Tibor Kiss

Bedingungen für den Wegfall eines Artikels: Distribution und Interpretation von Präposition-Nomen-Kombinationen[1]

Abstract:
Kombinationen aus Präposition und artikelloser Nominalprojektion, deren syntaktischer Kopf ein zählbares Substantiv im Singular ist, fristeten lange Zeit ein Schattendasein in der Grammatikschreibung. Sie wurden ignoriert oder als Ausnahmen beschrieben, obwohl sie offenkundig regelhaft gebildet werden. Im vorliegenden Aufsatz verwenden wir computerlinguistische Verfahren, insbesondere *Annotation Mining* und logistische Regression, um die syntaktische Distribution dieser Kombinationen zu charakterisieren und anhand zweier Präpositionen *(ohne* und *unter)* detailliert die Realisationsbedingungen zu bestimmen.

1. Einleitung

Präpositionen werden in vielen Sprachen mit Nominalphrasen kombiniert. Viele Präpositionen gestatten jedoch auch eine Kombination mit einem artikellosen Substantiv oder einer artikellosen Substantivgruppe, wie die Beispiele in (1) zeigen.

(1) auf parlamentarische Anfrage, bei absolut klarer Zielsetzung, in untertreibender Anspielung, mit schwer beladenem Rucksack, ohne mündliche Vorwarnung, unter sanfter Androhung

Die Beispiele in (1) sind bemerkenswert, denn die Substantive werden als *zählbar* analysiert und dürften infolge dieser Eigenschaft eigentlich nicht ohne Artikel realisiert werden. So stellt der Duden (2005, S. 337) in der Regel 442 fest, dass *„Substantive mit Merkmalkombination ‚zählbar' plus Singular ... grundsätzlich immer ein Artikelwort bei sich [haben], und wenn es als letzte Möglichkeit der indefinite Artikel ist."*
Deutlich wird anhand der Beispiele in (1) auch, dass es sich hierbei nicht um Paarbildungen aus Präposition und Substantiv handelt: Pränominale Modifikation ist in vielen Fällen möglich, teilweise auch durchaus komplex. Dieser Eindruck wird noch dadurch verstärkt, dass postnominale Komplemente nicht ausgeschlossen sind:

(2) Riess-Passer unterstrich die Forderung nach Stilllegung des grenznahen Atomkraftwerks Temelin in Tschechien; (NZZ, AUSLAND, 30.11.2001)

Das Beispiel in (2) zeigt darüber hinaus, dass die fraglichen Konstruktionen nicht nur adverbial verwendet werden können, sondern auch in regierten Kontexten erscheinen – die Präposition *nach* wird in (2) durch *Forderung* regiert.
In der Folge bezeichne ich Kombinationen wie die in (1) und (2) als Präposition-Nomen-Kombinationen (PNKen). PNKen wurden in der Grammatikschreibung lange ignoriert oder als Ausnahmen charakterisiert; in den letzten Jahren hat sich die Position aber sowohl aus typologischer als auch aus grammatiktheoretischer Perspektive gewandelt. Himmelmann (1998, S. 316) hat beobachtet, dass PNKen in vielen Sprachen auftreten, deren Grammatiken

[1] Die vorliegenden Ergebnisse entstammen dem Projekt „Erfassung und Analyse syntaktischer und semantischer Eigenschaften von Präposition-Substantiv-Sequenzen unter Verwendung computerlinguistischer Analyseverfahren", das seit 2009 freundlicherweise von der DFG gefördert wird. Die Ergebnisse hätten ohne die Mitarbeit von Antje Müller, Claudia Roch, Tobias Stadtfeld, Katja Keßelmeier und Jan Strunk nicht erzielt werden können. Ich danke ihnen herzlich. Durch die Einladung zur IDS Jahrestagung 2010 hatte ich die Gelegenheit, Teile des vorliegenden Aufsatzes auf der IDS Jahrestagung 2010 zu präsentieren. Ich danke Stefan Engelberg und den Organisatoren für die Einladung, Anke Holler, Kristel Proost und Stefan Engelberg für ihre Kommentare und den Zuhörern für die Diskussionsbeiträge.

verlangen, dass zählbare Substantive im Singular durch einen Artikel ergänzt werden müssen. Stvan (1998) bietet für das Englische eine Analyse an, die auf den semantischen Eigenschaften des Substantivs basiert; Baldwin et al. (2006) entwickeln eine Typologie der Konstruktion für das Englische und stellen beiläufig fest, dass es sich zumindest in Teilen um eine produktive – im syntaktischen Sinne also regelhafte – Konstruktion handelt, die durch Ausnahmen nicht erfasst werden kann.[2] Während der Duden (2005) die Konstruktion als Ausnahme beschreibt, zeigen Kiss (2007) und Dömges et al. (2007) anhand eines quantitativen Produktivitätsmaßes, dass es sich um eine produktive und somit regelhafte Kombination handelt, die somit auch durch die Grammatik erfasst werden muss.

PNKen stellen so die Grammatikschreibung vor eine Herausforderung: Es besteht einerseits weitgehender Konsens darüber, dass die Kombination in vielen Sprachen regelhaft ist; andererseits ist aber auch klar, dass die zugrunde liegenden Regeln offensichtlich nicht dem Typ entsprechen, der etwa die Kombination einer DP mit einem transitiven Verb steuert.[3] Introspektive Urteile zur Kombinierbarkeit von Präposition und artikelloser Nominalgruppe werden nur zögerlich getroffen. Sprecher sind ebenfalls sehr zögerlich, neue PNKen zu bilden. Dem steht wiederum die quantitative Produktivität des Prozesses gegenüber – eigentlich ein Widerspruch!

Zur Identifikation der Grammatikalitätsbedingungen der PNKen wähle ich das *Annotation Mining* (vgl. Chiarcos et al. 2008). Es bietet sich als exploratives Analyseverfahren an, in dem auf die Identifikation oder Elizitation von Grammatikalitätsurteilen verzichtet wird. Grammatikalitätsbedingungen werden aus umfangreichen Sprachdaten abgeleitet, die wiederum durch Merkmale auf den unterschiedlichen linguistischen Ebenen annotiert sind (Lexikon, Syntax, Semantik, globaler Kontext). Die annotierten Daten werden Klassifikationsverfahren aus der inferentiellen Statistik unterworfen, im vorliegenden Fall insbesondere der binären logistischen Regression (Harrell 2001), die in Abs. 4.2 vorgestellt werden.

Anhand von zwei Pilotstudien zur Syntax der Präpositionen *ohne* und *unter* möchte ich das Verfahren vorstellen und erläutern, warum die Ansätze von Stvan (1998) und Baldwin et al. (2006) grundsätzlich in die richtige Richtung weisen, aber dann eben doch nicht auf das Deutsche übertragbar sind. Die Modelle zeigen, dass es unabhängig von der Interpretation der Präpositionen strukturelle Faktoren gibt, die die Weglassbarkeit des Artikels zumindest partiell bedingen.

2. Zur Grammatik von Präposition-Nomen-Kombinationen

Die hier verwendete Definition von PNKen ist gegenüber anderen Definitionen durch eine interne und eine externe Bedingung eingeschränkt. Obwohl diese Einschränkungen eigentlich zwingend mitgedacht werden müssen, zeigen etwa die Untersuchungen von Le Bruyn et al. (2009) – in denen weder die interne noch die externe Bedingung eine Rolle spielt – dass es erforderlich ist, die Bedingungen explizit zu machen. Die interne Bedingung sieht vor, dass unter PNKen nur solche Kombinationen aus einer Präposition und einer artikellosen Nominalprojektion fallen, *in denen das Substantiv ein zählbares Substantiv im Singular ist.* Weder PPen mit Massentermen noch PPen mit nackten Pluralia fallen so unter die PNKen. Da im Deutschen Artikel bei Massentermen nicht obligatorisch sind – Indefinita sind sogar untersagt

---

[2] Studien zu PNKen im Englischen bezeichnen die Konstruktion als *determinerless PPs* (Baldwin et al. 2006) oder *bare PPs* (Le Bruyn et al. 2009).

[3] Ich gehe im Folgenden davon aus, dass vollständige Nominalprojektionen, die einen Artikel enthalten, DPen sind. Die Festlegung ist im vorliegenden Kontext schon deswegen nützlich, weil so Nominalprojektionen mit Artikel notationell von Nominalprojektionen ohne Artikel unterschieden werden können. Eine Festlegung auf eine bestimmte Analyse der Nominalprojektion ist damit nicht verbunden.

– und bei Pluralia zumindest optional, entstehen durch den Wegfall des Artikels in diesen Phrasen keine Probleme. Betrachtete man die Distribution des Artikels bei Pluralia und Massentermen in anderen Sprachen, etwa den romanischen Sprachen (vgl. Espinal und McNally 2010), so würden andere Definitionen erforderlich.
Für das Deutsche liegt zunächst aber die Tücke im Detail: Zum einen setzt die Definition ein Konzept der Zählbarkeit voraus, zum anderen handelt es sich hierbei bestenfalls um eine notwendige Definition, nicht aber um eine hinreichende, denn es kann nicht jede Präposition mit jedem Substantiv in einer PNK kombiniert werden.
Die externe Bedingung setzt voraus, dass es sich bei der Präposition in der PNK um eine Präposition handelt, die überhaupt DP-Objekte besitzen kann. Auch diese Annahme ist nicht trivial. Sie führt dazu, dass wir nicht alle Präpositionen des Deutschen untersuchen, sondern nur einen Teilbereich der Präpositionen, nämlich solche, die DP-Objekte besitzen und in PNKen auftreten können. Diese Präpositionen sind in (3) aufgeführt; in Abs. 2.3 werde ich anhand der Präposition *per* das Verhalten von Kombinationen aus Präposition und Substantiv erläutern, die dieser Bedingung nicht genügen. Die in (3) aufgeführten Präpositionen erfüllen das externe Kriterium, allerdings nicht mit jedem zählbaren Substantiv im Singular, wie anhand von (4) verdeutlicht werden kann.[4]

(3) an, auf, bei, dank, durch, für, gegen, gemäß, hinter, in, mit, mittels, nach, neben, ohne, seit, über, um, unter, vor, während, wegen

(4) *Riess-Passer wehrt sich *gegen Forderung* nach Stilllegung des grenznahen Atomkraftwerks Temelin in Tschechien.

Das Beispiel in (4) ist eine Variation des Beispiels in (2), bei der auch auf den wohl tatsächlich obligatorischen Artikel vor *Forderung* verzichtet wurde. Auch semantisch sehr nahe Substantive können in PNKen nicht einfach ausgetauscht werden, wie anhand der Substantive *Voraussetzung* und *Prämisse* in (5) und (6) gezeigt werden kann.

(5) Auch Philipp Egli besteht auf einer eigenen Handschrift – unter der Voraussetzung/der Prämisse des Einverständnisses des Ensembles.

(6) Auch Philipp Egli besteht auf einer eigenen Handschrift – unter Voraussetzung/*Prämisse des Einverständnisses des Ensembles.

Während *Voraussetzung* mit und ohne Artikel als Komplement von *unter* erscheinen kann, ist bei *Prämisse* nur die Realisation in einer PP möglich.
Der Kontrast zwischen (5) und (6) könnte durch zwei Muster erklärt werden: entweder durch einen Rekurs auf eine entsprechende Definition der Zählbarkeit, oder durch den Nachweis tatsächlich fehlender Regelhaftigkeit der Konstruktion. Diese Punkte wollen wir in den folgenden Abschnitten diskutieren.

2.1 Das Problem der Zählbarkeit

Die Dudenregel 442 (Duden 2005, S. 337) besagt, dass zählbare Substantive im Singular im Deutschen durch einen Artikel begleitet werden müssen. Daraus folgt nicht nur, dass die Beispiele in (1) und (2) ungrammatisch sein sollten. Die Regel setzt auch voraus, dass die Eigenschaft der Zählbarkeit definiert wird. Zählbarkeit könnte als lexikalische Abbildung von Worten auf Denotate verstanden werden. Sind die Denotate zählbar, dann ist auch das Wort zählbar. Diese Definition geht zumindest auf Jespersen (1924) zurück und fand auch in der gene-

[4] Das Beispiel (4) wird dann nicht als ungrammatisch empfunden, wenn es als Überschrift eines Zeitungsartikels verwendet würde. In Überschriften findet man häufig PNKen, die außerhalb von Überschriften als deutlich abweichend bewertet werden.

rativen Grammatik Rückhalt. Sie führt aber zu einer Vielzahl von Problemen. Hier ist nicht der Raum, um dies auch nur andeutungsweise zu erläutern, es soll nur erwähnt werden, dass es nach dieser Definition keine Synonyme S_1 und S_2 in einer Sprache geben sollte, bei denen S_1 zählbar ist, S_2 aber nicht.[5] Darüber hinaus kommt diese Definition in Schwierigkeiten, wenn ein zählbares Wort nicht zählbar verwendet wird, man vergleiche hierzu die Verwendung von *Fahrzeug* vs. *Fahrzeuge* in (7).

(7) Hier bekommen Sie mehr Fahrzeug/Fahrzeuge für ihr Geld.

Allan (1980) hat eine alternative Sichtweise des Konzepts der Zählbarkeit in der Sprachwissenschaft etabliert, die in jüngeren Arbeiten von Borer (2005) und Bale und Barner (2009) auch formal expliziert wurde. Nach dieser Auffassung ist Zählbarkeit keine lexikalische Eigenschaft, sondern eine Eigenschaft nominaler *Phrasen,* die durch *kontextuelle* formale Merkmale zugewiesen wird, etwa die Pluralmorphologie oder die Realisation eines Artikels. Dieser Idee folgend ist ein Nomen zunächst einmal weder zählbar noch nicht zählbar. Wird dieses Nomen in einen entsprechenden Kontext eingesetzt, ist die resultierende Phrase aber zählbar oder nicht. Für das Beispiel (7) bedeutet dies, dass *Fahrzeug,* wenn es weder flektiert noch durch einen Artikel begleitet wird, als Massenterm analysiert wird. Das Pluralmorphem bildet hingegen einen Kontext für Zählbarkeit, ebenso würde ein solcher Kontext durch einen singularischen oder pluralischen Artikel geschaffen. Für die Syntax der PNKen ergibt sich nun allerdings eine merkwürdige Konsequenz: Da hier die relevanten Substantive in Kontexten realisiert werden, in denen weder ein Artikel noch ein Pluralmorphem vorliegt, wird die Dudenregel 442 schon deswegen nicht verletzt, weil ein Substantiv in einer PNK nach dieser Definition nicht zählbar ist. Dies bedeutet aber auch, dass jedes Substantiv in einer PNK realisiert werden können müsste, weil hier nie ein Kontext für Zählbarkeit vorliegt. Konsequenterweise sollten sich PNKen also wie PPen verhalten und ihre besondere Betrachtung wäre überflüssig. Dass die Schlussfolgerung nun wieder voreilig ist, folgt aus zwei Beobachtungen: Zum einen haben wir bereits gesehen, dass nicht jedes Substantiv in einer PNK realisiert werden kann; zum anderen würde dies auch bedeuten, dass jedes Substantiv in einer PNK die Interpretation eines Massenterms erhält. Dies ist aber ebenfalls falsch, wie man an Beispielen wie (8) und (9) sieht.

(8) Milosevic unterschrieb auch unter ø/der/einer Androhung von NATO-Bombardementen nicht.

In (8) können wir beobachten, dass bei den Realisationen als PP (DP mit definitem oder indefinitem Artikel) oder PNK die gleiche Interpretation hervorgerufen wird. Es ist dann die plausiblere Überlegung, dass bei einer Realisation mit Artikel tatsächlich Zählbarkeit vorliegt und somit Bedeutungsgleichheit auch bei einer Realisation ohne Artikel. Die gleiche Argumentation gilt auch für die mit *durch* eingeleitete Phrase in (9).

(9) Ursprünglich war der Artikel als Verbot der Beleidigung jeder Religion eingeführt worden, der Diktator engte ihn jedoch nach 1980 auf die Beleidigung des Islam ein, und er verschärfte ihn durch ø/eine/die Androhung der Todesstrafe. (NZZ AUSLAND, 25.02.1995, Originalbeleg *mit* definitem Artikel)

---

[5] Der Leser erwartet hier sicherlich eine Stellungnahme, aus der hervorgeht, ob es nun entsprechende Synonyme gibt oder nicht. Anhand von Beispielen wie *Fußbekleidung vs. Schuhe* könnte man diese Frage nun positiv beantworten. Allerdings ist zu berücksichtigten, dass das Konzept der Synonymie nicht präzise definiert ist. Es wird jedoch auch niemand bestreiten wollen, dass es sehr schwer werden würde, Modelle zu konstruieren, in denen die Aussagen (i) und (ii) unterschiedliche Wahrheitswerte besitzen:

(i) Du hast aber viele Schuhe.

(ii) Du hast aber viel Fußbekleidung.

Es zeigt sich hier, dass die kontextuelle Definition der Zählbarkeit ebenso problematisch ist wie die lexikalistische Position. Darüber hinaus ist es notwendig zu erläutern, was im Rahmen einer automatischen Klassifikation eigentlich unter einem *zählbaren Substantiv* zu verstehen ist, wenn weder gänzlich auf den Typ noch gänzlich auf die einzelnen Vorkommen rekurriert werden kann.

Die Zählbarkeitsklassifikation basiert auf zwei statistischen Klassifikatoren, die einerseits das Verhältnis von Plural und Singularvorkommen betrachten und andererseits syntaktische Kontexte (die durch die POS-Tags vor dem zu klassifizierenden Nomen identifiziert werden, vgl. Abs. 3.). Diese Klassenbildung erfolgt nach einem Vorschlag von Stadtfeld (in Vorb.), basierend auf den Arbeiten von Gillon (1999) und Barner und Snedeker (2004). Stadtfeld unterscheidet fünf Zählbarkeitsklassen (I = zählbar, II = Pluralia tantum, III = echt ambig, IV und V = nicht zählbar), die nach drei Kriterien identifiziert werden können.

Das erste Kriterium – der Messmodus – erfasst, ob ein Singularvorkommen eines Nomens mit *mehr* hinsichtlich einer definierten Quantität verglichen werden kann. Für echt zählbare Substantive (Klasse I) ergibt dies keinen Sinn, für nicht-zählbare Substantive der Klasse IV ist ein Vergleich hinsichtlich Quantität von Individuen oder einer anderen Messgröße möglich, für nicht-zählbare Substantive der Klasse V ist ein Vergleich nur hinsichtlich einer nicht auf Individuen quantifizierenden Messgröße möglich.

(10) a. ??New York hat mehr Hochhaus als Berlin.
b. Paul hat mehr Besteck als Peter.
c. Paul hat mehr Reis als Peter.

Das Beispiel (10a) ist auch dann inakzeptabel, wenn vorher bekannt ist, dass das Gesamtgewicht der Hochhäuser in New York höher ist als das Gewicht aller Hochhäuser Berlins (Beispiele dieses Typs findet man eigentlich nur in Werbekontexten). Das Beispiel (10b) kann wahr sein, wenn es tatsächlich insgesamt mehr individuelle Besteckelemente gibt oder wenn Paul mehr Bestecktypen als Peter besitzt. Das Beispiel (10c) setzt voraus, dass die Messgröße das Gesamtvolumen ist. Individuen sind also bei Klasse V ausgeschlossen, führen aber bei Klasse IV zu einer Typ- bzw. Behälterinterpretation.

Der zweite Test überprüft verborgene Typlesarten bei Pluralrealisationen. Auch hier erfolgt eine Einbettung des Nomens unter *mehr,* das Nomen wird nun aber im Plural realisiert. Bei Massentermen der Klasse V ist der Test nicht anwendbar, weil sie keine Pluralrealisation besitzen, bei Massentermen der Klasse IV ergibt die explizite Realisation einer Numeralkonstruktion eine äquivalente Lesart (11), bei zählbaren Substantiven der Klasse I ist diese Äquivalenz nicht gegeben (12).

(11) a. Paul hat mehr Weine als Peter.
b. Paul hat mehr Sorten Wein als Peter.

(12) a. Ulrich vertreibt mehr Weingläser als Paul.
b. Ulrich vertreibt mehr Sorten Weinglas als Paul.

Der letzte Test betrifft die Verwendung des indefiniten Artikels in Kontexten der Form *indefiniter Artikel + N ist ein Hyperonym.* Bei zählbaren Substantiven der Klasse I ist der indefinite Artikel obligatorisch (13), die Klassen IV und V gestatten die Realisation des indefiniten Artikels nicht (14), es kommt zu Typlesarten, die im vorgegebenen Kontext blockiert sind, weil hier allgemeine Eigenschaften das Prädikat bilden.

(13) a. Ein Auto ist ein Fahrzeug.
b. *Auto ist ein Fahrzeug.

(14) a. Stahl ist eine Metalllegierung.

b. ??Ein Stahl ist eine Metalllegierung.

(15) a. Besteck ist ein Werkzeug.

b. ??Ein Besteck ist ein Werkzeug.

Ein wesentlicher Vorteil dieses feineren Klassifikationsansatzes ist die Identifikation echt mehrdeutiger Substantive, die in der bislang nicht diskutierten Klasse III erfasst werden. Diese Substantive sind für eine binäre Klassifikation insofern problematisch, als sie Eigenschaften von *zählbaren* und *nicht-zählbaren* Nomina aufweisen. So gestatten Substantive wie *Kuchen* oder *Fisch* die Verwendung des indefiniten Artikels (16a), was für die Zählbarkeit spricht, zugleich kann aber der Artikel auch weggelassen werden, wie (16b) zeigt.

(16) a. Ein Kuchen ist ein Grundnahrungsmittel.

b. Kuchen ist ein Grundnahrungsmittel, meinte schon Marie-Antoinette.

PNKen, in denen das Substantiv der Klasse III angehört, werden zunächst aus der Analyse ausgeschlossen, bis die Charakteristika dieser Klasse genauer bestimmt sind. Erfreulicherweise enthält diese Klasse wenige Elemente.

In der folgenden Analyse (vgl. Abs. 3. und 4.) bezeichne ich die in den PNKen auftretenden Substantive als *zählbar,* weil sie der Klasse I und als *nicht-zählbar,* wenn sie den Klassen IV oder V zugeordnet werden konnten. Eine binäre Opposition wird damit nicht behauptet.

2.2 Produktivität

Wir können somit davon ausgehen, dass zählbare Substantive in PNKen realisiert werden, ein Verweis auf eine mögliche Analyse als Massenterme ist falsch und hilft nicht weiter. Aus dieser Annahme würde folgen, dass Nomina in PNKen grundsätzlich als Massenterme interpretiert würden – und somit PNKen auch grundsätzlich grammatisch sein sollten, wobei eine Unakzeptabilität dann entstünde, wenn eine Massenterminterpretation inkompatibel wäre.

Ein weiterer Ausweg könnte natürlich darin bestehen, die Konstruktionen insgesamt als Ausnahmen zu analysieren. Dies ist etwa der Stand der Dinge im Duden: die Regel 395 zählt die Ausnahmen auf, die die Realisation eines zählbaren Substantivs ohne Artikel gestatten. Kurz gesagt: die hier bislang vorgestellten, zweifelsfrei grammatischen Beispiele zählen nicht dazu.

Diese Feststellung soll aber nicht genügen. Basierend auf dem morphologischen Produktivitätsmaß *P* aus Baayen (2001) und dessen Verallgemeinerung für syntaktische Kontexte in Evert (2004) legen Dömges et al. (2007) eine quantitative Analyse der Produktivität von PNKen mit *unter* vor, aus der hervorgeht, dass es sich bei PNKen um regelhafte Kombinationen handelt. Ohne die technischen Details dieser Analyse hier zu diskutieren, kann der Ansatz wie folgt zusammengefasst werden: Wenn man davon ausgehen würde, dass eine syntaktische Kombination tatsächlich eine aufzählbare Ausnahme ist, dann müsste es auch endlich viele Instanzen dieser Ausnahme geben. Dies wiederum würde bedeuten, dass bei einem genügend großen Korpus irgendwann einmal beobachtet werden könnte, dass das Vokabular für diese spezielle Konstruktion nicht mehr anwächst (während bei regelhaften, nicht aufzählbaren Konstruktionen ein stetiges Vokabularwachstum beobachtbar ist). Wenn wir nun eine Präposition als Beispiel auswählen, dann besteht das fragliche Vokabular aus Substantiven, die mit dieser Präposition kombiniert werden können. In diesem Sinne wächst das Vokabular von *unter Androhung* auf *unter sanfter Androhung* nicht an, weil wir hier das Substantiv *Androhung* bereits gesehen haben. Anders ist dies, wenn wir zunächst *unter Androhung* und dann *unter Auswertung* oder auch *unter Gewaltandrohung* beobachten würden.

Konkretisiert auf eine Präposition wie *unter* bedeutet dies, dass zu irgendeinem Zeitpunkt t_1 die Wahrscheinlichkeit, dass zu Folgezeitpunkten t_{1+n} noch neue Instanzen von *un-*

ter+Substantiv beobachtet werden können, unter einen kritischen Wert fällt, wenn der Prozess nicht produktiv ist.
Diese Annahme kann nun von beobachteten Daten auf nicht beobachtete Daten, also von bereits sehr umfangreichen empirischen Korpora auf wesentlich umfangreichere nicht beobachtete Korpora übertragen werden, wenn man sog. LNRE-Modelle *(large number of rare events,* Baayen 2001) zur Vorhersage nicht beobachteter Daten verwendet. Dömges et al. (2007) haben gezeigt, dass das plausibelste Modell für *unter+Substantiv* dasjenige Modell ist, das unendlich viele Instanzen dieser Kombination vorhersagt. Somit ist die Kombination regelhaft, auch der Weg über die Ausnahme ist verbaut.

2.3 Eine Analyse des Englischen, die im Deutschen nicht funktioniert

Für PNKen im Englischen schlägt Stvan (1998) vor, dass es einen Prozess der N-Selektion gibt. Nach dieser Analyse werden PNKen durch die in ihnen enthaltenen Substantive lizenziert, wie anhand der Substantive *school, jail* und *prison* in (17) illustriert werden kann.

(17) from school, at school, in jail, from jail, in prison, from prison …

Baldwin et al. (2006) legen Stvans Analyse zugrunde und argumentieren, dass es sich bei den hier realisierten Substantiven häufig um solche handelt, die auch in anderen syntaktischen Kontexten ohne Artikel realisiert werden können, obwohl die Artikelrealisation eigentlich obligatorisch ist. Als Beispiel mag hier das Substantiv *school* in (18) dienen, das als Subjekt eines kategorischen Satzes artikellos realisiert wird. Baldwin et al. (2006) bezeichnen solche Substantive als *defective nouns* und charakterisieren sie semantisch als *institutional nouns.*

(18) School is over.

Stvan (1998) beobachtet nun weiterhin, dass bei einer Realisation eines Substantivs in einer PNK eine pragmatische Bedeutungserweiterung erfolgt. So bedeutet das Beispiel in (19a) nicht allein, dass Marys Gatte im Gefängnis verortet ist, sondern vor allem, dass Marys Gatte in diesem Gefängnis eine Haftstrafe verbüßt; (19b) bedeutet nicht nur, dass John in einer Schule lokalisiert werden kann, sondern, dass eine institutionell geprägte kontextuell näher zu explizierende Beziehung zwischen John und der Schule besteht.

(19) a. Mary’s husband is in prison.
b. John is at school.

Baldwin et al. (2006) schlagen neben PNKen mit N-Selektion auch PNKen mit P-Selektion vor. Sie charakterisieren solche PNKen durch semantische Selektionsregeln die den Typ des nominalen Komplements einschränken, wobei sie implizit davon ausgehen, dass durch die semantische Selektion auf das Substantiv auch die Polysemie der Präposition reduziert wird (zur Polysemie siehe Abs. 3). Diese Selektion ergibt dann produktive Muster wie die in (20) und (21).

(20) by train, by plane, by bus, by pogo stick, by hydro-foil, …

(21) on disc, on CD, on DVD, on tape, on stick, on memory card, …

Es stellt sich dann natürlich die Frage, ob die Vorschläge von Stvan (1998) und Baldwin et al. (2006) vom Englischen auf das Deutsche übertragen werden können.
Zunächst ist hier festzuhalten, dass es produktive Muster des Typs (20) und (21) auch im Deutschen gibt, aber gerade nicht bei PNKen, so wie ich sie definiert habe. Eine Übersetzung von *by* ins Deutsche würde wahrscheinlich die Präposition *per* ergeben, die dann in der Tat Muster wie in (22) zeigt.

(22) per Zug, per Flugzeug, per Bus, per Skateboard, …

Hier greift nun die externe Bedingung der Definition der PNKen: Die Präposition *per* besitzt im Deutschen überhaupt keine DP-Objekte, d.h. Objekte mit realisiertem Artikel. Somit handelt es sich bei den Beispielen in (15) nicht um PNKen. Muster des Typs (20) werden im Deutschen durch DPen mit definitem Artikel realisiert, so wie man in (23a) sehen kann. Diese Kombination ist auch ohne Präposition verwendbar, wie (23b) zeigt, aber niemals ohne Artikel (24).[6]

(23) a. Fahren wir mit dem Bus/mit dem Zug/mit dem Skateboard?
b. Nehmen wir den Bus/den Zug/das Skateboard?

(24) a. *Fahren wir mit Bus/mit Zug/mit Skateboard?
b. *Nehmen wir Bus/Zug/Skateboard?

In den Beispielen in (23) wird keineswegs vorausgesetzt, dass ein Fahrzeug salient, einzigartig oder vorerwähnt wäre.

Die von Stvan (1998) beobachtete pragmatische Bedeutungserweiterung findet sich im Deutschen nicht oder kaum in PNKen, wohl aber in Verschmelzungsformen wie (25).

(25) Er ist im Gefängnis.

Dass PNKen nicht notwendigerweise mit Bedeutungsveränderungen korrespondieren, hatten wir bereits anhand der Beispiele in (8) und (9) verdeutlicht, auf die hier nochmals verwiesen werden soll. Wenn hier eine artikellose Phrase eine Bedeutung besitzt, die nahezu identisch zu einer Phrase mit Artikel ist, dann kann eine pragmatische Bedeutungserweiterung kein konstitutiver Bestandteil von PNKen im Deutschen sein.

Auch nicht beobachten lassen sich *defekte Nomina.* Beispiele des Typs (18) verlangen im Deutschen einen definiten Artikel. Bereits in Hinsicht auf die sog. *institutional nouns (school, church)* stellen Baldwin et al. (2006, S. 170) unter Verweis auf Himmelmann (1998) fest, dass *„the fact that institutional nouns can occur without determiners in these environments is, however, a peculiarity of English; related Germanic languages such as German or Swedish require the definite article here.“*

(26) Die Schule ist aus.

Während wir für die hier diskutierten Beispiele wohl zurecht behaupten können, dass eine einfache Übertragung der Ansätze von Stvan (1998) und Baldwin et al. (2006) auf das Deutsche nicht möglich ist, sollte die generelle Idee, dass die Semantik der beteiligten Elemente eine konstitutive Rolle spielt, weiter überprüft werden. Tatsächlich haben wir in unserer Analyse die Einbeziehung der Semantik der Präposition und der Semantik des Substantivs vorgesehen und es wird sich zeigen, dass beide eine Rolle spielen.

Wir schließen aber zunächst unsere Betrachtung mit der Schlussfolgerung, dass Analysen dem Phänomenbereich nicht gerecht werden, die entweder das Substantiv als nicht zählbar oder PNKen insgesamt als Ausnahmen klassifizieren. Ebenso wenig ist eine einfache Übertragung der Analysen aus dem Englischen auf das Deutsche möglich. Und schließlich muss auch nochmals betont werden, dass sich die Konstruktionen der reinen Introspektion entziehen. Eine Alternative bietet hier das Annotation Mining.

---

[6] Hier mag man einwenden, dass *Fahren wir doch mit Bus und Bahn!* grammatisch ist und somit der o.g. Behauptung widerspricht. Der Grammatikalität dieses Beispiels liegt allerdings eine andere Regularität zugrunde: Mit *oder* oder *und* verknüpfte artikellose Aufzählungen von Substantiven sind nach Präpositionen des Deutschen grundsätzlich möglich. Obwohl dieser Teilbereich der Grammatik der Präpositionen noch nicht gut untersucht ist, so kann man doch festhalten, dass Aufzählungen zu einer Pluraldenotation führen und somit die Beschränkungen für artikellose Pluralia im Deutschen gelten. Ich denke, dass ein Beispiel wie etwa *Ich fuhr mit Bus und Bahn.* nicht wahr sein kann, wenn ich tatsächlich nur mit dem Bus (oder nur mit der Bahn) gefahren bin.

3. Annotation Mining

Die vorliegende Analyse basiert auf einem Zeitungskorpus, den Jahrgängen von 1993 bis 1999 der Neuen Zürcher Zeitung mit einem Umfang von ca. 230 Millionen Wörtern. Die gesamte Annotation basiert auf einem XML-Stand-Off-Format, Daten und Annotate sind also strikt voneinander getrennt, was sowohl die Extraktion einzelner Annotationsebenen als auch ihr Hinzufügen deutlich vereinfacht. Als Werkzeug für die manuelle Annotation insbesondere der Präpositionsbedeutungen verwenden wir MMAX2 (Müller und Strube 2006).

Da es das Ziel der vorliegenden Analyse ist, konstitutive Bedingungen für die Realisation von PNKen insbesondere auch in Abgrenzung zur Realisation in der Form einer PP zu identifizieren, werden nicht nur PNKen annotiert, sondern auch PPen, für die gilt, dass Präposition und Substantiv in einer PNK aufgetreten sind, und PPen, die ein zählbares Substantiv einbetten, das nicht in einer PNK mit derselben Präposition aufgetreten ist.

Für diese Daten sehen wir die folgenden Annotationen vor:

Lexikalische Ebene: Part-of-Speech-Tags (kategoriale Etikettierung), Flexionsmorphologie, Derivationsmorphologie von Substantiven, Zählbarkeit von Substantiven, Interpretation von Substantiven durch GermaNet-Tops (d.h. Wortfelder), Interpretation von Präpositionen, Kompositabildung bei Substantiven.

Syntaktische Ebene: Einbettungsmodus (Adjunkt oder Komplement) der PP/PNK, syntaktische Komplemente des Nomens, pränominale Modifikatoren des Nomens.

Globale Ebene: Ist die Phrase Bestandteil einer Überschrift, eines Titels oder eines Zitats? Ist die Phrase idiomatisch?

In Überschriften, Titeln und Zitaten könnten PNKen nicht aufgrund syntaktischer oder semantischer Bedingungen realisiert werden, sondern weil Überschriften und Titel typische Orte für extragrammatische Kürzungen sind. Rein idiomatische PNKen mögen Kombinationsregeln folgen, die von den generellen Bedingungen für die Realisation von PNKen abweichen. Die Annotation dient auf der globalen Ebene also wesentlich dazu, bei der Analyse einzelne Fälle auszuschließen.

Große Teile der Annotation erfolgen automatisch mit der Hilfe der folgenden Werkzeuge: Der Regression Forest Tagger (Schmid und Laws 2008) wird verwendet für die kategoriale Etikettierung und die flexionsmorphologische Analyse (der Tagger enthält die morphologische Analysekomponente SMOR, vgl. Schmid 2004), der Tree Tagger (Schmid 1995) wird für die flache syntaktische Verarbeitung (Chunk Parsing) verwendet. Derivationsmorphologische Kategorien werden aus IMSLex (Lezius et al. 2000) abgeleitet.

Die Analyse verwendet zwei Ressourcen, um die Interpretation der Substantive festzulegen. Die erste Ressource ist GermaNet, (Kunze und Lemnitzer 2002), die deutsche Version des WordNet (vgl. http://wordnet.princeton.edu/). Wir verwenden die 23 Top-Level-Kategorien und jedes Substantiv wird mit allen Kategorien annotiert, denen es angehört. Top-Level-Kategorien entsprechen Wortfeldern. Darüber hinaus verwenden wir das Lexikon HaGenLex (Hartrumpf et al. 2003). HaGenLex kodiert spezifische sortale Informationen, die in einer formalen Ontologie strukturiert sind.

Die Klassifikation der Zählbarkeit erfolgt so, wie in Abs. 2.1 bereits beschrieben wurde. Zählbare Substantive sind solche, die der Klasse I angehören, nicht-zählbare Substantive sind solche, die den Klassen IV bzw. V angehören.

Durch automatische Annotationsverfahren liegen somit Informationen vor, die bei den Nomina von der internen Struktur über die syntaktische Distribution bis hin zur sortalen Klassifikation reichen. Leider liegt ein entsprechendes Verfahren für die Interpretation der Präpositionen nicht vor. Obwohl der Grad der Polysemie der Präpositionen nahezu ein Gemeinplatz ist,

sind bislang keine Versuche unternommen worden, Kategorieninventare *(Sense Tags)* für Präpositionsinterpretationen zu entwickeln. Sowohl die manuelle als auch die automatische Annotation setzen aber ein vordefiniertes Kategorieninventar voraus. Dieses Inventar muss natürlich die Polysemie der Präpositionen abbilden, darf aber nicht so komplex sein, dass die manuelle Annotation unmöglich wird. So legt etwa Schröder (1986) eine umfangreiche merkmalsbasierte Bestimmung der Interpretation von Präpositionen vor, die aber auf ca. 200 Merkmalen basiert und daher nicht für die Annotation eingesetzt werden kann. Dennoch ist Schröders Lexikon eine wertvolle Ressource.

Unsere Analyse basiert neben Schröder (1986) auf drei weiteren, im wesentlichen gebrauchsorientierten Ressourcen zur Beschreibung der Präpositionsinterpretationen des Deutschen: der Analyse von Helbig und Buscha (2001), der Darstellung im Duden „Deutsch als Fremdsprache“ (2002) und der Arbeit von Durrell und Brée (1993) zur Interpretation temporaler Präpositionen.

Zu diesen Ressourcen ist zu sagen, dass sie sehr heterogen und teilweise eher beispielorientiert als regelbasiert sind. Vor einer Anwendung ist also zunächst eine Angleichung und kritische Überprüfung der genannten Inventare erforderlich. Dies führt häufig zu einer Erweiterung von Interpretationsmerkmalen für einzelne Präpositionen und nicht selten zu einer Neudefinition von Teilen des Inventars. Das Annotationsschema für Präpositionsinterpretationen wird in Müller et al. (2010) detailliert beschrieben, ich möchte es an dieser Stelle anhand der wesentlichen Merkmale charakterisieren:

- Es gibt ein Inventar von insgesamt 27 Merkmalen, wobei die fünf Merkmale *spatial, temporal, kausal, modal* und *Vorhandensein* in einer Merkmalshierarchie organisiert sind, also Unterkategorien besitzen. Die Unterkategorien für temporale Merkmale entstammen der Analyse von Durrell und Brée (1993), die Untermerkmale der Kategorie *spatial* der Analyse von Schröder (1986) – zu *Vorhandensein* vgl. die Abs. 4.3 und 5.
- Annotatoren bewegen sich mittels Entscheidungsbaumverfahren durch die jeweiligen Hierarchien und haben auch die Möglichkeit, anstelle eines maximal spezifischen Interpretationstyps ein Hyperonym als Interpretation anzugeben. In den Analysen in Abs. 4 machen wir nur von den jeweiligen Obertypen Gebrauch.
- Zusätzlich zu den fünf Bedeutungen mit Unterbedeutungen gibt es 22 Bedeutungen ohne Unterbedeutung: *Zustand, komitativ, Quantitätsdifferenz, Beteiligung, Unterordnung, Zuordnung, Wechselbeziehung, Rangfolge, Über-/Unterschreitung, Bezugspunkt, Thema, Empfänger, Stellvertretung/Ersatz, Austausch, komparativ, restriktiv, kopulativ, adversativ, distributiv, Stellungnahme, Agens, Realisation.* Im Gegensatz zu den Merkmalen mit Untertypen mögen viele der hier genannten Merkmale eher abstrakt erscheinen. Wir werden die Merkmale anhand der Fallstudien in Abs. 4 genauer erläutern.
- Es besteht die Möglichkeit der Kreuzklassifikation, so dass bestimmte Eigenschaften der Interpretation nicht als Merkmale dargestellt werden; dies gilt insbesondere für die Repräsentation *direktionaler* Interpretationen. Direktionalität sollte bei sämtlichen lokalen Bedeutungsvarianten verfügbar sein, würde aber bei der Mehrzahl von Präpositionen mit einer lokalen Interpretation zu einer vorhersagbaren Hinzufügung des Merkmals *direktional* führen, das zudem nahezu immer aus dem Kasus des Objekts abgeleitet werden kann.
- Pro Präposition liegen zwischen einer und siebzehn Bedeutungen vor.

Das Vorkommen der Präpositionen im NZZ-Korpus in PNKen und PPen (PPen mit N aus PNK und PPen mit N nicht in PNK) ist im Folgenden dargestellt; die Präpositionen sind hierbei nach ihrer Polysemie geordnet:

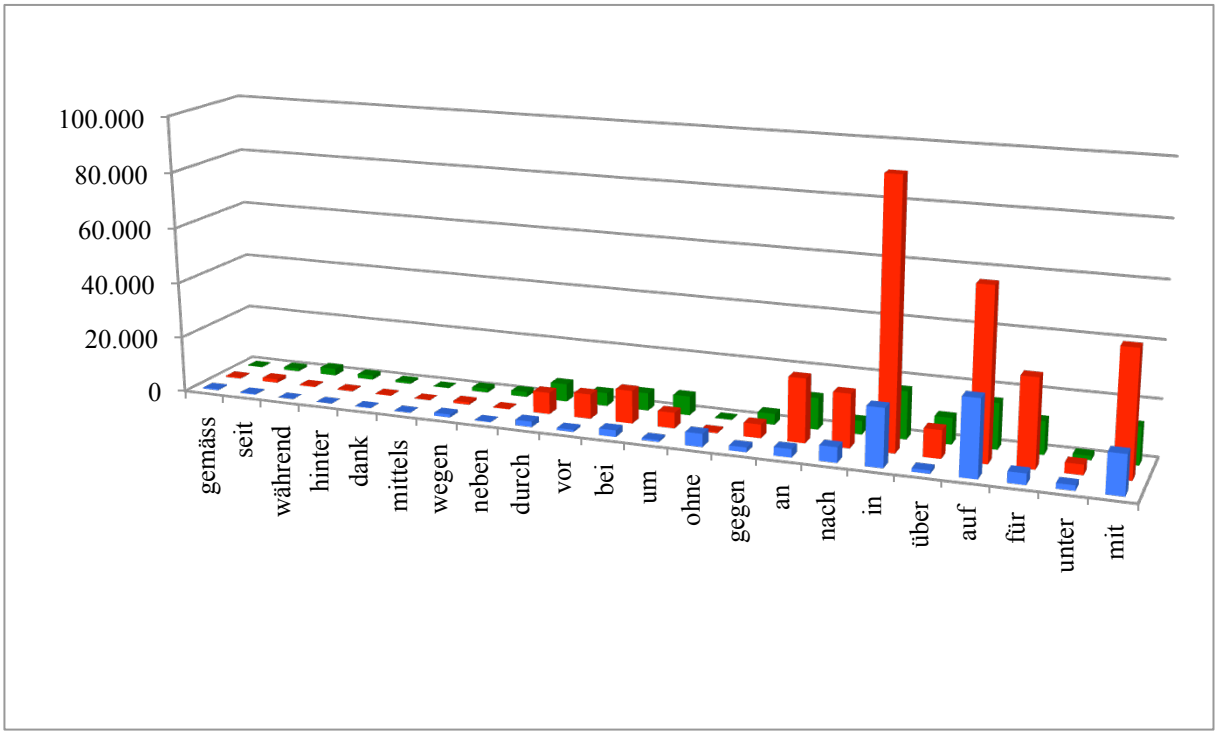
(27) Vorkommen von PNK und PP

	Bedeutungen	Unterbed.	PNK	PP (N in PNK)	PP (N nicht in PNK)
gemäß	1	1	455	418	177
seit	1	2	297	1.298	1.065
während	1	3	24	356	2.507
hinter	1	8	61	343	1.532
dank	2	3	171	246	741
mittels	2	3	195	75	154
wegen	2	3	1.091	886	1.409
neben	3	4	164	253	1.790
durch	5	6	1.970	7.674	6.311
vor	5	15	786	8.756	4.688
bei	5	24	2.340	11.679	6.280
um	6	7	701	5.549	6.765
ohne	6	11	4.632	473	218
gegen	6	16	1.687	4.831	3.857
an	6	29	2.916	22.563	11.174
nach	7	11	5.364	18.942	4.750
in	7	20	20.425	93.434	16.871
über	7	17	1.168	9.880	9.275
auf	7	29	27.011	59.938	16.131
für	8	10	3.885	31.029	11.612
unter	10	22	1.864	3.605	1.497
mit	11	18	13.981	43.833	13.262
Summe			91.188	326.061	122.066

Die Tabelle verdeutlicht, dass nicht alle Präpositionen zwingend polysem sind – und somit eine P-selektionale Analyse, die in jedem Fall die Polysemie der Präposition voraussetzt, nicht tragen würde. Die nicht-polysemen Präpositionen sind *gemäß* und *während.* Die größte Gruppe besteht aus Präpositionen, die zwischen zwei und sechs Bedeutungen besitzen. Schließlich gibt es auch eine Gruppe hochpolysemer Präpositionen, die sieben und mehr Interpretationen besitzen.

Betrachtet man die Gruppe bis zur Präposition *neben* (mit drei Bedeutungen), so ließe sich der Schluss formulieren, dass niedrige Polysemie auf der Typ-Ebene auch ein geringeres Vorkommen der Präposition nach sich zieht. Diese Annahme bestätigt sich allerdings nicht, wenn die weiteren Präpositionen betrachtet werden. So gehört *unter* zu den hochpolysemen Präpositionen, es werden aber verhältnismäßig wenig PNKen/PPen mit *unter* realisiert. Die graphische Darstellung der Tabelle (27) in (28) verdeutlicht dies nochmals. Die erste Reihe markiert hier die Häufigkeit der PNKen, die zweite die der PPen mit denselben Substantiven und die dritte die der PPen mit anderen zählbaren Substantiven.

(28) Vorkommen von PNK und PP (blau: PNK, rot: PP mit N in PNK, grün: PP mit N nicht in PNK)



Die Präposition *ohne* besitzt die Besonderheit, dass hier mehr Vorkommen in PNKen zu beobachten sind als in PPen.

4. Pilotstudie: Bedingungen für die Realisation artikelloser Phrasen als Komplemente von *ohne* und *unter*

4.1 Grundannahmen

Auch eine Analyse, die auf dem Annotation Mining aufsetzt, kann nicht gänzlich unabhängig von introspektiven Urteilen durchgeführt werden. Allerdings kommt introspektiven Bewertungen hier eine indirekte und vermittelnde Rolle zu. Annotation Mining bedeutet ja nichts anderes, als Rohdaten entweder automatisch oder manuell mit Annotationen zu versehen, was wiederum voraussetzt, dass für die relevanten Ebenen linguistische Annotationsschemata entwickelt werden. Es ist nun sehr unwahrscheinlich, dass die Entwicklung eines solchen Schemas – das ja eigentlich eine Mikrogrammatik ist – ohne jedes introspektive Urteil erfolgt. Die Zuordnung einzelner Elemente zu einem Schema kann ohne introspektive Befragung und insbesondere ohne einen bereits vorliegenden Kriterienkatalog nicht erfolgen. Ein zweiter Faktor betrifft die relative Unabhängigkeit der Annotationsschemata. Diese werden typischerweise unabhängig voneinander und auch von unterschiedlichen Personen entwickelt. Somit werden hier introspektive Urteile, wenn diese bei der Entwicklung eine Rolle spielten, auf eine breite Basis gestellt. Gerade vor dem Hintergrund dieser Überlegungen ist das Annotation Mining das Mittel der Wahl, wenn es um explorative Analysen geht. Die einleitenden Überlegungen haben gezeigt, dass PNKen regelhaft sind. Die explorative Analyse dient nun dazu, möglichst objektiv (d.h. auf eine breite empirische Basis gestützt) die Bedingungen dieser Regelhaftigkeit zu identifizieren. Insbesondere, wenn es sich um multikausale Bedingungen handelt, ist ein solches Vorgehen einer traditionellen, durch Introspektion geleiteten Analyse überlegen.

Für den vorliegenden Fall bieten sich nun Analysetechniken an, die es gestatten, die zugrunde liegende Fragestellung möglichst genau nachzuvollziehen. Eine solche Analysetechnik ist die binäre logistische Regression (Harrell 2001). Die Verletzung der Dudenregel 442 erfährt dabei die folgende Deutung: Die Realisation eines Artikels darf als regelhafte Grundbedingung gelten, d.h. wenn keine anderen Bedingungen erfüllt sind, ist die Realisation des Artikels der Normalfall. Wir betrachten den Wegfall des Artikels als vorherzusagende Eigenschaft und verwenden die im Annotation Mining identifizierten Merkmale von PNKen und PPen als Prädiktoren. Eine binäre logistische Regression identifiziert nun in den Merkmalen solche, die die Wahrscheinlichkeit eines Artikelwegfalls anheben und solche, die eine Anhebung absenken, so dass insgesamt eine Wahrscheinlichkeit für den Artikelwegfall in Abhängigkeit der vorhandenen Eigenschaften formuliert werden kann.
In den Analysen ist somit die abhängige Variable der Faktor DET, der die beiden Werte *no* und *yes* annehmen kann. Die folgenden Eigenschaften werden durch 92 Merkmale repräsentiert:

- Welche Interpretation besitzt die Präposition?
- Zu welchen Wortfeldern wird das Substantiv gezählt?
- Welche interne syntaktische Struktur besitzt das nominale Komplement: Wird es adjektivisch modifiziert, besitzt es ein Komplement, welche Kategorie besitzt das Komplement?
- Handelt es sich beim Kopf des nominalen Komplements um ein Kompositum oder nicht?
- Handelt es sich beim Nomen um ein abgeleitetes Nomen?
- Welche externe syntaktische Einbettung für die PNK/PP liegt vor: Handelt es sich um einen adverbialen Modifikator oder ein regiertes Komplement?

Bevor die Daten auf diese Weise analysiert wurden, war zunächst eine Bereinigung der Daten erforderlich: PNKen, die in Überschriften, Titeln und Zitaten vorkamen, wurden ebenso wenig für die Analyse berücksichtigt wie idiomatische Ausdrücke. Der Grund für die Elimination von PNKen in Überschriften, Titeln und Zitaten ist, dass diese Textstrukturen die extragrammatische Verknappung eines Textes in besonderem Maße fördern und somit hier auftretende PNKen möglicherweise eine grammatische Analyse verzerren würden. Ähnliches gilt auch für idiomatische Ausdrücke, die die Form einer PNK, aber ebenso häufig auch die Form einer PP annehmen. Deswegen wurden auch sie eliminiert. Insgesamt liegen den beiden Pilotanalysen die Daten in der Tabelle (29) zugrunde:[7]

(29)

Präposition	Σ	PP	PNK
ohne	3.750	591	3.159
unter	5.181	4.334	857

Zur Bewertung der eigentlichen Analyse ist es zunächst erforderlich, eine automatische Klassifikation durchzuführen, die davon ausgeht, dass alle Merkmale, die relevant sein könnten, auch tatsächlich relevant sind (einen sog. *full model fit)*. Einem *full model fit* liegen im vorliegenden Fall 92 Merkmale zugrunde, die die o.g. Eigenschaften abbilden. Auf der Basis dieses *full model fits* werden Merkmale, deren Relevanz als gering eingeschätzt wird, entweder au-

[7] Die Daten in der Tabelle (29) unterscheiden sich nicht nur deswegen von den Daten der Tabelle (27), weil PNKen in bestimmten Textstrukturen nicht berücksichtigt worden sind. Weiterhin ausgeschlossen werden automatisch, aber falsch annotierte Daten (etwa aufgrund von Tippfehlern), unvollständige Sätze, Postpositionen, die falsch als P gechunkt wurden, Zählbarkeitsfehler wegen Synkretismus, nicht erkannter Pluralform oder nominalisiertem Infinitiv, PPen in Funktionsverbgefügen, PP mit postnominalem Adverb, Verbpartikel, die fälschlich als Präposition klassifiziert und eine fehlerhafte syntaktische Analyse nach sich ziehen, sowie Numerativkonstruktionen.

tomatisch oder manuell verworfen und darauf erneut ein Modell entwickelt. Ein Merkmal wird dann verworfen, wenn es einen Modellkoeffizienten besitzt, der sich mit einer bestimmten Wahrscheinlichkeit nicht von 0 unterscheidet. Wenn ein Modellkoeffizient (= Gewichtung des Klassifikationsmerkmals) den Wert 0 besitzt, dann bedeutet dies nichts anderes, als dass das betreffende Merkmal für die vorliegende Analyse keine Bedeutung besitzt. Bei einer automatischen Analyse, etwa durch *fast backward elimination,* geht man davon aus, dass nur Merkmale erhalten bleiben, für deren Koeffizienten mit einer bestimmten Wahrscheinlichkeit (typischerweise wenigstens 95 %) angenommen werden kann, dass sie von 0 verschieden sind (vgl. Baayen 2008, S. 186). Dies kann aber dazu führen, dass einflussreiche Merkmale eliminiert werden, weswegen Harrell (2001, S. 56) auch von einem automatischen Verfahren abrät. Als Faustregel kann man angeben, dass ein Prädiktor entfernt wurde, wenn die Wahrscheinlichkeit, dass der Wert ungleich 0 ist, niedriger lag als 20 %. Für Prädiktoren, die nicht in den Bereich zwischen 20 und 95 % fielen, wurde jeweils ein Modell ohne und mit dem Prädiktor konstruiert und die Modellparameter wurden miteinander verglichen.

4.2 Binäre logistische Regression

Für die binäre logistische Regression gehen wir davon aus, dass es eine dichotome abhängige Variable gibt, beispielsweise eine Kategorie mit zwei Werten, die intern durch die Werte 0 und 1 kodiert werden kann. Die unabhängigen Variablen können metrische oder kategoriale Werte besitzen. Im vorliegenden Modell ist die abhängige Kategorie die Zugehörigkeit zu den Kategorien PNK und PP; die unabhängigen Variablen besitzen kategoriale Werte, aus denen wir ableiten wollen, ob ein Artikel realisiert wird (= PP) oder nicht (= PNK). Es sind die in Abs. 4.1 vorgestellten Merkmale.

Die binäre logistische Regression bestimmt allerdings nicht, ob eine Phrase zu einer der beiden Kategorien gehört, sondern gibt die Wahrscheinlichkeit an, mit der die Phrase zu den genannten Kategorien gehört. Man spricht hier von logistischer Regression, weil im Gegensatz zur linearen Regression nicht die gewichtete Summe der Prädiktoren gebildet wird, sondern die logistische Funktion der gewichteten Summe der Prädiktoren: $logit(p) = log(p/1–p)$.Die gewichtete Summe der Prädiktoren kann Werte zwischen $+\infty$ und $-\infty$ annehmen. Die inverse *logit*-Funktion bildet diese Werte aber auf das Intervall [0; 1] ab und stellt somit sicher, dass die resultierende Funktion tatsächlich eine Wahrscheinlichkeitsfunktion ist, die eine S-Form annimmt. Im Modell liegt somit eine Abbildung von Eigenschaften auf 92 Merkmale vor, von denen diejenigen Merkmale als Prädiktoren verwendet werden, die die Chancen beeinflussen, dass ein Vorkommen der einen oder anderen Kategorie zugeordnet wird.

In den folgenden Modellen wird jeweils die Wahrscheinlichkeit für die Realisation eines Artikels bestimmt. Die unabhängigen Variablen (= Prädiktoren) können hierbei positive und negative Werte annehmen. *Negative* Werte sprechen *gegen* die Realisation eines Artikels, *positive* Werte *für* die Realisation. Betrachten wir ein vereinfachtes Beispiel, in dem es nur einen Prädiktor gibt, der den Wert -1,386 annimmt. Für einen solchen Wert ergibt die Anwendung des inversen *logit* den Wert 0,2, d.h. die Wahrscheinlichkeit für die Weglassbarkeit des Artikels liegt in der betreffenden Konstruktion bei $1 – 0{,}2 = 0{,}8$.

4.3 Logistische Modelle für den Wegfall des Artikels bei *ohne* und *unter*

Das in Abs. 4.1 beschriebene Verfahren der Elimination von Prädiktoren führte zu Modellen für den Wegfall des Artikels bei den Präpositionen *ohne* und *unter,* die auf 13 (für *ohne)* bzw. 22 (für *unter)* Merkmalen aus den ursprünglichen 92 Merkmalen basieren.

4.3.1 Das Modell für *ohne*

Wir betrachten zunächst das Modell für *ohne,* das in (30) dargestellt wird.

(30) Koeffizienten für ein logistisches Modell für den Wegfall des Artikels bei *ohne*

	Koeffizient	**S.E.**	**p**
ACHSENABSCHNITT	-2,40	0,11	0,000
UNG-NOMINALISIERUNG	-1,36	0,19	0,000
ADJ IN N’	1,14	0,12	0,000
KAUSAL	1,21	0,13	0,000
KOMITATIV	2,28	0,52	0,000
BETEILIGUNG	3,40	0,49	0,000
VORHANDENSEIN	-0,78	0,15	0,000
DEP-S	5,08	1,05	0,000
DEP-NP	2,97	0,17	0,000
DEP-PP	2,20	0,15	0,000
GN-RELATION	-1,03	0,41	0,011
GN-ATTRIBUT	-1,35	0,30	0,000
GN-EREIGNIS	-0,84	0,14	0,000
GN-ARTEFAKT	-0,41	0,16	0,008

Die Tabelle in (30) gibt die relevanten Faktoren mit ihren Koeffizienten, d.h. Gewichtungen an, den Standardfehler der Schätzung *(engl. standard error,* d.h. *S.E.)* und die Wahrscheinlichkeit, dass die Gewichtung nur zufällig von 0 abweicht. Der Standardfehler kann genutzt werden, um nach der Daumenregel (Koeffizienz ± S.E. × 2) 95-%-Konfidenzintervalle für die Koeffizienten zu bestimmen. Selbst bei den Werten für GN-RELATION und GN-ARTEFAKT macht diese Berechnung deutlich, dass die Konfidenzintervalle 0 nicht einschließen.

Würden die Intervalle den Wert 0 einschließen, dann wäre es durchaus denkbar, dass der Wert eines Koeffizienten *tatsächlich* 0 beträgt, was bedeuten würde, dass dieser Koeffizient für den Wegfall eines Artikels keine Rolle spielt. Der Wert für den Achsenabschnitt kann in jedem Fall vernachlässigt werden: er ergibt sich als Artefakt aus der Modellierung der jeweils binären Werte für die unabhängigen Variablen und der Form der internen Kodierung dieser Werte.

Die identifizierten Eigenschaften können wie folgt interpretiert werden:

Die Merkmale KAUSAL, KOMITATIV und BETEILIGUNG (zur Definition dieser Eigenschaften siehe Abs. 5) besitzen positive Werte. Sie verschieben somit die Gewichte zugunsten einer Realisation eines Artikels. Eine erste naheliegende Interpretation dieser Werte könnte lauten: Wenn eine Präposition die Merkmale KAUSAL, KOMITATIV oder BETEILIGUNG besitzt, dann ist die Realisation einer NP (mit Artikel) wahrscheinlicher als die Realisation einer Nominalprojektion ohne Artikel. Beispiele für die Zuweisungen dieser Interpretationen sind in (31) bis (33) gegeben.[8]

[8] Die Interpretation KAUSAL subsumiert die Interpretation KONDITIONAL, die in Bsp. (31a) vorliegt.

(31) KAUSAL:

a. Sämtliche kurdischen Politiker sind davon überzeugt, dass *ohne einen Machtwechsel* in Bagdad die Kurdenfrage des Iraks nicht zu lösen sei. (NZZ, AUSLAND, 28.01.1993)

b. Und für eigentliche Verhandlungen hat Frau Kumaratunga nach eigenen Aussagen *ohne absolute Parlamentsmehrheit* noch kein klares Mandat. (NZZ, AUSLAND, 14.09.1994)

(32) KOMITATIV:

a. Ein mobiles Einsatzkommando überwältigte nach Polizeiangaben den aus Tunesien stammenden Geiselnehmer, als er *ohne das Kind* den Gerichtssaal verließ. (NZZ, VERMISCHTE MELDUNGEN, 14.07.1993)

b. Kessler, der *ohne Anwalt* vor Gericht erschien, betonte in seinem umfangreichen Plädoyer, seine anerkanntermaßen scharfe Kritik richte sich nicht gegen die Juden als Religionsgemeinschaft. (NZZ, ZÜRICH UND REGION, 11.03.1998)

(33) BETEILIGUNG:

a. Die abschließende Verhandlung der Nevada State Athletic Commission fand *ohne den Angeklagten* statt, dem außerdem die Kosten des Verfahrens aufgebrummt wurden. (NZZ, SPORT, 10.07.1997)

b. (keine Bsp. im Korpus gefunden)

Die Beispiele verdeutlichen – zumindest für die Merkmale KAUSAL und KOMITATIV, dass die Bedingungen für die Realisation nicht absolut zu werten sind, und insbesondere auch nicht isoliert. Ansonsten dürften die Belege (31b) und (32b) gar nicht existieren.
Das Merkmal VORHANDENSEIN besitzt demgegenüber einen negativen Wert. Liegt diese Interpretation vor, dann ist die Wahrscheinlichkeit der Realisation einer Nominalprojektion ohne Artikel höher als die Wahrscheinlichkeit der Realisation einer NP (mit Artikel). Beispiele finden sich in (34).

(34) VORHANDENSEIN:

a. Was weiter als positives Zeichen zu deuten wäre, Fragezeichen betreffend Stabilität und Solidität gegen die Portugiesen zum Trotz, ist die Serie von *acht Partien ohne Niederlage,* die längste seit 1924. (NZZ, SPORT, 02.04.1993)

b. Die Anklage wirft dem ersten von drei Angeklagten, einem 32jährigen *Mann ohne Beruf,* die Mitwirkung an allen drei Tötungsdelikten vor. (NZZ, STADT UND KANTON ZÜRICH, 14.04.1993)

Wir sehen hier also zwei unterschiedliche Interpretationsgruppen für die Präposition *ohne.* Besitzt die Präposition die Interpretationen KAUSAL, KOMITATIV oder BETEILIGUNG, wird die Realisation eines Artikels bevorzugt. Bei der Interpretation VORHANDENSEIN hingegen wird die Weglassbarkeit des Artikels suggeriert. In Abs. 5 werden wir auf diese Verteilung zurückkommen und fragen, ob die hier dargestellte Verteilung der Interpretationen der Präposition *ohne* nicht eine zugrunde liegende Eigenschaft dieser Konstruktion verbirgt.

Die mit DEP beginnenden Merkmale spezifizieren syntaktische Komplemente des Nomens in einer PNK, d.h. selegierte Sätze, postnominale NPen und postnominale PPen. Die drei Merkmale suggerieren, dass die Realisation eines Komplements die Realisation eines Artikels fördert. Das Merkmal ADJ IN N' legt fest, ob das Nomen in einer PNK adjektivisch modifiziert wurde. Das Merkmal besitzt wie DEP-S, DEP-NP und DEP-PP einen positiven Koeffizienten, d.h. fördert die Realisation eines Artikels. Allen vier Merkmalen ist gemein, dass sie die syntaktische Komplexität der Nominalprojektion erhöhen, so dass man den Schluss ziehen könn-

te, dass eine höhere syntaktische Komplexität der Nominalprojektion auch die Wahrscheinlichkeit eines Artikelwegfalls deutlich absenkt. Wir werden aber bei der Analyse von *unter* sehen, dass eher die Realisation eines Komplements als die syntaktische Komplexität die Wahrscheinlichkeit eines Artikelwegfalls beeinflusst.
Deverbale Nomina, die mit *–ung* gebildet werden, unterstützen ebenfalls den Wegfall des Artikels. Hier bedarf es eine genaueren Untersuchung bzw. auch der Integration von Verfahren zur Disambiguierung *ung*-nominalisierter Substantive, so wie dies etwa in Eberle (2010) vorgeschlagen wird. Es ist allerdings festzuhalten, dass eine Präferenz des Artikelwegfalls bei deverbalen Nomina eine übergreifende Eigenschaft zu sein scheint – sie begegnet uns erneut im Modell für *unter*.
Bei den Merkmalen GN-RELATION, GN-ATTRIBUT, GN-EREIGNIS und GN-ARTEFAKT handelt es sich um eine Zuordnung der Substantive zu Wortfeldern aus GermaNet (vgl. Kunze und Lemnitzer 2002). Diese Merkmale bezeichnen also semantische Eigenschaften der Substantive. Nomina, die mehr als einem Wortfeld angehören, sind polysem oder homonym. Für solche Substantive wird in der vorliegenden Klassifikation keine Disambiguierung durchgeführt. Da die Zuordnung typbasiert erfolgt, besitzt ein mehrdeutiges Nomen in jeder PNK bzw. PP *jede* Zuordnung aus GermaNet. Die Definition der für das Modell für *ohne* relevanten Kategorien in WordNet ist in (35) gegeben.

(35) a. *relation:* an abstraction belonging to or characteristic of two entities or parts together
b. *attribute:* an abstraction belonging to or characteristic of an entity
c. *event:* something that happens at a given place and time
d. *artefact:* a man-made object taken as a whole

Beispiele für Substantive des Wortfelds *relation* sind in (36a) gegeben, Beispiele des Wortfelds *attribute* in (36b), des Wortfelds *event* in (36c) und des Wortfelds *artefact* in (36d).

(36) a. Kameradschaft, Heiratsantrag
b. Tonhöhe, Gefühllosigkeit
c. Diffusion, Konjunkturprogramm
d. Klosteranlage, Schutzvorrichtung

Diese Wortfelder unterstützen den Wegfall des Artikels, was erneut am negativen Vorzeichen erkennbar ist. Dies ist auch deswegen interessant, weil unter diese Wortfelder keineswegs nur Abstrakta fallen, die in der Literatur ebenso oft wie fälschlich den Massentermen zugeordnet werden (vgl. etwa Bale und Barner 2009), sondern auch Konkreta. Dass diese Annahme nicht korrekt sein kann, sieht man an Substantiven wie *Gefühllosigkeit* oder *Kameradschaft,* die jeweils als zählbar klassifiziert werden können.
Diese Merkmale können als Selektionsrestriktionen interpretiert werden für substantivische Komplemente von *ohne,* wenn *ohne* die Interpretation VORHANDENSEIN besitzt. Bei mehr als 50 % der Vorkommen von *ohne* mit der Interpretation VORHANDENSEIN wird ein Substantiv ausgewählt, das in wenigstens eines der o.g. Wortfelder fällt, wobei 51,5 % der Substantive das Merkmal GN-EVENT besitzen und nur 35,3 % der Substantive das Merkmal GN-ARTEFACT.

4.3.2 Das Modell für *unter*

In (37) findet sich das Modell für *unter,* das mit seinen 22 Merkmalen nahezu doppelt so viele relevante Merkmale aufweist wie das Modell für *ohne.* Dass hier doppelt so viele Merkmale erforderlich sind, liegt sicherlich insbesondere daran, dass *unter* wesentlich mehr Bedeutun-

gen als *ohne* aufweist und damit mehr Interpretationen der Präposition und auch mehr Wortfelder der Substantive berücksichtigt werden müssen.

(37) Koeffizienten für ein logistisches Modell für den Wegfall des Artikels bei *unter*

	Koeffizient	**S.E.**	**p**
ACHSENABSCHNITT	-0,4379	0,1657	0,008
UNG-NOMINALISIERUNG	-0,8346	0,2259	0,000
ADJ IN N'	-1,0177	0,1432	0,000
KOMPOSITUM	2,1719	0,2538	0,000
REGIERT	1,9894	0,3017	0,000
SPATIAL	2,3237	0,2044	0,000
KAUSAL	1,3047	0,2272	0,000
UNTERORDNUNG	3,0529	0,2559	0,000
ZUORDNUNG	3,4228	0,1861	0,000
ÜBER-/UNTERSCHREITUNG	4,4186	0,3677	0,000
DEP-S	8,4717	4,0734	0,037
DEP-NP	0,8551	0,1436	0,000
DEP-PP	0,3043	0,2170	0,161
GN-GRUPPE	0,5241	0,2563	0,041
GN-KOMMUNIKATION	-0,9149	0,1443	0,000
GN-ORT	2,2704	0,6208	0,000
GN-RELATION	-2,1161	0,6022	0,000
GN-BESITZ	-0,8482	0,3665	0,021
GN-ATTRIBUT	-2,2847	0,2741	0,000
GN-ARTEFAKT	0,4169	0,1601	0,009
GN-MENSCH	1,8870	0,4999	0,000
HL-AD	-1,0253	0,1888	0,000
HL-AS	-1,4214	0,3804	0,000

Hier fällt nun allerdings als erstes auf, dass alle Koeffizienten für die Präpositionsbedeutungen ein positives Vorzeichen besitzen. Es werden zwar fünf der zehn Oberbedeutungen von *unter* im Modell berücksichtigt, aber diese Oberbedeutungen legen jeweils eine Blockade des Artikelwegfalls nahe. Es gibt keine Bedeutungen, die unmittelbar *für* den Artikelwegfall stehen, so wie etwa VORHANDENSEIN für *ohne.* Hier zeigt sich ein Muster, dass ich in Kiss (2007) bereits für spatiale Interpretationen beobachtet habe. Hier stellte ich fest, dass spatiale Interpretationen bei PNKen mit *unter* sehr selten auftreten, um nicht zu sagen nie. Vergleicht man den Koeffizienten des Merkmals SPATIAL mit den Werten für die anderen Bedeutungen, so wird erwartet, dass insbesondere UNTERORDNUNG, ZUORDNUNG und ÜBER-/UNTERSCHREITUNG die Realisation von PNKen blockieren. Wir finden also weniger lizenzierende Merkmale *für* den Artikelwegfall als Bedeutungen, die den Wegfall behindern.

Einige Muster aus dem Modell für *ohne* sind auch im Modell für *unter* wieder vorhanden. Insbesondere ist beobachtbar, dass die Merkmale DEP-S, DEP-NP und DEP-PP erneut gegen den Wegfall des Artikels sprechen.[9] Das Merkmal ADJ IN N' besitzt bei *unter* allerdings einen

[9] Die Irrtumwahrscheinlichkeit bei Dep-PP liegt oberhalb von 0,05. Dieser Faktor wurde dennoch weiter im Modell gehalten, weil sich die Modellparameter insgesamt verschlechtern, wenn auf diesen Faktor verzichtet

negativen Koeffizienten. Dies bedeutet, dass das Vorhandensein von Adjektiven in nominalen Komplementen von *ohne* die Weglassbarkeit des Artikels erschwert, dieselbe Bedingung bei Komplementen von *unter* aber den Wegfall des Artikels begünstigt.
Darüber hinaus scheint zunächst auch die externe Distribution der Phrase eine Rolle zu spielen: Das Merkmal REGIERT bestimmt, ob die PP bzw. PNK von einem Substantiv oder Verb regiert wird. Es spielt bei *ohne* keine Rolle, während es bei *unter* den Wegfall des Artikels erschwert. Der Grund hierfür kann aber auch sein, dass es nur sehr wenige Verben oder Nomina gibt, die die Präposition *ohne* regieren. Präpositionalobjekte, deren Kopf die Präposition *unter* ist, sind dagegen nicht so selten. Im vorliegenden Korpus werden nur 1,2 % der Vorkommen von *ohne* überhaupt regiert, aber 3,6 % der Vorkommen von *unter,* wobei *unter* ja auch insgesamt häufiger vorkommt. Nichtsdestotrotz kann man festhalten, dass Präpositional*objekte* mit *unter* den Wegfall des Artikels nicht unterstützen.
Weiterhin können wir beobachten, dass nominale Komplemente von *unter,* die als Komposita analysiert werden können, den Wegfall des Artikels ebenfalls nicht unterstützen. Es wird allerdings weiterer Untersuchungen bedürfen, um herauszufinden, warum dies so ist. Schließlich erkennen wir beim Modell für *unter* ebenso wie beim Modell für *ohne* den Einfluss der Nominalisierung durch *ung.* Nominalisierungen unterstützen den Wegfall des Artikels.

Die einzelnen Merkmale für die Wortfelder des Substantivs zeigen ein heterogenes Bild, das zu weiteren Untersuchungen einlädt: Es ist zunächst einmal auffällig, das GN-ORT einen hohen positiven Wert besitzt. GN-ORT dürfte Ausdruck einer Selektionsrestriktion bei *spatialen* Interpretationen von *unter* sein. Andere GN-Merkmale unterstützen hingegen den Wegfall des Artikels. Das Modell für *unter* macht aber nicht nur Gebrauch von Wortfeldern aus GermaNet, sondern auch von der ontologischen Klassifikation in HaGenLex. Die beiden relevanten Merkmale HL-AS (HaGenLex, abstrakt und statisch) und HL-AD (HaGenLex, abstrakt und dynamisch) sind Beschreibungen für *abstrakte* bzw. *abstrakte Sachverhalte* aus der Sortenhierarchie von HaGenLex (vgl. Hartrumpf et al. 2003). Beispiele für die beiden Gruppen finden sich in (38).

(38) a. abstrakte dynamische Sachverhalte (Merkmal HL-AD): Lauf, Diebstahl, Veränderung
b. abstrakte statistische Sachverhalte (Merkmal HL-AS): Umstand, Notlage, Auslastung

Wie auch anhand der Beispiele in (38) erkennbar ist, liegen abstrakte Nomina vor, die aber zählbar sind. Wir sehen somit einen weiteren Grund, an der Annahme zu zweifeln, dass Abstrakta nicht zählbar sein dürfen. Substantive, die als abstrakte statische oder dynamische Konzepte in HaGenLex charakterisiert sind, fördern den Artikelwegfall bei *unter.*

4.4 Die Güte der Modelle

Die Güte logistischer Modelle kann einerseits anhand bestimmter Kennzahlen bemessen werden, andererseits auch durch unterschiedliche Evaluationsverfahren bestimmt werden. Harrell (2001) empfiehlt für die Evaluation logistischer Modelle die sog. *bootstrap validation,* die im Gegensatz zu anderen Validierungsverfahren, insbesondere im Gegensatz zur *crossvalidation* den Vorteil hat, dass alle Daten zum Training *und* zum Testen verwendet werden können. Mittels *bootstrap validation* kann bestimmt werden, ob das Modell zu sehr an den gesehenen Daten orientiert ist (ein *overfit* vorliegt) oder ob hier plausibel von gesehenen auf nicht gese-

wird. Harrell (2001, S. 56) weist darauf hin, dass die Elimination von Faktoren aufgrund einer Überschreitung des Signifikanzniveaus keineswegs immer ratsam ist.

hene Daten generalisiert werden kann und somit ein Modell mit Vorhersagekraft vorliegt. Bei einer *bootstrap validation* wird aus einem Datensatz mit *n* unterschiedlichen Elementen durch Auswahl mit Zurücklegen eine Menge von Datensätzen mit *m (= n)* **nicht-**unterschiedlichen Elementen gebildet. Eine Auswahl mit Zurücklegen bedeutet hier, dass ein Datensatz in die Stichprobe übernommen wird, aber auch danach nochmals für die Stichprobe zur Verfügung steht. Die Trainingsmenge enthält also genau so viele Elemente wie der ursprüngliche Datensatz, aber durch das Zurücklegen werden einzelne Daten häufiger als einmal in der Trainingsmenge erscheinen. Die Testmenge besteht dann gerade aus denjenigen Elementen des ursprünglichen Datensatzes, die nie ausgewählt wurden. Ebenso wie bei einer *crossvalidation* wird dieser Prozess mehrfach wiederholt. Erfreulicherweise legen bereits die reinen Kenndaten der beiden Modelle den Schluss nahe, dass hier Modelle mit Vorhersagekraft entwickelt wurden und diese Annahme wird auch durch die *bootstrap validation* bestätigt.

(39) Grundlegende Kenndaten der Modelle

	Model L.R.	p	C	D_{xy}
ohne	1.063,5	0	0,876	0,753
unter	2.245,6	0	0,937	0,874

Der Wert *Model L.R. (=Model Likelihood Ratio)* gibt an, wie sehr sich die Vorhersagekraft des Modells von einem Modell mit willkürlich gewählte Prädiktoren (bzw. eben ohne jede Prädiktoren) verändert. Unter *p* wird erneut die Wahrscheinlichkeit bestimmt, mit der diese Veränderung nur zufällig sein könnte – sie liegt bei 0.
Besonders relevant sind die hohen Werte für *C* und *Somers* D_{xy} in (39). Somers D_{xy}-Wert bestimmt das Verhältnis von konkordanten zu nicht-konkordanten Übereinstimmungen bei einem Modell, dessen abhängige Variable die Werte 0 und 1 annehmen kann. Eine Übereinstimmung ist konkordant, wenn im Modell die vorhergesagten Wahrscheinlichkeiten mit den tatsächlichen Klassenzugehörigkeiten korrespondieren. Der Wert für D_{xy} entspricht $(C - 0{,}5) \times 2$. C und D_{xy} beschreiben eigentlich denselben Sachverhalt, aber D_{xy} ist eine Abbildung des *C*-Wertes auf eine Wahrscheinlichkeitsskala. Ein willkürliches Modell ohne Vorhersage besitzt einen *C*-Wert von 0,5 ($D_{xy} = 0$). Werte für *C* über 0,8 ($D_{xy} = 0{,}6$) legen den Schluss nahe, dass das Modell Vorhersagekraft besitzt; wir sehen bei dem Modell für *unter* sogar einen Wert deutlich über 0.9.
Das Ergebnis der *bootstrap validation* für die beiden Modelle ist in (40) dargestellt, hierbei wurden jeweils 200 Iterationen durchgeführt, d.h. 200 Trainingssätze wurden erstellt, mit den Modellparametern trainiert und mit den verbleibenden Daten evaluiert.

(40) *Bootstrap validation* der Modelle

a. *ohne*

	Alle Daten	Training	Test	Optimismus	Korrektur
D_{xy}	0,7526	0,7570	0,7500	0,0070	0,7456
E_{max}	0,0000	0,0000	0,0096	0,0096	0,0096

b. *unter*

	Alle Daten	Training	Test	Optimismus	Korrektur
D_{xy}	0,8736	0,8744	0,8692	0,0052	0,8684
E_{max}	0,0000	0,0000	0,0055	0,0055	0,0055

Die Zeilen geben darüber Auskunft, dass der ursprüngliche D_{xy}-Wert für das Modell für *ohne* bei 0,7526 lag; die Differenz zwischen dem Training bei der *bootstrap validation* und dem Test betrug allerdings (0,757 – 0,750) = 0,007. Dieser Wert charakterisiert den Optimismus des Modells, d.h. denjenigen Anteil an der Vorhersagekraft, der eine zu starken Ausrichtung an die Trainingsdaten geschuldet ist. Entsprechend muss der ursprüngliche Wert um dieses Maß an Optimismus nach *unten* korrigiert werden, so dass sich nach Validierung ein Wert von 0,7456 ergibt. Entsprechendes gilt in der zweiten Zeile für die durchschnittliche maximale Fehlerrate. Im Ursprungsmodell liegt diese bei 0, weil mit allen Daten trainiert wurde. Der Testfall zeigt aber eine durchschnittliche Veränderung von 0,0096, so dass dieser Wert nun als Fehlerrate angesetzt wird. Die Ergebnisse für *unter* wiederholen die für *ohne.*

Vor dem Hintergrund dieser Evaluation erfahren die Koeffizienten der o.g. Modelle die folgende Interpretation: Je höher ein Wert ausfällt, desto einflussreicher ist der Wert, desto deutlicher wird die Wahrscheinlichkeit für die Weglassbarkeit des Artikels in Richtung des Vorzeichens des Koeffizienten modifiziert. So ist etwa der Wert für DEP-S in beiden Modellen sehr hoch und positiv, so dass wir davon ausgehen können, dass die Realisation eines sententialen Komplements des Nomens die Weglassbarkeit des Artikels extrem erschwert. Liegen aber – etwa bei *ohne* keine Dependenten des Substantivs vor, dafür aber die entsprechenden Interpretationen von Präposition und Substantiv, dann ist die Weglassbarkeit des Artikels sehr wahrscheinlich. Wir wollen im folgenden Abschnitt der Frage nachgehen, ob es sich nun tatsächlich um rein semantische Faktoren handelt oder hier zugrunde liegende strukturelle Faktoren eine Rolle spielen könnten.

5. Externe Faktoren beim Artikelwegfall mit *ohne?*

Eine Interpretation der Präposition *ohne* – VORHANDENSEIN – unterstützt die Weglassbarkeit des Artikels, während die Interpretationen KAUSAL, KOMITATIV und BETEILIGUNG die Weglassbarkeit eher erschweren.

Hier stellt sich die Frage, ob die Verteilung aus intrinsischen Faktoren der Interpretationen erklärt werden kann, oder ob nicht doch andere, verborgene Faktoren eine Rolle spielen könnten. Wenn wir die Interpretationsmöglichkeiten von *ohne* insgesamt betrachten, fällt auf, dass die Interpretationen zumindest in Teilen ein syntaktisches Korrelat besitzen, somit also eine Abbildung zwischen Interpretationen, syntaktischen Korrelaten und der Artikelweglassbarkeit denkbar ist.

Hierzu erscheint es zunächst sinnvoll, die Bedeutung der Präposition *ohne* zusammen mit der Bedeutung der Präposition *mit* zu betrachten. Das Annotationsschema für Präpositionsbedeutungen trägt den Besonderheiten dieser beiden Präpositionen – und der Tatsache, dass zumindest einige Interpretationen der einen in Opposition zu Interpretationen der anderen zu charakterisieren sind – dadurch Rechnung, dass bestimmte Kategorien nur von diesen Präpositionen besetzt werden können.

Die Präposition *ohne* teilt mit der Präposition *mit* die Interpretation VORHANDENSEIN mit den Untertypen SYNTHETISCH und ANALYTISCH. Während diese Interpretation bei *mit* angibt, dass etwas vorhanden ist, drückt sie bei *ohne* einen Mangel aus. Die Binnendifferenzierung von VORHANDENSEIN, SYNTHETISCH und ANALYTISCH kann leichter anhand von *mit* erläutert werden:

VORHANDENSEIN bezeichnet im Schema das Vorhandensein (bzw. bei *ohne:* Fehlen) einer Sache, eines Merkmals oder einer Eigenschaft; SYNTHETISCH bezeichnet das Vorhandensein einer Sache, eines Merkmals oder einer Eigenschaft, die als einer anderen zugehörig betrach-

tet werden kann, ohne im Sinne einer Mereologie Teil des anderen zu sein; man kann diese Art der Zugehörigkeit auch als eine betrachten, die durch synthetische Urteile erfasst werden kann, z.B. *eine Flasche mit Verzierung.* Demgegenüber beschreibt ANALYTISCH eine Zugehörigkeit, die tatsächlich als Teil-Ganzes-Beziehung beschrieben werden kann bzw. entsprechend durch ein analytisches Urteil ausgedrückt wird. Die analytische Form der Urteile hat den Charakter des Pleonastischen und deswegen findet man typische analytische Interpretationen von *mit* nur dann, wenn das Teil in bestimmter Weise modifiziert wird, wie die Kontraste in (41) zeigen:

(41) a. eine Hand mit Fingern vs. eine Hand mit *drei* Fingern
b. ein Auto mit Motor vs. ein Auto mit Dieselmotor

Für *ohne* gilt dies natürlich nicht in gleichem Maße, weil das Fehlen eines ansonsten mitgedachten Teils einen höheren Informationsgehalt besitzt als das Vorhandensein dieses mitgedachten Teils:

(42) a. ein Auto ohne Motor
b. eine Hand ohne Daumen

Die drei Merkmale VORHANDENSEIN, SYNTHETISCH und ANALYTISCH können die Präpositionen *mit* und *ohne* nur annehmen, wenn sie nominale Phrasen modifizieren.
Zusätzlich zu diesen Merkmalen besitzt *ohne* die in (43) aufgeführten Interpretationen:

(37) KAUSAL, KONDITIONAL, ART UND WEISE, BEGLEITUMSTAND, BETEILIGUNG, KOMITATIV, INSTRUMENTAL

Diese Interpretationen treten typischerweise auf, wenn *ohne* entweder eine verbale Projektion modifiziert oder eine nominale Projektion, deren Kopf ein Ereignisnominalisierung ist. Weiterhin muss hierbei berücksichtigt werden, dass die Merkmale BETEILIGUNG und KOMITATIV ebenso wie das oben bereits angesprochene Merkmal VORHANDENSEIN nur durch die Präpositionen *ohne* und *mit* besetzt werden kann. Wir können somit sagen, dass zumindest eine Kernbedeutung der Präpositionen *ohne* und *mit* durch die Merkmale VORHANDENSEIN, BETEILIGUNG und KOMITATIV bestimmt wird, wobei die Verteilung dieser Merkmale entweder tatsächlich einer externen syntaktischen Distribution (welches Element wird syntaktisch modifiziert) oder einer durch die syntaktische Distribution vermittelten Interpretationsbedingung (wird ein Ereignis oder etwas Dinghaftes modifiziert) geschuldet ist.
Diese beiden Bedingungen können für die Interpretation BETEILIGUNG wie folgt illustriert werden: In (44) liegen verbale Modifikationen vor, in (45) nominale Modifikationen mit ereignisartigem Charakter.

(44) a. Die Hauptverhandlung wird am 8. Februar notfalls auch ohne den Angeklagten fortgesetzt. (NZZ, AUSLAND, 28.01.1993)
b. Der Viertelmeilen-Sprint, einer der anfänglich vermuteten Höhepunkte, entwickelte sich ohne den großen Abwesenden von Zürich, Michael Johnson, zum erwarteten US-Duell zwischen Antonio Pettigrew und Jerome Young, dem Zweit- und dem Drittschnellsten der Saison. (NZZ, SPORT, 12.08.1999)

(45) a. Nur akademische Miles-Davis-Forscher dürften von diesem detaillierten Einblick in die Werkstatt profitieren und stundenlang abgebrochene Versuche oder gar Durchläufe ohne den Solisten mitverfolgen. (NZZ, PHONO-SPEKTRUM, 02.10.1996)
b. Denn eine Siegerehrung ohne die Amerikanerin ist letztlich doch wohl (noch) ebenso die Ausnahme von der Regel wie die Durchführung einer Abfahrt in zwei Läufen. (NZZ, SPORT, 16.12.1995)

Es scheint also insgesamt so zu sein, dass die einen Interpretationen von *ohne* nur in verbalen syntaktischen Kontexten realisiert werden können (bzw. in nominalen syntaktischen Kontexten, in denen das Substantiv ereignisartig interpretiert werden muss), während die anderen in nominalen Kontexten realisiert werden. Nominale Kontexte scheinen nun die Artikelweglassbarkeit zu bevorzugen, verbale Kontexte (bzw. Nomina mit Ereignislesart) hingegen blockieren die Artikelweglassbarkeit. Wenn wir nun nochmals das Modell für die Artikelweglassbarkeit bei *ohne* in (30) betrachten, so könnte die Präferenz für die Weglassbarkeit bei der Interpretation VORHANDENSEIN bzw. die Präferenz für die Artikelrealisation bei den Interpretationen KAUSAL, KOMITATIV und BETEILIGUNG auch so gedeutet werden, dass die Präposition *ohne* dann ein artikelloses Komplement besitzen kann, wenn die Präposition eine Nominalprojektion modifiziert, aber nicht, wenn sie eine verbale Projektion modifiziert. Dies muss nicht notwendigerweise bedeuten, dass die Analyse der PNK konstruktional sein muss, denn die verbale bzw. nominale Modifikation könnte formal auch durch den semantischen Typ des zu modifizierenden Elements erfasst werden, was dann wiederum auch den Einschluss von Daten wie (45) gestatten würde.
Da nun wiederum die Bedeutungen VORHANDENSEIN, KOMITATIV und BETEILIGUNG diese Verteilung aufweisen und nur als Bedeutungen für die Präpositionen *ohne* und *mit* verwendet werden, läge der Schluss nahe, auch das Annotationsschema für die Interpretationen so zu modifizieren, dass etwa das Merkmal [±EREIGNIS] als Kreuzklassifikator verwendet wird, um das Merkmal VORHANDENSEIN von den anderen beiden zu trennen, und entsprechend das Schema nochmals zu vereinfachen.

6. Schlussfolgerungen

Die grundlegende Idee hinter dem *Annotation Mining* ist, eine möglichst hohe Zahl *potentiell* relevanter linguistischer Eigenschaften bei einer möglichst hohen Zahl einzelner Daten zu repräsentieren und aus den Daten durch Klassifikationsverfahren diejenigen Eigenschaften zu extrahieren, die für die syntaktische Distribution etwa der PNKen relevant sein können. Anhand der Präpositionen *ohne* und *unter* haben wir zeigen können, welche Merkmale die Weglassbarkeit des Artikels bedingen. Für die Präposition *ohne* konnten wir zunächst ableiten, dass bestimmte Interpretationen der Präposition für die Weglassbarkeit des Artikels beim Komplement sprechen, andere hingegen dagegen. Eine genauere Betrachtung zeigte dann aber bereits, dass strukturelle Faktoren (was wird modifiziert) hier eine Rolle spielen könnten. Strukturelle Faktoren müssen neben den interpretatorischen Faktoren ohnehin angesetzt werden, um zu erklären, welchen Einfluss die Realisation eines postnominalen Komplements auf die Weglassbarkeit des Artikels hat.
Für die Präposition *unter* konnten wir diesen letzten Aspekt erneut sichtbar machen, so dass man von einer allgemeinen strukturellen Eigenschaft sprechen kann: Besitzt ein Substantiv ein postnominales Komplement, dann wird die Weglassbarkeit des Artikels erschwert, wobei dies insbesondere gilt, wenn das Komplement ein Satz ist.
Besonders auffällig ist im Vergleich der Präpositionen, dass bei *ohne* eine Kernbedeutung den Artikelwegfall erleichtert, eine andere hingegen den Wegfall erschwert, während es bei *unter* nur Interpretationsmerkmale gibt, den die Artikelwegfall erschweren. Vor einer Betrachtung weiterer Präpositionen ist es sicherlich zu früh, diese Beobachtung zu werten. Dennoch bietet sich hier eine Spekulation an: Bei *ohne* stehen die Kernbedeutungen somit für bzw. wider den Artikelwegfall (möglicherweise auch in Verbindung mit weiteren strukturellen oder interpretativen Bedingungen), bei *unter* ist aber ähnliches (noch) nicht zu beobachten. Ist dies viel-

leicht der Grund dafür, dass der Artikelwegfall bei *ohne* schon der Regelfall ist, bei *unter* aber noch die Ausnahme?

Literatur:

Allan, Keith. 1980. Nouns and countability. *Language* 56(3), 541-567.

Baayen, R. Harald. 2001. *Word Frequency Distributions.* Dordrecht: Kluwer.

Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Pratical Introduction to Statistics Using R.* Cambridge: Cambridge University Press.

Baldwin, Timothy, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger und Ivan Sag. 2006. In search of a systematic treatment of determinerless PPs. In Patrick Saint-Dizier (eds.), *Syntax and Semantics of Prepositions*. Springer, Dordrecht, 163-179.

Bale, Alan und David Barner. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. *Journal of Semantics 26,* 217-252.

Barner, David und Jesse Snedeker (2004). *Quantity judgements and individuation: evidence that mass nouns count.* 25 Francis Avenue, Cambridge: Laboratory for Development Studies, Harvard University, Shannon Hall.

Borer, Hagit. 2005. *Structuring Sense, Vol. I: In Name Only*. Oxford: Oxford University Press.

Chiarcos, Christian, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz und Manfred Stede. 2008. A flexible framework for integrating annotations from different tools and tagsets. *Traitement Automatique des Langues.* Special Issue Platforms for Natural Language Processing. ATALA, 49 (2).

Dömges, Florian, Tibor Kiss, Antje Müller und Claudia Roch. 2007. Measuring the Productivity of Determinerless PPs. In Costello, F., J. Kelleher und M. Volk (eds): *Proceedings of the 4th ACLSIGSEM Workshop on Prepositions,* Prag, 31-37.

Duden 2005. *Duden. Die Grammatik*. Duden Band 4. Bibliographisches Institut & F.A. Brockhaus AG, Mannheim.

Duden 2002. *Duden*. *Deutsch als Fremdsprache*. Bibliographisches Institut & F.A. Brockhaus AG, Mannheim.

Durell, Martin und David Brée. 1993. German temporal prepositions from an English perspective. In Cornelia Zelinsky-Wibbelt (ed.)*, The Semantics of Prepositions. From Mental Processing to Natural Language Processing*. De Gruyter, Berlin/New York, 295-325.

Eberle, Kurt. 2010. *–ung* Nominalizations of Verbs of Saying in German: Events and Propositions. *Proceedings of Chronos IX,* Paris.

Espinal, M. Teresa und Louise McNally. 2010. Bare singular nominals and incorporating verbs in Spanish and Catalan. *Journal of Linguistics.*

Stefan Evert. 2004. A Simple LNRE Model for Random Character Sequences. In: *Proceedings of the 7mes Journees Internationales d'Analyse Statistique des Donnees Textuelles,* 411-422.

Gillon, Brendan S. 1999. *The Lexical Semantics of English Count and Mass Nouns.* Department of Linguistics, McGill University Montreal.

Harell jr., Frank E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer: New York.

Hartrumpf, Sven, Hermann Helbig und Rainer Osswald. 2003. The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment. *Traitement automatique des langues* 44(2), 81-105.

Helbig, Gerhard und Joachim Buscha. 2001. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Leipzig, Langenscheidt.

Himmelmann, Nikolaus. 1998. Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology 2*, 315–353.
Jespersen, Otto. 1924. *The philosophy of grammar*. London: Allen & Unwin.
Kiss, Tibor. 2007. Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *Zeitschrift für Sprachwissenschaft 26 (2),* 317-345.
Kunze, Claudia und Lothar Lemnitzer. 2002. GermaNet - representation, visualization, application. *Proc. LREC 2002*, main conference, Vol V., 1485-1491.
Le Bruyn, Bert, Henriëtte de Swart and Joost Zwarts. 2009. *Bare PPs across languages*. Workshop on Bare nouns, Paris.
Lezius, Wolfgang, Stefanie Dipper und Arne Fitschen. 2000. IMSLex – Representing Morphological and Syntactic Information in a Relational Database. In: Heid, Ulrich, Stefan Evert, Egbert Lehmann und Christian Rohrer (Hrsg.). *Proceedings of the 9th EURALEX International Congress.* Stuttgart, 133-139.
Müller, Antje, Katja Keßelmeier, Claudia Roch, Tobias Stadtfeld, Jan Strunk und Tibor Kiss. 2010. Creating a Feature Space for the Annotation of Preposition Senses in German. *Proceedings of Linguistic Evidence 2010,* Tübingen.
Müller, Christoph und Michael Strube. 2006. Multilevel annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, Joybrato Mukherjee, (eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods.* Peter Lang, Frankfurt a.M., 197-214.
Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT Workshop*, Dublin.
Schmid, Helmut, Arne Fitschen und Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC 2004*, Lissabon, 1263-1266.
Schmid, Helmut und Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING 2008*, Manchester.
Schröder, Jochen. 1986. *Lexikon deutscher Präpositionen*. Leipzig, VEB Verlag Enzyklopädie.
Stvan, Laurel S. 1998. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Ph.D. thesis, Northwestern University, Evanston/ Chicago, IL.