Die Konstruktion von Lexika für die maschinelle syntaktische Analyse

Norbert BRÖKER Stefanie DIPPER

1. Einleitung

In diesem Beitrag diskutieren wir einige Fragen zur Konstruktion eines Lexikons für Unifikationsgrammatiken. Lexika in realistischer Größenordnung und Beschreibungstiefe können aus Korpora extrahiert werden, sind jedoch mit einem grundsätzlichen Problem behaftet: Das Lexikon enthält nur Subkategorisierungsrahmen, die tatsächlich in den Korpora aufgetreten sind, ist also lückenhaft. Seltener gebrauchte Subkategorisierungsrahmen fehlen in diesen Lexika.

Zur Ergänzung fehlender und Verbesserung existierender Lexikoneinträge wollen wir linguistische Generalisierungen über Subkategorisierungsrahmen ausnutzen, von denen nicht alle mit traditionellen Lexikalischen Regeln beschrieben werden können. In diesem Artikel schlagen wir dazu eine Architektur vor, die Lexikalische Regeln generellerer Art unterstützt.

2. Szenario

Die vorliegende Arbeit entstand im Rahmen von Pargram, einem Kooperationsprojekt zwischen Xerox PARC, XRCE Grenoble und der Universität Stuttgart zur Erstellung von parallelen LFG-Grammatiken für Englisch, Französisch und Deutsch für die maschinelle Übersetzung (s. `http://www.parc.xerox.com/istl/groups/nltt/pargram/`).Die deutsche Grammatik nutzt halbautomatisch erstellte abstrakte Subkategorisierungslexika (ECKLE 1999), die aus Zeitungskorpora extrahiert wurden und im TSNLP-Format vorliegen (ESTIVAL et al. 1995). Daraus werden vollautomatisch LFG-Lexika abgeleitet, die jedem Lemma eine Menge von Subkategorisierungsrahmen zuordnen. Momentan enthält das Subkategorisierungslexikon etwa 14.000 Lemmata, denen durchschnittlich 2 Subkategorisierungsrahmen (und maximal 23 für *entscheiden*) zugeordnet werden. Die manuelle Bearbeitung verbietet sich aufgrund dieser Größenordnung.

Obwohl der LFG-Formalismus (und auch das verwendete Grammatikentwicklungssystem XLE) traditionelle Lexikalische Regeln bereitstellt, sollen in dem hier vorgestellten Ansatz Diathesen wie z.B. Passiv schon bei der Erstellung des LFG-Lexikons behandelt werden, und nicht erst bei der Strukturanalyse durch Lexikalische Regeln. Traditionelle Lexikalische Regeln sind damit überflüssig.

Zur Illustration zeigen wir einen Ausschnitt der Ausgangs- und Zieldaten für das Verb *geben*. Das Ausgangsformat verwendet eine an TSNLP angelehnte Notation: Ein Subkategorisierungsrahmen wird als kommaseparierte Folge von Komplementen geschrieben, die die Funktion (z.B. `subj`, `obj`, `v-comp`, etc.) und die Kategorie des Komplements (z.B. `NP_nom`, `PP_mit`, etc.) angeben. Das Zielformat enthält in diesem Beispiel eine Disjunktion (`{... | ...}`), die durch Passivierung entsteht. Komplemente werden durch Bindestrich getrennt; ihre Funktion wird durch die Position kodiert. Eigenschaften des Subkategorisierungsrahmens als Ganzes (wie z.B. Passiv) werden durch zusätzliche Attribute beschrieben.

Ausgangsformat:

geben	(subj(NP_nom),obj(NP_acc),iobj(NP_dat))

Zielformat:

geben	{@(NPnom-NPacc-NPdat geben)
	| @(null-NPnom-NPdat-PASSIVE geben) }

Unser Modell operiert nicht wie traditionelle Lexikalische Regeln auf einzelnen Subkategorisierungsrahmen, sondern auf Mengen von Subkategorisierungsrahmen. Die folgenden Abschnitte motivieren und illustrieren diesen Ansatz: Wir diskutieren Klassen von Operationen über Lexikoneinträgen, die mehrere Subkategorisierungsrahmen eines Lemmas simultan betrachten müssen. Diese Operationen führen dazu, daß in Lexikoneinträgen Rahmen hinzugefügt, modifiziert oder gelöscht werden.

3. Erweiterung der Menge von Subkategorisierungsrahmen

Ein Beispiel hierfür sind Verben mit satzförmigen Komplementen. In ECKLE (1999) wird z.B. gezeigt, daß ein Verb, das einen zu-Infinitiv,

einen Verb-zweit-Satz und einen ob-Satz subkategorisiert, auch immer einen daß-Satz und einen wh-eingeleiteten Satz einbetten kann.

Ein anderes Beispiel stellen Verben mit obligatorischen Lokativ-Ergänzungen dar. Z.B. erfordert *sich befinden* eine Ortsangabe in Form eines Adverbs (*dort*) oder einer PP (*auf/unter/bei der Brücke*). In diesen Fällen subkategorisiert das Verb nicht eine bestimmte Präposition (wie das bei *bestehen auf* der Fall ist), sondern eine ganze Klasse von Präpositionen: nämlich alle Präpositionen, die eine lokative Bedeutung haben. Es ist zu erwarten, daß nicht zu jedem Verb mit lokativer Ergänzung Belege für alle lokativen Präpositionen gefunden werden können. Gleichzeitig aber stellen diese Präpositionen eine geschlossene Klasse dar. Daher liegt folgendes Vorgehen nahe: Weist ein Verb eine signifikante Anzahl von Subkategorisierungsrahmen auf, die sich nur in der Wahl der Präposition unterscheiden, und handelt es sich dabei um lokative Präpositionen, dann werden analog Rahmen mit den noch nicht abgedeckten lokativen Präpositionen hinzugefügt.

Ein naheliegender Schritt ist dann, alle lokativen Präpositionen zu einer abstrakteren Kategorie zusammenzufassen, so daß die linguistische Generalisierung explizit kodiert ist. Dann würden in diesem Beispiel mehrere Rahmen kollabieren. Hier wird deutlich, daß nur solche Eigenschaften automatisch aus Korpora extrahiert werden können, die aufgrund der Oberflächenform identifizierbar sind. Dagegen ist bei Präpositionen nicht anhand eines isolierten Belegs feststellbar, ob es sich um den Typ *sich befinden auf* oder *bestehen auf* handelt. Der hier vorgestellte Ansatz erlaubt es also, auf der Basis vorhandener Rahmen auf abstraktere Eigenschaften zu schließen.

4. Kollabierung von mehreren Rahmen

Neben einer Erweiterung der Menge der Subkategorisierungsrahmen kann es auch sinnvoll sein, mehrere Rahmen zu kollabieren. Ein relevanter Fall ist z.B. das Auftreten des Satzkorrelats *es*. In vielen Verblexika – automatisch und manuell erstellten – findet sich Information dazu, ob ein Satzkorrelat üblich, möglich oder nicht möglich ist: **(es) unterbinden, daß*, *(es) wissen, daß*, *(*es) herausfinden, daß*. *Herausfinden* subkategorisiert dem-

nach nur einen daß-Satz, *unterbinden* einen daß-Satz mit Korrelat, und *wissen* subkategorisiert alternativ einen daß-Satz mit oder ohne Korrelat (was verschiedenen Subkategorisierungsrahmen entspricht). Wie BERMAN et al. (1998) zeigen, ist das Vorkommen des Satzkorrelats *es* jedoch weitgehend von pragmatischen Faktoren, d.h. von Diskurseigenschaften, bestimmt. Das Satzkorrelat ist demnach nicht Teil des Subkategorisierungsrahmens, sondern prinzipiell bei satzförmigen Objekten möglich.

Somit kollabieren die zwei genannten Rahmen bei *wissen* zu einem abstrakteren. Verben, denen nur einer dieser Rahmen zugeordnet wird, erhalten ebenfalls den abstrakteren Rahmen; zusätzlich werden hier die besonderen Implikaturen, die mit einem markierten Gebrauch verbunden sind, in einem diskursrelevanten Merkmal repräsentiert und stehen somit für die semantische Analyse und die Generierung zur Verfügung.

5. Präferenzbildung

Am Beispiel der Satzkorrelate haben wir schon gezeigt, daß neben der Subkategorisierung weitere Informationstypen relevant sein können. Eine weitere Anwendung ist die Auswahl präferierter Lesarten bei gleichzeitiger Erkennung auch seltener Konstruktionen. Häufig werden rein statistische Verfahren zur Bewertung von Alternativen verwendet; ein anderer Weg ist die Adaption der Optimalitätstheorie (PRINCE & SMOLENSKY 1998; BRESNAN in Vorb.; JOHNSON in Vorb.). Hierbei handelt es sich um eine linguistische Theorie zur Auswahl aus mehreren Kandidaten bei der Generierung, die mit der Vergabe von symbolischen Präferenzmarken und deren Anordnung in einer Optimalitätshierarchie arbeitet.

Eine mögliche Präferenz ist die Bevorzugung der ditransitiven Lesart gegenüber der transitiven Lesart mit einer komplexen NP für *Er gibt das Buch der Frau*. Beide Analysen sind prinzipiell grammatisch zulässig, aber im konkreten Kontext (d.h. bei Vorliegen einer entsprechenden konkurrierenden Analyse) ist die Präferenz klar.[1]

[1] Für unser Argument ist die Richtung der Präferenz unwichtig; die Existenz einer bevorzugten Lesart ist ausreichend für das Auftreten dieses Problems.

Dieser Ansatz führt dann zu Problemen, wenn mehrere Präferenzen formuliert werden, ohne daß immer eine konkurrierende Analyse existiert.[2] Es ist z.B. nicht ausreichend, alle ditransitiven Lexikoneinträge zu bevorzugen, da in diesem Fall auch dann Präferenzmarken vergeben werden, ohne daß die konkurrierende Analyse überhaupt möglich ist (weil der betreffende Verbeintrag keinen transitiven Rahmen enthält). In diesen Fällen können sich wegen des Fehlens des Vergleichspartners unerwünschte Effekte bei der Ordnung der Alternativen ergeben, die im Effekt dazu führen, daß die gewünschte Analyse doch unterdrückt wird bzw. eine unerwünschte Alternative weiterhin als eine der besten gilt (FRANK et al. 1998).

Solche Effekte können nur unterbunden werden, indem sichergestellt wird, daß Präferenzen nur dann vergeben werden, wenn konkurrierende Analysen wirklich auftreten. In diesem Fall ist die Betrachtung mehrerer Subkategorisierungsrahmen eines Verbs nötig, um die Präferenzmarke nur dann zu vergeben, wenn sowohl ditransitive als auch transitive Lesarten vorliegen.

6. Implementation

Einige kurze Anmerkungen zur Implementation müssen an dieser Stelle genügen; weitere Informationen sind unter der URL `http://www.ims.uni-stuttgart.de/~nobi/projects/dingo/www/` zu finden.

Die Ableitung des LFG-Lexikons aus dem Subkategorisierungslexikon muß vollautomatisch möglich sein. Eine manuelle Bearbeitung ist wegen des Umfangs sowie der Notwendigkeit der wiederholten Ableitung des LFG-Lexikons (nach Erweiterungen, Verfeinerungen, Löschungen im Subkategorisierungslexikon) unakzeptabel. Wegen der regelmäßigen Verbesserung des Subkategorisierungslexikons muß man davon ausgehen, daß weitere Subkategorisierungsrahmen hinzukommen oder bestehende verfeinert werden. Die Abbildung muß also modifizierbar sein. Aus die-

[2] Dies ist ein wesentlicher Unterschied zur Anwendung der Optimalitätstheorie in der Generierung, wo die Existenz (mindestens) einer Alternative immer gesichert ist.

sem Grund und wegen der Komplexität der Abbildung ist ein dem Linguisten plausibles logisches Modell der Abbildung bereitzustellen.

Wir verwenden ein (am Institut entwickeltes) Werkzeug, das manuell erstellte Entscheidungsbäume für die Beschreibung der Abbildungsfunktion nutzt. Ein Knoten im Graph stellt eine Wahlmöglichkeit dar; von diesem Knoten ausgehende Kanten zu weiteren Knoten realisieren eine bestimmte Auswahl. Die Kanten sind mit Bedingungen und Aktionen annotiert, die die Überquerbarkeit der Kante beschränken und dabei durchzuführende Veränderungen beschreiben. Die Aktionen, die auf einem Pfad vom Start- bis zu einem Endknoten gefunden werden, beschreiben die Veränderungen, die am jeweiligen Eingabeobjekt vorzunehmen sind.[3]

Für die Konstruktion des LFG-Lexikons (vgl. Abbildung 1) werden aus dem Subkategorisierungslexikon alle Subkategorisierungsrahmen zu einem Lemma entnommen und zu einer Liste zusammengefaßt. Das Lemma und diese Liste werden dem Konversionsverfahren unterworfen. Die Ergebnisse dieser Konversion werden als Einträge dem LFG-Lexikon hin-

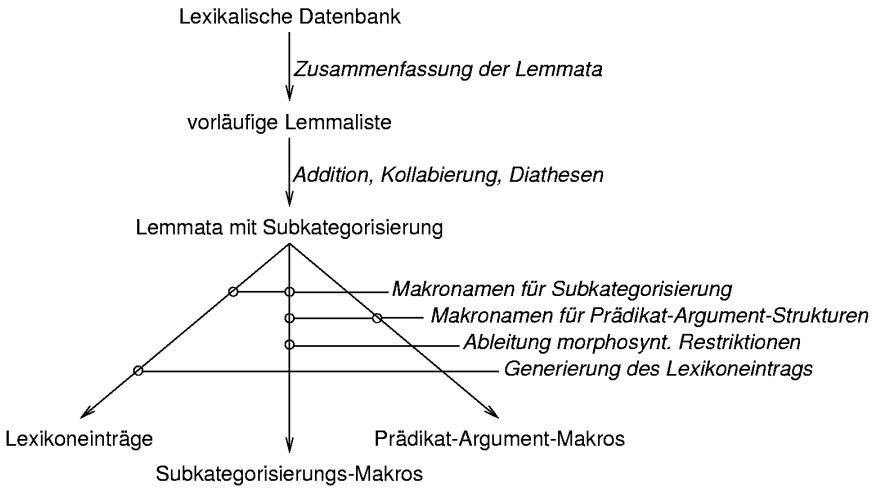

Abbildung 1: Ablaufdiagramm

[3] Weitere Anwendungen dieses Werkzeugs finden sich z.B. in der Akquisition von morphologischen Eigenschaften unbekannter Wörter (ähnlich dem Verfahren in LEZIUS 1996).

zugefügt. Dabei werden nicht nur einzelne Lexikoneinträge konstruiert: Die so erzeugten Lexikoneinträge enthalten Generalisierungen durch Makroaufrufe. Diese Makros beschreiben die Subkategorisierung in zwei Ebenen und werden ebenfalls automatisch definiert (vgl. das Beispiel in Abbildung 2[4]). Eine Ebene ist sprachspezifisch und enthält morphosyntaktische Restriktionen (im Falle des Deutschen unter anderem die Kasuszuweisung). Die zweite Ebene kodiert die sprachunabhängige Prädikat-Argument-Struktur. Die Aufrufhierarchie der Makros ist in Abbildung 3 dargestellt.

Ausgangsformat:

```
geben        (subj(NP_nom),obj(NP_acc),iobj(NP_dat))
```

Zielformat:

```
geben        {@(NPnom-NPacc-NPdat geben)
             | @(null-NPnom-NPdat-PASSIVE geben) }
```

Subkategorisierungsmakros:

```
NPnom-NPacc-NPdat(_stem)      =      @(SUBJ-OBJ-OBJ2 _stem)
                                     @(ASSIGN-CASE OBJ acc)
                                     @(ASSIGN-CASE OBJ2 dat).
null-NPnom-NPdat-PASSIVE(_stem) = @(NULL-SUBJ-OBJ2 _stem)
                                     @(ASSIGN-CASE OBJ2 dat)
                                     @(PASSIVE).
```

Prädikat-Argument-Makros:

```
SUBJ-OBJ-OBJ2(_stem)          =      PRED='_stem<SUBJ,OBJ,OBJ2>'.
NULL-SUBJ-OBJ2(_stem)         =      PRED='_stem<NULL,SUBJ,OBJ2>'.
```

Abbildung 2: Generierte Lexikoneinträge und Makrodefinitionen

7. Relation zu Lexikalischen Regeln

Die traditionelle Auffassung von Lexikalischen Regeln ist, daß sie aus bestehenden Lexikoneinträgen neue generieren. Lexikalische Regeln sind 'kontextfrei' in dem Sinne, daß nur eine Subkategorisierungsalternative betrachtet wird, die entweder umgeschrieben oder in Alternativen aufgespalten wird. Das klassische Beispiel ist die Passivierung, die aus einem transitiven Subkategorisierungsrahmen (`subj(NPnom), obj(NPacc)`) die Existenz eines weiteren (`subj(NPnom)`) ableitet.

---

[4] Die Notation ist an XLE angelehnt; formale Parameter werden durch einen Unterstrich _ gekennzeichnet, @(...) stellt einen Makroaufruf dar.

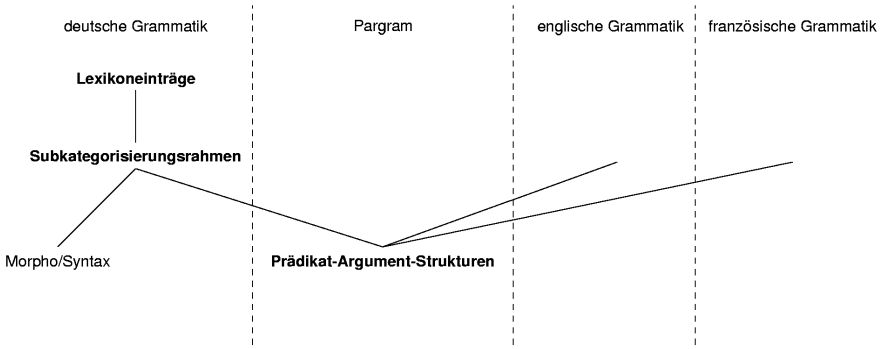

Abbildung 3: Aufrufhierarchie der Makros

Die hier beschriebenen Regeln sind genereller als Lexikalische Regeln, wie sie von LFG oder HPSG postuliert werden. Sie erweitern sowohl (praktisch) die Anwendungen (z.B. auf Präferenzbestimmung) als auch (formal) die Domäne (Mengen von Subkategorisierungsrahmen anstelle einzelner Rahmen). Sie entsprechen daher eher Redundanzbeziehungen, die im endgültigen Lexikon zwar bestehen, aber nicht produktiv abgeleitet werden.

8. Zusammenfassung

Bei der Konstruktion eines computerlinguistischen Lexikons ist es oft notwendig, nicht nur lokal einen Subkategorisierungsrahmen eines Lemmas zu betrachten, sondern mehrere Subkategorisierungsrahmen des Lemmas in Betracht zu ziehen, um die Konsistenz und Präzision des Lexikons sicherzustellen. Wir haben verschiedene Beispiele für die Anwendung dieses Modells gegeben.

Aufgrund der technischen Gegebenheiten zielt unsere Implementation auf die Generierung aller Subkategorisierungsalternativen. Der Prozeß läßt sich aber in zwei Schritte aufgliedern; erstens die Bestimmung der Veränderungen am Lexikoneintrag (Addition/Löschung von Subkategorisierungsrahmen, Präferenzmarken, etc.) und zweitens deren Repräsentation. Lexika, die mehr Struktur aufweisen (z.B. durch Vererbungsbeziehungen), können ebenfalls auf diese Weise konstruiert werden.

Weitere Arbeiten müssen in zwei Richtungen zielen: Zum ersten sind weitere Regeln empirisch zu validieren und in die Lexikonkonstruktion zu integrieren. Beispiele hierfür sind das Passiv mit *bekommen*, produktive Verbpräfigierung, Diathesen u.a. Längerfristig müßte dabei auch untersucht werden, inwieweit sich Alternationen von Subkategorisierungsrahmen auf abstraktere Eigenschaften zurückführen lassen. Idealerweise ließen sich alle Rahmen eines Verbs aus einem Bündel von solchen Eigenschaften ableiten, sodaß nicht mehr eine Liste von Subkategorisierungsrahmen Ausgangspunkt der Lexikon-Erzeugung ist (wie momentan), sondern eine abstraktere (semantische) Beschreibung der Lemmata (LEVIN 1993). Der softwaretechnische Vorteil des hier vorgestellten Ansatzes ist, daß diese Eigenschaften sukzessive eingeführt werden können, sobald (Teil-)Erkenntnisse über begrenzte Domänen bekannt sind. Sie tragen sofort zur schrittweisen Verbesserung des Lexikons bei.

Andererseits ist auf der Modellierungsseite zu fragen, ob die Subkategorisierungsrahmen als Menge ausreichend strukturiert sind. Es ist z.B. denkbar, daß für die Präferenzbestimmung die vollständige Derivationsgeschichte der Subkategorisierungsrahmen wichtig ist. Momentan ist diese Information nicht repräsentiert, sodaß Präferenzen nur globaler Art sein können oder nur über einen einzelnen Derivationsschritt formuliert werden können.

Literatur

BERMAN et al. (1998): J. B., S. DIPPER, C. FORTMANN, J. KUHN, Argument Clauses and Correlative *es* in German: Deriving Discourse Properties in a Unification Analysis. In: *Proceedings of the LFG98 Conference*, Brisbane, Australia. `http://www-csli.stanford.edu/publications/LFG3/lfg98-toc.html` .

BRESNAN, J. (in Vorb.): Optimal Syntax. In: J. DEKKERS, F. VAN DER LEEUW and J. VAN DE WEIJER (eds.), Optimality Theory: Phonology, Syntax and Acquisition. Oxford University Press.

ECKLE, J. (1999): Linguistisches Wissen zur Lexikonakquisition aus deutschen Textcorpora. Dissertation Universität Stuttgart.

ESTIVAL et al. (1995): D. E., K. FALKEDAL, S. LEHMANN, H. COMPAGNION, L. BALKAN, D. ARNOLD, F. FOUVRY, J. KLEIN, J. BAUR, K. NETTER, S. OEPEN, S. REGNIER-

PROST, E. DAUPHIN, V. LUX, The Construction of Test Material. TSNLP Report (WP 3.1). http://cl-www.dfki.uni-sb.de/tsnlp/publications.html#wp3.1 .

FRANK et al. (1998): A. F., T.H. KING, J. KUHN, J. MAXWELL, Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In: *Proceedings of the LFG98 Conference*, Brisbane, Australia. http://www-csli.stanford.edu/publications/LFG3/lfg98-toc.html .

JOHNSON, M. (in Vorb.): Optimality-theoretic Lexical Functional Grammar. http://www.cog.brown.edu/~mj/papers/cuny98.ps.gz .

LEVIN, B. (1993): English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press.

LEZIUS, W. (1996): Morphologiesystem Morphy. In: R. HAUSSER (ed.), Linguistische Verifikation. Dokumentation zur ersten Morpholympics 1994. Tübingen: Niemeyer. http://www-psycho.uni-paderborn.de/lezius/paper/molympic.ps .

PRINCE, A., SMOLENSKY, P. (1998): Optimality Theory. The MIT Press.