

# 15. Korpusanalyse

## 15.1 Implementation und Applikation von Grammatiksystemen

### 15.1.1 Bestandteile eines Grammatiksystems

- formaler Algorithmus
- linguistische Methode

### 15.1.2 Optionen für die Wortformerkennung

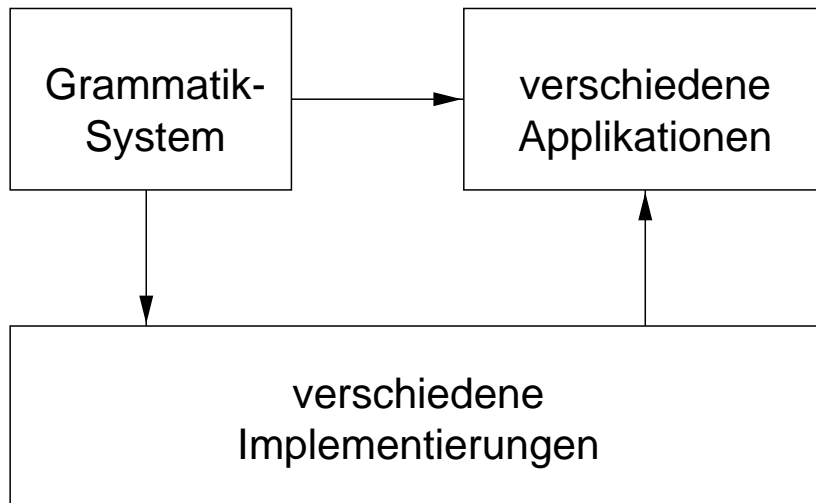
- *formaler Algorithmus:*
  - C-Grammatik (Sektion 7.4)
  - PS-Grammatik (Sektion 8.1)
  - LA-Grammatik (Sektion 10.2)
- *linguistische Methode:*
  - Wortform-Methode
  - Morphem-Methode (siehe Sektion 13.5)
  - Allomorph-Methode

### 15.1.3 Minimalstandard eines wohldefinierten Grammatiksystems

Ein Grammatiksystem ist nur dann wohldefiniert wenn es zugleich

- in einer gegebenen *Implementation* verschiedene *Applikationen*, und
- in einer gegebenen *Applikation* verschiedene *Implementationen* erlaubt.

### 15.1.4 Grammatiksystem, Implementation und Applikation



### 15.1.5 Unterschiedliche Realisierungen von LA-Morph

1988 in LISP (Hausser & Todd Kaufmann)

1990 in C (Hausser & Carolyn Ellis)

1992 in C, 'LAMA' (Norbert Bröker)

1994 in C, 'LAP' (Gerald Schüller)

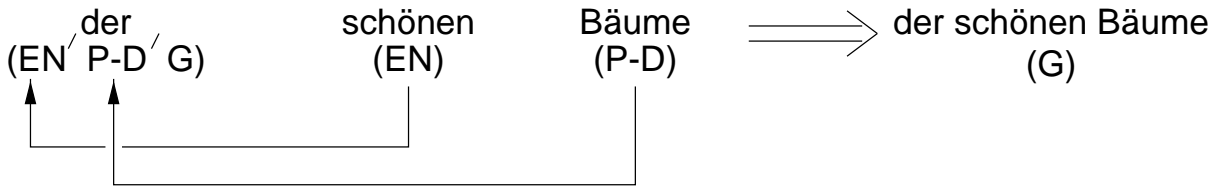
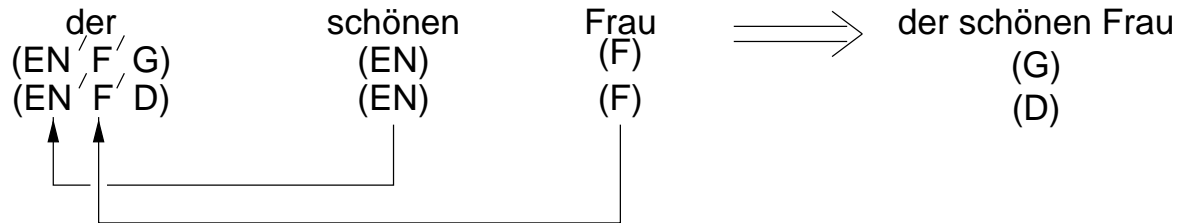
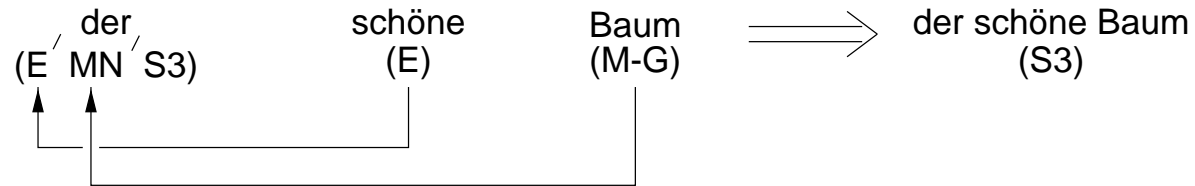
1995 in C, 'Malaga' (Björn Beutel)

### 15.1.6 Gemeinsame Strukturprinzipien verschiedener LA-Morph-Implementierungen

- Spezifikation der Allo- (siehe 14.1.1) und der Kombi-Regeln (siehe 14.4.1) auf der Grundlage von Mustern, die auf die Eingaben abgepaßt werden.
- Abspeicherung der analysierten Allomorphe in einer Trie-Struktur und ihr linksassoziativer Lookup mit paralleler Verfolgung alternativer Zerlegungen (siehe Sektion 14.3).
- Modulare Trennung von Motor, Regelkomponenten und Lexikon, die einen problemlosen Austausch der Teile, etwa bei der Anwendung auf neue Domänen oder Sprachen, erlaubt.
- Verwendung desselben Motors und desselben Algorithmus für die Kombi-Regeln der Morphologie-, Syntax- und Semantikkomponente.
- Verwendung derselben Regelkomponenten für Analyse und Generierung in Morphologie, Syntax und Semantik (vgl. Sektion 10.4).

## 15.2 Subtheoretische Varianten

### 15.2.1 Kombinatorik der Artikelform der



## 15.2.2 Abhängigkeit der Adjektivendung vom Artikel

der	schöne	Baum	$\Rightarrow$	der schöne Baum (S3)	(siehe 15.2.1)
(ER / ein / MN / S3)	(ER)	(M-G)	$\Rightarrow$	ein schöner Baum (S3)	

## 15.2.3 Exhaustive versus distinktive Kategorisierung (bei Ableitung von der schönen Frauen)

der	schönen	Frauen						
6	19	5	·	4	·	1		Multiplikation exhaustiver Lesarten
	114		+	20	=	134		Zahl der geordneten Eingabepaare

der	schönen	Frauen						
3	1	2	·	1	·	1		Multiplikation distinktiver Lesarten
	3		+	2	=	5		Zahl der geordneten Eingabepaare

### 15.2.4 Lesartendarstellung über Lexikoneinträge

[der (E' MN' S3) DEF-ART]

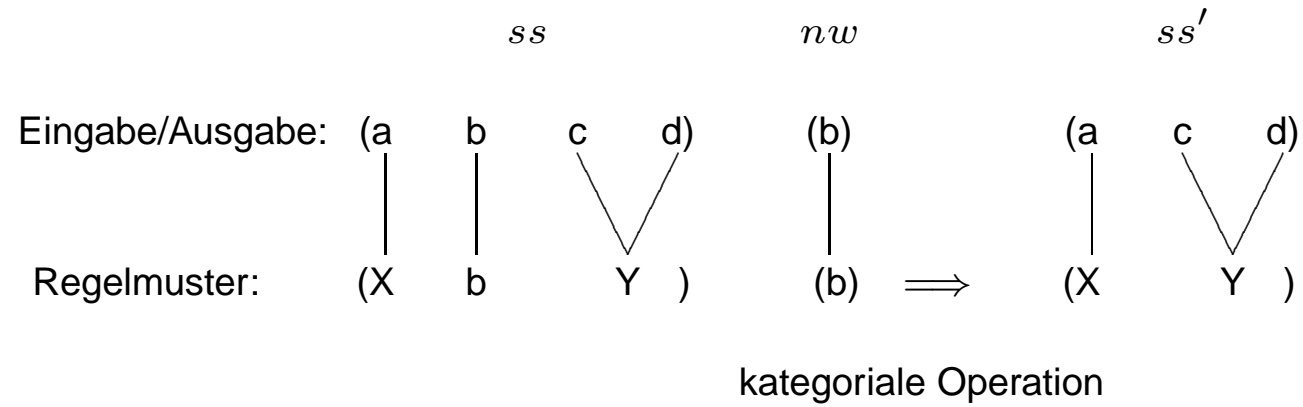
[der (EN' F' G&D) DEF-ART]

[der (EN' P-D' G) DEF-ART]

### 15.2.5 Lesartendarstellung über Multikats

[der ((E' MN' S3) (EN' F' G&D) (EN' P-D' G)) DEF-ART]

## 15.2.6 Listenbasierte Musterabpassung (LAP)



### 15.2.7 Merkmalbasierte Musterabpassung (Malaga)

$$\begin{array}{l}
 \text{input-output: } \begin{array}{c} ss \\ \left[ \begin{array}{l} mm1 = a \\ mm2 = b \\ mm3 = c \\ mm4 = d \end{array} \right] \end{array} \\
 \\
 \text{rule pattern: } \begin{array}{c} \left[ \begin{array}{l} mm2 = b \\ X \end{array} \right] \end{array}
 \end{array}
 \begin{array}{c}
 nw \\
 \left[ mm5 = b \right] \\
 \\
 \left[ mm5 = b \right] \implies [X]
 \end{array}
 \begin{array}{c}
 ss' \\
 \left[ \begin{array}{l} mm1 = a \\ \\ mm3 = c \\ mm4 = d \end{array} \right] \\
 \\
 [X]
 \end{array}$$

kategoriale Operation

## 15.3 Bau von Korpora

### 15.3.1 Die 15 Genres des Brown- und des LOB-Korpus

	Brown	LOB
A Presse: Reportagen	44	44
B Presse: Kommentare	27	27
C Presse: Rezensionen	17	17
D Religion	17	17
E Handwerk, Handel und Freizeit	36	38
F Trivialliteratur	48	44
G Literatur, Biographien, Essays	75	77
H Regierungsdokumente etc.	30	38
J Geistes- und naturwissenschaftliche Schriften	80	80
K Erzählungen allgemein	29	29
L Kriminalromane	24	24
M Science-fiction	6	6
N Abenteuer- und Wildwestgeschichten	29	29
P Liebesromane	29	29
R Humor	9	9
Gesamt	500	500

### 15.3.2 Kučera & Francis' Desiderata der Korpuskonstruktion

1. Exakte Spezifikation der verwendeten Sprachtexte, so daß sich die Benutzer einen genauen Begriff von der Zusammensetzung des Materials machen können.
2. Vollständige Synchronizität: nur Texte aus einem einzigen Kalenderjahr werden verwendet.
3. Die verschiedenen Genres werden in einem vorgegebenen Größenverhältnis zueinander gefüllt, wobei die individuellen Textbeispiele nach dem Zufallsprinzip ausgewählt werden (*random sampling*).
4. Formale Spezifikation der im Korpus enthaltenen Informationen und automatischer Zugriff auf sie.
5. Genaue und vollständige Beschreibung der elementaren statistischen Eigenschaften des Korpus und seiner Komponenten (Genres), mit der Möglichkeit, die Analyse auf Erweiterungen des Korpus auszudehnen.

### 15.3.3 Schwierigkeiten bei der Konstruktion eines repräsentativen balancierten Korpus

Bei dem Begriff "Genre" handelt es sich nicht um ein wohldefiniertes Konzept. Die bisher aufgestellten Genres basieren auf reiner Intuition. Bisher gibt es keine empirische Nachweise für ein existierendes System von Genre-Unterscheidungen.

N. Oostdijk 1988

### 15.3.4 Opportunistische Korpora, Referenzkorpus und Monitorkorpora

## 15.4 Auswertung von Korpora

### 15.4.1 Definition des Begriffs Rang

Die Position einer Wortform in der Frequenzliste

### 15.4.2 Definition des Begriffs Frequenzklasse (F-Klasse)

F-Klasse  $=_{def}$  [Frequenz der Types # Anzahl der Types]

Die Zahl der F-Klassen ist kleiner als die Anzahl der Ränge. Im BNC gibt es z. B. 655 270 Ränge aber nur 5 301 F-Klassen (0.8% der Ränge).

Daher sind F-Klassen für eine klare Darstellung der Type/Token-Distribution besser geeignet als die sonst allgemein üblichen Ränge.

**Menge der Types:** Menge der unterschiedlichen Wortformen eines Textes

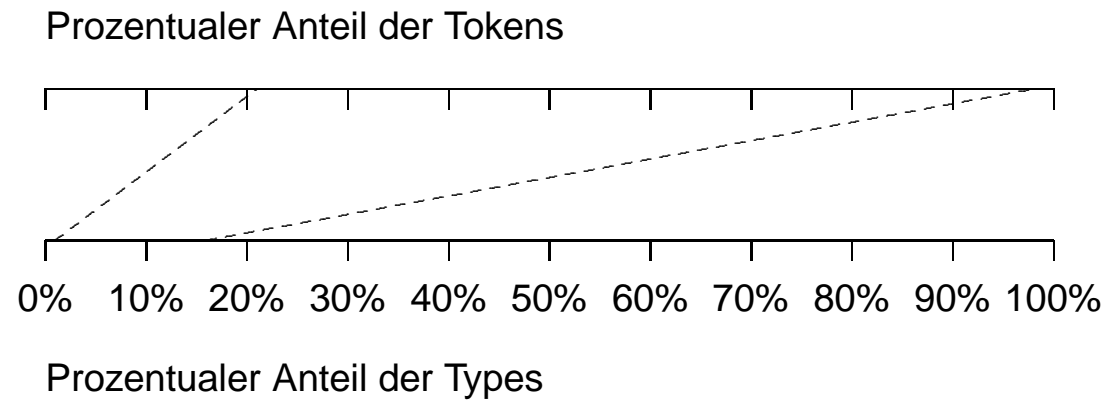
**Menge der Token:** Menge der tatsächlich vorkommenden Wortformen

**Type-Token-Beziehung:** Eine bestimmte Wortform (Type) manifestiert sich in konkreten Vorkommen (Token) in einem Korpus.

### 15.4.3 Type/Token-Verteilung im BNC (*oberflächenbasiert*)

F-class	start_r	end_r	types	tokens	types-%	tokens-%	
Anfang (Die ersten 9 F-Klassen)							
1 (the)	1	1	1	5776399	0.000152	6.436776	
2 (of)	2	2	1	2789563	0.000152	3.108475	
3 (and)	3	3	1	2421306	0.000152	2.698118	
4 (to)	4	4	1	2332411	0.000152	2.599060	
5 (a)	5	5	1	1957293	0.000152	2.181057	
6 (in)	6	6	1	1746891	0.000152	1.946601	
7 (is)	7	7	1	893368	0.000152	0.995501	
8 (that)	8	8	1	891498	0.000152	0.993417	
9 (was)	9	9	1	839967	0.000152	0.935995	
sums			9	19648696	0.001368 %	21.895 %	
Mitte (9 Stichproben)							
1000	1017	1017	1	9608	0.000152	0.010706	
2001	2171	2171	1	4560	0.000152	0.005081	tokens
3000	3591	3591	1	2521	0.000152	0.002809	per
3500	4536	4536	1	1857	0.000152	0.002069	type:
4000	5907	5910	4	5228	0.000607	0.005826	1307
4500	8332	8336	5	4005	0.000758	0.004463	801
4750	10842	10858	17	9367	0.002579	0.010438	551
5000	16012	16049	38	11438	0.005764	0.012746	301
5250	44905	45421	517	26367	0.078420	0.029381	51
Ende (Die letzten 9 F-Klassen)							
5292	108154	114730	6577	59193	0.997620	0.065960	9
5293	114731	122699	7969	63752	1.208763	0.071040	8
5294	122700	132672	9973	69811	1.512736	0.077792	7
5295	132673	145223	12551	75306	1.903775	0.083915	6
5296	145224	161924	16701	83505	2.533260	0.093052	5
5297	161925	186302	24378	97512	3.697732	0.108660	4
5298	186303	225993	39691	119073	6.020456	0.132686	3
5299	225994	311124	85131	170262	12.912938	0.189727	2
5300	311125	659269	348145	348145	52.807732	0.387946	1
Summen			551 116	1 086 559	83.595012 %	1.210778 %	

### 15.4.4 Korrelation von Type und Token-Häufigkeit



### 15.4.5 Semantische Signifikanz

Je höher die Frequenz umso niedriger die semantische Signifikanz.

Beispiele: der, ist, mit oder zu

Je niedriger die Frequenz, desto höher die semantische Signifikanz:

Beispiele:

Abbremsung, Babyflaschen, Campingplatz, ...

(Hapax Legomena aus dem Limas-Korpus – 1 062 624 laufende Wortformen)

audiophile, butternut, customhouse, dustheap, ...

(Hapax Legomena aus dem BNC – ca 100 000 000 laufende Wortformen)

**Hapax Legomena:** Wortformen die in einem Korpus nur ein einziges Mal vorkommen.

Wortformen, die im Lexikon stehen, aber im BNC überhaupt nicht vorkommen:

aspheric, bipropellant, dynamotor...

## 15.4.6 Zipfs Gesetz

$$\text{Frequenz} \cdot \text{Rang} = k \quad (\text{Konstante})$$

GEORGE K. ZIPF (1902–1950)

## 15.4.7 Illustration von Zipfs Gesetz an Beispielen aus dem BNC

Wortform	Rang	·	Frequenz	=	Produkt
the	1	·	5 776 399	=	5 776 399
and	2	·	2 789 563	=	5 579 126
...					
was	9	·	839 967	=	7 559 703
...					
holder	3 251	·	2 870	=	9 330 370

## 15.5 Statistisches Tagging

### 15.5.1 Anfang der Frequenzliste des Brown-Corpus

69971-15-500	THE	21341-15-500	IN
36411-15-500	OF	10595-15-500	THAT
28852-15-500	AND	10099-15-485	IS
26149-15-500	TO	9816-15-466	WAS
23237-15-500	A	9543-15-428	HE

Der Eintrag 9543-15-428 HE bedeutet beispielsweise, daß die Wortform “HE” 9 543mal im Brown-Corpus vorkommt und dabei in allen 15 Genres und in 428 der 500 Teiltex te vertreten ist.

### 15.5.2 Statistisches Tagging

Statistisches Tagging basiert darauf, daß zunächst die Wortformen eines kleinen Teilkorpus (*core corpus*) in Handarbeit kategorisiert werden – oder eine halbautomatische Kategorisierung zumindest nachträglich sorgfältig ediert und korrigiert wird. Die für die Klassifikation der Wortformen verwendeten Kategorien werden *tags* oder *labels* genannt. Ihre Summe bezeichnet man als *tagset*.

Nach dem manuellen Tagging des Teilkorpus werden mit Hilfe von *Hidden Markov Models* (HMMs) die Wahrscheinlichkeiten der Übergänge von einem Tag zum nächsten geschätzt.

Mit dem so trainierten HMM werden dann für das alle Token des Gesamtkorpus Tags ermittelt.

### 15.5.3 Teilmenge des *basic (C5) tagset (BNC)*

AJ0 Adjective (general or positive) (e.g. good, old, beautiful)

CRD Cardinal number (e.g. one, 3, fifty-five, 3609)

NN0 Common noun, neutral for number (e.g. aircraft, data, committee)

NN1 Singular common noun (e.g. pencil, goose, time, revelation)

NN2 Plural common noun (e.g. pencils, geese, times, revelations)

NP0 Proper noun (e.g. London, Michael, Mars, IBM)

UNC Unclassified items

VVB The finite base form of lexical verbs (e.g. forget, send, live, return)

VVD The past tense form of lexical verbs (e.g. forgot, sent, lived, returned)

VVG The -ing form of lexical verbs (e.g. forgetting, sending, living, returning)

VVI The infinitive form of lexical verbs (e.g. forget, send, live, return)

VVN The past participle form of lexical verbs (e.g. forgotten, sent, lived, returned)

VVZ The -s form of lexical verbs (e.g. forgets, sends, lives, returns)

### 15.5.4 Stichprobe aus getagtem BNC

We	PNP	for	PRP	.	PUN
have	VHB	this	DT0	So	AV0
reviewed	VVN	newsletter	NN1	,	PUN
our	DPS	series	NN0	look	VVB
corporate	AJ0	.	PUN	out	AVP
design	NN1	It	PNP	for	PRP
over	PRP	will	VM0	the	AT0
the	AT0	be	VBI	`	PUQ
past	AJ0	in	PRP	New	NP0
few	DT0	A4	ZZ0	Look	NP0
months	NN2	format	NN1	'	PUQ
and	CJC	but	CJC	not	XX0
as	PRP	similar	AJ0	to	TO0
November	NP0	to	PRP	miss	VVI
we	PNP	the	AT0	out	AVP
will	VM0	banner	NN1	on	PRP
be	VBI	for	PRP	what	DTQ
presenting	VVG	News	NN1	is	VBZ
a	AT0	for	PRP	happening	VVG
new	AJ0	Local	AJ0	in	PRP
design	NN1	Groups	NN2	Age	NN1

### 15.5.5 Alphabetische Wortformenliste (Stichprobe des BNC)

1 activ nn1-np0 1	8 activating aj0-nn1 6
1 activ np0 1	47 activating aj0-vvg 22
2 activa nn1 1	3 activating nn1-vvg 3
3 activa nn1-np0 1	14 activating np0 5
4 activa np0 2	371 activating vvg 49
1 activatd nn1-vvb 1	538 activation nn1 93
21 activate np0 4	3 activation nn1-np0 3
62 activate vvb 42	2 activation-energy aj0 1
219 activate vvi 116	1 activation-inhibition aj0 1
140 activated aj0 48	1 activation-synthesis aj0 1
56 activated aj0-vvd 26	1 activation. nn0 1
52 activated aj0-vvn 34	1 activation/ unc 1
5 activated np0 3	282 activator nn1 30
85 activated vvd 56	6 activator nn1-np0 3
43 activated vvd-vvn 36	1 activator/ unc 1
312 activated vvn 144	1 activator/ unc 1
1 activatedness nn1 1	7 activator/tissue unc 1
88 activates vvz 60	61 activators nn2 18
5 activating aj0 5	1 activators np0 1

Jeder Eintrag besteht aus (i) einer Ziffer, die die Frequenz (Zahl der Vorkommen) der Oberflächen-Tag-Kombination im ganzen Korpus angibt, (ii) der Oberfläche mit (iii) dem zugeordneten Tag und (iv) der Zahl der Teiltexthe, in denen diese Wortform mit dem angegebenen Tag gefunden wurde.

### 15.5.6 Fehlerraten beim statistischen Tagging

Leech (1995) gibt die Fehlerrate von CLAWS4 mit 1.7% an, was zunächst sehr gut erscheinen mag. Bedenkt man aber, daß die letzten (niederfrequentesten) 1.2% der Token 83.6% der Types ausmachen, dann könnte die Fehlerrate von 1.7% auch bedeuten, daß ca. 90% der Types nicht (korrekt) erkannt werden.

### 15.5.7 Schwächen des statistischen Tagging

1. Die Kategorisierung ist zu ungenau, um im regelbasierten Parsing Verwendung finden zu können.
2. Ein statistischer Tagger leistet weder Lemmatisierung noch Segmentierung.
3. Das Gesamtbild der Häufigkeitsverteilung wird durch ein künstliches Aufblähen der Typezahl verzerrt. (Einwand: das hängt nicht von der Methode des Tagging sondern von der Betrachtungsweise ab.)
4. Das Resultat eines statistische Taggers bleibt ist immer nur das Ergebnis probabilistischer Berechnungen.