JAMES KILBURY

Inkrementelle Identifikation von Sprachvarietäten

Einleitung: ein Gedankenspiel und ein Problem

Man stelle sich vor, welche Auskünfte die Reisenden an einem großen internationalen Flughafen wie Heathrow in London erhalten möchten.[1] Sie fragen nicht nur nach Reiseverbindungen, Abflugzeiten usw., sondern auch nach Dienstleistungen, Wechselstuben, Postschaltern und Wickelräumen. Personal, das die Fragen beantworten kann, ist teuer und muss ohnehin einen vernetzten Computer zur Verfügung haben. Ist es nicht nahe liegend, den menschlichen Vermittler einzusparen und dem Reisenden den direkten Zugang zum Computer zu geben?

Leider setzt die übliche Abfrage von Daten aus einer Datenbank voraus, dass der Benutzer mit einer speziellen, Programmiersprachen ähnlichen Anfragesprache wie z. B. SQL vertraut ist, in der Fragen an die Datenbank formuliert werden müssen. Gewöhnlich verfügen Flugreisende aber nicht über solch spezielle Kenntnisse. Das Computersystem könnte so programmiert werden, dass der Reisende Fragen in normalem geschriebenem Englisch oder Deutsch über eine Tastatur eintippt. Doch selbst diese Annäherung des Computers an die Bedürfnisse des Benutzers stellt eine Hürde dar, die man letztlich überwinden möchte. Ideal wäre es, wenn die Fragen direkt in ein Mikrofon gesprochen werden könnten, um vom Computersystem zuerst phonetisch identifiziert und anschließend in formelartige Anfragen an die Datenbank übersetzt zu werden. Die erste Aufgabe wird von der *automatischen Spracherkennung* geleistet, die z. B. auch in automatischen Diktiersystemen eingesetzt wird, während spezielle Programme für die inhaltliche Analyse von Sätzen und längeren Texten den zweiten, übersetzungsähnlichen Schritt erledigen.[2]

Technologien für die inhaltliche Analyse geschriebener Texte sind inzwischen hoch entwickelt, und auch die automatische Spracherkennung kann gute Ergebnisse für bestimmte Anwendungen liefern. In der Regel aber müssen Spracherkennungssysteme auf die Aussprache jedes einzelnen Benutzers trainiert oder „eingestimmt“ werden, bevor sie adäquat arbeiten können. Zu den völlig individuellen Eigenschaften der Stimme des einzelnen Sprechers kommen darüber hinaus die charakteristischen Merkmale seiner *Sprachvarietät* hinzu, z. B., ob er eine britische, amerikanische oder vielleicht eine australische Form des Englischen spricht.

Betrachten wir das Szenario für den Einsatz eines Informationssystems mit automatischer Spracherkennung am Flughafen Heathrow genauer. Das System wäre mit einem breiten Spektrum von Sprachvarietäten (bzw. Dialekten) aus allen Teilen der englischsprachigen Welt konfrontiert. Die Unterschiede zwischen diesen Varietäten betreffen alle linguistischen Ebenen, also nicht nur Aussprache, sondern z. B. auch Wortwahl und sogar

[1] Dieser Aufsatz ist aus einem Vortrag entstanden, den ich im Mai 2003 am University College Dublin und an der Oxford University gehalten habe.

[2] Vgl. Carstensen *et al.* (2001) für eine Einführung in solche computerbasierten Sprachtechnologien.

Satzbau. Manche Unterschiede sind wichtig, um eine Äußerung korrekt zu interpretieren, und können deswegen von den Entwicklern des Informationssystems nicht ignoriert werden. Im gegebenen Szenario gibt es jedoch keine Möglichkeit, das Spracherkennungssystem an den individuellen Sprecher anzupassen, sei es durch Training (etwa vorheriges Vorlesen von Testsätzen) oder durch gezielte Fragen des Systems („Wo kommen Sie her?“ oder „Welchen Schulabschluss haben Sie?“).

Jede Anpassung an den Sprecher müsste also vom System selbst ausgehen, was gegenwärtige Systeme nicht leisten. Dass eine solche Anpassung im Prinzip wünschenswert ist, hat mehrere Gründe: Wie bereits erwähnt, sind manche Unterschiede zwischen Sprachvarietäten wesentlich für eine korrekte Interpretation von Äußerungen. Auf jeden Fall ist die Aufgabe für das System einfacher, wenn die Varietät bereits bekannt ist, da der so genannte *Suchraum*, d. h. die Menge der möglichen Analysen, damit stark reduziert und die Verarbeitung effizienter gemacht werden kann. Es ist sogar denkbar, dass ein System Antworten in natürlicher Sprache generiert, die an die Varietät des individuellen Benutzers angepasst sind.

Wie ließe sich ein System konstruieren, das die Sprachvarietät eines Benutzers *automatisch* identifiziert? Um diese Frage beantworten zu können, müsste man die Varietäten einer Sprache zuerst *formal modellieren* können, um so ihre Beziehungen untereinander explizit zu erfassen. Erst nach dieser Modellierung könnte ein Algorithmus – also ein systematisches Verfahren – entwickelt werden, mit dem eine Varietät einem konkreten Sprecher zuzuweisen wäre. Vorschläge für eine Lösung sind bereits 1986 von mir gemacht worden,[3] aber verschiedene Fortschritte der letzten Jahre rechtfertigen den jetzigen Aufsatz mit neueren Überlegungen.

Ausgangspunkte für eine Lösung

Aus methodologischen Gründen ist es sinnvoll, eine Reihe von Annahmen über die Identifikation von Sprachvarietäten zu machen. Die vielleicht wichtigste Voraussetzung ist, dass Varietäten sich überhaupt durch *diskrete* Eigenschaften sinnvoll charakterisieren lassen. Sicherlich können wir die Verwendung von Sprache auch *quantitativ* beschreiben und z. B. festhalten, dass ein Sprecher eine gegebene Aussprache oder ein Wort mit einer bestimmten Häufigkeit benutzt; im Folgenden werde ich jedoch nur solche Eigenschaften berücksichtigen, die *absolut* gelten oder nicht, wie z. B., ob ein Muttersprachler des Deutschen *Teich* und *Teig* gleich oder verschieden ausspricht. Neue Eigenschaften können fortlaufend hinzukommen, solange der Benutzer weiterspricht. Wenn aber Eigenschaften einmal festgestellt worden sind, dann wird angenommen, dass sie fortan gelten; d. h., diese *inkrementelle* Ansammlung von Informationen über die Sprachvarietät ist *monoton* im mathematischen Sinne. Aus gewonnenen Informationen können weitere geschlossen werden; einzelne Eigenschaften sind aber mehr oder weniger informativ, indem sie stark oder nur wenig die verbleibende Menge von potenziellen Sprachvarietäten des Sprechers einschränken.

Neben diesen sehr allgemeinen Voraussetzungen gelten weitere Annahmen speziell über die Variation in natürlichen Sprachen. Obwohl englische und deutsche Beispiele verwendet werden, soll die Vorgehensweise als allgemein und auf beliebige Sprachen übertragbar

[3] Vgl. Kilbury (1986).

verstanden werden. Keine Unterscheidung wird hier zwischen *Varietät* und *Dialekt* gemacht; vielmehr sollen verschiedene Arten von Variation – u. a. geographische, soziale und stilistische – gleich behandelt werden. Wir wissen, dass zwei Menschen nie identisch sprechen und dass sie selbst im eigenen Sprechen viel Variation zeigen, aber als deskriptive Fiktion kann man homogene Varietäten voraussetzen, denen signifikante Gruppen von einzelnen Sprechern, so genannte *speech communities*, zugeordnet werden können. Diese Varietäten können dann durch endliche Bündel von Eigenschaften sinnvoll charakterisiert werden, die die Basis für eine Klassifikation bilden.

Variation ist auf allen deskriptiven Ebenen der Sprache zu beobachten. In der Aussprache können Unterschiede *phonetisch* (z.Ḃ. verschiedene *r*-Laute im englischen Wort *red*) oder *phonemisch* sein. Zu Letzteren gehören die Vokale von *pot* bzw. *part*, die als [ɒ] bzw. [ɑː] im britischen oder als [ɑː] bzw. [ɑːr] im amerikanischen Englisch erscheinen können; *putt* kann den Vokal [ʌ] oder – im irischen Englisch – [ʊ], den Vokal von *put*, enthalten. In diesen Fällen sind Verwechslungen bei Hörern möglich, die eine andere Varietät des Englischen sprechen. Unterschiede können auch die Verbindung an Wortgrenzen betreffen, so dass manche Sprecher einen *Linking-r*-Laut in *far away*, aber nicht in *far from* aussprechen, während andere den *r*-Laut immer aussprechen oder weglassen.

Lexikalische Unterschiede dagegen betreffen nicht die Aussprache im Allgemeinen, sondern Aspekte einzelner Wörter und anderer lexikalischer Einheiten und können phonemischer, morphologischer, syntaktischer oder semantischer Art sein. So kann *either* mit dem Vokal [aɪ] oder [iː] ausgesprochen werden. Das Verb *light* ‚anzünden' hat je nach Varietät die Vergangenheitsform *lighted* oder *lit*, während Fragen mit der Konstruktion *have you (got) something* oder *do you have something* gebildet werden können. Semantische Unterschiede bestimmen, ob man den Kofferraum bzw. die Motorhaube des Autos mit *boot* bzw. *bonnet* (im britischen Englisch) oder mit *trunk* bzw. *hood* (im amerikanischen Englisch) bezeichnet.

Die *explizite* Beschreibung sprachlicher Variation stellt eine große Herausforderung an die Linguistik dar. Einer der wenigen Ansätze, die die Voraussetzungen für die formale Erfassung im Rahmen eines computerlinguistischen Modells erfüllt, stammt von Ch.-J. N. Bailey, der beobachtet hat, dass „a great deal of linguistic variation patterns in an implicational manner"[4]. Ein Beispiel dafür liefert die Aussprache von geschriebenem *-du-* mit [dj] oder [dʒ] im Englischen: im britischen Englisch ist die Aussprache mit [dʒ] am stärksten in dem Wort *gradual* verbreitet, weniger in *individual* und am wenigsten in seltenen Wörtern wie *residual.* Wer jedoch [dʒ] in *duty* ausspricht, wird es auch in den anderen Wörtern verwenden. Diese Beziehungen lassen sich formal in einer *implikativen Skala* ausdrucken:

gradual ← *individual* ← *residual* . . . ← *duty*

Seit Bailey haben formale Aussagen in der Form von Implikationen eine wichtige Rolle auch in anderen Bereichen der Linguistik und Computerlinguistik übernommen. In der *Generalized Phrase-Structure Grammar*[5] werden sie benutzt, um Verallgemeinerungen über syntaktische Kategorien (d. h. Klassen von Wörtern und Phrasen) auszudrücken. Demzufolge enthält eine Beschreibung der englischen Syntax z. B. die folgenden *feature co-occurrence restrictions*:

[4] Bailey (1973: 28).
[5] Vgl. Gazdar *et al.* (1985).

[AGR] → [-N, +V]
‚Wenn eine Kategorie für Kongruenz[6] spezifiziert ist, dann ist sie ein Verb.'

[+INV] → [+AUX, FIN]
‚Wenn eine Kategorie invertiert[7] ist, dann ist sie ein finites Auxiliar[verb].'

Spätere computerlinguistische Arbeiten von Barg und Kilbury[8] setzen Implikationen indirekt ein, um Schlussfolgerungen über „neue" Wörter (d. h. bisher nicht belegte oder in einem Lexikon nicht enthaltene Wörter) in einem sprachverarbeitenden System zu erfassen. Vielen Deutschen ist z. B. das deutsche Wort *Sund* ‚Meerenge' und seine Flexion ungeläufig. Einzelne flektierte Formen wie *Sundes* geben partielle Informationen über den Flexionstyp, aber erst die Pluralform *Sunde* neben der Singularform *Sund* lässt eine eindeutige Identifikation zu. Die Auseinandersetzung mit diesem Problem schuf einen wesentlichen Teil der formalen und methodologischen Grundlagen für die in diesem Aufsatz beschriebene inkrementelle Identifikation von Sprachvarietäten. Ungelöst blieb aber, wie große Datenmengen systematisch verarbeitet werden können, um daraus gültige linguistische Verallgemeinerungen in der Form von Implikationen zu gewinnen. Diese Voraussetzung konnte erst durch die Ergänzung der bisherigen Methoden um die im Folgenden beschriebene *Formale Begriffsanalyse* erfüllt werden.

Formale Begriffsanalyse

Die von Ganter und Wille entwickelte Formale Begriffsanalyse[9] (FBA) ist eine mathematische Theorie, die die explizite Beschreibung von Individuen durch automatisch induzierte Klassifikationen oder logische Implikationen ermöglicht. Ausgangspunkt ist der *formale Kontext*, der intuitiv gesehen durch eine endliche Menge von *Gegenständen* (bzw. *Objekten* oder Individuen) gebildet wird, die jeweils durch eine endliche Anzahl von eindeutigen *Merkmalen* oder Eigenschaften charakterisiert sind. Aus mathematischer Sicht besteht ein Kontext aus einer Menge von Gegenständen, einer Menge von Merkmalen und einer binären Inzidenzrelation zwischen diesen Mengen. Dieser Begriff des formalen Kontextes ist extrem allgemein und mächtig, so dass er geeignet zu sein scheint, große Teile unseres Wissens über die Welt zu modellieren. In einem linguistischen Beispiel lassen sich die Vokale eines phonologischen Systems als Objekte mit phonologischen Eigenschaften als Merkmalen modellieren:

	mid	front	round
i		x	
e	x	x	
y		x	x
œ	x	x	x
u			x
o	x		x
a			

[6] ‚Kongruenz' (engl. *agreement*) bezieht sich z. B. auf die Übereinstimmung zwischen Subjekt und Verb in *John sleeps* gegenüber *people sleep*.

[7] ‚Inversion' bezieht sich auf die Wortstellung z. B. des Hilfsverbs in *Can John read?* gegenüber *John can read*.

[8] Vgl. Barg und Kilbury (2000).

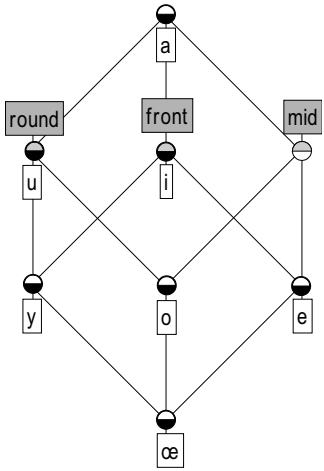
[9] Vgl. Ganter und Wille (1996) sowie Petersen (2004); die dazugehörige Software kann unter der URL http://sourceforge.net/projects/conexp (22.07.2004) heruntergeladen werden. Ich danke Wiebke Petersen, deren Dissertation die FBA behandelt, für ihre Hilfe bei der Vorbereitung dieses Abschnitts.

Ein *formaler Begriff* eines Kontextes *K* ist ein Paar (*A, B*) bestehend aus einer Menge von Objekten *A*,dem *Begriffsumfang* oder der *Extension*, und einer Menge von Merkmalen *B*, dem *Begriffsinhalt* oder der *Intension*. In einem formalen Begriff (*A, B*) muss gelten, dass sein Inhalt gerade aus den Merkmalen besteht, die auf alle Gegenstände des Umfangs zutreffen, und dass der Umfang alle Gegenstände umfasst, die die Merkmale des Inhalts gemeinsam haben. In unserem Beispiel hätte der Begriff ‚mittlerer vorderer Vokal' den Umfang {e, œ} und den Inhalt {mid, front}.

Ein *formaler Begriffsverband*[10] ist die Menge aller formalen Begriffe eines Kontextes, die durch die Relation $\leq$ zwischen Unter- und Oberbegriffen *partiell geordnet* werden. Ein Unterbegriff umfasst demnach weniger Gegenstände als ein Oberbegriff und wird durch mehr Merkmale charakterisiert:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$$

Der Verband selbst kann als eine *monotone multiple Vererbungshierarchie* aufgefasst werden. Für das linguistische Beispiel ergibt sich folgender formaler Begriffsverband:



Aus dieser Hierarchie lässt sich direkt ablesen, dass der Vokal [œ] die Merkmale *round* und *mid* von [o] erbt sowie das Merkmal *front* von [y] und [e]. Außerdem lässt sich an der Hierarchie erkennen, auf welche Gegenstände ein Merkmal zutrifft: *round* trifft auf [u], [o], [y] und [œ] zu.

Neben dem Begriffsverband definiert man die *Menge der gültigen Implikationen* eines formalen Kontextes *K* als die Menge aller Implikationen der Form *M*→*N* (mit *M* und *N* als Mengen von Merkmalen in *K*), die mit dem Kontext verträglich sind. Der Begriffsverband und die Implikationsmenge sind vollkommen äquivalent und enthalten genau die Informationen des Kontextes, aus dem sie automatisch induziert werden.

Die Methoden der FBA lassen sich bei der Modellierung von Varietäten des Englischen[11] direkt einsetzen. Als Beispiel betrachten wir den folgenden formalen Kontext, der

[10] Vgl. Maurer (1987) oder ähnliche Werke für die Erläuterung mathematischer Begriffe wie „Verband".
[11] Vgl. Trudgill und Hannah (2002) für eine deskriptive Zusammenfassung der Variation im Englischen.

Informationen über die Aussprache der Wörter *new* und *due*, die Bedeutung von *boot* als ‚Stiefel' oder ‚Kofferraum' sowie über die Vergangenheitsform des Verbs *light* in Varietäten *A1, . . . , B3* (den Gegenständen des Kontextes) erfasst:

	new	new	due	due	due	boot	boot	lit	lighted
	[nu:]	[nju:]	[du:]	[dju:]	[dʒu:]	(foot)	(car)	past	past
A1	x		x			x		x	?
A2		x		x		x		x	?
B1		x		x		x	x	x	?
B2		x		x	x	x	x	?	?
B3		x			x	x	x	?	?

Diese Daten zeigen z. B., dass alle Sprecher mit der Aussprache [dʒu:] das Wort *boot* auch in der Bedeutung ‚Kofferraum' benutzen und dass niemand das Wort so verwendet, der die Aussprache [nu:] hat. Die Inferenzen ergeben sich direkt aus der induzierten Menge der Implikationen oder durch die partielle Ordnung im Begriffsverband. Attraktiv an der Sichtweise des Begriffsverbandes ist deren leichte Integration in moderne Formalismen der Computerlinguistik, die als Nächstes hier erörtert werden.

Head-Driven Phrase-Structure Grammar

Die *Head-Driven Phrase-Structure Grammar*[12] (HPSG) ist ein moderner Grammatikformalismus, der in der Linguistik und Computerlinguistik starke Verbreitung gefunden hat. Auf Grund ihrer Homogenität wird sie dazu verwendet, Informationen auf allen linguistischen Ebenen zu kodieren. Sie kann auch als Basis praktischer Systeme, z. B. in dem am Anfang dieses Aufsatzes beschriebenen Szenario,[13] eingesetzt werden.

HPSG ist aus früheren Formalismen, insbesondere aus GPSG und PATR-II,[14] entstanden, in denen *ungetypte Merkmalsstrukturen* als grundlegende Datenstruktur für die Repräsentation sprachlicher Informationen dienen:

$$\begin{bmatrix} \text{category} : \text{noun} \\ \text{agreement} : \begin{bmatrix} \text{case} : \text{accusative} \\ \text{number} : \text{singular} \\ \text{gender} : \text{feminine} \end{bmatrix} \end{bmatrix}$$

In der obigen Merkmalsstruktur erscheinen category, agreement, case usw. als *Merkmale*, denen ein *Wert* zugewiesen wird; ein Merkmal zusammen mit seinem Wert bildet eine *Spezifikation*. Werte sind entweder *atomar*, wie der Wert noun des Merkmals category, oder *komplex*. Der Wert von agreement ist nicht atomar, sondern selbst eine Merkmalsstruktur und somit komplex. Verschiedene Merkmalsstrukturen können durch die partielle Ordnungsrelation der *Subsumption* in Beziehung zueinander gesetzt werden, was dem Begriffsverband der FBA direkt entspricht. Durch die monotone Operation der

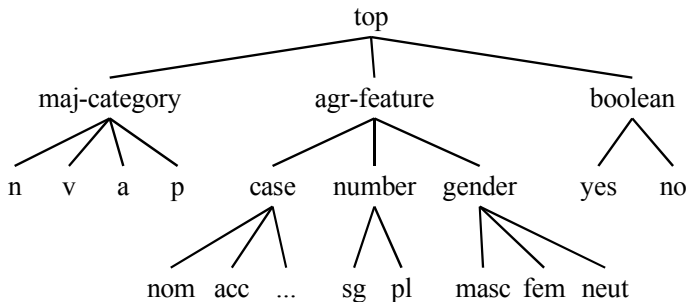
[12] Vgl. Pollard und Sag (1987) und (1994) sowie, für die formalen Grundlagen, Carpenter (1992).
[13] Vgl. Wahlster (2000).
[14] Vgl. Carstensen *et al.* (2001) und Shieber (1986) für einführende Darstellungen.

Unifikation kann aus zwei Merkmalsstrukturen eine neue gebildet werden, die die Informationen aus beiden enthält; vorausgesetzt wird, dass die unifizierten Merkmalsstrukturen in ihren Spezifikationen kompatibel sind und dass die Unifikation damit *gelingen* kann.

HPSG erweitert die ungetypten Merkmalsstrukturen um *Typen* wie *noun* und *agr* sowie um die atomaren Werte in der folgenden Merkmalsstruktur:

$$\begin{bmatrix} \textit{noun} & \\ \text{agreement} : & \begin{bmatrix} \textit{agr} \\ \text{case} : \textit{accusative} \\ \text{number} : \textit{singular} \\ \text{gender} : \textit{feminine} \end{bmatrix} \\ \text{count} : \textit{no} & \end{bmatrix}$$

Die Typen einer Beschreibung bilden eine *Typhierarchie*, die – bis auf ein fehlendes maximales Element für Inkompatibilität – wiederum einen Verband darstellt. Formalismen sind *wohlgetypt*, wenn „every feature that occurs is *appropriate* and takes an appropriate value", und *völlig wohlgetypt*, wenn sie wohlgetypt sind und „every feature which is appropriate [is] present".[15] Die folgende Typhierarchie, deren Typen durch Subsumption partiell geordnet sind, formalisiert syntaktische Informationen:



Fundamental für die HPSG ist das linguistische *Zeichen*, das mit einer getypten Merkmalsstruktur formal repräsentiert wird:

$$\begin{bmatrix} \textit{basic - lexical - sign} \\ \text{PHON} : \quad \textit{/hund /} \\ \text{MORPH} : \textit{ns} \\ \text{SYN} : \quad \textit{noun} \\ \text{SEM} : \quad \textit{'dog'} \end{bmatrix}$$

Dieses lexikalische Zeichen für das deutsche Wort *Hund* enthält Angaben über seine Aussprache (PHON), Flexion bzw. Pluralbildung (MORPH), syntaktische Kategorie (SYN)

[15] Carpenter (1992: 79).

und Bedeutung (SEM). Das Zeichen ist einfach (*basic*) und nicht abgeleitet, wie z. B. bei *Welt-an-schau-ung*.

Die Beschreibung von Varietäten in HPSG

Die herausragende Eigenschaft des Zeichens in der HPSG besteht darin, dass es die Repräsentation von linguistischen Informationen verschiedenster Art in einer homogenen Datenstruktur ermöglicht, wodurch auch Informationen über Sprachvarietäten erfasst werden können. Darüber hinaus bieten Subsumption und Vererbung in der Typhierarchie die Möglichkeit, Inferenzen über die Varietät eines Sprechers zu machen, insbesondere dann, wenn der Typ, der diese Varietät bezeichnet, durch Unifikationen während der linguistischen Analyse einer Folge von Äußerungen *verschärft* werden kann.

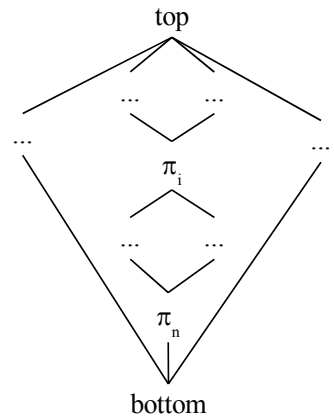
$$\begin{bmatrix} \textit{basic - lexical - sign} \\ \text{VARIETY} : \textit{br} \\ \text{PHON} : \quad \textit{/bu :t /} \\ \text{SYN} : \quad \textit{noun} \\ \text{SEM} : \quad \textit{'car compartment'} \end{bmatrix} \quad \begin{bmatrix} \textit{basic - lexical - sign} \\ \text{VARIETY} : \textit{top} \\ \text{PHON} : \quad \textit{/bu :t /} \\ \text{SYN} : \quad \textit{noun} \\ \text{SEM} : \quad \textit{'footgear'} \end{bmatrix}$$

Die hier angegebenen lexikalischen Zeichen erfassen Informationen über Variation bei der Verwendung des englischen Wortes *boot* in der Bedeutung ‚Kofferraum' oder ‚Stiefel'. In dem Teil der Typsignatur, der eine Klassifikation von Sprachvarietäten ausdrückt, subsumiert der Typ *top* alle Varietäten des Englischen; d. h., er sagt *nichts* über die spezielle Varietät eines Sprechers aus. Der Typ *br* dagegen ist informativ und subsumiert lediglich spezifischere Varietäten – hier vereinfacht – des britischen Englisch. Beide Zeichen sind somit in der Teilgrammatik eines britischen Sprechers enthalten, während die eines Amerikaners nur das zweite Zeichen hat.

Abschließend muss noch geklärt werden, wie die gesammelten Informationen über einen Sprecher bei der inkrementellen Analyse seiner Fragen zusammenkommen. Die Verarbeitung erfordert die Konstruktion einer Gesamtrepräsentation einer Frage in Form eines phrasalen Zeichens, das alle sich aus der Analyse ergebenden Informationen enthält. Verschiedene Teile einer Äußerung verraten Informationen unterschiedlicher Informativität bzw. Spezifität über die Varietät des Sprechers. Wir nehmen jedoch an, dass alle Informationen über die Varietät miteinander *kompatibel* sind[16] und damit durch Unifikation kombiniert werden können. Am Ende einer Analyse liegt die präziseste Hypothese über die Varietät vor, die die Äußerungen des Sprechers erlaubt. Während der Analyse können die gewonnenen Informationen aber bereits einbezogen und genutzt werden. Diese Modellierung kann in einer Konvention formalisiert werden, wonach der Varietätstyp, d. h. der als Wert des Merkmals *VARIETY* erscheinende Typ eines komplexen Zeichens, mit dem entsprechenden Typ von jedem Zeichen unifiziert wird, das eine *Konstituente* des ersten bildet. Bei der Analyse folgender Teile der Äußerung setzt sich der Prozess fort. Damit ergibt sich die inkrementelle *Verschärfung* von Hypothesen über die Varietät, die modelliert werden soll.

[16] Für die Anwendung in einem praktischen System müsste man diese Annahme wohl aufweichen, aber die dafür erforderlichen – wohl statistischen – Methoden können hier nicht besprochen werden.

Die inkrementelle Identifikation einer Sprachvarietät stellt eine monotone Folge sukzessiver Approximationen $\pi_0, \ldots, \pi_i, \ldots, \pi_n$ mit *top* = π_0 im folgenden Verband dar:



Dabei erfasst das *Ideal* $I(\pi)$ (d. h. die Menge aller Typen *x* derart, dass $x \subseteq \pi$) alle Varietätseigenschaften, die in der bisher verarbeiteten Eingabe belegt oder daraus inferiert worden sind. Der *Filter* $F(\pi)$ (d. h. die Menge aller Typen *x* bis auf *bottom* für Widerspruch derart, dass $\pi \subseteq x$) erfasst alle noch offenen Eigenschaften möglicher spezifischerer Varietäten. Die Varietätsgrammatik $G(\pi)$ ist dann die Teilmenge von Zeichen aus der Gesamtgrammatik *G*, die mit einem der Varietätstypen in $I(\pi) \cup (F(\pi) \backslash bot)$ spezifiziert sind.

Es bleibt nun, daran zu erinnern, wie der sehr komplexe Teil der Typsignatur konstruiert werden kann, der eine Klassifikation der Sprachvarietäten bildet und alle Varietätstypen enthält. Die *manuelle* Erstellung solcher Klassifikationen aus den Daten über Variation übersteigt die Fähigkeiten und Ressourcen von Linguisten. Oben haben wir jedoch bereits gesehen, dass die elementaren empirischen Daten über Eigenschaften verschiedener Sprachvarietäten in einem formalen Kontext aufgelistet werden können. Daraus kann wiederum ein Begriffsverband mit der für die FBA implementierten Software *automatisch* erstellt werden, und dieser Verband kann dann direkt in die HPSG-basierte Beschreibung einer Sprache integriert werden. Damit wird ein wesentlicher Schritt für die von Bailey anvisierte *pan-lectal* (d. h. parallel alle Varietäten einer Sprache erfassende) *grammar*[17] geleistet.

Vorteile des Ansatzes und offene Fragen

Der hier geschilderte Ansatz bringt eine Reihe von Vorteilen mit sich:

- Er erfasst wichtige Aspekte sprachlicher Variation – insbesondere solche, die mit hierarchischen Beziehungen darstellbar sind – in einem computerlinguistischen Modell.
- Die Modellierung passt direkt in den verbreiteten Formalismus der HPSG, ohne dass dieser geändert werden muss.

[17] Vgl. Bailey (1973).

- Die Anwendungsdomäne der HPSG wird damit um die Variation und – potenziell – um den Sprachwandel erweitert.
- Die inkrementelle Identifikation von Sprachvarietäten erscheint in diesem Licht als ein Nebenprodukt des Analysevorgangs.
- Er nutzt die gewonnenen Informationen über die Sprachvarietät eines Sprechers, anstatt diese lediglich als Verrauschung der Eingabe zu behandeln.
- Zumindest potenziell bietet er die Möglichkeit praktischer Anwendung in einem System für Spracherkennung mit einer Reduktion des Suchraums und damit einem Gewinn an Effizienz der Verarbeitung.

Einige Fragen bleiben noch völlig offen:

- Es ist unklar, wie fremde Varietäten (von nicht muttersprachlichen Sprechern) und die Vermischung von Varietäten behandelt werden können.
- Viel Arbeit wäre erforderlich, um die auf die Anforderungen der HPSG zugeschnittenen phonetischen bzw. phonologischen Repräsentationen für eine Sprache detailliert auszuarbeiten.[18]
- Methodologisch muss noch geklärt werden, wie die systematische Sammlung relevanter Daten für eine solche Grammatik des Englischen zu bewältigen wäre.

Die bisher erzielten Ergebnisse lassen jedoch eine Fortsetzung dieser Untersuchungen als sinnvoll erscheinen.

Literatur

BAILEY, Charles-James N. *Variation and Linguistic Theory*. Washington, D. C., 1973.

BARG, Petra und James KILBURY. „Incremental Identification of Inflectional Types“, *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken (2000), 49-54.

BIRD, Steven und Ewan KLEIN. „Phonological Analysis in Typed Feature Systems“, *Computational Linguistics* 20 (1994), 455-491.

CARPENTER, Bob. *The Logic of Typed Feature Structures*. Cambridge 1992.

CARSTENSEN, Kai-Uwe, Christian EBERT, Cornelia ENDRISS, Susanne JEKAT, Ralf KLABUNDE und Hagen LANGER. *Computerlinguistik und Sprachtechnologie: Eine Einführung.* Heidelberg und Berlin 2001.

GANTER, Bernhard und Rudolf WILLE. *Formale Begriffsanalyse*. Berlin 1996.

GAZDAR, Gerald, Ewan KLEIN, Geoffrey PULLUM und Ivan SAG. *Generalized Phrase-Structure Grammar*. Oxford 1985.

KILBURY, James. „Language Variation, Parsing, and the Modelling of Users' Varieties“, *Proceedings of the 7th European Conference on Artificial Intelligence*. Bd. II. Brighton (1986), 29-32.

MAURER, Joseph. *Mathemecum. Begriffe, Definitionen, Sätze, Beispiele*. Braunschweig 1987.

PETERSEN, Wiebke. „A Set-theoretical Approach for the Induction of Inheritance Hierarchies“, *Electronic Notes in Theoretical Computer Science* 53 (2004), 1-13. http://www.elsevier.nl/locate/entcs/volume53.html (17.05.2004).

POLLARD, Carl und Ivan SAG. *Information-based Syntax and Semantics*. Stanford 1987.

POLLARD, Carl und Ivan SAG. *Head-driven Phrase-Structure Grammar.* Stanford 1994.

[18] Hierzu liefern Bird und Klein (1994) mit der *declarative phonology* wichtige Grundlagen.

SHIEBER, Stuart M. *An Introduction to Unification-based Approaches to Grammar.* Stanford 1986.

TRUDGILL, Peter und Jean HANNAH. *International English. A guide to the varieties of Standard English*. London 2002.

WAHLSTER, Wolfgang. *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin u. a. 2000.