem und Vorkommen. In diesem Kapitel wird der Aufbau dieses Systems mit den zugrundeliegenden Datenbank-Tabellen skizziert. Es ist anzumerken, dass aus Gründen der Einfachheit, Fremdschlüssel, die eigentlich auf Attribute der Tabellen der Mutter-DB des DCS zeigen müssten, nicht angegeben wurden, um nicht alle referenzierten Tabellen besprechen zu müssen, die ja für das Verständnis von SanSemAn nicht notwendig sind.

6.1 Tabellen-Schemata

Das Gerüst des [SanSemAn](#) arbeitet auf vier Tabellen, welche sich in einer MySQL-Datenbank befinden, die wiederum auf der dem DCS-Projekt zugrundeliegenden Datenbank beruht und mit Daten aus dieser gefüllt ist. Im Folgenden sind die Schemata der vier Relationen als SQL-Kommando, sowie eine kurze Beschreibung der darin definierten Attribute (Spalten) gegeben:

1. **ontology**

```

1 CREATE TABLE ontology (
2   IDWortlisteOpencyc int(11) NOT NULL,
3   Word varchar(255) NOT NULL,
4   Annotation varchar(255) NOT NULL,
5   IDWortliste int(11) NOT NULL,
6   WordWortliste varchar(300) NOT NULL,
7   PRIMARY KEY (IDWortlisteOpencyc),
8   KEY (IDWortliste)
9 );
```

Schema 2: **ontology**

In der Tabelle **ontology** sind alle möglichen Bedeutungen, die für ein bestimmtes Lexem auftreten können, gespeichert. Das Attribut **IDWortlisteOpencyc** ist Primärschlüssel¹³¹

¹³¹**IDWortlisteOpencyc** zeigt eigentlich auf die Tabelle **WortlisteOpencyc** der Mutter-DB.

und repräsentiert eine Bedeutung aus dem Bedeutungs-Inventar [OpenCyc](#), **Word** ist ein charakteristischer Überbegriff der jeweiligen Bedeutung, welche in **Annotation** genauer spezifiziert wird. **IDWortliste** ist eine numerische ID, welche jedem Sanskrit-Lexem zugewiesen ist, das in Umschrift in HTML-Notation durch das Attribut **WordWortliste** repräsentiert wird¹³².

2. references

```

1 CREATE TABLE 'references' (
2   ID int(11) NOT NULL AUTO_INCREMENT,
3   IDTEA int(11) NOT NULL,
4   Text varchar(500) NOT NULL,
5   IDW int(11) NOT NULL,
6   PRIMARY KEY (ID),
7   KEY (IDW, IDTEA)
8 );
```

Schema 3: **references**

Die Relation **references**¹³³ beinhaltet pro Attribut-Tupel (Datensatz) genau ein Vorkommen eines bestimmten Sanskrit-Lexems, welches in dem Attribut **IDW** mit seiner einzigartigen ID gespeichert wird¹³⁴. Der Primärschlüssel **ID** wird in der späteren Anwendung zur Darstellung der jeweiligen zu annotierenden Textstelle verwendet und ist ein fortlaufender Index. In **Text** ist der Kontext des Vorkommens in HTML gespeichert. Das eigentliche Vorkommen ist per HTML-Tag mit gelbem Hintergrund für die Darstellung vorbereitet. **IDTEA** dient zur Verknüpfung mit der Tabelle, welche die von den Annotatoren zugewiesenen Bedeutungen speichert. Der Verbund-Schlüssel (**IDW, IDTEA**) stellt sicher, dass pro Textstelle und Lexem nur ein Tupel vorhanden ist.

3. ontology_reference

```

1 CREATE TABLE ontology_reference (
2   ID int(11) NOT NULL,
3   IDTEA int(11) NOT NULL,
4   IDUser int(11) NOT NULL,
5   IDWortlisteOpencyc int(11) NOT NULL,
```

¹³²Die Attribute **IDWortliste** und **WordWortliste** sind eigentlich Fremdschlüssel, die auf eine Wortliste der Mutter-DB zeigen.

¹³³Hier wird der Name in Hochkommata angegeben,

weil es sonst zu Problemen mit dem SQL-Schlüsselwort **references** kommt.

¹³⁴**IDW** müsste als Fremdschlüssel auf die Wortliste der Mutter-DB zeigen.

```

6      FOREIGN KEY (IDTEA) REFERENCES 'references'(IDTEA)
7              ON DELETE CASCADE ON UPDATE CASCADE,
8      FOREIGN KEY (IDUser) REFERENCES users(ID)
9              ON DELETE CASCADE ON UPDATE CASCADE,
10     PRIMARY KEY (ID, IDUser)
11 );

```

Schema 4: `ontology_reference`

In `ontology_reference` werden die zugewiesenen Bedeutungen gespeichert. In der aktuellen Version der Web-Anwendung SanSemAn entsprechen die Werte des Attributs `ID` den Werten des Attributs `ID` der Tabelle `references`¹³⁵. `IDTEA` dient ebenfalls der Verknüpfung von `ontology_reference` und `references`. `IDWortlisteOpencyc` speichert die zugewiesene Bedeutung aus der Menge der möglichen Bedeutungen und der Primärschlüssel (`ID, IDUser`) stellt sicher, dass der Annotator, mit der in `IDUser` referenzierten ID aus der Tabelle `users`, nur eine Zuweisung pro Lexem und Vorkommen machen kann¹³⁶.

4. `users`

```

1 CREATE TABLE users (
2     ID int(11) NOT NULL auto_increment,
3     Name varchar(255) NOT NULL,
4     ShortName varchar(50) NOT NULL,
5     NextAnnotationSardula int(11) NOT NULL default '0',
6     NextAnnotationJana int(11) NOT NULL default '0',
7     PRIMARY KEY ('ID')
8 );

```

Schema 5: `users`

In `users` sind die zur Annotation zugelassenen Benutzer hinterlegt. Das Attribut `Name` speichert den Anzeigenamen des Benutzers, `ShortName` den für den späteren Login benötigten Benutzernamen. Die aktuelle Version von SanSemAn verwendet im Umfang

¹³⁵Auf die Bedingung eines Fremdschlüssels wurde jedoch absichtlich verzichtet, um bei Bedarf die Verknüpfung der beiden Tabellen auch über `IDTEA` als Fremdschlüssel zu lösen, und somit die notwendige Übereinstimmung der Ordnung der IDs zwischen den beiden Tabellen zu um-

gehen.

¹³⁶In einem echten System wäre die Zuweisung mehrerer Bedeutungen jedoch wünschenswert. Aus Gründen der Einfachheit wird dies hier aber unterbunden. Vgl. auch Seite 38 in Kapitel 5.4 und Seite 57 in Kapitel 7.3.

dieser Arbeit nur zwei Lexeme: *jana* und *śārdūla*. Aus diesem Grund sind entsprechend die Attribute **NextAnnotation(*)** vorhanden, in welchen statisch die Position (**ID** aus **references**) des nächsten zu annotierenden Vorkommens gespeichert wird. Sollte SanSemAn tatsächlich in einem größeren Umfang zum Einsatz kommen, sind diese 'statischen' Attribute in eine Tabelle der Form **positions (IDUser, IDW, POS)** zu überführen und die Anwendungslogik ist entsprechend anzupassen.

6.2 Die Benutzer-Schnittstelle

Um einer möglichst breiten Schicht von Anwendern eine plattformunabhängige Anwendung zur Annotation der ambigen Sanskrit-Lexeme bereitzustellen, bietet das Internet natürlich optimale Voraussetzungen. Die Entwicklung des SanSemAn wurde daher in der weitverbreiteten Sprache PHP durchgeführt. Mit der Möglichkeit der einfachen Vermischung serverseitiger Skript-Befehle und HTML-Tags und der guten Implementation von MySQL-Befehlen, ist PHP ein ideales Mittel zur Umsetzung dieses Systems.

Auf der Startseite **index.php**, die im Internet¹³⁷ aufzurufen ist, besteht die Möglichkeit sich als Annotator mit einem vorher eingerichteten Benutzernamen anzumelden, ist der eingegebene Name vorhanden, leitet die Seite zu **aannotate.php** weiter. Der Anmeldevorgang ist SESSION-basiert, d. h. auf dem Server wird die Verwendung der globalen **\$_SESSION** Variablen freigeschaltet, hierin werden die vorher aus der DB abgefragten relevanten Information gespeichert und sind somit zu jeder Zeit für das System verfügbar, solange bis ein TIMEOUT oder das Skript **logout.php** die **SESSION** zerstört und somit den Zugriff auf geschützte Seiten verhindert. Dass eine Seite (in diesem Fall ein Skript) nach außen geschützt ist, wird durch Einbindung des Skriptes **auth.php** sicher gestellt, das im Falle entsprechender, nicht gesetzter **SESSION**-Variablen auf die Startseite weiterleitet.

In einem im linken Bereich der Benutzeroberfläche befindlichen Menü hat der Benutzer jederzeit die Möglichkeit zwischen den Aktionen *Annotieren*, *Vergeichen* und *Logout* zu wählen. Entscheidet man sich zu annotieren, wird das zu annotierende Lexem abgefragt und die Stelle, an der fortgefahren werden soll, wird aus der DB ermittelt. Dann zeigt **annotate.php** die entsprechende Textstelle mit gelb hinterlegtem Ziel-Lexem an und stellt in einer Tabelle die **3** von Benutzer mit ID 1 am häufigsten zugewiesenen Bedeutungen dar. Mit Hilfe von in der rechten Spalte befindlichen Radio-Buttons, kann sich

¹³⁷Stand August 2011: <http://sanseman.asyavamasya.com>.

der Benutzer für eine Bedeutung entscheiden. Hat er schon vorher für diese Textstelle und dieses Lexem eine Bedeutung zugewiesen, wird die entsprechende Bedeutung in der Tabelle mit gelbem Hintergrund dargestellt. Klickt er den Knopf 'Bedeutung zuweisen!' wird die durch den Radio-Button ausgewählte Bedeutung und weitere relevante Daten an **save.php** übergeben. **save.php** fügt (**insert**) dann die Bedeutung, die Benutzer-ID, die ID des Lexems und den Wert der oben erwähnten IDTEA in die Tabelle **ontology_reference** ein, sofern kein entsprechender Datensatz vorhanden ist, ansonsten wird die Aktion **update** analog ausgeführt. Wurde keine Bedeutung übergeben, wird der Wert auf **NULL** gesetzt. Anschließend inkrementiert **save.php** den Eintrag **NextAnnotation(*)** des aktuellen Benutzers und leitet zu **annotate.php** zurück.

Möchte der Benutzer vergleichen, wird erneut das zu vergleichende Lexem abgefragt. Nun berechnet **compare.php** für welche Vorkommen des Lexems unterschiedliche Bedeutungen zugewiesen wurden, sollten von einem Benutzer noch keine Zuweisungen gemacht worden sein, werden diese ignoriert. Anschließend werden die Positionen (IDs) der jeweiligen Vorkommen in dem Array **\$_SESSION['ids']** gespeichert und das Vorkommen aus **\$_SESSION['ids']** mit dem Index 0 wird dargestellt. Außerdem sind hier alle Bedeutungen, die für das Lexem in **ontology** hinterlegt sind, zuweisbar. Klickt man 'zurück' oder 'weiter' übergibt **compare.php** den aktuell angezeigten Index an **save.php**, welches den Anzeige-Index-Zähler **\$_SESSION['POS']** entsprechend de- oder inkrementiert und zu **compare.php** zurück leitet. Das Klicken auf 'Bedeutung zuweisen!' passt entsprechend per **update** die zugewiesene Bedeutung an.

Der im Zuge dieser Arbeit entwickelte SanSemAn soll hier lediglich, wie schon oben erwähnt, der Schaffung einer von WSD-Algorithmen statistisch verwertbaren Trainings- und Vergleichsmenge zu disambiguierender Zielwörter dienen. Daher sei nochmals auf seinen Prototypen-Status hingewiesen. Für einen professionellen Einsatz sind unbestritten noch etliche Veränderungen von Nöten, die jedoch einen für diese Arbeit angemessenen Aufwand übersteigen würden.

7 Auswertung der Annotationen mit SanSemAn

Die genauere Betrachtung der Textauszüge aus dem **DCS** während der Annotation und bei den beiden Korrektur-Durchläufen hat weitere Fragestellungen mit sich gebracht, insbesondere hinsichtlich der Behandlung der Konsistenz und dem damit verbundenen Pool an zugrunde liegenden Bedeutungen: a) Bei der Bedeutung *jana*₁₀ (relative) soll-

te in einer weiteren Studie philologisch und etymologisch überprüft werden, ob *jana* tatsächlich eine eigenständige Bedeutung als 'Verwandter' haben kann oder ob sich eine solche erst in Komposita wie *svajana* oder *bandhujana* durch die semantische Wirkung von *sva* oder *bandhu* ergibt. b) Die Plural-Verwendung von Singular-Formen von *jana*₄ (people) ist bekannt. Ob sich allerdings diese Plural-Verwendung auch auf Singular-Vorkommen von *jana* in einer Bedeutung wie *jana*₁₀ (national) übertragen lässt, sollte geprüft werden. Oft wurde bei einer Singular-Form von *jana* mit *jana*₁₀ annotiert, wenn im Satz-Fenster Wörter für König, Reich oder Namen derselben vorkamen, aus dem Kontext aber hervorging, dass nicht ein einzelner Bürger, sondern alle Bürger gemeint waren. c) Bei IDTEA 410723¹³⁸ ist die Rede von einem *ikṣvākujana*. Hier sollte geklärt und dementsprechend konsistent annotiert werden, ob der Fokus mehr auf die Tatsache gelenkt werden soll, dass ein Mitglied der Familie der Ikṣvāku-s ein Verwandter derselben ist oder ob, aufgrund der Tatsache, dass ein Mitglied der Königsfamilie auch ein Bürger des Reiches ist, der Augenmerk auf 'Bürger' fallen sollte. Oder ob sogar aus dem weiteren Kontext hervorgeht, dass es sich gar nicht um ein Mitglied der Familie handelt, sondern nur um einen Bürger der Regentschaft der Ikṣvāku-s, bzw. ob, sofern es sich doch um ein Mitglied der Familie handelt, ein solches überhaupt als Bürger bezeichnet werden kann.

Diese Fragen können aber im Umfang dieser Arbeit, um einen angemessenen Rahmen nicht zu sprengen, nicht eingehend behandelt werden.

7.1 Inter-Annotator-Übereinstimmung

Als Zielwörter für das WSD-System, welches im nächsten Kapitel besprochen wird, wurden die beiden polysemen Substantive *śārdūla* und *jana* gewählt, deren mögliche Bedeutungen aus der Tabelle **ontology**¹³⁹ in den Tabellen 2 und 3 aufgeführt werden.

Wie den Tabellen zu entnehmen ist, gibt es durchschnittlich 11,5 Bedeutungen pro Lexem. Es wurden 148 Vorkommen von *śārdūla* und 271 Vorkommen von *jana*, also insgesamt 419 Vorkommen der beiden Lexeme, aus dem Epos *Rām.* von den beiden Annotatoren Oliver Hellwig und Jonas Soiné, die Philologen und mit der Sprache des Sanskrit vertraut sind, mit Hilfe des Systems **SanSemAn** mit Bedeutungen versehen.

In die weiteren statistischen Auswertungen, sofern eine Erhöhung der Anzahl der

¹³⁸Vgl. mit der Übersetzung von IDTEA 410723 auf Seite 55. ¹³⁹Vgl. Tabellen-Schema von **ontology** in Kapitel 6.1.

Tabelle 2: Bedeutungen von *jana*

	oCycID	Wort	Annotation
1	970	life form	any living entity
2	971	man	the generic use of the word to refer to any human being; “it was every man for himself”
3	972	person	a human being; “there was too much for one person to do”
4	973	people	(plural) any group of human beings (men or women or children) collectively; “old people”; “there were at least 200 people in the audience”
5	974	race	people who are believed to belong to the same genetic stock; “some biologists doubt that there are important genetic differences between races of human beings”
6	29685	man	an adult male person (as opposed to a woman); “there were two women and six men on the bus”
7	29686	Jana	name of a man
8	29687	national	a person who owes allegiance to that nation; “a monarch has a duty to his subjects”
9	29688	commoner	a person who holds no title
10	108719	relative	a person related by blood or marriage; “police are searching for relatives of the deceased”; “he has distant relations back in New Jersey”
11	114684	group	any number of entities (members) considered as a unit
12	116036	attendant	one who attends or waits on another

möglichen Bedeutungen überhaupt eine Auswirkung auf das zu berechnende Maß hat, fließen jedoch nur die tatsächlich verwendeten Bedeutungen mit ein¹⁴⁰, welche in Tabelle¹⁴¹ 4 aufgeführt sind. Es wird also bei allen berechneten Maßen davon ausgegangen, dass für *śārdūla* 3 Bedeutungen zur Verfügung standen und für *jana* entsprechend 7, also insgesamt sind 10 Bedeutungen mit einem Durchschnitt von 5,0 Bedeutungen pro Wort zu berücksichtigen.

In Abbildungen 1 und 2 werden die relativen Verteilungen der Bedeutungen von *jana*

¹⁴⁰Dies leistet auch der Tatsache Genüge, dass in der aktuellen Version von SanSemAn bei dem Prozess der Annotation nur die 3 häufigsten Bedeutungen zur Verfügung stehen.

¹⁴¹Die Indizes der Einträge in Spalte **Wort** in Tabelle 4 beziehen sich auf die Zeilennummern aus Tabellen 2 und 3.

Tabelle 3: Bedeutungen von *śārdūla*

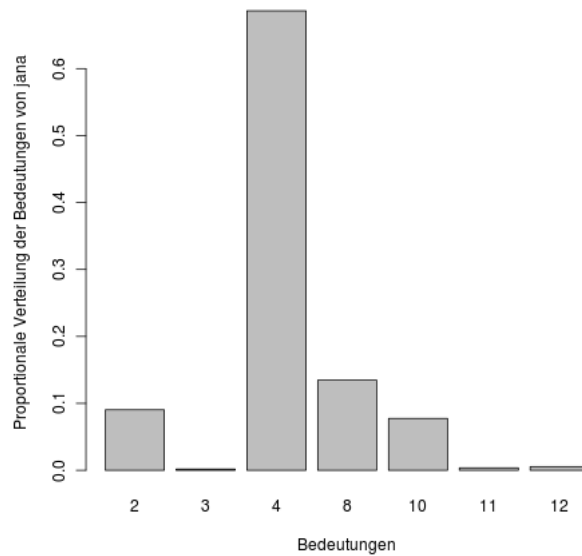
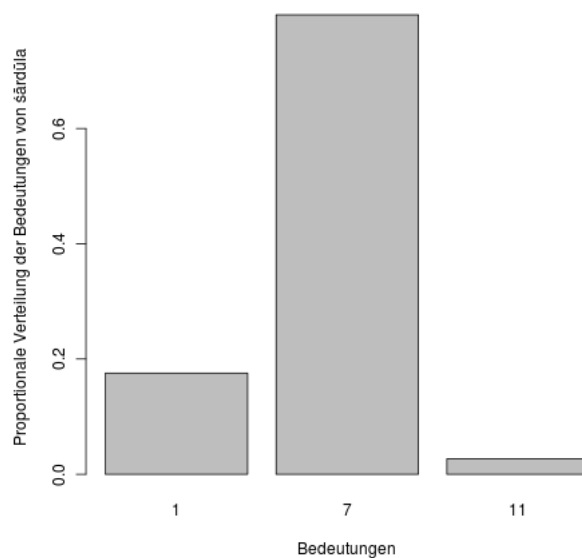
	oCycID	Wort	Annotation
1	23633	tiger	large feline of forests in most of Asia having a tawny coat with black stripes; endangered
2	23634	lion	large gregarious predatory feline of Africa and India having a tawny coat with a shaggy mane in the male
3	23635	panther	a leopard in the black color phase
4	23636	leopard	large feline of African and Asian forests usually having a tawny coat with black spots
5	23637	Śarabha	name of a fabulous animal
6	23638	Śārdūla	a kind of bird
7	23639	best	the person who is most outstanding or excellent; "he could beat the best of them"
8	23640	leader	a person who rules or guides or inspires others
9	23641	Plumbago zeylanica	a kind of shrub
10	23642	Śārdūla	name of two metres
11	23643	Śārdūla	name of a Rākṣasa; a spy of Rāvaṇa

Tabelle 4: Tatsächlich verwendete Bedeutungen von *jana* und *śārdūla*

Bezeichner	Bedeutung	oCycID
<i>jana</i> ₂	man	971
<i>jana</i> ₃	person	972
<i>jana</i> ₄	people	973
<i>jana</i> ₈	national	29687
<i>jana</i> ₁₀	relative	108719
<i>jana</i> ₁₁	group	114684
<i>jana</i> ₁₂	attendant	116036
<i>śārdūla</i> ₁	tiger	23633
<i>śārdūla</i> ₇	best	23639
<i>śārdūla</i> ₁₁	Śārdūla	23643

und *śārdūla*, welche in Tabelle 4 aufgeführt sind, dargestellt.

Die prozentuale Übereinstimmung der beiden Annotatoren (ITA) liegt bei 86,4% für

Abbildung 1: Proportionale Verteilung der Bedeutungen von *jana*Abbildung 2: Proportionale Verteilung der Bedeutungen von *sārdūla*

jana, 100% für *sārdūla* und bei 91,2% insgesamt¹⁴². Diese hohen Werte im Vergleich zu

¹⁴²Im ersten Durchlauf der Annotation wurden 37 der 271 Vorkommen von *jana* von den Annotatoren mit unterschiedlichen Bedeutungen versehen. Im weiteren Verlauf dieser Arbeit sind

prozentuale Übereinstimmungen, sofern nicht anders angegeben, immer auf den ersten Durchlauf der Korrektur der Annotationen bezogen.

z. B. 60%–70% im SENSEVAL-3 Projekt¹⁴³, lassen sich einerseits darauf zurückführen, dass der Grad der Ambiguität von *śārdūla* mit 3 zur Verfügung stehenden Bedeutungen nicht sehr hoch war und das große Verhältnis von 148 : 271 ($\approx 5 : 9$) Vorkommen von *śārdūla* zu *jana* somit die schlechtere ITA von *jana* stark beeinflusste. Andererseits ist die ITA von *jana* selbst noch weit über den Werten, die aus den SENSEVAL-Projekten bekannt sind. Dies ist auf eine Berechnung derselben zurückzuführen, die zeitlich nach dem ersten Korrekturlauf der Annotationen stattgefunden hat und natürlich auf die Tatsache, dass hier nur zwei Nomen, eines davon mit geringer Ambiguität und keine Verben oder Adjektive annotiert wurden. Der künstliche Charakter dieses Experimentes sollte bei einer Bewertung dieser Daten nicht vernachlässigt werden.

7.2 Konsistenz der Annotationen

7.2.1 Bedeutungen von *śārdūla*

Die Interrater-Reliabilität bzw. ITA bei dem Nomen *śārdūla* ist mit 100%iger Übereinstimmung sehr hoch. Ähnlich hohe Übereinstimmungen wurden auch bei anderen Annotations-Durchläufen für manche Wörter festgestellt. Schon in SENSEVAL-1 berichtet KILGARIFF [36] eine Übereinstimmung von 95,5% für die Wörter *onion*, *generous*, *sack*, *shake*. Auch in SENSEVAL-2 kam es zu fast 100%iger Präzision der Annotationen von echten Nomen und Komposita¹⁴⁴. Wie oben erwähnt, wurden 148 Vorkommen des Lexems *śārdūla* aus dem Rām. willkürlich gewählt und von den beiden Annotatoren Oliver Hellwig und Jonas Soiné mit Hilfe von SanSemAn annotiert. Die drei tatsächlich verwendeten Bedeutungen *śārdūla*_{1,7,11} sind in den kurzen Kontext-Auszügen von in der Regel 7–9 Lexemen, gut erkennbar. Beide Annotatoren weisen konsequent die korrekten Bedeutungen zu.

Wie sich in Abbildung 2 erkennen lässt, ist Bedeutung 7 (*śārdūla*₇ 'Bester') mit 79,7% und einer Amplitude von über 60% zu Bedeutung 1, die häufigste. Es handelt sich hier um eine Verwendung von *śārdūla* als Hinterglied in einem Tatpuruṣa-Kompositum, welches keine großen Zweifel bezüglich seiner Semantik zulässt. Beispiele für die Verwendung als 'Bester' sind unter anderem: *kapiśārdūla* (Bester der Affen), *muniśārdūla* (Bester der Seher), *naraśārdūla* (Bester der Menschen), *hariśārdūla* (Bester

¹⁴³Vgl. MIHALCEA U. A. [47, Kap. 2.4].

¹⁴⁴Genauer: MWEs – müssen nicht, aber können Komposita sein; ein Lexem, welches aus einer Gruppe von zwei oder mehreren Lexe-

men besteht und dessen Eigenschaften nicht zwangsläufig aus den einzelnen Komponenten vorhersehbar sind. Vgl. KILGARIFF [37] und MIHALCEA U. A. [47].

der Gelb-Braunen¹⁴⁵), *rākṣasaśārdūla* (Bester der Rākṣasa-s¹⁴⁶), *plavagaśārdūla* (Bester der Affen), *vānaraśārdūla* (Bester der Affen), *manujaśārdūla* (Bester der Menschen¹⁴⁷), *ikṣvāku-śārdūla* (Bester der Ikṣvāku-s¹⁴⁸), *rājaśārdūla* (Bester der Könige), *nṛpaśārdūla* (Bester der Könige) und *puruṣaśārdūla* (Bester der Menschen). Im Folgenden werden einige Kontext-Auszüge mit entsprechendem Vorkommen und Übersetzung aufgeführt:

tvam caiva naraśārdūla sahāsmābhir gamiṣyasi

Und nur Du, Oh Bester der Menschen, wirst mit uns gehen!

DCS IDTEA 331824

rāmas tu muniśārdūlam uvāca sahalakṣmaṇaḥ

Rāma jedoch, von Lakṣmaṇa begleitet, sprach zu dem Besten der Seher...

DCS IDTEA 339294

Bedeutung 1 (*śārdūla*₁ 'Tiger') ist mit einer Verteilung von 17,5% auf Platz 2. In dieser Bedeutung tritt das Lexem meist alleinstehend, innerhalb des Hinterglied eines Tatpuruṣa-Kompositums, als Dvandva oder innerhalb des Vorderglied eines Bahuvṛīhi-Kompositums auf. Als Vorderglied in einem Dvandva ist in diesem Test-Corpus z. B. folgendes Vorkommen mit dieser Bedeutung vorhanden:

IDTEA 514095 *mṛgamārjāraśārdūlān* (Wild, Wildkatzen und Tiger¹⁴⁹); als Hinterglied in einem Tatpuruṣa z. B. IDTEA 533581 *śārdūlamṛgasamghuṣṭam* (Lärm von Tigern und Wildtieren¹⁵⁰) und IDTEA 566574 *siṃhaśārdūlajuṣṭāḥ* (Von Löwen und Tigern Gemochte¹⁵¹). In folgender Textstelle zeigt sich die Verwendung im Vorderglied eines Bahuvṛīhi-Kompositums:

teṣāṃ śārdūladarpāṇāṃ mahāsyānāṃ mahaujasām

Deren, die große Mäuler, große Stärke und den Stolz der Tiger hatten, ...

DCS IDTEA 462572

Und in dieser Textstelle ist die alleinstehende Verwendung als 'Tiger' zu erkennen:

śārdūlenāmiṣasyārthe mṛgarājāṃ yathā hatam

Wie den zum Zwecke der Nahrung vom Tiger getöteten König des Wilds...

¹⁴⁵Hier: Affen.

¹⁴⁶Hier: Dämonen.

¹⁴⁷*manuja* (Von Manu geboren).

¹⁴⁸Name einer königlichen Linie im [Mbht.](#)

¹⁴⁹Akk. Pl. m. des Dvandva-Kompositums.

¹⁵⁰Akk. Pl. m. des Tatpuruṣa.

¹⁵¹Nom. Pl. m./f. des Tatpuruṣa.

DCS IDTEA 527304

Am seltensten taucht das Lexem in Bedeutung 11 (*śārdūla*₁₁ 'Śārdūla, Name eines Dämons') mit einer Verteilung von nur 2,7% (insgesamt 4 mal) auf. Statistisch lässt sich in unseren Text-Auszügen diese Bedeutung immer dann ableiten¹⁵², wenn das Lexem alleine steht und im Kontext Dämonen (Rākṣasa-s) oder Namen von Dämonen (z. B. Rāvaṇa) auftauchen. Nur bei nachfolgender Passage, kann die Bedeutung ohne Kenntnis eines größeren Kontextes, schwerer erschlossen werden – geht man davon aus, dass im Rām. Tiger nicht angesprochen werden¹⁵³, jedoch sehr wohl ein Dämon, nämlich Śārdūla angesprochen werden kann, erscheint es zumindest möglich (sogar nicht unwahrscheinlich) diese Bedeutung abzuleiten:

jātoḍvego 'bhavat kiṃcic chārdūlaṃ vākyaṃ abravīt

Zu einem geworden, dessen Furcht entstanden war, sprach er eine Rede zu Śārdūla.

DCS IDTEA 1312098

Aus den beiden nächsten Textstellen geht die Bedeutung als Name klar hervor:

tadā rākṣasaśārdūlaṃ śārdūlo bhayavihvalaḥ

..., dann (sagte) Śārdūla, erregt durch Angst, zum Besten der Dämonen ...

DCS IDTEA 1311686

śārdūlasya mahad vākyaṃ athovāca sa rāvaṇaḥ

Zur großen Rede des Śārdūla sprach dann Rāvaṇa ...

DCS IDTEA 1311809

7.2.2 Bedeutungen von jana

Bei *jana* zeigte sich die Annotation mit [SanSemAn](#) deutlich schwieriger als bei *śārdūla*, selbst nach einer Glättung¹⁵⁴ der unterschiedlich zugewiesenen Bedeutungen, verblieben noch 37 Unstimmigkeiten (nach einem zweiten Durchlauf immer noch 28). Die [ITA](#) von 86,4% (89,7% im zweiten Durchlauf) lässt sich unter anderem auf die problematische Granularität¹⁵⁵ der möglichen Bedeutungen zurückführen. Dennoch ist der Wert von 86,4% im Vergleich zu anderen Annotations-Unternehmungen wie SENSEVAL-1 bis

¹⁵²Der Mensch zieht aus dem Kontext ähnliche Schlüsse, außer bei IDTEA 1312098, welches später noch behandelt wird.

wie beispielsweise dem [Pañcat.](#) keineswegs angenommen werden dürfte.

¹⁵³Was verifiziert werden müsste und in einem Text

¹⁵⁴Der oben erwähnte erste Korrektur-Durchlauf.

¹⁵⁵Dazu mehr in Kapitel 7.3.

SENSEVAL-3, bei welchen durchschnittliche Übereinstimmungen¹⁵⁶ von ca. 67% erzielt wurden, recht hoch und zeigt eine akzeptable bis sehr gute Übereinstimmung an. Ein ähnliches Vorgehen¹⁵⁷ bei SENSEVAL-1, wo nur professionelle Lexikographen annotierten und vor der Berechnung der ITA kritische Fälle besprechen konnten, führte ebenfalls zu einer ITA von über 80%.

Annotationen mit *jana*₄

Die am häufigsten zugewiesene Bedeutung ist mit 68,6% im ersten Durchlauf *jana*₄ (people). In dieser Bedeutung tritt das Lexem in den unterschiedlichsten morphologischen Erscheinungen auf: oft alleinstehend im Plural oder Singular und in den untersuchten Textauszügen in sämtlichen Kasus, außer im Vokativ. Die Annotationen bei den alleinstehenden und im Hinterglied von Komposita stehenden Plural- und Singular-Formen sind bis auf einige Ausnahmen konsistent, bei denen es hauptsächlich Überschneidungen mit *jana*₂ (man) gibt. So wählt A1¹⁵⁸ bei IDTEA 533257 *jana*₂ und A2 *jana*₄ als Bedeutung:

mantrapūtena haviṣā hutvā mantravido janāḥ

Die Mantra-kennenden Menschen (Leute) opferten durch eine
Mantra gereinigte Opfergabe und ...

DCS IDTEA 533257

Bei IDTEA 356719 ist für das Lexem, welches hier in einem Tatpuruṣa auftaucht, Ähnliches der Fall:

siktām candanatoyaiś ca śiraḥsnātajanair vṛtām

... die Besprenkelte und mit Menschen (Leuten) Umgebene, deren
Köpfe mit Sandelwasser gewaschen wurden, ...

DCS IDTEA 356719

Bei IDTEA 414520 annotiert A1 *jana*₈ (national) und A2 *jana*₄ (people):

pratyuvāca janāḥ sarvaḥ śrīmadvākyaṃ anuttamam

Alle Bürger (Leute) erwiderten der unübertroffenen, prächtigen
Rede ...

DCS IDTEA 414520

¹⁵⁶Vgl. auch MIHALCEA U. A. [47], KILGARIFF [36] ¹⁵⁸Im Folgenden steht abkürzend A1 für Oliver Hellwig und A2 für Jonas Soiné.

¹⁵⁷Vgl. AGIRRE [1, S. 87].

Und bei IDTEA 373244 weist A1 *jana₄* und A2 *jana₂* zu:

adhiruhya janaḥ śrīmān udāsīno vyalokayat

Der prächtige (Leute) Mensch erhob (sich), und betrachtete abseits sitzend ...

DCS IDTEA 373244

Die folgenden Passagen sind Beispiele, bei welchen die beiden Annotatoren übereinstimmend die Bedeutung *jana₄* zuwiesen:

tām adya sītāṃ paśyanti rājamārgagatā janāḥ

Jetzt sehen die auf dem Königsweg gegangenen Leute Sītā.

DCS IDTEA 373319

ikṣvākuvaṃśaprabhavo rāmo nāma janaiḥ śrutāḥ

Der, dessen Ursprung in der Familie der Ikṣvāku-s liegt, nämlich Rāma, wird von den Leuten gehört.

DSC IDTEA 318780

tatastu sā lakṣmaṇarāmapālītā mahācamūr hṛṣṭajanā yaśasvinī

Da jedoch, (Verb) die von Lakṣmaṇa und Rāma beschützte, ruhmreiche Riesenarmee von (freudig) erregten Männern ...

DCS IDTEA 1755769

tataḥ pauraṇaḥ sarvaḥ śrutvā rāmābhiṣecanam

Dann hörten alle Stadtbewohner von der Krönung von Rāma und ...

DCS IDTEA 356148

tadā hy ayodhyānilayaḥ sastrībālābalo janaḥ

Denn da (Verb) die Leute, deren Heim in Ayodhyā ist, mitsamt Frauen, Mädchen und Jungen, ...

DCS IDTEA 355941

Die Vorkommen des Lexems mit der Bedeutung *jana₄* im Vorderglied eines Kompositums sind bis auf einige Abweichungen auch konsequent zugewiesen. Auch hier sind die Überschneidungen meist mit *jana₂*. So wie in folgenden Passagen, in welchen das Lexem im Vorderglied eines Tatpuruṣa-Kompositums auftritt, A1 die Bedeutung 'people' und A2 'man' zu:

tvayā ca saha gantavyaṃ mayā gurujanājñayā

Und mit Dir muss ich aufgrund der Weisung der Lehrer-Leute
(des Lehrer-Menschen) gehen.

DCS IDTEA 370648

In der Passage mit IDTEA 462434 annotiert hingegen A1 mit *jana₂* und A2 mit *jana₄*:

apayāhi janasthānāt tvaritaḥ sahabāndhavaḥ

Verlasse schnell Janasthāna¹⁵⁹, mitsamt Gefolge!

DCS IDTEA 462434

In der Passage IDTEA 2013395, ebenfalls im Vorderglied eines Tatpuruṣa-Kompositums, stimmen die beiden Annotatoren in der zugewiesenen Bedeutung *jana₄* beispielsweise überein:

tato madhyaṃ janaughānāṃ praviśya munipuṃgavaḥ

Dann, nachdem er die Mitte der Menschenmengen betreten hat,

(Verb) er als Held der Seher ...

DCS IDTEA 2013395

In den Annotationen mit *jana₄* ist somit durchaus eine konsequente Linie zu erkennen. Die wenigen Überschneidungen mit anderen Bedeutungen sind auf Probleme der Granularität und wahrscheinlich auch auf Übersetzungsfehler bzw. „Vertipper“ zurückzuführen.

Annotationen mit *jana₈*

Die zweit häufigste Bedeutung des Lexems ist mit 13,5% im ersten Korrektur-Durchlauf *jana₈* (national). Die Annotationen mit dieser Bedeutung sind in den untersuchten Textpassagen, relativ zu den anderen Bedeutungen (abgesehen von *jana₁₀*), am wenigsten konsistent. In fast allen Fällen der Nicht-Übereinstimmung hat einer der beiden Annotatoren *jana₈* zugewiesen: Überschneidungen gibt es meist mit *jana₄* (people) und *jana₁₀* (relative). Auch diese Überschneidungen werden im Kapitel über die Granularität nochmals angesprochen. Auffällig ist die inkonsequente Annotation bei manchen Vorkommen der Bigramme *sva jana* und *jana sarva*. Entscheidet sich A1 bei diesen Bigrammen, aufgrund des Vorkommens von Wörtern für König, Reich oder Namen derselben im Satz-Fenster, für *jana₈*, so wählt A2 meist¹⁶⁰ die entsprechende Bedeutung

¹⁵⁹Nach MONIER-WILLIAMS [50] heißt *janasthāna* ¹⁶⁰Außer bei IDTEA 472077, wo für *svajana* die Bedeutung *jana₈* gewählt werden muss; siehe unten.

analog zu dem Muster, *jana*₁₀ (relative) für *svajana* zu wählen. Solch eine Inkonsistenz ließe sich leicht durch genaue Spezifikation bezüglich der den Annotationen zugrunde liegenden Methodik aus dem Wege räumen. Syntaktisch lassen sich keine diese Bedeutung propagierenden Strukturen ableiten. IDTEA 414520 auf S. 52 und folgende Stelle, sind Beispiele für diese Inkonsistenz:

na cāyam ikṣvākujanaḥ prahr̥ṣṭaḥ pratibhāti me

Und nicht erscheint mir dieser Ikṣvāku-Bürger (-Verwandte) erfreut.

DCS IDTEA 410723

Eine weitere Unstimmigkeit bei den zugewiesenen Bedeutungen, die sich auf die Granularität zurückführen lässt, ist IDTEA 434751, bei der A1 *jana*₂ und A2 *jana*₈ annotiert:

rājyahetoḥ kathaṃ pāpam ācaret tvadvidho janaḥ

Wie soll der Dir ähnelnde Mensch (Bürger), eine Sünde um der Herrschaft willen begehen?

DCS IDTEA 434751

Beispiele für übereinstimmende Annotationen finden sich in den folgenden Passagen:

kva yāsyasi mahārāja hitvemaṃ duḥkhitaṃ janam

Wohin wirst Du gehen Oh König, nachdem Du diesen Bürger verlassen hast?

DCS IDTEA 413210

iyaṃ sarāṣṭrā sajanā dhanadhānyasamākulā

Diese, die ein Reich hat, die Bürger hat, die einen Überfluss an Geld und Korn hat...

DCS IDTEA 2228200

Im Verhältnis zur Anzahl der Annotationen von *jana*₄ zu *jana*₈ ist die Konsistenz der Annotationen von *jana*₈ natürlich geringer als bei *jana*₄. Da sich auch hier die Abweichungen meist auf Probleme der Granularität zurückführen und durch Rücksprache der Annotatoren schnell bereinigen lassen, ist auch hier eine konsequente Linie bei den Bedeutungszuweisungen zu erkennen.

Annotationen mit *jana*_{2,3,10,11,12}

Die Annotationen mit den Bedeutungen *jana*_{2,3,10,11,12} machen insgesamt ca. 17% aller Annotationen aus und werden hier deshalb zusammengefasst dargestellt.

Die bei IDTEA 2395113 übereinstimmende¹⁶¹ Verwendung von *jana*₁₁ (group), beruht auf einer Glättung durch die Korrekturläufe, die einmalige Verwendung kann als „Vertipper“ angesehen werden und ist klar der Vagheit der Begriffe und deren Granularität zuzuschreiben. Ähnliches gilt für die einmalige nicht übereinstimmende Verwendung von *jana*₃ (person).

Die zweimalige Verwendung von *jana*₁₂ (attendant) bei den IDTEAs 421979 und 2008372 erklärt sich durch den Kontext und bedarf keiner näheren Erläuterung¹⁶².

Auf die Verwendung von *jana*₂ (man) wurde weiter oben in Verbindung mit den Überschneidungen mit hauptsächlich *jana*₄ schon eingegangen. Hier sollen nur zwei übereinstimmende Passagen aufgeführt werden:

uktvā na bhetyam iti strījanam sa tato 'rjunah

Arjuna sprach zu der Frau „Fürchtet (Euch) nicht!“ ...

DCS IDTEA 1945831

yasya rāmah priyah putro jyeṣṭho gurujanapriyah

Wessen lieber, bester, von dem Lehrermenschen gemochter Sohn

Rāma ...

DCS IDTEA 576749

Die Bedeutung *jana*₁₀ (relative) wird von beiden Annotatoren konsequent in Verbindung mit dem Bigramm *svajana* zugewiesen. Die meisten der Überschneidungen mit *jana*₂ und *jana*₄ wurden auch weiter oben schon besprochen. So werden nun abschließend noch zwei übereinstimmende Passagen gezeigt:

nirguṇah svajanaḥ śreyān yaḥ paraḥ para eva saḥ

Welcher tugendfreie Verwandte der Beste ist, der ist der Beste.

DCS IDTEA 1389125

paurān svajanavan nityam kuśalam pariprcchati

Wie ein Verwandter fragt er die Stadtleute nach dem Wohlergehen.

¹⁶¹Es sei nochmals darauf hingewiesen, dass nach wäre.
der aktuellen Programmlogik von [SanSemAn](#),
diese Bedeutung gar nicht zuweisbar gewesen

¹⁶²Näheres wird im Kapitel 7.3 über die Granularität erwähnt.

7.3 Probleme aufgrund der Granularität

Die linguistische Granularität beschäftigt sich mit der semantischen Schärfe sprachlicher Zeichen. Im Bezug auf die Konzeption von Bedeutungs-Inventaren zum Zwecke der WSD und deren Auswertungen, wurde in Kapitel 5 die „Körnigkeit“ des öfteren erwähnt, insbesondere in den Fußnoten 81 und 83 sind weitere Verweise zu finden.

Bei der Annotation der Bedeutungen von *jana* lassen sich einige entstandene Schwierigkeiten auf die Granularität der Bedeutungen zurückführen. Hier sei nochmals ein Ausschnitt aus Tabelle 4 mit Erweiterungen gegeben:

Tabelle 5: Zwei Äste der WordNet-Hierarchie mit Bedeutungen von *jana*

Ast	Ebene	Bezeichner	Bedeutung	oCycID
1	3	<i>jana</i> ₁₁	group	114684
1	4	<i>jana</i> ₄	people	973
2	7	<i>jana</i> ₃	person	972
2	8	<i>jana</i> ₂	man	971
2	8	<i>jana</i> ₈	national	29687
2	8	<i>jana</i> ₁₀	relative	108719
2	10	<i>jana</i> ₁₂	attendant	116036

Es soll nicht zu detailliert auf die semantische Hierarchie der 7 Bedeutungen von *jana* eingegangen werden. Dennoch soll ein grober Zusammenhang dieser Begriffe anhand der in WordNet hinterlegten Hierarchie gegeben werden.

Alle Bedeutungen sind Hyponyme¹⁶³ des Synsets *entity*, welches den Ursprung der Hierarchie darstellt. Für die zu untersuchenden Bedeutungen lassen sich zwei verschiedene Hauptpfade identifizieren. Bedeutung 4 ist direktes Hyponym von Bedeutung 11, welche, nur getrennt durch das Synset *abstract entity*, direktes Hyponym von *entity* ist. Somit sind *jana*₄ und *jana*₁₁ einer weit grobkörnigeren Ebene als die restlichen Bedeutungen zuzuordnen:

$$\xrightarrow{1} \textit{entity} \xrightarrow{2} \textit{abstract entity} \xrightarrow{3} \textit{jana}_{11} \xrightarrow{4} \textit{jana}_4$$

¹⁶³Vgl. Hyponymie.

Der zweite Hauptpfad, welcher natürlich auch von dem Synset *entity* ausgeht, lässt 6 Knoten (Ebenen) aus, bis mit *jana₃* das nächste Hyponym erreicht wird. Dieses ist wiederum Hyperonym von *jana₂*, *jana₈* und *jana₁₀*, die sich alle auf Ebene 8 befinden. Den Schluss dieses Pfades bildet auf Ebene 10 *jana₁₂*:

$$\xrightarrow{1} \textit{entity} \xrightarrow{2\dots7} \textit{jana}_3 \xrightarrow{8} \textit{jana}_2, \textit{jana}_8, \textit{jana}_{10} \xrightarrow{9\dots10} \textit{jana}_{12}$$

Die einmalige Annotation mit *jana₁₁* wurde oben ja schon mehr oder weniger als statistischer Ausreißer erklärt. Diese Annotation ist zwar sicher nicht falsch, da aber die beiden Bedeutungen direkte Nachbarn der Hierarchie sind, ist auch eine Überscheidung der beiden verständlich. Auf die Aufgabe der WSD sollten so feine hierarchische Bedeutungsunterschiede keinen Einfluss haben¹⁶⁴.

Die unterschiedlichen Annotationen mit Elementen, welche dem zweiten angesprochenen Ast zuzuschreiben sind, sollten in einem echten Evaluierungssystem als Treffer gewertet werden, da dies alles Ko-Hyponyme der selben hierarchischen Ebene sind¹⁶⁵.

Problematischer sind die Annotationen, bei denen die Annotatoren nicht-übereinstimmend Elemente aus beiden Pfaden zugewiesen haben; das sind rund 82% der unterschiedlich zugewiesenen Bedeutungen. Doch auch hier lässt sich ein gewisser Trend erkennen: *jana₄* unterstreicht allgemein den kollektiven Charakter des Lexems in seiner entsprechenden grammatikalischen Funktion. Die Bedeutungen von *jana* hingegen, die Hyponyme von *jana₃* sind, spezifizieren eine bestimmte Aufgabe, der das einzelne Individuum nachkommt oder zumindest die Singularität der jeweils spezifizierten Bedeutung. Es spielen also Nuancen des Numerus in die Entscheidung für oder gegen eine Bedeutung dieser Gruppe hinein. Dem einen Annotator erscheint der eine Aspekt wichtiger, dem anderen der Andere – eine eventuelle Lösung für ein echtes WSD-System wäre die Möglichkeit, pro Vorkommen des zu annotierenden Lexems mehrere Bedeutungen zu weisen zu können und ein grob-körniges *Scoring* für die Evaluation zu verwenden. Somit ließe sich einerseits die spezielle Funktion des *jana* abbilden und andererseits wäre dem kollektiven Charakter von *jana₄* Ausdruck verliehen, sobald es nicht auszuschließen ist, dass es sich um mehrere dieser Art handelt.

¹⁶⁴ Vgl. AGIRRE [1, S. 9 u. Kap. 3].

¹⁶⁵ Vgl. Gleichung Nr. 6.14 auf Seite 38. IDE und WILKS in AGIRRE [1, S. 52] argumentieren, dass gröb-körnige Bedeutungen *die Bedeutungen* sind, die unterschieden werden sollten, weil selbst Menschen Probleme haben, sehr feinkörnige Bedeutungsunterschiede zu machen.

7.4 Cohens κ

COHEN [17] entwickelte ein Maß zur Berechnung der ITA der von zwei Beurteilern für $n \in \mathbb{N}$ Objekte in $z \in \mathbb{N}$ Kategorien getätigten Zuweisungen. Im Gegensatz zur rohen prozentualen Übereinstimmung der ITA, entfernt die κ -Statistik von dieser die Übereinstimmung durch Zufall und ist somit ein robusteres Maß der ITA. Cohens κ , oft auch mit κ -Koeffizient oder κ -Statistik referenziert, ist durch Gleichung (7.1) definiert:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \quad (7.1)$$

Die Urteilshäufigkeiten h_{ii} mit $1 \leq i \leq z$ können in einer $z \times z$ -Matrix bzw. Kontingenz-Tafel abgetragen werden, wobei das zweite i die Spaltenhäufigkeiten und das erste i die Zeilenhäufigkeiten repräsentiert. Die relative Häufigkeit bzw. Wahrscheinlichkeit (die Diagonale in der Kontingenz-Tafel¹⁶⁶), dass beide Annotatoren übereinstimmend bewerten, ist p_a und in Gleichung (7.2) dargestellt. Die Wahrscheinlichkeit, dass zwei zufällig urteilende Annotatoren übereinstimmend bewerten würden, ist p_e und in Gleichung (7.3) zu erkennen. Hierfür wird die Summe über die Produkte der Randsummen¹⁶⁷ gebildet und ins Verhältnis zum Quadrat aller beurteilten Objekte gesetzt:

$$p_a = \frac{\sum_{i=1}^z h_{ii}}{n} \quad (7.2)$$

$$p_e = \frac{1}{n^2} \cdot \sum_{i=1}^z \left(\sum_{k=1}^z h_{ki} \cdot \sum_{k=1}^z h_{ik} \right) \quad (7.3)$$

Tabellen 6, 7 und 8 stellen die jeweilige Kontingenz-Tafel der entsprechenden Verteilungen nach dem ersten Korrektur-Durchlauf dar. Die Annotationen von Oliver Hellwig sind auf der y-Achse und die von Jonas Soiné auf der x-Achse aufgetragen. Beispiel: In Tabelle 6 ist Spalte 972 leer, d.h. Oliver Hellwig hat Bedeutung 972 (Person) nie annotiert, Jonas Soiné hat in Zeile 972 Spalte 971 den auf 0,002 gerundeten Wert 0,002386635, welcher mit 419 (Anzahl aller untersuchten Vorkommen) multipliziert 1 ergibt, zusätzlich bedeutet das, dass A1 an dieser Stelle Bedeutung 971 (man) annotiert hat.

Die Summe der in roter Schrift dargestellten Verteilungen übereinstimmender Annotationen¹⁶⁸ ergeben die in Kapitel 7.1 auf Seite 47 schon erwähnten rohen prozentualen

¹⁶⁶In den Tabellen 6, 7, 8 in roter Schrift hervorgehoben.

¹⁶⁷Zeilensumme mal Spaltensumme.

¹⁶⁸Diese Werte sind die Anzahl der Annotations-

Paare (Kategorie1,Kategorie2) geteilt durch die Anzahl aller beobachteten, d. h. annotierten Vorkommen.

Tabelle 6: Kontingenz-Tafel von *jana* und *śārdūla* gemeinsam

	108719	114684	116036	23633	23639	23643	29687	971	972	973
108719	0.045	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.002
114684	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
116036	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
23633	0.000	0.000	0.000	0.062	0.000	0.000	0.000	0.000	0.000	0.000
23639	0.000	0.000	0.000	0.000	0.282	0.000	0.000	0.000	0.000	0.000
23643	0.000	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.000
29687	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.002	0.000	0.019
971	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.036	0.000	0.029
972	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000
973	0.000	0.000	0.002	0.000	0.000	0.000	0.012	0.012	0.000	0.406

Tabelle 7: Kontingenz-Tafel von *jana*

	108719	114684	116036	29687	971	972	973
108719	0.070	0.000	0.000	0.011	0.000	0.000	0.004
114684	0.000	0.004	0.000	0.000	0.000	0.000	0.000
116036	0.000	0.000	0.004	0.000	0.000	0.000	0.000
29687	0.000	0.000	0.000	0.103	0.004	0.000	0.030
971	0.000	0.000	0.000	0.000	0.055	0.000	0.044
972	0.000	0.000	0.000	0.000	0.004	0.000	0.000
973	0.000	0.000	0.004	0.018	0.018	0.000	0.627

Tabelle 8: Kontingenz-Tafel von *śārdūla*

	23633	23639	23643
23633	0.176	0.000	0.000
23639	0.000	0.797	0.000
23643	0.000	0.000	0.027

Übereinstimmungen. Die in türkis dargestellten Werte sind die Häufigkeiten abweichend zugewiesener Bedeutungen, die hier nur zur Berechnung der Zufallswahrscheinlichkeiten verwendet wurden.

Nachfolgend sind die berechneten¹⁶⁹ κ -Koeffizienten aufgeführt. Die Berechnung erfolgte in der Programmiersprache R mit den Prozeduren die in REVELLE [56] beschrieben sind.

Cohens κ für jana und śārdūla insgesamt: 0,87 bei 419 annotierten Vorkommen.

Cohens κ für jana: 0,73 bei 271 annotierten Vorkommen.

Cohens κ für shardula: 1,00 bei 148 annotierten Vorkommen.

Die κ -Statistik für *śārdūla* bedarf keiner Erklärung und der im Vergleich zu *jana* erhöhte Gesamtwert ist auf den Einfluss von *śārdūla* zurückzuführen. Die 37 nach dem ersten Korrektur-Durchlauf vorhandenen unterschiedlichen Annotationen erklären den niedrigeren Wert für *jana*. Im Vergleich zu der in MIHALCEA U. A. [47] berechneten 'micro- κ statistic' von 0,58 ist der Wert natürlich deutlich höher und zeugt von einer sehr großen Übereinstimmung, die sich auch schon vorher mehrfach zu erkennen gegeben hat.

8 Die WSD durch DL

In diesem Abschnitt werden die Ergebnisse der **Disambiguierung** des in Abschnitt 6 präparierten Trainings-Corpus durch ein etwas modifiziertes Verfahren der **DL** nach YAROWSKY [66] besprochen. Bevor jedoch die Schilderung dieses Algorithmus, seiner Anwendung und Evaluierung vorgenommen wird, sollen einige Anmerkungen zum finalen Trainings-Corpus gemacht werden.

8.1 Das finale Trainings-Corpus

Um eine einheitliche, konsistente Klassifizierung durch entsprechend gewählte *features* zu gewährleisten, wurde das mit **SanSemAn** in mehreren Durchläufen erstellte Corpus nochmals einer Glättung unterzogen. Dies bezieht sich in erster Linie auf die Passagen mit Vorkommen von *jana*, da hier, wie oben schon erwähnt, verschiedene, der Granularität geschuldete Probleme auftraten. Bei der Disambiguierung von *jana* wurden nur lexikalische *features* verwendet. Die Annotationen wurden dahingehend angepasst, dass sich z. B. das *feature* [-1w sva] → *jana*₁₀ (relative) durchgehend in allen Annotationen

¹⁶⁹Auch hier erfolgte die Berechnung nach dem ersten Korrektur-Durchlauf.

erkennen lässt, ungeachtet der feineren semantischen Nuancen. Bei der Verteilung der Bedeutungen von *sārdūla* haben sich natürlich keine Veränderungen ergeben. Zu der Verteilung der Bedeutungen von *jana* lässt sich anmerken, dass nach dem 1. Korrekturdurchlauf die Bedeutungen *jana*₃ und *jana*₁₁ noch zugewiesen waren, in der finalen Version jedoch nicht mehr vorhanden sind. Generell hat *jana*₄ weniger Zuweisungen erhalten und die Differenz im Wesentlichen auf *jana*_{2,8,10} übertragen. In Abbildung 3 ist eine Gegenüberstellung der Verteilungen nach dem 1. Korrekturlauf (grau schraffiert) und der finalen Version (grau) dargestellt:

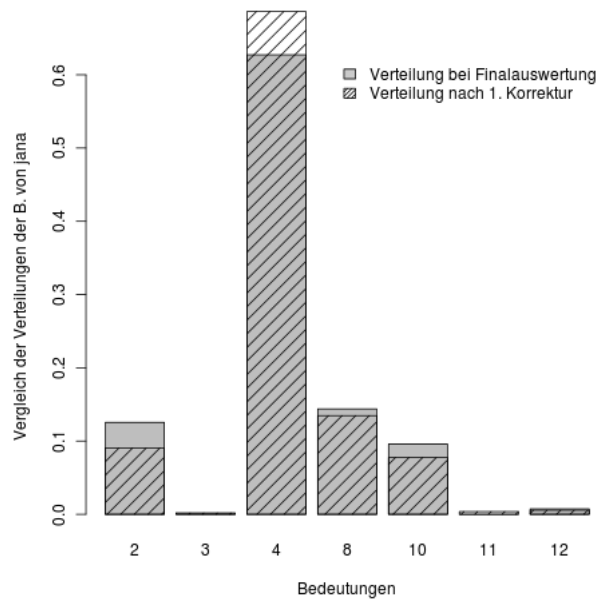


Abbildung 3: Proportionale Verteilung der Bedeutungen von *jana* nach dem 1. Durchlauf und bei der finalen Version

8.2 Der DL-Algorithmus nach Yarowsky

Schon in Kapitel 5.2 auf den Seiten 30ff wurde ein definitorischer Überblick der Beschaffenheit von *decision lists* nach AGIRRE [1, S. 170f] gegeben.

In YAROWSKY [66] beschreibt der Autor genau diesen Algorithmus zur WSD von Homographen. In seinem Fall werden Ambiguitäten bezüglich verschiedener Akzente von französischen und spanischen Lexemen aufgelöst, z. B. *marche/marché* (Schritt-/Markt) oder *sabana/sábana* (Grassteppe, Bettuch). Er erreicht damit eine durchschnittliche Exaktheit der WSD von 97% bei einer Disambiguierung von 13 Homographen. Im

Folgenden sind die einzelnen Schritte des Algorithmus aufgeführt. Weiter oben wird eine 'modifizierte' Verwendung desselben angedeutet, da für die kleine Trainingsmenge von 419 Sätzen der Algorithmus schon nach deutlich weniger Aufwand sehr gute Ergebnisse gezeigt hat. Deshalb wurden einige Schritte, die Yarowsky schildert, ausgelassen. Am Ende der Beschreibung jedes Schrittes wird erklärt, wie der jeweilige Schritt auf das Vorhaben dieser Arbeit angewendet wurde:

Schritt 1: Die Ambiguitäten der Akzent-Muster identifizieren

Die meisten Wörter im Französischen oder Spanischen haben ohnehin nur ein Akzentmuster und müssen somit auch nicht disambiguiert werden. Zur Feststellung der Verteilung der möglichen Akzentuierungen eines Zielwortes sind alle Vorkommen desselben mit entsprechender Akzentuierung innerhalb des Corpus zu zählen und tabellarisch mit prozentualer Verteilung darzustellen.

Diesen Schritt auf das in dieser Arbeit unternommene Experiment übertragen habend, erhält man die in Abbildungen 2 und 3 ersichtlichen Verteilungen.

Schritt 2: Trainings-Kontext sammeln

Für ein bestimmtes Lexem der ambigen Homographen, die in Schritt 1 identifiziert wurden, werden $\pm k$ Wörter Kontext für jedes Vorkommen des Lexems in dem Corpus gesammelt, das vorgefundene Akzent-Muster wird gespeichert und anschließend der Akzent aus dem Vorkommen entfernt – so erhält man eine Trainingsmenge.

Für den in dieser Arbeit verfolgten Zweck ist dieser Schritt dadurch erfüllt, dass die durch das Attribut **IDTEA**¹⁷⁰ identifizierbaren Sätze (Kontext-Passagen) ohnehin nicht mit Bedeutungen annotiert und somit immer zur Klassifikation zur Verfügung stehen¹⁷¹.

Schritt 3: Verteilung der Kollokationen berechnen

Für eine vordefinierte Menge an Kollokationen¹⁷² wie z. B.

¹⁷⁰Definition in Schema 3 (**references**).

¹⁷¹Tatsächlich wurde zur Klassifizierung nicht die Tabelle **references** verwendet. Sie wurde vielmehr in eine genauere Tabelle **referenzen** überführt, die die einzelnen **IDTEA**-s in Lexeme, Positionen im Satz und POS-Tags auf-

spaltet. Dazu später mehr.

¹⁷²YAROWSKY [66] verwendet den Begriff hier im weiteren Sinne von umgebenden lexikalischen Einheiten und er soll nicht ausschließlich nur idiomatische oder nicht-kompositionelle Verbindungen implizieren.

- **+1w** Das Lexem direkt rechts vom Zielwort
- **+1w+2w** Das Bigramm rechts vom Zielwort
- **-1w** Das Lexem direkt links vom Zielwort

werden die Verteilungen für jede ambige Form des Zielwortes berechnet. Diesem Vorgehen liegt die Annahme der ungleichen Verteilung von bestimmten Kollokationsmustern hinsichtlich des zu klassifizierenden Lexems zugrunde. Wenn zusätzlich syntaktisch/morphologische Informationen zur Verfügung stehen, kann es sinnvoll sein, diese mit einzubeziehen und die Menge an *features* entsprechend zu erweitern.

Dieser Schritt wurde wegen der kleinen Trainingsmenge in soweit vereinfacht, dass auf eine statische Festlegung von Kollokationen und die damit verbundene Sichtung von entsprechend verteilten Kontext-Lexemen verzichtet wurde. Statt dessen, ergaben sich schon während der Phase der Vereinheitlichung der Annotationen signifikante *features*, die dann direkt in *decision lists* umgesetzt wurden. In den Tabellen 9 und 10 sind die für die Disambiguierung der Lexeme *jana* und *śārdūla* relevanten *features* und deren Verteilung bezüglich der verwendeten Bedeutungen dargestellt. Hier ist auffällig, dass sich *feature*-Mengen ergaben, die absolut charakteristisch für die jeweilige Bedeutung sind. Kennzeichnend hierfür ist die Tatsache, dass ein *feature* immer nur in Verbindung mit **einer** Bedeutung auftritt, d. h. die verbleibenden Felder einer Zeile beinhalten (fast) immer den Wert 0. Die starke Tendenz eines Wortes, nur eine Bedeutung pro Kollokation zu zeigen, wird schon von YAROWSKY [66, S. 92] berichtet.

Tabelle 9: Verteilung der *features* für *śārdūla*

ID	<i>feature</i>	<i>śārdūla</i> ₁	<i>śārdūla</i> ₇	<i>śārdūla</i> ₁₁
1	Hinterglied Komp.	2	117	0
2	Satz <i>rāvaṇa</i>	0	0	2
3	-1w <i>rākṣasa</i>	0	0	1
4	+1w <i>vākya</i>	0	0	1

Tabelle 10: Verteilung der *features* für *jana*

ID	<i>feature</i>	<i>jana</i> ₂	<i>jana</i> ₄	<i>jana</i> ₈	<i>jana</i> ₁₀	<i>jana</i> ₁₂
1	-1w <i>tapasvin</i>	1	0	0	0	0

2	-1w muni	1	0	0	0	0
3	-1w jānapada m.	0	0	6	0	0
4	-1w jānapada adj.	0	0	3	0	0
5	-1w bhaktimant	0	0	1	0	0
6	-1w duḥkhita	0	0	1	0	0
7	-1w ikṣvāku	0	0	1	0	0
8	-1w nātha	0	0	1	0	0
9	-1w guru	2	0	0	0	0
10	-1w shUra	1	0	0	0	0
11	-1w sakhI	0	5	0	0	0
12	-1w vividhā	0	0	1	0	0
13	-1w sa	0	0	3	0	0
14	-1w paura	0	0	6	0	0
15	-1w prakṛti	0	0	2	0	0
16	-1w ārya	1	0	0	0	0
17	-1w suhṛd	21	0	0	0	0
18	-1w strī	0	2	0	0	0
19	-1w sva	0	0	0	21	0
20	-1w para	0	0	0	1	0
21	-1w bandhu	0	0	0	5	0
22	-1w bhṛtya	0	0	0	0	1
23	+1w kashcid	1	2	0	0	1
24	+1w parikleśa	1	0	0	0	0
25	+1w janapada	0	0	4	0	0
26	+1w ogha	0	14	0	0	0
27	+1w shrīmant	1	0	0	0	0
28	+1w rājan	0	0	2	0	0
29	-2w samagra	0	0	0	0	1
30	-2w prakṛta	0	0	1	0	0
31	-2w jānapada m.	0	0	1	0	0
32	-2w-1w rājan pūj	0	0	1	0	0
33	-2w-1w vyathita sant.	0	0	1	0	0
34	-2w-1w rājya gata	0	0	1	0	0
35	-1w madvidha	1	0	0	0	0
36	-1w tvadvidha	2	0	0	0	0

37	-2w-1w mantra vid	1	0	0	0	0
38	-1w rakṣas	0	0	1	0	0

Dass derart „bedeutungsspezifische“ *features* festzustellen sind, lässt sich außerdem zu einem großen Teil darauf zurückführen, dass nur ein sehr kleines Trainings-Corpus von 419 Sätzen verwendet wurde.

Schritt 4: Verteilungslisten nach Log-Likelihood in decision lists sortieren

Für jede Kollokation und jeweils zwei verschiedene Akzentmuster wird das Verhältnis des **Log-Likelihood** berechnet, welches in Gleichung 8.1 dargestellt ist.

$$\text{LogL} = \left| \log \left(\frac{\text{Pr}(\text{Accent_Pattern}_1 | \text{Collocation}_i)}{\text{Pr}(\text{Accent_Pattern}_2 | \text{Collocation}_i)} \right) \right| \quad (8.1)$$

Anschließend werden die Verteilungslisten nach absteigendem LogL in *decision lists* sortiert.

Wenn es sich um Bedeutungen und *features* handelt, ist die obige Gleichung anzupassen. Dann werden natürlich nicht Akzentmuster und Kollokationen berücksichtigt, sondern entsprechend Wortbedeutungen und *features*. In unserem Fall führt jedoch eine Berechnung des LogL lediglich bei Tabelle 9 ID 1 und bei Tabelle 10 ID 23¹⁷³ überhaupt zu verschiedenen Werten, weil bei diesen Zeilen für *Bedeutung*₂ Werte ungleich 0 zur Verfügung stehen. Bei allen anderen Berechnungen ergäbe sich im Nenner 0, eine Division durch 0 ist in der hier verwendeten Arithmetik jedoch nicht definiert. Dies wäre durch die von YAROWSKY [66, S. 91 Fn. 6] vorgeschlagene Addition eines α von z. B. 0,15 zu umgehen, allerdings würde auch das zu identischen LogL-s führen, da somit der Zähler immer $1 + \alpha$ und der Nenner immer $0 + \alpha$ wäre.

Daher wurde in den *decision lists*, die im Laufe dieser Arbeit generiert wurden, auf eine Berechnung der LogL-s verzichtet. Stattdessen wurden für die Implementierung nützliche Substitut-Werte eingesetzt. Die Verteilungslisten 9 und 10 wurden in die *decision lists* 11 und 12 überführt und sind nachfolgend aufgeführt:

Tabelle 11: *decision list* für *śārdūla*

¹⁷³Dieses *feature* wurde wegen seiner geringen Aussagekraft nicht in die DL übernommen.

ID	feature	LogL	Classification
1	Hinterglied Komp.	1.77	23639
2	Satz rāvaṇa	1.00	23643
3	-1w rākṣasa	0.88	23643
4	+1w vākya	0.88	23643
5	default	0.00	23633

Tabelle 12: *decision list* für *jana*

ID	feature	LogL	Classification
1	-1w tapasvin	0.88	971
2	-1w vividhā	0.88	29687
3	-1w sa	0.88	29687
4	-1w paura	0.88	29687
5	-1w prakṛti	0.88	29687
6	+1w janapada	0.88	29687
7	+1w rājan	0.88	29687
8	-2w prakṛta	0.88	29687
9	-2w jānapada m.	0.88	29687
10	-2w-1w rājan pūj	0.88	29687
11	-2w-1w vyathita sant.	0.88	29687
12	-2w-1w rājya gata	0.88	29687
13	-1w rakṣas	0.88	29687
14	-1w sva	0.88	108719
15	-1w para	0.88	108719
16	-1w bandhu	0.88	108719
17	-1w bhṛtya	0.88	116036
18	-2w samagra	0.88	116036
19	-1w nātha	0.88	29687
20	-1w ikṣvāku	0.88	29687
21	-1w muni	0.88	971
22	-1w guru	0.88	971
23	-1w shūra	0.88	971
24	-1w ārya	0.88	971

25	-1w suḥṛd	0.88	971
26	+1w parikleśa	0.88	971
27	+1w shrīmant	0.88	971
28	-1w madvidha	0.88	971
29	-1w tvadvidha	0.88	971
30	-2w-1w mantra vid	0.88	971
31	-1w duḥkhita	0.88	29687
32	-1w bhaktimant	0.88	29687
33	-1w jānapada adj.	0.88	29687
34	-1w jānapada m.	0.88	29687
35	+1w ogha	0.88	973
36	-1w strī	0.88	973
37	-1w sakhī	0.88	973
38	-1w prākṛta	0.88	29687
39	default	0.00	973

Bei der *decision list* für *śārdūla* entspricht die absteigende Reihenfolge der *features* den prozentualen Verteilungen derselben. Tatsächlich kommt die Bedeutung *śārdūla*₇ (best) am häufigsten vor und ist statistisch am wahrscheinlichsten. Die nächsten drei *features* greifen die vier Vorkommen des Lexems in der Bedeutung *śārdūla*₁₁ (Śārdūla) ab. Das *default-feature* berücksichtigt die weit häufiger zugewiesene Bedeutung *śārdūla*₁ (tiger). Die Implementierung erwies sich jedoch mit dieser Logik am einfachsten, d. h. wenn keines der vorigen *features* eine Klassifizierung aktiviert, wird Bedeutung 23633 zugewiesen, was genau den Vorkommen des Lexems mit dieser Bedeutung entspricht. In einem echten System müssten natürlich die entsprechenden *features* auch für *śārdūla*₁ in die *decision list* mit aufgenommen werden.

Bei der *decision list* für *jana* verläuft das Verfahren fast identisch, allerdings werden die mehr oder weniger gleich verteilten *features* 1–37 zuerst getestet und bei Aktivierung wird entsprechend klassifiziert. Die häufigste Bedeutung *jana*₄ (people) wird, sofern vorher nicht anders klassifiziert wurde, als *default* zugewiesen.

Schritt 5: Optionale Optimierung – Pruning und Interpolation

Unter *Pruning* versteht man in der Informatik und dem maschinellen Lernen generell Methoden, die zur Vereinfachung und Optimierung von beispielsweise Entscheidungs-

oder Suchregeln führen. In Bezug auf DL gibt es zwei Arten von *Pruning*: 1. *features*, die gar nie erreicht werden können, werden aus der *decision list* entfernt, das löst das Problem der *redundancy by subsumption*. Gäbe es z. B. in der *decision list* von *śārdūla* an Stelle 2 eine Regel wie [-1w NOMEN]→ 23643, mit einer selbst definierten Klasse **NOMEN**, würde diese immer vor Regel 3 treffen und selbige somit redundant erscheinen lassen; durch *Pruning* könnte sie entfernt werden. 2. *features*, die auf einer Überanpassung an das Entwicklungs-Corpus beruhen und später zu mehr inkorrekten Klassifizierungen führen als zu korrekten, können auch durch *Pruning* entfernt werden.

Als weiteren Schritt zur Optimierung des Algorithmus schlägt Yarowsky die Interpolation der LogL-Verhältnisse, die für die gesamte Datenmenge berechnet wurden (globale Wahrscheinlichkeiten)¹⁷⁴ und der LogL-Verhältnisse, die zu einem bestimmten Zeitpunkt während der Abarbeitung der DL, nachdem alle höherrangigen Regeln nicht gegriffen haben, über der *verbleibenden* Trainingsmenge berechnet werden (verbleibende Wahrscheinlichkeiten)¹⁷⁵, vor. Als interpolierende Funktion gibt er 8.2 an:

$$\beta \times global + \gamma \times residual \quad (8.2)$$

Nach größerer Signifikanz und Kardinalität der jeweiligen Trainingsmenge (*global* oder *verbleibend*), soll entsprechend β oder γ dominieren.

Auf beide Möglichkeiten zur Optimierung wurde im Umfang dieser Arbeit verzichtet. Aufgrund der kleinen Trainingsmenge wurden weder eigene Klassen definiert und in *features* verwendet, noch konnte auf eine Überanpassung geprüft werden. Ein *Pruning* erschien somit nicht als indiziert. Da angenommen wurde, dass die LogL-s einer gegebenen Kollokation bzw. eines *features* auf jeder Zeile der *decision list* funktional äquivalent sind, kann bei der Interpolation immer $\beta = 1, \gamma = 0$ gelten¹⁷⁶ und somit werden nur globale Wahrscheinlichkeiten benutzt, die ja in diesem Fall ohnehin fast alle identisch sind.

Schritt 6: decision lists für allgemeine Klassen von Ambiguitäten erstellen

Für Typen ähnlicher Ambiguität können sich auch ähnliche *decision lists* ergeben, die auf gleichen Evidenzen beruhen. Yarowsky führt das Beispiel der Subjunktiv/Futur Unterscheidung zwischen *-ara/ará* im Spanischen an, welches durch nahe adverbielle Be-

¹⁷⁴Yarowsky verwendet den Begriff *global probabilities*.

¹⁷⁶Vgl. YAROWSKY [66, S. 92].

¹⁷⁵Yarowsky verwendet den Begriff *residual proba-*

stimmungen der Zeit klassifiziert wird. Er empfiehlt eine generelle *decision list* für alle *-ara/ará*-Ambiguitäten zu erstellen. Dies geschieht im Wesentlichen durch Anwendung der Schritte 2–5, mit der Ausnahme, in Schritt 2 nur eine in der Trainingsmenge ausgeglichene Anzahl von Instanzen von *-ara/ará* zu wählen. Anschließend wird die Exaktheit der allgemeinen Liste mit der der spezifischen Liste verglichen und wenn Sie ähnlich gute Ergebnisse erzielt, wird sie verwendet.

Dieser Schritt wurde für diese Arbeit nicht umgesetzt, da fast ausschließlich sehr spezifische *features* zu erkennen waren und die beiden generierten *decision lists* schon ausgezeichnete Ergebnisse lieferten.

Schritt 7: Die decision lists verwenden

Sobald die *decision lists* akquiriert wurden, können diese verwendet werden, um Ambiguitäten in neuen Kontexten durch Klassifizierung aufzulösen. Für jedes Vorkommen eines ambigen Wortes wird in der entsprechenden *decision list* nach dem ersten *feature* bei gegebenem Kontext gesucht das trifft und eine entsprechende Klassifizierung wird zugewiesen. Falls keine Regel greift, wird ein *default*-Wert zurück gegeben.

Aus einer statistischen Perspektive wird das erste *feature* in der entsprechenden Liste mit größter Wahrscheinlichkeit das Zielwort disambiguieren. Dieses Verfahren ist unverfänglich und in Bayes'scher Wahrscheinlichkeitstheorie begründet.

Entgegen anderen Verfahren wie NB, neuronale Netze, etc., die auch weitere, weniger wahrscheinlichere Kontext-*features* berücksichtigen, verwendet der Algorithmus der DL *nur* das erste aktivierte Kontext-*feature* und lässt spätere *features*, die möglicherweise auch aktiviert werden würden, außer Acht.

Tests haben jedoch gezeigt, dass ein solches Verfahren sehr gute Ergebnisse erzielt und eine Miteinbeziehung weiterer Evidenzen keine erkennbaren Vorteile mit sich bringt¹⁷⁷. Der geringere Rechenaufwand sorgt hingegen für bessere Laufzeiten.

Dieser Schritt wurde natürlich für diese Arbeit umgesetzt. Die Klassifizierungen wurden anhand der in der Datenbank-Tabelle 4 gespeicherten IDTEA-s durchgeführt und als neue Datensätze mit dem Attribut **IDUser** = 3, sowie entsprechender **ID**, **IDTEA** und **IDWortlisteOpencyc** eingefügt. Damit stehen sie für eine nachfolgende Evaluierung der Exaktheit des entwickelten Systems der WSD zur Verfügung.

¹⁷⁷Siehe YAROWSKY [66, S. 93].

8.3 Evaluierung der WSD durch DL

Der oben beschriebene Algorithmus wurde auf das mit [SanSemAn](#) annotierte Corpus¹⁷⁸ angewendet. Die relevanten *features* sind den Tabellen (*decision lists*) [11](#) und [12](#) zu entnehmen.

Da die interne Speicherung aller relevanter Daten in mySQL-Tabellen erfolgte, wurden die einzelnen Regeln entsprechend mit SQL-Kommandos, welche innerhalb von PHP-Schleifen auf die einzelnen Text-Passagen angesetzt wurden, getestet. So wurde für einen bestimmten Satz, der in einzelne Wörter mit POS-Tag aufgespaltet als Ausgabe einer SQL-Anfrage vorlag, je nach Zielwort, die relevante *decision list*, nach LogL in absteigender Reihenfolge sortiert, abgearbeitet. Für viele Vorkommen für das Lexem *jana* gab es nur eine spezifische Regel, die genau auf dieses Vorkommen abgestimmt wurde, da sich keine allgemeineren Strukturen erkennen ließen, das Zielwort zu disambiguieren. Ein solches Verfahren ist sehr zeitaufwändig, insbesondere dann, wenn es auf ein größeres Corpus angewendet worden wäre.

Hier ging es lediglich darum, überhaupt funktionierende *decision lists* für die Sanskrit Lexeme zu generieren und Erfahrungen über dieses unerforschte Gebiet zu sammeln.

Allerdings soll den hier erzielten Ergebnissen nicht aller Wind aus den Segeln genommen werden. Trotz der aufwändigen und spezialisierten *feature*-Mengen, konnte eine sehr große Exaktheit des Systems erreicht werden.

Eine prozentuale ITA zwischen den vereinheitlichten Annotationen von A1 und A2 und den Annotationen die vom Klassifikator getätigt wurden von 97,9%, grenzt an perfekte Übereinstimmung. Die κ -Statistik ist mit 0,97 fast identisch und sehr gut¹⁷⁹. Insgesamt wurden 9 Instanzen mit einer falschen Bedeutung annotiert. Die Grundlinie der Exaktheit, bei Zuweisung der häufigsten Bedeutung, liegt bei 68,5%¹⁸⁰.

3 Kategorien für *sārdūla* wurden falsch zugewiesen: bei den DCS IDTEA-s 497806, 512209 und 514095. Ersteres Vorkommen ist im Vokativ Sg. m. und alleinstehend, der Klassifikator wies *sārdūla*₁ (tiger) zu. Solche Fälle (beispielsweise alleinstehende Vorkommen) wurden in der Implementierung durch den *default*-Wert abgegriffen, ohne explizit ausformuliert worden zu sein. Dies weicht natürlich von dem Algorithmus ab, führt auf diesem künstlichen Niveau aber zum selben Ergebnis. Die menschlichen Annotatoren

¹⁷⁸Die Besprechung dieser Annotationen erfolgte in [Kapitel 7](#). ¹⁸⁰132 von 419 Bedeutungen weichen von den beiden häufigsten Bedeutungen *jana*₄ (people) und *sārdūla*₇ (best) ab.

¹⁷⁹Auch YAROWSKY [[66](#)] erzielte eine ITA von 97%.

entschieden sich bei dieser IDTEA für *śārdūla*₇ (best). Bei den beiden letzteren Vorkommen steht das Zielwort im Hinterglied eines Kompositums, das führt somit regelgerecht zu einer Klassifizierung von *śārdūla*₇ (best). Die menschliche Zuweisung war in beiden Fällen *śārdūla*₁ (tiger), da es sich bei IDTEAS 512209 um ein Tatpuruṣa, das einen Tiger spezifiziert handelt und bei IDTEA 514095 um ein Dvandva, in dem es ebenfalls um einen Tiger geht.

Für *jana* wurden 6 Kategorien falsch zugewiesen: bei den DCS IDTEA-s 386091, 404373, 415623, 418085, 1983457 und 1180654. Bei den ersten 5 Vorkommen wurde vom Klassifikator immer der *default*-Wert *jana*₄ (people) zugewiesen und von den menschlichen Annotatoren *jana*₈ (national). Dies erklärt sich dadurch, dass für den Menschen die Verwendung von *jana*₈ indiziert erscheint, da auf einer höheren semantischen Ebene die Satzstruktur oder mehrere signifikante Lexeme diese Verwendung rechtfertigen. Für die Maschine wurden genau solche *features* noch nicht kodiert.

Die letzte falsche Zuweisung des Klassifikators (*jana*₄) wurde vom Menschen mit *jana*₂ (man) bewertet, weil das Lexem im Singular steht und im Satzfenster *madvidha* auftaucht, dies deutet eine allgemeinere, generische Verwendung von *jana* an, welche ebenfalls noch nicht als *feature* kodiert wurde.

8.4 Ausblick

Abschließend kann angemerkt werden, dass das hier konstruierte System durchaus funktioniert und erfolgversprechend ist. Da bisher kaum Veröffentlichungen über das Thema WSD in Verbindung mit Sanskrit bekannt sind, sollte diese Arbeit untersuchen, ob sich diese Aufgabe überhaupt bewerkstelligen lässt.

Natürlich ist das oben aufgebaute Experiment stark vereinfacht und in dieser Art sicher auch nicht auf *echte* Anwendungen übertragbar, doch wurde gezeigt, dass eine maschinelle Wortbedeutungsdisambiguierung von digitalen Sanskrit-Daten mit guten Ergebnissen durchführbar ist.

In weiteren Experimenten sollten andere Algorithmen, insbesondere auf größeren Trainings-Corpora, getestet werden. Der hier verwendete DL-Algorithmus wurde strikt auf Datenbank-Tabellen aufgebaut. Weit eleganter wäre eine Kombination von Datenbank-Tabellen und programmierspracheninternen Speichermöglichkeiten.

Am wichtigsten erscheint jedoch, größere bedeutungsannotierte Sanskrit-Corpora zu

erstellen. Dafür sprechen die Überspezifiziertheit der *features* und eventuell entstehende Probleme bei deren Anwendung auf größere Corpora. Die Annotationen mit SanSemAn haben gezeigt, dass es unumgänglich ist, den Annotatoren die Möglichkeit zu geben, pro Vorkommen eines Zielwortes mehrere Bedeutungen zu annotieren. Dementsprechend muss auch über ein ausgeklügeltes Scoring-System¹⁸¹ nachgedacht werden.

Die Entwicklung eines solchen Systems, von einem fundierten DB-Design, über benutzerfreundliche Oberflächen, bis hin zu tatsächlichen Annotations-Unternehmen, sind ohne Frage sehr zeitintensiv.

Trotz der zeitaufwändigen Prognosen für zukünftige Forschungen, erscheinen die damit verbundenen Ergebnisse sehr vielversprechend und könnten klassischen Indologen die textkritische Arbeit sicher enorm erleichtern.

Glossar

AB AdaBoost. 33, 35

Ambiguität Lat. *ambiguitās* 'Doppelsinn'; engl. *ambiguity*; frz. *ambiguïté*; auch: Ambivalenz, Amphibolie, Mehrdeutigkeit. 5, 6

CBC Clustering by Committee Algorithmus. 36

CL Computer Linguistik; im Englischen oft auch mit NLP, Natural Language Processing bezeichnet. 1, 11, 36, 37, 76

DCS Digital Corpus of Sanskrit. <http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/>. 2, 24, 40, 44, 50–53

Diachronie Nach Saussure die „historische Sprachwissenschaft“. Die Achse der Aufeinanderfolge. 2, 11

Disambiguierung Beseitigung von Mehr- bzw. Zweideutigkeit, die durch lexikalische, grammatische oder syntaktische Ambiguität verursacht wird. 1, 61

DL Decision List Algorithmus. 31, 35, 37, 61, 66, 69, 72

DP Determinansphrase. 12

¹⁸¹Hier gab es ja nur eine Vergleichsbedeutung, ohne Berücksichtigung der Bedeutungshierarchien. Vgl. auch Kapitel 7.3.

- DSO** Das DSO-Corpus wurde von einem Team der Defence Science Organisation in Singapur zusammengestellt. Es ist ein bedeutungsannotiertes Corpus, welches englische Texte aus den Brown und Wall Street Journal Corpora enthält. Die Annotationen mit WordNet 1.5 *synsets* wurden von Linguistik-Studenten der Universität Singapur vorgenommen. 35
- EST** Erweiterte Standard-Theorie. Die erweiterte [Standard-Theorie](#) versucht eine inhaltliche Abgrenzung von Ambiguitäten und Vagheit bzw. Ambiguitäten und Nicht-Ambiguitäten in der [Standard-Theorie](#) auszudrücken. 5
- GS** Die generative Semantik stützt sich auf die „Aspects“ von CHOMSKY [15], ist jedoch eine stark davon abweichende Grammatiktheorie, die den Erzeugungsprozess von Sätzen nicht mit syntaktischen Strukturen beginnt, um diese semantisch zu interpretieren, sondern semantische Strukturen als direkt erzeugbar betrachtet. 5, 6
- gTG** Eine generative Transformationsgrammatik ist eine um Transformationsregeln erweiterte generative Grammatik \rightarrow PSG nach Chomsky, die den taxonomischen Strukturalismus überwindet. Eine gTG ist ein Tupel $gTG = (G, T, P)$. G ist eine allg. Regelgrammatik, T eine endl. Menge von Transformationen und P ein Programm für die Reihenfolge von T . 3-5, 76
- HAL** Hyperspace analouge to Language Algorithmus. 36
- Hyponymie** Hyponymie ist eine semantische Relation zwischen lexikalischen Einheiten: Im extensionalen Sinn ist die Bedeutung von \langle Rose \rangle inkludiert in der Bedeutung von \langle Blume \rangle , d.h. die Extension von \langle Blume \rangle ist weiter als die von \langle Rose \rangle . Im intensionalen Sinn ist das umgekehrt: hier impliziert \langle Rose \rangle den Inhalt von \langle Blume \rangle und die Intension ist entsprechend größer. So kann H. als Implikation definiert werden: $\text{Rose} \rightarrow \text{Blume}$; damit sind \langle Rose \rangle und \langle Tulpe \rangle Ko-Hyponyme oder artgleiche Begriffe des Hyperonyms \langle Blume \rangle . Vergleiche dazu auch LEWANDOWSKI [41, S. 395]. 9, 10, 57, 76
- IC-Analyse** Immedient constituent analysis; Konstituentenanalyse; ihr liegt die Einsicht zugrunde, dass zwischen Satz und Wort Zwischenstufen in hierarchischer Anordnung liegen, die Satzglieder, die den Satz konstituieren. 4

- ITA** Inter-Tagger-Agreement. Auch: IAA Inter-Annotator-Agreement. Wenn nicht anders angegeben, ist hier schlicht die prozentuale Übereinstimmung gemeint. [39](#), [47](#), [49](#), [51](#), [59](#)
- KI** Künstliche Intelligenz. [1](#), [11](#), [25](#)
- kNN** k-Nearest Neighbor Algorithmus. [33](#), [35](#)
- langue** Sprachtheoretischer Begriff Saussures: das überindividuelle und konventionelle Sprachsystem als Inventar bzw. „Wörterbuch“ von Zeichen und Regeln, das der aktual-konkreten Rede \rightarrow [parole](#) zugrunde liegt. [2](#)
- LSA** Latent Semantic Analysis Algorithmus. [36](#)
- MWE** Multi-Word Expressions. [49](#)
- NB** Naïver Bayes Algorithmus. [32](#), [35](#), [70](#)
- Onomasiologie** Namenskunde, Bezeichnungslehre, Begriffsforschung. [1](#)
- OpenCyc** OpenCyc ist ein Bedeutungs-Inventar zur maschinellen Auswertung von Alltagswissen. <http://www.opencyc.org>. [41](#)
- parole** Nach Saussure: Sprechen, Rede; individueller Gebrauch des Sprachsystems zu konkret aktueller Verständigung. [2](#), [75](#)
- POS** Part-Of-Speech. Wortart. [35](#)
- PP** Präpositionalphrase. [12](#)
- PSG** Phrasenstrukturgrammatik. Eine PSG ist ein Tupel, mit $PSG = (V, E, S, P)$. V ist das Vokabular, E das Endvokabular, S das Startsymbol und P die Menge der Produktionen der Form $X \rightarrow Y$. [3](#), [74](#)
- SanSemAn** Der **SanskritSemAnnotator**. Ein Multi-Annotatoren-System zur semantischen Annotation. Im August 2011 unter <http://sanseman.asyavamasya.com>. [1](#), [40](#), [45](#), [49](#), [51](#), [56](#), [61](#), [71](#)
- Semi-Thue-System** Ein STS ist ein Tupel (Σ, S) mit dem Alphabet Σ und einer Menge S von Substitutionen mit $S \subseteq \Sigma^* \times \Sigma^*$. Für die auf Σ^* definierte, einschriftige Ableitungsrelation \Rightarrow_s gilt: $w_1 \Rightarrow_s w_2$, genau dann, wenn es Wörter $\alpha, \beta \in \Sigma^*$

gibt und $u \rightarrow v \in S$, so dass gilt: $w_1 = \alpha u \beta$ und $w_2 = \alpha v \beta$. Somit ist α Präfix beider Wörter und β analog Suffix. 3, 4

Standard-Theorie Die Standardtheorie (vgl. CHOMSKY [15]) ist eine Erweiterung der gTG , um syntaktisch-semantische Beziehungen zwischen Sätzen, um die Erzeugung von Sätzen besser darstellen zu können. 6, 74

SVM Support Vector Machines. 34, 35

Synchronie Bei Saussure ein Sprachzustand (im Gegensatz zum Entwicklungszustand einer Sprache). Die Achse der Gleichzeitigkeit. Im Zustand der S. ist auch der Zustand der D. enthalten. 2, 11

VP Verbalphrase. 12

WordNet Das Princeton WordNet <http://wordnet.princeton.edu> ist eine der meist genutzten lexikalischen Ressourcen des Englischen in der CL. Wörter werden in Mengen von Synonymen gruppiert (*synsets*), kurze Definitionen gegeben und die semantischen Relationen zu anderen *synsets* werden repräsentiert. WordNet liefert eine semantische Hierarchie (vgl. **Hyponymie**) und ein semantisches Netzwerk von Wortbeziehungen. 28, 30, 57

WSD Word sense disambiguation. Wortbedeutungsdisambiguierung. 1, 14, 25, 37, 57, 62

Sanskrit Abkürzungen

ŚB Śatapatha Brāhmaṇa. 16

ṚV Ṛg Veda. 15

AV Atharva Veda. 15

Br. Up. Bṛhadāraṇyakopaniṣad. 15

Bg. Bhagavadgītā. 19

Mīmāṃsā Mīmāṃsā-darśana. 17

Mādh. Mādhava. Früher fälschlich als Sāyaṇa bekannt. Ein prominenter Kommentator des ṚV. 20

Manu. Manu-Smṛti. 17, 21

Mbht. Mahābhārata. 14, 16, 21, 50

Nir. Nirukta des Yāska. 20

Pañcat. Pañcatantra oder Trantrākhyāyika. 51

Rām. Rāmāyaṇa des Vālmīki. 45, 49, 51

Stellenverzeichnis

ŚB 1.5.2.19, 18

ŚB 1.6.3.41, 16

ṚV 1.101.5, 15

ṚV 1.24.11, 15

ṚV 10.121.9, 22

ṚV 10.85.9, 21

ṚV 3.35.4, 19

ṚV 3.8.9, 16

ṚV 4.3.6, 16

ṚV 4.48.1, 22

ṚV 5.22.4, 22

ṚV 5.37.4, 21

Amara 2.6.38, 22

AV 12.1.36, 18

AV 19.19.8, 15

AV 2.15.3, 23

AV 3.15.4, 19

Br. Up. 1.4.10, 15

Bg. 2.47, 20

Mīmāṃsā 1.4.23, 17

Mādh. zu ṚV 1.137.1, 20

Manu. 1.46, 21

Manu. 2.181, 17

Manu. 2.21, 23

Mbht. 3.130.3, 23

Mbht. 3.130.4, 23

Mbht. 6.113.41, 21

Mbht. 8.161.3, 16

Nir. 2.5, 20

Abbildungsverzeichnis

1	Proportionale Verteilung der Bedeutungen von <i>jana</i>	48
2	Proportionale Verteilung der Bedeutungen von <i>śārdūla</i>	48
3	Proportionale Verteilung der Bedeutungen von <i>jana</i> nach dem 1. Durchlauf und bei der finalen Version	62

Code-Schnipsel

1	Lesk-Algorithmus	26
2	<code>ontology</code>	40
3	<code>references</code>	41
4	<code>ontology_reference</code>	41
5	<code>users</code>	42

Tabellenverzeichnis

1	Klassifizierungs-Beispiel von 'know' durch <i>decision list</i>	31
2	Bedeutungen von <i>jana</i>	46
3	Bedeutungen von <i>śārdūla</i>	47
4	Tatsächlich verwendete Bedeutungen von <i>jana</i> und <i>śārdūla</i>	47
5	Zwei Äste der WordNet-Hierarchie mit Bedeutungen von <i>jana</i>	57
6	Kontingenz-Tafel von <i>jana</i> und <i>śārdūla</i> gemeinsam	60
7	Kontingenz-Tafel von <i>jana</i>	60
8	Kontingenz-Tafel von <i>śārdūla</i>	60
9	Verteilung der <i>features</i> für <i>śārdūla</i>	64
10	Verteilung der <i>features</i> für <i>jana</i>	64
11	<i>decision list</i> für <i>śārdūla</i>	66
12	<i>decision list</i> für <i>jana</i>	67

Literatur

- [1] AGIRRE, ENEKO (HRSG.): *Word sense disambiguation : algorithms and applications*. Berlin : Springer, 2007 (Text, Speech and Language Technology ; 33)
- [2] AGIRRE, ENEKO ; MARTÍNEZ, David: Learning class-to-class selectional preferences. In: *Proceedings of the Conference on Natural Language Learning*. Toulouse, 2001, S. 15–22
- [3] AMARASIṂHA ; PĀDHYE, NARHAR G. (HRSG.): *Nāmaṅgānuśāsanaṃ*. 2. Ed. Poona : Oriental Book Agency, 1969 (Poona oriental series; 69)
- [4] BEEH, VOLKER: Eine Verallgemeinerung der rewrite rule. In: *Deutsche Sprache. Zeitschrift für Theorie, Analyse und Dokumentation*. München, 1973, S. 1–15
- [5] BLOOMFIELD, MAURICE (HRSG.): *Hymns of the Atharva-Veda : together with extracts from the ritual books and the commentaries*. Oxford : Clarendon, 1897 (The sacred books of the East ; 42)
- [6] BUCK, CARL D.: *A dictionary of selected synonyms in the principal Indo-European languages : a contribution to the history of ideas; with the co-operation of colleagues and assistants*. Chicago : Univ. Press, 1949
- [7] BUITELAAR, PAUL: The SENSEVAL-II Panel on Domains, Topics and Senses. In: *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-II)*. Toulouse, 2001
- [8] BUITENEN, JOHANNES A. B. v. (HRSG.): *The Mahābhārata*. Bd. 3: 4: The book of Virāṭa. 5: The book of the effort. Chicago : Univ. of Chicago Pr., 1978
- [9] BUITENEN, JOHANNES A. B. v.: *The Bhagavadgītā in the Mahābhārata : text and translation*. Chicago : Univ. of Chicago Press, 1981
- [10] BUSSMANN, CLAUDIA (HRSG.): *Lexikon der Sprachwissenschaft*. 4., durchges. u. bibliogr. erg. Aufl. Stuttgart : Kröner, 2008
- [11] BÖHTLINGK, RUDOLF / v.: *Sanskrit-Wörterbuch*. St. Petersburg : Kaiserl. Akad. d. Wissenschaften, 18XX
- [12] BÜHLER, GEORG (HRSG.): *The laws of Manu*. Delhi [u.a.] : Motilal Banarsidass, 1975 (The sacred books of the East ; 25)

- [13] CHOMSKY, NOAM: *Syntactic structures*. s-Gravenhage : Mouton, 1957 (Ianua linguarum : Series minor ; 4)
- [14] CHOMSKY, NOAM: *Current issues in linguistic theory*. The Hague [u.a.] : Mouton, 1964 (Ianua linguarum : Series minor ; 38)
- [15] CHOMSKY, NOAM: *Aspects of the theory of syntax*. Cambridge, Mass. : MIT Pr., 1965 (Special technical report / Massachusetts Institute of Technology, Research Laboratory of Electronics ; 11)
- [16] CHOMSKY, NOAM: *The logical structure of linguistic theory*. New York : Plenum Press, 1975
- [17] COHEN, JACOB: A Coefficient of Agreement for Nominal Scales. In: *Educational and Psychological Measurement* 20 (1960), Nr. 1, S. 37
- [18] EDMONDS, PHILIP: Lexical disambiguation. In: BROWN, KEITH (HRSG.): *Encyclopedia of language & linguistics*. 2. ed. Amsterdam : Elsevier, 2006
- [19] EGGELING, JULIUS: *The satapatha-brāhmaṇa according to the text of the Mādhyandina school*. Oxford : Clarendon Press, 1882–1900 (Sacred book of the East)
- [20] FRIES, NORBERT: *Ambiguität und Vagheit : Einführung und kommentierte Bibliographie*. Tübingen : Niemeyer, 1980 (Linguistische Arbeiten ; 84)
- [21] GARBE, RICHARD (HRSG.): *Die Bhagavadgītā : aus dem Sanskrit übersetzt; mit einer Einleitung über ihre ursprüngliche Gestalt, ihre Lehren und ihr Alter*. 2. verb. Aufl. Leipzig : Haessel, 1921
- [22] GELDNER, KARL F.: *The Harvard Oriental Series*. Bd. 33–36: *Der Rig-Veda I–IV*. Cambridge : Harvard University Press, 1951
- [23] GENTZEN, GERHARD: Untersuchungen über das logische Schließen I. In: *Mathematische Zeitschrift* Bd. 39. Berlin, Heidelberg : L. Lichtenstein, 1934, S. 176–210
- [24] GENTZEN, GERHARD: Untersuchungen über das logische Schließen II. In: *Mathematische Zeitschrift* Bd. 39. Berlin, Heidelberg : L. Lichtenstein, 1935, S. 405–431
- [25] GRASSMANN, HERMANN: *Wörterbuch zum Rig-Veda*. Wiesbaden : Harrassowitz Verlag, 1996. – 6., überarbeitete und ergänzte Auflage von Maria Kozińska

- [26] GREIMAS, ALGIRDAS J.: *Strukturelle Semantik : methodologische Untersuchungen*. Braunschweig : Vieweg, 1971 (Wissenschaftstheorie, Wissenschaft und Philosophie ; 4)
- [27] HALLIDAY, MICHAEL ; HASAN, RUQAIYA: *Cohsion in English*. London : Longman, 1976
- [28] HAMM, FRITZ (HRSG.) ; KAMP, HANS (HRSG.) ; SOLSTAD, TORGRIM (HRSG.) ; ROSSDEUTSCHER, ANTJE (Hrsg.): *Disambiguation and Reambiguation*. Stuttgart : Universitätsbibliothek der Universität Stuttgart, 2009 (SinSpeC ; 6)
- [29] HELLWIG, OLIVER: A computational framework for linguistic research in post-Vedic Sanskrit. In: *14th World Sanskrit Conference*. Kyoto, 2009
- [30] HELLWIG, OLIVER: SanskritTagger, a stochastic lexical and POS tagger for Sanskrit. In: *Sanskrit Computational Linguistics*. Berlin : Springer Verlag, 2009 (First and Second International Symposia), S. 266–277
- [31] HUET, GÉRARD: Lexicon-directed Segmentation and Tagging of Sanskrit. In: *Themes and Tasks in Old and Middle Indo-Aryan Linguistics*. Delhi : Motilal Banarsidass, 2006, S. 307–325
- [32] JACKENDOFF, RAY: *Semantic interpretation in generative grammar*. 2. print. Cambridge, Mass. [u.a.] : MIT Pr., 1975 (Studies in linguistics series ; 2)
- [33] JÄGER, LUDWIG: L. Jäger: Saussure Kritik ohne Text-Kritik? In: *Zeitschrift für germanistische Linguistik* 5, 1977, S. 298–312
- [34] JÄGER, LUDWIG: L. Jäger: F. de Saussures semiologische Begründung der Sprachtheorie. In: *Zeitschrift für germanistische Linguistik* 6, 1978, S. 18–30
- [35] KĀMBOJA, JIYĀ L.: *Semantic change in Sanskrit*. 1. ed. Delhi : Nirman Prakashan, 1986. – Teilw. zugl.: Dehli, Univ., 1973
- [36] KILGARIFF, ADAM: Senseval: An exercise in evaluating word sense disambiguation programs. In: *Proceedings of the European Conference on Lexicography (EURALEX)*. Liège, 1998, S. 174–176
- [37] KILGARIFF, ADAM: English lexical sample task description. In: *Proceedings of Senseval-2: Second international Workshop on evaluating Word Sense Disambiguation Systems*. Toulouse, 2001, S. 17–20

- [38] KUPFER, KATHARINA: *Die Demonstrativpronomina im Rigveda*. Europäische Hochschulschriften. 2002 (Linguistik 244 Reihe 21)
- [39] LEACOCK, CLAUDIA ; CHODOROW, MARTIN ; MILLNER, GEORGE A.: Using corpus statistics and WordNet relations for sense identification. In: *Computational Linguistics* 24(1) (1998), S. 147–165
- [40] LESK, MICHAEL: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: *Proceedings of the ACM-SIGDOC Conference*. Toronto, Canada, 1986, S. 24–26
- [41] LEWANDOWSKI, THEODOR: *UTB ; 1518*. Bd. 1: *Linguistisches Wörterbuch*. 6. Aufl., unveränd. Nachdr. d. 5., überarb. Aufl. Heidelberg : Quelle & Meyer, 1994
- [42] LEWANDOWSKI, THEODOR: *UTB ; 1518*. Bd. 2: *Linguistisches Wörterbuch*. 6. Aufl., unveränd. Nachdr. d. 5., überarb. Aufl. Heidelberg : Quelle & Meyer, 1994
- [43] LEWANDOWSKI, THEODOR: *UTB ; 1518*. Bd. 3: *Linguistisches Wörterbuch*. 6. Aufl., unveränd. Nachdr. d. 5., überarb. Aufl. Heidelberg : Quelle & Meyer, 1994
- [44] MACDONELL, ARTHUR A.: *A Vedic Grammar for Students*. Low Prive Publications, 1997
- [45] MAYRHOFER, MANFRED: *Etymologisches Wörterbuch des Altindoarischen*. Bd. I-III. Heidelberg : Carl Winter – Universitätsverlag, 1992
- [46] MELAMED, I. D. ; RESNIK, PHILIPP: Tagger evaluation given hierarchical tag sets. In: *Computers and teh Humanities* 34(1–2) (2000), S. 79–84
- [47] MIHALCEA, RADA ; CHKLOVSKI, TIMOTHY ; KILGARRIFF, ADAM: The SENSEVAL-3 English Lexical Sample Task. In: *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, 2004
- [48] MIHALCEA, RADA ; MOLDOVAN, DAN: A method for word sense disambiguation of unrestricted text. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Maryland, 1999, S. 152–158
- [49] MIHATSCH, WILTRUD: *Lexical data and universals of semantic change*. Tübingen : Stauffenburg, 2004 (Stauffenburg-Linguistik ; 35)

- [50] MONIER-WILLIAMS, MONIER: *A Sanskrit-English dictionary etymologically and philologically arranged with special reference to cognate indo-european languages.* Oxford Clarendon Press, 1960
- [51] MYLIUS, KLAUS: *Wörterbuch Sanskrit-Deutsch.* 6. Aufl. Leipzig [u.a.] : Langenscheidt, Verl. Enzyklopädie, 1999
- [52] RADFORD, ANDREW (HRSG.): *Linguistics : an introduction.* 2. ed. Cambridge : Cambridge Univ. Press, 2009
- [53] RAJAN, MANDYAM ANANDAMPILLAI S.: *Sanskrit and computer-based linguistics : proceedings of the Seminar on Knowledge Representation in Sanskrit & Allied Semantic Theories; Bangalore, March 6 - 7, 1993 = Saṃvit.* Melkote : Seminar on Knowledge Representation in Sanskrit & Allied Semantic Theories, 1993 (The Academy of Sanskrit Research series ; 27). – Beitr. teilw. engl., teilw. Sanskrit
- [54] RESNIK, PHILIP: *Selection and information: A class-based approach to lexical relationships,* University of Pennsylvania, Dissertation, 1993
- [55] RESNIK, PHILIP: Selectional preference and sense disambiguation. In: *Proceedings of ACL Workshop on Tagging Text with Lexical Semantics, Why, What and How?* Washington, 1997, S. 52–57
- [56] REVELLE, WILLIAM: *psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University (Veranst.), 2010. – R package version 1.0-93
- [57] RICHTER, ELISE: *Über Homonymie.* Festschrift für Universitätsprofessor Hofrat Dr. Paul Kretschmer. Beiträge zur Griechischen und Lateinischen Sprachforschung. Wien; Leipzig, 1926. – 167–201 S
- [58] ROLFES, EUGEN (HRSG.): *Philosophische Bibliothek.* Bd. 8f: *Aristoteles: Kategorien/ Lehre vom Satz. und Porphyrius: Einleitung in die Kategorien.* Übersetzt und mit einer Einleitung versehen von Eugen Rolfes. Unveränd. Neuausg. d. 2. Aufl. von 1928. Hamburg : Meiner, 1958
- [59] SAUSSURE, FERDINAND D. ; BALLY, CHARLES (HRSG.): *Grundfragen der allgemeinen Sprachwissenschaft.* 3. Aufl. / mit einem Nachw. von Peter Ernst. Berlin : de Gruyter, 2001 (De-Gruyter-Studienbuch)

- [60] SCHUMACHER, KINGA: Four Methods for Supervised Word Sense Disambiguation. In: SPRINGER (HRSG.): *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB) June 27-29, 2007, CNAM, Paris, France. Springer*. Springer, 6 2007 (LNCS 4592), S. 317–328
- [61] TARSKI, ALFRED: The Semantic Conception of Truth and the Foundations of Semantics. In: *Philosophy and Phenomenological Research*. Berkeley : University of California, 1944
- [62] ULLMANN, STEPHEN: *Semantics : an introduction to the science of meaning*. Oxford : Blackwell, 1962
- [63] WACKERNAGEL, JAKOB: *Altindische Grammatik*. Bd. 2,2: Die Nominalsuffixe. Göttingen : Vandenhoeck & Ruprecht, 1954
- [64] WELLS, RULON S.: Immediate Constituents. In: *Lg* 23, 1947, S. 81–117
- [65] WOLSKI, WERNER: *Schlechtbestimmtheit und Vagheit, Tendenzen und Perspektiven : methodologische Untersuchungen zur Semantik*. Tübingen : Niemeyer, 1980 (Germanistische Linguistik 28)
- [66] YAROWSKY, DAVID: Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Las Cruces, 1994, S. 88–95

- Guttenbergerklärung -

Hiermit erkläre ich gemäß §22 Abs. 5 der Prüfungsordnung für die Magisterstudiengänge der Philosophischen Fakultät der Universität Tübingen, dass die Arbeit von mir selbständig verfaßt wurde und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet wurden.

Jonas Soiné

Stuttgart, den 8. August 2011