Zur Morphologie und Semantik von Nominalkomposita

Stefan Langer
Centrum für Informations- und Sprachverarbeitung
Universität München
Oettingenstr. 67
D 80538 München
stef@cis.uni-muenchen.de

English abstract

The treatment of German compound nouns in electronic lexicography is confronted with two major difficulties: correct segmentation and semantic interpretation. For the correct segmention of non-restricted noun compounds a complete encoding of connecting elements — which are called compounding suffixes in this article — is needed. These have to be encoded as morphological features of the potential first constituents. The first part of this paper contains a description of the methods and the results of the encoding of compounding suffixes for a complete lexicon of German non-complex nouns. In the second part, corpus statistics of relations between semantic classes and compound heads are presented, and it will be shown that they they are useful for the disambiguation of polysemous constituents and the determination of relations between them.

Zusammenfassung

Die Behandlung von deutschen Komposita in der elektronischen Lexikographie läßt sich in zwei Schritte gliedern — die Segmentierung und die Zuordnung lexikalischer Information. Eine korrekte Zerlegung ist nur möglich, wenn die Formen der in Komposita auftretenden Erstglieder erkannt werden können, die nicht immer einer Flexionsform entsprechen, und insbesondere im Falle von nominalen Erstgliedern nicht ohne weiteres herleitbar sind. Um diese Kompositionsformen zu erfassen wurden für ein umfassendes elektronisches Lexikon sämtliche Fugenelemente (im Folgenden als Kompositionssuffixe bezeichnet) bei Nomina unter Hinzuziehung von Korpusdaten kodiert. An die morphologische Kodierung schloß sich eine statistische Untersuchungen zur Semantik zweigliedriger Nominalkomposita an. Die aus dem Korpus extrahierten Daten zu den Selektionseigenschaften von Kompositaköpfen ermöglichen es, polyseme Köpfe und Erstglieder teilweise zu disambiguieren, sowie die Relationen zwischen Kompositagliedern zu erkennen.

1 Einführung

1.1 Hintergrund

Komposition ist ein extrem produktives Muster der deutschen Wortbildung. In großen Korpora lassen sich ohne weiteres mehrere Millionen verschiedene Komposita finden. Die große Zahl von Komposita können mit realistischem Aufwand nicht manuell lexikographisch erfaßt werden. Doch auch die automatische Zuweisung von morphologischen oder semantischen Eigenschaften ist nicht unproblematisch. Bereits die automatische Segmentierung ist nicht fehlerfrei vorzunehmen. Besondere Schwierigkeiten macht die Segmentierung von Komposita, wenn es sich beim Erstglied um ein Nomen handelt, da dessen Form nicht in jedem Falle einer Flexionsform entspricht (Augst 1975). Eine Analyse, die nur Stämme und Flexionsformen, sowie ein nicht-paradigmisches *-s* femininer Erstglieder berücksichtigt, kann zwar recht brauchbare Resultate erzielen (Lezius 1996), aber viele Komposita können auf diese Weise nicht segmentiert werden, und es werden Formen als Erstglieder zugelassen, die als solche nie auftreten können. Zur Vermeidung von Unter- bzw. Übergenerierung ist nötig, alle und nur die Formen von Nomina zu erfassen, die als Erstglieder in Komposita eingehen. Eine solche Form der Kodierung wurde bereits früher beschrieben in einer IBM-Studie (Rackow 1992, 11) zur Behandlung von Komposita in einem System zur maschinellen Übersetzung (LMT). Hier werden 28 Kompositionssuffixe bei deutschen Nomina aufgezählt. Ein ähnlicher Ansatz wird für das Morphologieanalysesystem GERTWOL beschrieben (Haapalainen 1995), und liegt auch der Kodierung potentieller Erstglieder für das vollständige elektronische Wörterbuch CISLEX (Guenthner/Maier 1996) (Guenthner 1996) zugrunde, die im ersten Teil des vorliegenden Artikels vorgestellt wird.

Die lexikographische Erfassung dieser Regularitäten ermöglicht eine oder mehrere morphologisch korrekte Segmentierungen; dies ist notwendige Voraussetzung für eine semantische Analyse. Bei Determinativkomposita — die die Mehrzahl der Komposita im Deutschen ausmachen — können zahlreiche Relationen zwischen den Gliedern auftreten, die jedoch in der Lexemform nicht in Erscheinung treten. Während die meisten Komposita von muttersprachlichen Sprechern des Deutschen eindeutig interpretiert werden können — d. h. die Relation zwischen den Gliedern wird erkannt — bringt die Erkennung der nicht-ausformulierten Relation Schwierigkeiten für die automatische Interpretation mit sich. Im zweiten Teil dieser Arbeit wird gezeigt werden, daß die statistische Analyse der Semantik von

Komposita Ergebnisse liefert, die zur Erkennung von Relationen und zur Disambiguierung polysemer Bestandteile herangezogen werden können.

2 Morphologische Kodierung

2.1 Zur Terminologie

Unter einem „Fugenelement“ (so u.a. Fleischer/Barz (1992, 136), Grube (1976) und Fuhrhop (1995)), auch als „Fugenmorphem“ (Augst 1975, 1300) oder „Fugenzeichen“ (Drosdowski 1984, 450) bezeichnet, wird ein Wortbildungselement bezeichnet, das scheinbar als Interfix zwischen den beiden Gliedern eines Kompositums auftritt. Es ist aber seit langem bekannt, daß gegen eine Auffassung als Interfix die Abhängigkeit der Form dieses Wortbildungselements vom Erstglied spricht (Fanselow 1981, 10). Der Begriff „Fugenelement“ ist also (wie auch der Begriff „connecting element“ im Englischen) schlecht gewählt. An sich handelt es sich bei der Kombination Erstglied plus (sogenanntes) Fugenelement um nichts anderes, als eine besondere Form eines Nomens, wie es als Erstglied in Komposita eingeht. Dies wird weitgehend auch in der Forschungsliteratur erkannt (Fleischer/Barz 1992, 138), ohne daß sie jedoch die entsprechenden terminologischen Konsequenzen gezogen werden. Mir scheint jeglicher Begriff, der ein Kompositum aus „Fuge“ plus X ist, irreleitend, weshalb im folgenden Artikel durchgängig eine andere Terminologie verwendet wird. Ich spreche stets vom **Kompositionssuffix** des Erstglieds eines Kompositums. Aus der Kombination des nominalen Erstgliedes und einem Kompositionssuffix entsteht die **Kompositionsform** eines Nomens.

Klärungsbedürftig ist ferner das Verhältnis zwischen **Kompositionsformen** und **Flexionsformen**. Zahlreiche Kompositionsformen sind homograph mit der Genitiv-Singular-Form des Lexems. Wie die Forschung schon seit langem festgestellt hat (Fleischer/Barz 1992), sprechen gegen die Interpretation der mit dem Genitiv homographen Kompositionsformen als Flexionsformen gewichtige Gründe.

Bei den als Erstgliedern auftretenden pluralhomonymen Formen gestaltet sich die Entscheidung für oder gegen die Identifizierung mit der Pluralform schwieriger. Der Hauptgrund hierfür ist, daß sich nicht vielen Fällen bei mehreren Kompositionsformen eines Nomens, von denen nur eine der Pluralform entspricht, eine klare semantische Differenz insofern festmachen läßt, als hier pluralische Bedeutung des Erstgliedes zugrundeliegt (Weinrich 1993, 973f) . Allerdings korrespondieren Pluralform bzw. Singularform

keineswegs immer mit Plural- oder Singularbedeutung, und eine Abgrenzung zu den Fällen, in denen die semantische Differenz auf die Pluralbedeutung zurückzuführen ist, fällt äußerst schwer. Aus diesem Grund behandeln Grube (1976) und Rackow (1992) konsequent alle als Erstglied auftretenden Formen eines Lexems als Kompositionsformen, und machen keinen Unterschied zwischen pluralischen und nicht-pluralischen Erstgliedern, eine Vorgehensweise, die auch der hier vorgestellten Kodierung zugrundeliegt.

2.2 Datenquellen

Gedruckte Lexika der deutschen Sprache führen keineswegs systematisch Kompositionsformen auf. Die diesbezüglichen Angaben beschränken sich auf die Auflistung häufiger oder lexikalisierter Komposita, die natürlich dann die entsprechenden Kompositionsformen enthalten. Allerdings bietet die Liste der Kompositionsmorpheme in Augst (1975) eine beindruckende Abdeckung der Daten. Dennoch ist im Sinne des hier verfolgten Ansatzes unvollständig, da zahlreiche Pluralmorpheme, die ich ebenfalls als Kompositionssuffixe auffasse, nicht berücksichtigt werden. Aufgrund der fehlenden Daten aus Lexika mußte für die Kodierung auf ein Korpus zurückgegriffen werden, das 190 000 unterschiedliche Komposita enthielt. Sie entstammten größtenteils einem Textkorpus, das aus der Süddeutschen Zeitung gewonnen war, teilweise jedoch auch aus Fachtexten.

Die Kodierung war jedoch nicht vollständig korpusbasiert, denn von zahlreichen selteneren einfachen Nomina treten schlichtweg keine Komposita in Korpora auf. Nicht auftretende Nomina wurden soweit möglich in Analogie zu den häufigeren Nomina kodiert.

2.3 Verifizierung der Kodierung

Nach erfolgter Kodierung wurden die ermittelten Kompositionsformen mithilfe eines Segmentierungprogramms einem Testlauf auf dem Korpus unterzogen. Es traten zwei unterschiedliche Phänomene in Erscheinung, die auf Fehler bei der Kodierung der Kompositionsformen hinweisen konnten: Ein Kompositum konnte nicht zerlegt werden oder es wurde falsch zerlegt.

Eine Reihe von Wörtern wurden nicht zerlegt. Hierbei sind nur die Fälle interessant bei denen im Korpus eine falsche oder fragliche Kompositionsform auftauchte, insbesondere das fehlende *-s*-Suffix vor *s*, wie etwa in *Sitzungs-saal*. Da die Kompositionsformen nicht vollständig normiert sind, läßt sich nur schwer entscheiden, wo eine ungewöhnliche Form und wo ein Fehler vorliegt. Wir entschlossen uns, die fraglichen Belege, wenn sie allzu

zweifelhaft erschienen und auch die konsultierten Lexika keinen Anhaltpunkt für ihre Richtigkeit boten, als Schreibfehler einzustufen. Schreibfehler bei Kompositionsformen sind sehr häufig, u.a., weil sie von gängigen Rechtschreibkorrekturprogrammen nicht erfaßt werden.

Neben nicht zerlegten Komposita kamen auch zahlreiche Fälle falscher Zerlegung vor. Völlige Unplausibilität hatte meist den Grund, daß eine Kompositionsform kodiert war, die nur noch in bestimmten lexikalisierten Komposita auftrat, es sich also um eine synchron nicht mehr produktive Kompositionsform handelt. Ein Beispiel ist die Segmentierung des Wortes *Windstau* als *Winds-tau*: Die Kompositionsform *Winds-* kommt nur in *Winds-braut* vor, und ist nicht mehr produktiv. Um unproduktive Kompositionsformen für die Segmentierung neuer Komposita auszuschließen und damit die Anzahl falscher Segmentierungen zu reduzieren, wurden die entsprechenden Kodes als nicht produktiv markiert, und die Komposita ins Lexikon aufgenommen.

Die häufigsten Fälle falscher Segmentierung waren allerdings solche, in denen gegen morphologische Regularitäten nicht verstoßen wurde, sondern extreme semantische Interpretationsschwierigkeiten bestanden (wie etwa in *Antrags-teller*). Diese Fehler sind unabhängig von der morphologischen Kodierung der Kompositionsformen. Sie müßten durch Methoden der semantischen Analyse eliminiert werden.

2.4 Struktur des Bestands an Kompositionssuffixen

Nach der Verifizierung und der Korrektur offensichtlicher Fehler ergab sich ein Bestand von 68 Kompositionssuffixen. Tabelle 1 zeigt alle Kompositionssuffixe, die häufiger als zehnmal in der Datei der einfachen Nomina auftraten. Die Operationen bezeichnen:

Ø Null-Operation;

$\ominus$ Trunkierung um nachfolgende Zeichenkette;

$\oplus$ Konkatenation mit nachfolgender Zeichenkette;

“ Umlautung;

Es zeigt sich, daß ein Großteil der Kompositionsformen mit nur fünf Suffixen gebildet werden, wobei das bei weitem häufigste das Ø-Suffix ist, gefolgt von $\oplus s$ und dem $\oplus$*(e)n*-Suffix. Dann folgen in der Häufigkeitsauflistung drei Formen von Kompositionssuffixen, die als Pluralmorpheme von Fremdwörtern auftreten. Andere Suffixe (so das $\oplus$*er* und $\oplus$*es*-Suffix), sind dagegen relativ selten.

Anzahl der Lemmata	**Suffix**	**Beispiel**
22759	Ø	*Kohlsuppe*
9637	⊕s	*Staatsfeind*
5307	⊕n	*Soziologenkongreß*
4316	⊕en	*Straußenei*
2610	⊕nen	*Wöchnerinnenheim*
618	⊖us ⊕en	*Aphorismenschatz*
348	⊖um ⊕en	*Museenverwaltung*
255	⊖um ⊕a	*Aphrodisiakaverkäufer*
122	⊖e	*Kirchhof*
95	⊖a ⊕en	*Madonnenkult*
87	⊕e	*Hundehalter*
73	“ ⊕e	*Gänseklein*
59	⊖on ⊕en	*Stadienverbot*
43	⊕es	*Geisteshaltung*
38	“ ⊕er	*Blätterwald*
33	⊖en	*Südwind*
28	⊖on ⊕a	*Pharmakaanalyse*
25	⊕er	*Geisterstunde*
19	⊕ien	*Prinzipienreiter*
11	⊖e ⊕i	*Carabinierischule*

Tabelle 1: Häufige Kompositionssuffixe im Lexikon

3 Selektionspräferenzen in Nominalkomposita

3.1 Motivation

Die statistische Untersuchung der Semantik von Nominalkomposita anhand der semantischen Klassen im CISLEX (Langer 1996) (Langer/Maier/Oesterle 1996) wurde aus mehreren Gründen vorgenommen:

- Bei mehreren morphologisch korrekten Zerlegungen eines komplexen Nomens sollte die semantisch plausiblere erkannt werden.
- Zur Beschreibung von Nominalbedeutung gehört auch die Beschreibung der Semantik von Nominalkomposita. Diese kann nicht ausschließlich manuell kodiert werden.
- Für die aufgestellten semantischen Klassen im Lexikon wurde eine distributionelle Evaluationsbasis benötigt.

3.2 Zur Kompositasemantik

Bezüglich der Selektion des Erstgliedes zeigen Köpfe von Komposita Präferenzen. Beinahe jedes Nomen im Deutschen kann allerdings mit beinahe jedem anderen kombiniert werden, um ein interpretierbares Kompositum zu bilden. Es ist von daher zu erwarten, daß die Auswahl der Köpfe bezüglich ihrer Erstglieder sich nicht in hundertprozentiger Selektion einer oder mehrerer semantischer Klassen durch einen bestimmten Kopf ausdrückt; vielmehr ist es wahrscheinlich, daß beinahe jede häufigere Klasse bei jedem häufigeren Kopf auftritt.

Neben der Disambigierung der Kompositabestandteile ist die Feststellung der Relation zwischen den Gliedern notwendig zur Bedeutungsermittlung. Zwar läßt sich kaum eine abschließende Liste möglicher Relationen zwischen Kompositagliedern aufstellen (Meyer 1993, 7-9), eine Auflistung bestimmter wichtiger Haupttypen ist aber durchaus möglich, und wurde auch bereits vorgenommen, so etwa in Fanselow (1981).

3.3 Die statistische Untersuchung

Die vorgenommene Untersuchung zeigt Korrelationen zwischen semantischen Klassen im CISLEX und Köpfen in zweigliedrigen Komposita. Die Untersuchung wurde folgendermaßen strukturiert:

1. Alle Erstglieder der segmentierten Komposita im Korpus wurden mit den semantischen Klassen getaggt.Trat ein polysemes Lexem als Erstglied auf, wurden alle semantischen Klassen berücksichtigt.

2. Die semantischen Klassen aller Erstglieder wurden gezählt. Dabei wurden auch die in der semantischen Hierarchie übergeordneten Klassen berücksichtigt.

3. Für Köpfe k mit Häufigkeit f(k) wurden nun ebenfalls alle semantischen Klassen aller Erstglieder gezählt.

Die Selektionspräferenzen wurden folgendermaßen ermittelt: Es gibt eine gewisse durchschnittliche Frequenz einer Klasse bei allen Erstgliedern im ganzen Korpus. Bei einem speziellen Kopf wäre diese Klasse bei zufälliger Verteilung über den Korpus mit einer gewissen Häufigkeit zu erwarten. Wird diese Häufigkeit beträchtlich über- bzw. unterschritten, heißt das, daß der entsprechende Kopf die Klasse präferiert selegiert oder eben mit ihr inkompatibel ist. Das verwendete Maß für die Selektionspräferenzen ist die Transinformation (mutual information) (Church/Hanks 1990), für die

vorliegende Untersuchung nach folgender Formel berechnet:

$$I(s;k) = log \frac{f_k(s)/\Sigma f_k(S)}{f(s)/\Sigma f(S)}$$

TI ist der Transinformationswert. Ist er Null, so liegt keine Selektion für die Klasse s beim Kopf k vor, ist sie größer Null, selegiert der Kopf diese Klasse präferiert, ist sie kleiner Null, gibt es Inkompatibilitäten zwischen dem Kopf und der Klasse.

s ist die semantische Klasse, deren Transinformation TI bezüglich des Kopfes k ermittelt werden soll.

f ist die Häufigkeit von Token der semantischen Klasse bei einem bestimmten Kopf ($f_k(s)$), bzw. im Gesamtkorpus (f(s)); die Summenzeichen summieren über alle semantischen Klassen(S) bei einem Kopf bzw. im gesamten Korpus.

Nachfolgend einige Einzelergebnisse, die verschiedene Aspekte der Semantik von Nominalkomposita und die Bedeutung dieser Untersuchung für die semantische Kodierung demonstrieren. Es finden sich in den Statistiken folgende Werte:

l: lokale Häufigkeit einer semantischen Klasse bei allen Erstgliedern eines Kopfes

g: Gesamthäufigkeit einer semantischen Klasse bei allen Erstgliedern des Korpus

erw: bei zufälliger Verteilung erwarteter Wert für die lokale Häufigkeit

TI: Transinformation

In den Beispielen treten folgende Klassendeskriptoren auf:
ABS: Bildungsabschlüsse, AKT: Aktionen, ASP: Sportarten, DIS: Diskursobjekt, EIG: Eigenschaft, ERE: Ereignis, FES: Feste, FRU: Früchte, GED: Druckerzeugnisse, GMI: Genußmittel, KLE: Kleidung, KSF: Kleiderstoffe, KTE: Körperteil, MIN: Musikinstrument, NHG: Nahrungsgrundstoffe, NTI: Nutztier, PBA: Bäume, PBL: Blumen, PFL: Pflanzen, SFS: Feststoffe, SPO: Sport, STI: Säugetiere, TIE: Tiere, VEK: Verkehrsmittel, VOG: Vögel, WER: Werkzeug, WET: Wettererscheinungen, WIS: Wissenschaften, ZUS: Zustände.

3.4 Identifizierung von Relationen

```
Komposita auf -fell: 60 Auftreten im Korpus
  TIE: l:  36  g:  31226  erw: 11   TI :  1.125
  KTE: l:   9  g:   5301  erw:  1   TI :  1.512
  STI: l:  33  g:   2404  erw:  0   TI :  3.602
```

```
  NTI: l:   9  g:   1056  erw:  0   TI :  3.126
  MIN: l:   3  g:    855  erw:  0   TI :  2.238
  ERE: l:   0  g:  41141  erw: 15   TI :  - INF
```

Dieses Beispiel zeigt sehr deutlich eine Selektion derjenigen Erstglieder durch den Kopf *-fell*, die in Meronymierelation zu ihm stehen. Über die Hälfte der Erstglieder fällt in die semantische Klasse SÄUGETIERE. Diese präferierte Selektion zeigt sich auch bei den NUTZTIEREN und — via Vererbung — an dem in der Taxonomie übergeordneten Knoten TIER. Eine weitere selegierte Klasse sind die MUSIKINSTRUMENTE (*Trommelfell*, *Paukenfell* ...). Zudem zeigt sich eine eindeutige negative Selektionspräferenz: Als Erstglied zu *-fell* taucht im gesamten untersuchten Korpus nie ein EREIGNIS auf; eine Zufallsverteilung würde für diesen Kopf zu 15 Komposita mit einem solchen Erstglied führen.

```
-schutz: 242
  WET: l:   7  g:   1428  erw:  1   TI :  1.400
  VOG: l:   6  g:   1212  erw:  1   TI :  1.410
```

Die beiden statistisch herausfallenden Klassen verdeutlichen exemplarisch zwei unterschiedliche Relationen in Komposita mit *-schutz*: Der Schutz vor WETTERERSCHEINUNGEN und der Schutz von VÖGELN. Erstere Gruppe entspricht einer semantischen Selektion des Verbs *schützen* bezüglich seines Präpositionalkomplements *schützen vor*, letztere entspricht der Relation zwischen dem zugrundeliegenden Verb und seinem Akkusativobjekt.

3.5 Polysemie des Zweitglieds

```
-blatt: 279
 PFL: l:  41  g:   5738  erw:  8   TI :  1.526
 PBA: l:  12  g:   1698  erw:  2   TI :  1.515
 GED: l:  17  g:   3843  erw:  5   TI :  1.046
 WER: l:   8  g:   1169  erw:  1   TI :  1.483
 DAK: l:  11  g:   2197  erw:  3   TI :  1.170
```

In diesem Beispiel zeigt sich deutlich die Polysemie des Zweitglieds.

- *Blatt* als PFLANZENTEIL (sehr häufig zusammen mit BÄUMEN)
- *Blatt* als Teil eines DRUCKWERKES
- *Blatt* als Teil diverser WERKZEUGE

- *Blatt* im Sinne von 'Zeitung' im Zusammenhang mit DISKURSOBJEKTEN.

DAK bezeichnet eine Klasse, in die Diskursobjekte verschiedener Art eingeordnet wurden (*Nachrichten, Witz, Propaganda*); sie erweist sich hier trotz ihrer Uneinheitlichkeit als selektionsrelevant - ebenso bei anderen Lexemen wie *Brief*, *Formel* u. v. a. In der Statistik schlagen sich die Bedeutungsvarianten von *Blatt* in *Rotorblatt* oder *Ruderblatt* nicht nieder. Für diese Spezialbedeutung ist keine Reihenbildung erkennbar.

```
-ball: 147
  ASP: l:  12  g:   3443  erw:  2   TI :  1.460
  FES: l:   8  g:   1121  erw:  0   TI :  2.176
  PBL: l:   3  g:    297  erw:  0   TI :  2.524
  ABS: l:   2  g:    175  erw:  0   TI :  2.647
```

Auch hier zeigt sich deutlich die Polysemie des Zweitgliedes.

- *Ball* in der Bedeutung 'Ball für Ballspiele' selegiert als Erstglied SPORTARTEN (*Tennisball* etc.).
- In der Bedeutung TANZVERANSTALTUNG selegiert das Nomen als Erstglieder FESTE (*Silvesterball* etc.), BLUMEN (*Magnolienball* etc.) und AUSBILDUNGSABSCHLÜSSE (*Abitursball*, *Maturaball*).

Obwohl diese Klassen im Korpus im Zusammenhang mit diesem Kopf absolut gesehen nicht übermäßig häufig sind, fallen sie doch statistisch sehr stark heraus.

3.6 Köpfe mit schwach ausgeprägten Selektionspräferenzen

```
-problem: 322
  ERE: l:  95  g:  31671  erw: 46   TI :  0.725
  ZUS: l:  64  g:  14595  erw: 21   TI :  1.104
  EIG: l:  26  g:   5122  erw:  7   TI :  1.251
```

Die wenig spezifischen Selektionspräferenzen des Zweitglieds *-problem* zeigen sich darin, daß nur eine Gruppe von in der Hierarchie weit oben liegenden Klassen, namentlich EREIGNISSE, ZUSTÄNDE, und EIGENSCHAFTEN, sich in der Statistik niederschlägt, und dies zudem mit nicht allzu hohen TI-Werten. Aus der Statistik ergibt sich in diesem Fall kaum ein

Hinweis auf die Semantik der Zweitglieds: weder liegen Indizien für Polysemie des Lexems *Problem*, noch Hinweise auf klar abgrenzbare Typen von Relationen zwischen Erst- und Zweitglied vor.

Ein ebenso unspezifisches Profil wie für *-problem* ergibt sich für das Zweitglied *-produktion*. Im Endeffekt läßt sich hier nur eine Präferenz für Konkreta im weiteren Sinne (STOFFE und KONKRETA) ermitteln. Die zunächst vielleicht erwartete Präferenz für ARTEFAKTE ergibt sich aus den Zahlen nicht - dies läßt sich auf die große Zahl von Komposita mit *-produktion* zurückführen, die KONKRETA als Erstglied haben, die nicht den ARTEFAKTEN zugerechnet werden können (*Eisenproduktion*, *Getreideproduktion* etc.).

3.7 Ergebnisse der statistischen Auswertung

Als wichtigstes Ergebnis kann festgehalten werden: Die Kombinatorik von Nominalkomposita ist nicht völlig beliebig; es handelt sich aber beim semantischen Selektionsverhalten der Köpfe tatsächlich nur um Präferenzen, nicht um kategorische Selektionsrestriktionen. Bezüglich dieser Selektionspräferenzen läßt sich festhalten:

- Unterschiedliche Relationen zwischen Erst- und Zweitglied spiegeln sich bei in der präferierten Selektion mehrerer unterschiedlicher Klassen durch die Köpfe wider. Köpfe, die in einer unspezifischen Relation zu ihren Erstgliedern stehen (eine Relation, die sich etwa als 'in bezug auf' paraphrasieren ließe), wie *-frage*, *-problem*, *-idee*, zeigen gering ausgeprägte Selektionspräferenzen, die stets nur in bezug auf Klassen deutlich werden, die in der Taxonomie weit oben stehen. Feinere Klassen werden nicht statistisch auffallend als Erstglieder selegiert. Diese Eigenschaften könnten zur automatischen Erkennung von Relationen herangezogen werden.

- Polysemie des Kopfes spiegelt sich häufig in den Selektionspräferenzen bezüglich der Erstglieder wieder. Das typische Selektionsprofil eines polysemen Zweitgliedes ist dabei folgendes: Es werden mehrere relativ spezifische, klar voneinander abgrenzbare semantische Klassen selegiert, die keinen unmittelbaren gemeinsamen Mutterknoten haben. Die kleinste gemeinsame übergeordnete Klasse der selegierten Klassen wird nicht selegiert. Diese Ergebnisse legen nahe, daß eine automatische Erkennung von Polysemie für eine Reihe von Köpfen und ihre Disambiguierung aufgrund der semantischen Klasse des Erstglieds möglich ist.

- Viele Kompositaköpfe zeigen eindeutige Selektionspräferenzen bezüglich ihrer Erstglieder. Diese Eigenschaft ließe sich zur Disambiguierung polysemer Erstglieder heranziehen.
- Bei deverbalen Nomina lassen sich in vielen Fällen im Selektionsverhalten der Rektionskomposita die Selektionspräferenzen der zugrundeliegenden Verben wiedererkennen. Diese Eigenschaft könnte als Kriterium zur Erkennung von Verbkomplementen in der Syntaxanalyse herangezogen werden.

Aus einigen der genannten Punkte ergaben sich zudem direkte Konsequenzen für die semantische Kodierung. So wurde der Bestand an Klassen aufgrund der Ergebnisse der statistischen Untersuchung neu beurteilt und erweitert; bestimmte Nomina wurden aufgrund der erkannten Selektionspräferenzen neu kodiert.

4 Zusammenfassung

Trotz der Bedeutung der Kompositabildung in der Wortbildung des Deutschen sind Kompositionssuffixe bei Nomina bisher nur teilweise lexikographisch erfaßt. Eine korrekte Zerlegung von Nominalkomposita ist allerdings Voraussetzung für die weitere Analyse. Somit war es für das elektronische Wörterbuch CISLEX nötig, die Kompositionssuffixe für sämtliche Simplizia im Deutschen zu kodieren. Es zeigte sich daß relativ viele (68) verschiedene Kompositionsuffixe auftreten, wenn man jede Abweichung von der Nominativ-Singular Form bei Erstgliedern in Komposita berücksichtigt. Von diesen Kompositionssuffixen treten allerdings nur neun bei mehr als 100 Lemmata auf.

Die Kodierung machte die korrekte Zerlegung der Komposita in einem Korpus möglich. Das Korpus der zerlegten Komposita wurde dann zur Determination semantische Selektionspräferenzen der Köpfe zweigliedriger Komposita herangezogen. Es konnte gezeigt werden, daß diese Statistiken es prinzipiell erlauben, sowohl polyseme Erstglieder als auch Zweitglieder zu disambiguieren, und die Relationen zwischen den Gliedern zu bestimmen. Zudem deuten bestimmte, formal beschreibbare Selektionseigenschaften auf eventuelle Polysemie eines Kopfes hin.

Die Ergebnisse dieser Untersuchungen legen es nahe, die gefundenen Disambiguierungstrategien zu implementieren, und für ein größeres Korpus auf ihren quantitativen Erfolg zu testen. Während es unwahrscheinlich ist, daß dies bereits ebenso gute Ergebnisse liefert wie die manuelle Klassifi-

kation, so könnten diese Methoden doch als eine Grundlage für weitere Disambiguierungsschritte dienen — etwa unter Einbeziehung des Kontextes — und eine erste semantische Einordnung der Gesamtbedeutung eines Kompositums ermöglichen.

Danksagungen

Diese Arbeit wurde teilweise durch ein Doktorandenstipendium der DFG im Rahmen des Graduiertenkollegs SIL unterstützt.

References

Augst, Gerhard. 1975. Lexikon zur Wortbildung. Morpheminventar. Tübingen: Narr.

Augst, Gerhard. 1975a. Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache. Tübingen: Narr.

Church, Kenneth W., Patrick Hanks. 1990. Word Association Norms, Mutual Information and Lexicography. Computational LInguistics 16 (1), 22-29.

Drosdowski, Günter. 1984. Duden Grammatik der deutschen Gegenwartssprache. Mannheim u.a.: Dudenverlag. (= Duden Band 4).

Fanselow, Gisbert. 1981. Zur Syntax und Semantik von Nominalkomposita. Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung des Deutschen. Tübingen: Niemeyer.

Fleischer, Wolfgang, Irmhild Barz. 1992. Wortbildung der deutschen Gegenwartssprache. Tübingen: Niemeyer.

Fuhrhop, Nanna. 1995. Fugenelemente. In: Lang, Ewald und Gisela Zifonun (Hrsg.). Deutsch — typologisch. (= IdS-Jahrbuch 1995). Berlin/New York. S. 525-550.

Grube, Henner. 1976. Die Fugenelemente in Neuhochdeutschen appelativischen Komposita. In: Sprachwissenschaft 1, S. 187-222.

Guenthner, Franz 1996. Electronic Lexica and Corpora Research at CIS. International Journal of Corpus Linguistics 1(2).

Guenthner, Franz, Petra Maier-Meyer 1996. Überblick ber das CISLEX-Wörterbuchsystem. Tagungsband des Workshops Lexikon und Text, Tübingen 1994.

Haapalainen, Mariikka. 1995. GERTWOL. Ein System zur automatischen Wortformenerkennung deutscher Wörter. Elektronisches Dokument (WWW), http://www.lingsoft.fi/doc/gertwol/intro.

Langer, Stefan. 1996. Selektionsklassen und Hyponymie im Lexikon. München: CIS-Berichte.

Langer, Stefan, Petra Maier, Jürgen Oesterle. 1996. CISLEX — an Electronic Dictionary for German. Its Structure and a Lexicographic Application. Proceedings of COMPLEX 96.

Lezius, Wolfgang 1996. Morphologiesystem Morphy. In: Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1996. Tübingen: Niemeyer.

Meyer, Ralf 1993. Compound Comprehension in Isolation and Context.The contribution of coneptual and discourse knowledge to the comprehension of German novel noun-noun compounds. Tübingen: Niemeyer.

Rackow, Ulrike. 1992. On the Treatment of Compounds in Machine Translation. A Study. Heidelberg: IBM. (= IWBS Report).

Rackow, Ulrike, Ido Dagan, Ulrike Schwall. 1992. Automatic Translation of Noun Compounds. In: ICCL: Proceedings of COLING 92. Nantes: GETA, S. 1249-1253.

Weinrich, Harald. 1993. Textgrammatik der deutschen Sprache. Mannheim u.a.: Dudenverlag.