Lexikalisch kontrollierte Kreuzabhängigkeiten und Kongruenzmorphologien

Uwe Mönnich
SFB 340
Universität Tübingen
Seminar für Sprachwissenschaft
Wilhelmstr. 113
D-72074 Tübingen
uwe.moennich@uni-tuebingen.de

Overview

Introduction

Mild context-sensitivity: Swiss German

(1)a. *Jan säit das mer em Hans es huus hälfed aastriiche.*

b. John said that we helped Hans (to) paint the house.

(2)a. *Jan säit das mer d'chind em Hans es huus lönd hälfed aastriiche.*

b. John said that we let the children help Hans paint the house.

The NP's and the V's of which the NP's are objects occur in cross–serial order. *D'chind* is the object of *lönd*, *em Hans* is the object of *hälfe*, and *es huus* is the object of *aastriiche*. Furthermore the verbs mark their objects for case: *hälfe* requires dative case, while *lönd* and *aastriiche* require accusative. It appears that there are no limits on the length of such constructions in grammatical sentences of Swiss German.

Shieber intersects Swiss German with the regular language:

(3)a. *Jan säit das mer (d'chind)* (em Hans)* händ wele (laa)* (hälfe)* aastriiche.*

b. John said that we (the children)* (Hans)* the house wanted to (let)* (help)* paint.

The result of the intersection looks as follows:

(4) *Jan säit das mer (d'chind)^n (em Hans)^m händ wele (laa)^n (hälfe)^m aastriiche.*

As is well known, this language is not context–free.

Informal characterization of mild context-sensitive languages (Joshi):

a) Limited crossed dependencies

b) Constant growth

c) Polynomial parsing

Formal characterization of mild context-sensitivity: Parikh-Languages

Definition 1 *Let $V = \{a_i \mid 1 \leq i \leq n\}$ be a finite vocabulary. The Parikh-mapping P_v is the function from V^* into $\mathbb{N}^n$ defined by $P_v(x) = \langle |x|_{a_1}, \ldots, |x|_{a_n} \rangle$, where $|x|_{a_i}$ is the number of occurrences of a_i in x.*

Definition 2 *$M \subseteq \mathbb{N}^n$ is said to be a* **linear** *set if there is a constant $c \in \mathbb{N}^n$ and a finite set of periods $P \subseteq \mathbb{N}^n$ such that M is the set of all elements x in $\mathbb{N}^n$ of the form*

$$x = c + \sum_{i=1}^{k} n_i x_i,$$

with x_i in P and $n_i \in \mathbb{N}^n$.

Definition 3 *A subset of $\mathbb{N}^n$ is called* **semilinear** *if it is a finite union of linear sets.*

Definition 4 *Given a language $L \subseteq V^*$ over a finite vocabulary V. It is said to be a* **Parikh-language** *if its Parikh-mapping $P_v(L)$ is a semilinear subset of $\mathbb{N}^n$ where $|V| = n$.*

Fact

All known non-context-free natural language structures give rise to Parikh-languages.

Example

$L = \{a^n b^m c^n d^m\} \subseteq$ Swiss German
$L = L(c, P)$ with $c = 0$ and
$P = \{\langle 1,0,1,0\rangle, \langle 0,1,0,1\rangle\}$

Kongruenzmorphologie

Example: Alt-Georgisch, Boeder 1995

tkuenda	*micemul*	*ars*	*cnob-ad*
zu=Euch	gegeben	ist	Wissen-Adv

saidumlo-j	*igi*	*sasupevel-isa*
Geheimnis-Nom	Art=Nom	Königreich-Gen

m-is	*gmrt-isa-isa-j*
Art-Gen	Gott-Gen-Gen-Nom

"Euch gegeben ist das Wissen um das Geheimnis des Königreiches Gottes"

Extrapolierte Struktur des Beispiels:

N-j N-isa-j ... N-isa^{n-1}-j V

Parikh-Bild der Struktur:

$$\langle n, n, n(n-1)/2, 1\rangle$$

wobei $a_1 = N$
$a_2 = j$ (Nom-Suffix)
$a_3 = isa$ (Gen-Suffix)
$a_4 = V$

$\Rightarrow$ Das Alt-Georgische ist keine Parikh-Sprache (Kracht & Michaelis, 1996)

Morphologische Grammatiken

Substitution und Paarbildung

$(1 \text{ sieht } 2) \bullet \langle \text{Maria}, \text{Peter} \rangle$: Maria sieht Peter

Definition 5 *Eine* **(Halbring-) Morphologie** *ist eine Algebra* $\mathbb{M} = (M, +, \bullet, ', \pi)$, *die folgenden Gleichungen genügt, wobei* $+$ *und* $\bullet$ *zweistellige Operationen,* $'$ *eine einstellige und* π *eine nullstellige Operation sind:*

1. $x \bullet (y \bullet z) = (x \bullet y) \bullet z$
2. $x + (y + z) = (x + y) + z$
3. $x + y = x + z$ *impliziert* $y = z$
4. $(x + y) \bullet z = (x \bullet z) + (y \bullet z)$
5. $(x \bullet y)' = x \bullet y'$
6. $\pi \bullet \pi = \pi$
7. $\pi \bullet (x + y) = \pi \bullet x$
8. $x' \bullet (\pi \bullet y + z) = x \bullet (z + \pi \bullet y)$, *und* $x' \bullet \pi = x \bullet \pi$
9. $\pi' \neq \pi$

Die Elemente von M *werden* **Ausdrücke** *genannt,* $\bullet$ *bindet stärker als* $+$.

Definition 6 *Ein Ausdruck ist* **abgeschlossen**, *wenn* $x \bullet y = x$ *für alle* y.
Der **Grad** *eines abgeschlossenen Ausdrucks ist* 0, *sonst ist der Grad entweder unendlich oder das kleinste* n, *so daß*

$$x \bullet (1 + 2 + \cdots + n) = x$$

(wobei $\pi = 1$, $\pi' = \pi^1$, $\pi^{n+1} = \pi^{n'}$*).*
Die **Dimension** *von* x, *wenn nicht unendlich, ist das kleinste* m, *so daß*

$$(1 + 2 + \cdots + m) \bullet x = x.$$

Ausdrücke der Dimension 1 *sind* **Phrasen**, *abgeschlossene Phrasen* **Formeln**.
Eine kleinste Menge von Phrasen, die eine Morphologie erzeugt, ist ein **Vokabular**, *seine Elemente werden* **Morpheme** *genannt.*
Hier werden nur **lokal-finite** *Morphologien betrachtet: für jedes* x *gibt es natürliche Zahlen* m *und* n *so daß gilt*

$$(1 + 2 + \cdots + m) \bullet x = x = x \bullet (1 + 2 + \cdots + n).$$

Definition 7 *In einer* **linearen Morphologie** *besteht* M *aus allen* n*-Tupeln von Elementen aus* $(A \cup \mathbb{N})^+$, *wobei* A *das* **Alphabet von Symbolen** *sei,* $\mathbb{N}$ *die Menge der natürlichen Zahlen,* $A \cap \mathbb{N} = \emptyset$, $(A \cup \mathbb{N})^+ = (A \cup \mathbb{N})^* - \varepsilon$, *d.h. die Menge aller nichtleeren endlichen Ketten über* $A \cup \mathbb{N}$.
n *werde die Dimension für* x *in* M *genannt.*
Für x *und* y *der Dimensionen* r *und* s *werde definiert:*

(a) $x' =$ *jedes* n *in* x *wird durch* $n{+}1$ *ersetzt.*

(b) $x + y =$ *das* $(r + s)$*-Tupel, das durch Aneinanderreihen entsteht.*

(c) $x \bullet y =$ *das r-Tupel, bei dem jedes* k *in* x *durch* $y_{\overline{k}}$ *ersetzt ist, wobei* $\overline{k} \equiv k(\text{mod } s)$.

Definition 8 Lukasiewicz Morphologien *sind solche linearen Morpologien über einer Menge A, die von einem Vokabular V erzeugt werden, dessen Morpheme von der Form (a) oder $(a1 \dots n)$ sind für $a \in A$, wobei gilt, daß*

$$a \neq b \text{ falls } (a) \text{ und } (b1 \dots n) \text{ in } V,$$

und daß

$$a \neq b \text{ falls } (a1 \dots n) \text{ und } (b1 \dots m) \text{ in } V$$

für $n, m \in \mathbb{N}$.

Definition 9 Faktorisierung von Phrasen. *Sei eine Menge V von Phrasen in einer Morphologie gegeben, dann wird die Menge der V-Faktorisierungen rekursiv definiert durch:*

(i) wenn x ein abgeschlossenes Element von V ist oder ein $n \in \mathbb{N}$, dann ist (x) eine V-Faktorisierung,

(ii) wenn x in V den Grad n hat und $F_1, \dots, F_n$ V-Faktorisierungen sind, dann ist das n-Tupel $(x, F_1, \dots, F_n)$ eine V-Faktorisierung.

Das **Produkt** $\overline{F}$ *einer V-Faktorisierung F ist rekursiv definiert durch:*

(I) wenn x ein abgeschlossenes Element von V ist oder ein $n \in \mathbb{N}$, dann ist $(\overline{x}) = x$,

(II) wenn $(x, F_1, \dots, F_n)$ eine V-Faktorisierung ist, dann $(\overline{x, F_1, \dots, F_n}) = x \bullet (\overline{F}_1, \dots, \overline{F}_n)$.

Bemerkung

Es ist leicht zu sehen, daß jede Phrase in einer Lukasiewicz Morphologie genau eine V-Faktorisierung hat, d.h. $x = \overline{F}$. Morphologien, für die das gilt, heißen **monotektonisch**. Nach einem bekannten Satz aus der Allgemeinen Algebra sind die monotektonischen Morphologien genau die freien Morphologien. Außerdem ist jede Morphologie das homomorphe Bild einer freien Morphologie und damit einer Lukasiewicz Morphologie.

Definition 10 *Sei A ein endliches Alphabet. Unter einem* **Vokabular** *V verstehen wir eine endliche Menge $V \subseteq (A \cup \mathbb{N})^*$ von Ketten über der Vereinigung von A mit der Menge der natürlichen Zahlen.*

Definition 11 *Sei Σ die Signatur, die nur die Symbole C für Substitution und K für Paarbildung enthält. Die Menge der* **Terme** $T_\Sigma(V)$ *über einem endlichen Vokabular V ist dann folgendermaßen definiert:*

(a) $V \subseteq T$
(b) Wenn t_1 und t_2 in T sind, dann auch $C(t_1, t_2)$ und $K(t_1, t_2)$.

Eine morphologische Grammatik ist dann in Analogie zu einer kontext-freien Grammatik definiert, mit dem Unterschied, dass die Rolle von Ketten über dem Terminal- und Hilfsvokabular von den eben definierten Termen übernommen wird.

Definition 12 *Eine* **morphologische Grammatik** *G ist ein Quadrupel $G = \langle N, V, P, S\rangle$, wobei N eine endliche Menge von Hilfssymbolen, V ein Vokabular, P eine endliche Menge von Produktionen $P \subseteq N \times T(V \cup N)$ und $S \in N$ das Startsymbol bezeichnen.*

Unter der von einer morphologischen Grammatik G erzeugten (Term-)Sprache $L(G)$ verstehen wir die Menge der aus S herleitbaren Terme, in denen keine Hilfssymbole vorkommen:

$$L(G) = \{t \in T(V) \mid S \Rightarrow^* t\}$$

Mit einer solchen Termsprache ist in naheliegender Weise eine Kettensprache $\eta(L)$ verbunden, die aus L entsteht, nachdem alle Substitutionen ausgeführt sind:

$$\eta(L) = \{x \in A^* \mid x = \eta(t) \,\&\, t \in L\}$$

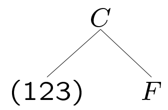
Example: L. Chance

$$\eta(L) = \{a^n b^n c^n\}$$

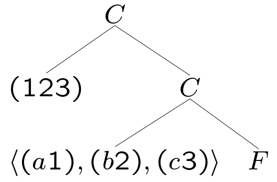
$$A = \{a, b, c\}$$

$$V = \{(123), (a), (b), (c), (a1), (b2), (c3)\}$$

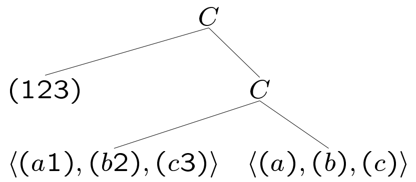
$$P = \left\{ \begin{array}{lll} S & \rightarrow & C(123)F \\ F & \rightarrow & CKK(a1)(b2)(c3)F \mid \\ & & KK(a)(b)(c) \end{array} \right\}$$



t_1



t_2



t_3

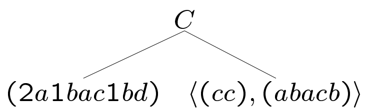
$$\eta(t_3) = a^2b^2c^2$$

Example: Alt-georgische Kongruenz, L. Chance

$$\eta(L) = \{abd, abacbd, abacbac^2bd, abacbac^2bac^3bd, \dots\}$$

$$\begin{array}{lcl} a & := & \text{Nomen} \\ b & := & \text{Nom-Suffix} \\ c & := & \text{Gen-Suffix} \\ d & := & \text{Verb} \end{array}$$

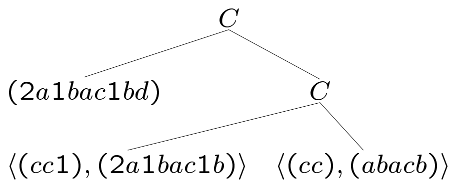
$$P = \left\{ \begin{array}{lll} S & \to & (abd) \mid \\ & & (abacbd) \mid \\ & & (abacbaccbd) \mid \\ & & C(2a1bac1bd)R \\ R & \to & CK(cc1)(2a1bac1b)R \mid \\ & & K(cc)(abacb) \end{array} \right\}$$



t_1

$$\eta(t_1) = \mathbf{abacb}a\mathbf{cc}bac\mathbf{cc}bd$$

t_3



t_2

$$\eta(t_2) = \langle (cccc), (abacbaccbacccb) \rangle$$
$$\eta(t_3) = \mathbf{abacbac^2bac^3b}ac^4bac\mathbf{c}^4bd$$

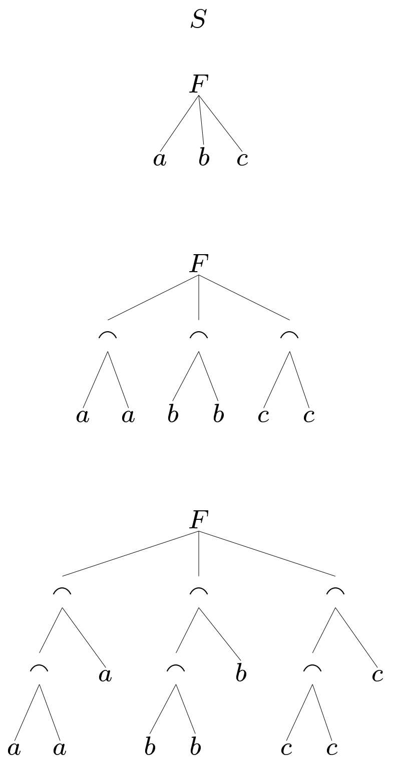
Makro-Charakterisierung von Morphologischen Grammatiken

Example ($a^nb^nc^n$)

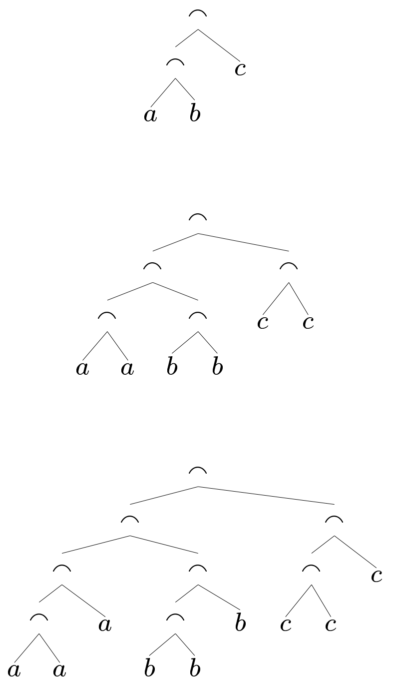
$$\begin{array}{lcl} \Sigma_0 & = & \{a, b, c\} \\ \mathcal{F}_0 & = & \{S\} \end{array} \qquad \begin{array}{lcl} \Sigma_2 & = & \{\frown\} \\ \mathcal{F}_3 & = & \{F\} \end{array}$$

$$P = \left\{ \begin{array}{ll} S & \to F(a, b, c) \\ F(x_1, x_2, x_3) & \to \\ & F(\frown(a, x_1), \frown(b, x_2), \frown(c, x_3)) \mid \\ & \frown(\frown(x_1, x_2), x_3)) \end{array} \right\}$$

Example ($a^nb^nc^n$)



Example ($a^nb^nc^n$)



Definition 13 *A* **Context–free tree grammar** *is a triple $G = \langle \Sigma, \mathcal{F}, P \rangle$, where Σ is a finite ranked alphabet of* terminals, *$\mathcal{F}$ is a finite ranked alphabet of* nonterminals *disjoint from Σ and P is a finite set of rules of the form*

$$F(x_1, \dots, x_m) \rightarrow t \qquad (n \geq 0)$$

*where $F \in \mathcal{F}_m$ and $t \in T(\Sigma \cup \mathcal{F}, X_m)$.**

*X is here assumed to be a denumerable set of variables $X = \{x_1, x_2, \dots\}$. $X_0 = \emptyset$, $X_m = \{x_1, \dots, x_m\}$.

Definition 14 *Let now $t_1, t_2 \in T(\Sigma \cup \mathcal{F}, X_m)$. t_2 is* **directly derivable** *from t_1*

$$t_1 \Rightarrow t_2$$

if there is a tree $u \in T(\Sigma \cup \mathcal{F}, X_{n+1})$ containing exactly *one occurrence of x_{n+1}, a corresponding rule $F(x_1, \dots, x_m) \rightarrow t$, and $s_1, \dots, s_m \in T(\Sigma \cup \mathcal{F}, X_n)$ such that*

$$t_1 = u[x_1, \dots, x_n, F(s_1, \dots, s_m)]$$

$$t_2 = u[x_1, \dots, x_n, t[s_1, \dots, s_m]]$$

t_2 is obtained from t_1 by replacing an occurrence of a subtree $F(s_1, \dots, s_m)$ by the tree $t[s_1, \dots, s_m]$.

For $m, n \geq 0$, $t \in T(\Sigma, X_m)$ and $t_1, \dots, t_m \in T(\Sigma, X_n)$

$$t[t_1, \dots, t_m]$$

denotes the result of **substituting** *t_i for x_i in t. Observe that $t[t_1, \dots, t_m]$ is in $T(\Sigma, X_n)$.*

As is customary $\overset{*}{\Rightarrow}$ denotes the transitive-reflexive closure of $\Rightarrow$.

Definition 15 *For* $m \geq 0$ *and* $s \in T(\Sigma \cup \mathcal{F}, X_m)$ *we define*

$$\mathcal{L}(G, s) = \{t \in T_\Sigma(X_m) \mid s \overset{*}{\Rightarrow} t\}$$

to be the **context–free tree language** *generated by* G *from* s.*

*Note that what we have introduced is the *unrestricted* mode of derivation. By imposing conditions on the form of the s_i and u in the definition of a direct derivation the resulting classes of languages generated by G vary with these conditions.

Makro Grammatik der Alt-Georgischen Kongruenz

$$\begin{array}{lcllcl} \Sigma_0 & = & \{a, b, c, d\} & & (\Sigma_2 & = & \{\frown\}) \\ \mathcal{F}_0 & = & \{S\} & \mathcal{F}_1 = \{S'\} & \mathcal{F}_2 & = & \{R\} \end{array}$$

$$P = \left\{ \begin{array}{lcl} S & \to & abd \mid abacbd \mid abacbac^2bd \mid \\ & & R(c^2, abacb) \\ S'(x) & \to & x \\ R(x_1, x_2) & \to & R(c^2x_1, x_2ax_1bacx_1b) \mid \\ & & S`(x_2ax_1bacx_1bd) \end{array} \right\}$$

$$S$$

$$\begin{array}{c} R \\ c^2 \quad abacb \end{array}$$

$$\begin{array}{c} R \\ c^2\mathbf{c^2} \quad \mathbf{abacb}a\mathbf{c^2}bac\mathbf{c^2}b \end{array}$$

$$\begin{array}{c} S` \\ \mathbf{abacbac^2bacc^2b}a\mathbf{c^2c^2}bac\mathbf{c^2c^2}bd \end{array}$$

Envoi

Makro-Sprachen $\cong$ Indizierte Sprachen

Parsing	polynomial
Wachstum	beschränkt: $\{2^{2^n}\} \notin$ IL