15. Corpus analysis

15.1 Implementation and application of grammar systems

15.1.1 Parts of a grammar system

- Formal algorithm
- Linguistic method

15.1.2 Options for grammar system of word form recognition

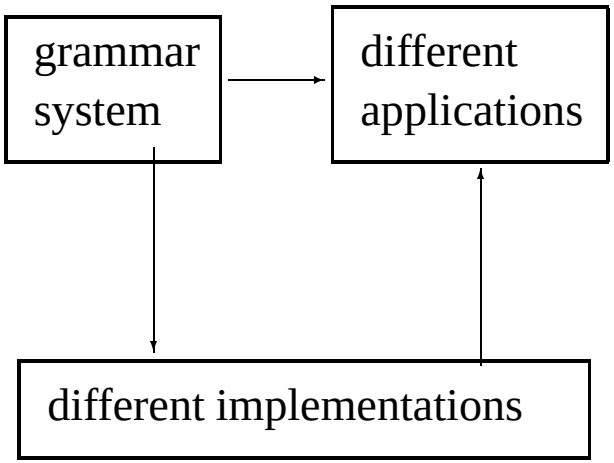
- Formal algorithm:
  C- (Section 7.4), PS- (Section 8.1), or LA-grammar (Section 10.2).
- Linguistic method:
  Word form, morpheme, or allomorph method (cf. Section 13.5).

15.1.3 Minimal standard of well-defined grammar systems

A grammar system is well-defined only if it simultaneously allows

1. different *applications* in a given *implementation*, and
2. different *implementations* in a given *application*.

15.1.4 Modularity of a grammar system



15.1.5 Different implementations of LA-morphology

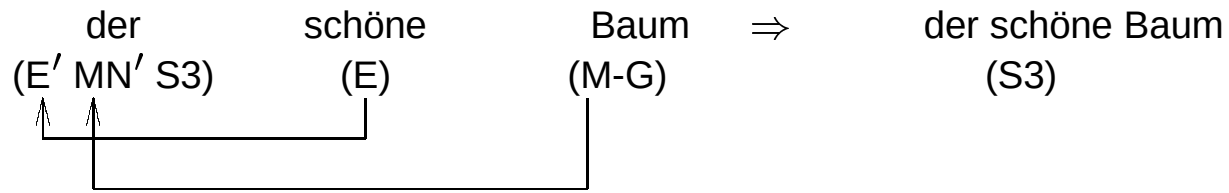
1988 in LISP (Hausser & Todd Kaufmann)
1990 in C (Hausser & Carolyn Ellis)
1992 in C, ‘LAMA’ (Norbert Bröker)
1994 in C, ‘LAP’ (Gerald Schüller)
1995 in C, ‘Malaga’ (Björn Beutel)

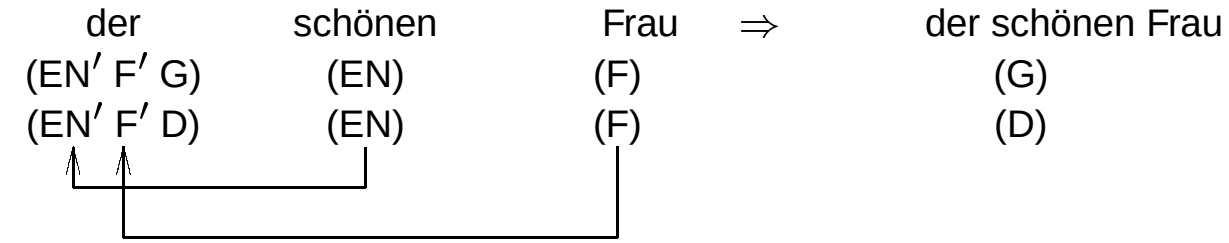
15.1.6 Structural principles common to different LA-Morph implementations

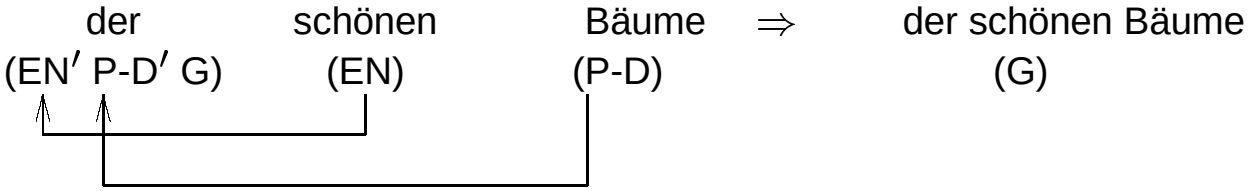
- Specification of the allo- (cf. 14.1.1) and the combi-rules (cf. 14.4.1) on the basis of patterns which are matched onto the input.
- Storage of the analyzed allomorphs in a trie structure and their left-associative lookup with parallel pursuit of alternative hypotheses (cf. Section 14.3).
- Modular separation of motor, rule components, and lexicon, permitting a simple exchange of these parts, for example in the application of the system to new domains or languages.
- Use of the same motor and the same algorithm for the combi-rules of the morphological, syntactic, and semantic components during analysis.
- Use of the same rule components for analysis and generation in morphology, syntax, and semantics.

15.2 Subtheoretical variants

15.2.1 Combinatorics of the German determiner der

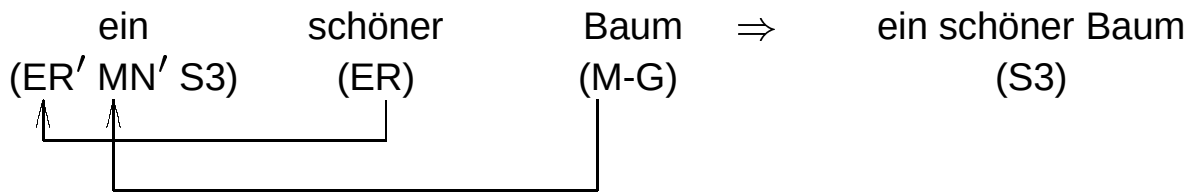




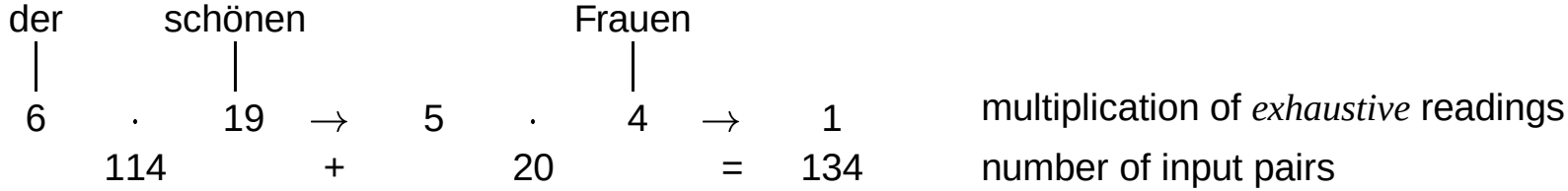


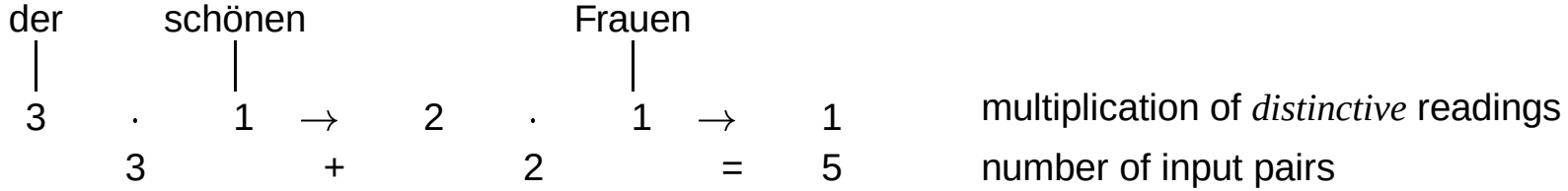
15.2.2 Agreement of adjective-ending with determiner

der schöne Baum (cf. 15.2.1)



15.2.3 Exhaustive versus distinctive categorization in deriving der schönen Frauen





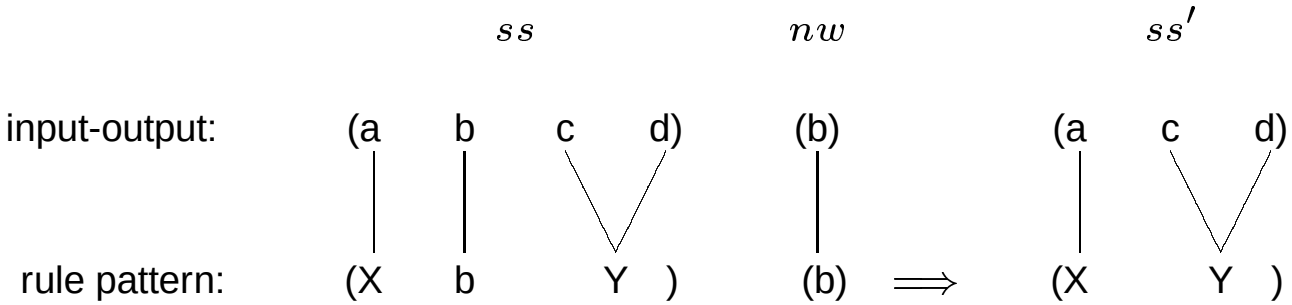
15.2.4 Representing lexical readings via different entries

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

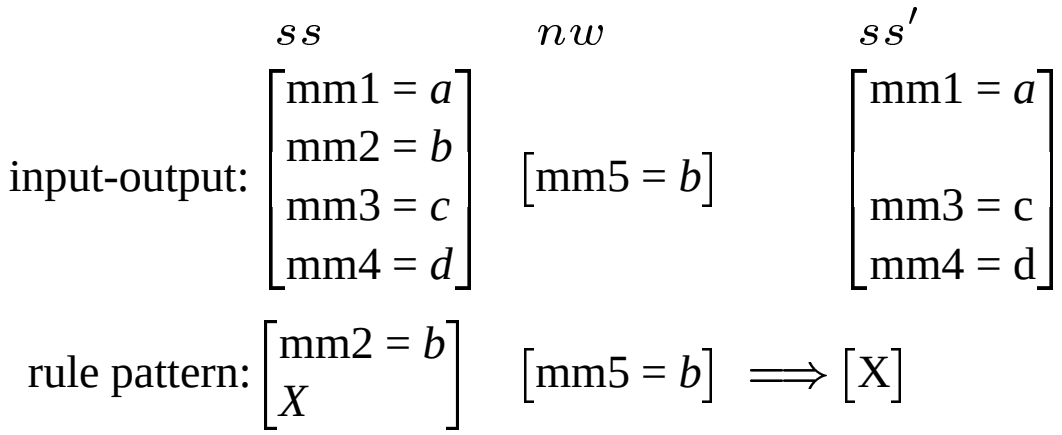
15.2.5 Representing lexical readings via multicats

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

15.2.6 List-based matching (LAP)



15.2.7 Feature-based matching (Malaga)



15.3 Building corpora

15.3.1 Text genres of the Brown and the LOB corpus

	Brown	LOB
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trade, and hobbies	36	38
F Popular lore	48	44
G Belle lettres, biography, essays	75	77
H Miscellaneous (government documents, foundation records, industry reports, college catalogues, industry house organ)	30	38
J Learned and scientific writing	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
Total	500	500

15.3.2 Kučera & Francis' desiderata for the construction of corpora

1. Definite and specific delimitation of the language texts included, so that scholars using the Corpus may have a precise notion of the composition of the material.
2. Complete synchronicity; texts published in a single calendar year only are included.
3. A predetermined ratio of the various genres represented and a selection of individual samples through a random sampling procedure.
4. Accessibility of the Corpus to automatic retrieval of all information contained in it which can be formally identified.
5. An accurate and complete description of the basic statistical properties of the Corpus and of several subsets of the Corpus with the possibility of expanding such analysis to other sections or properties of the Corpus as may be required.

15.3.3 Difficulties with achieving a representative and balanced corpus

> 'Genre' is not a well-defined concept. Thus genres that have been distinguished so far have been identified on a purely intuitive basis. No empirical evidence has been provided for any of the genre distinctions that have been made.
>
> N. Oostdijk 1988

15.4 Distribution of word forms

15.4.1 Definition of rank

The position of a word form in the frequency list

15.4.2 Definition of frequency class (F-class)

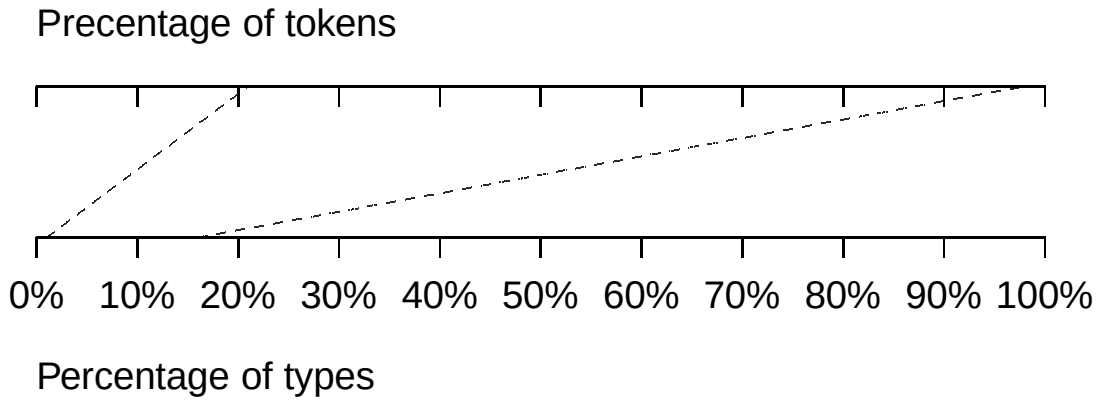
F-class $=_{def}$ [frequency of types # number of types]

There are much fewer F-classes in a corpus than ranks. In the BNC, for example, 655 270 ranks result in 5 301 F-classes. Thus, the number of the F-classes is only 0.8% of the number of ranks. Because of their comparatively small number the F-classes are well suited to bring the type-token correlation into focus.

15.4.3 Type-token distribution in the BNC (*surface-based*)

F-class	start_r	end_r	types	tokens	types-%	tokens-%	
beginning	(the first 9 F-classes)						
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	19 648 696	0.001368 %	21.895 %	
middle	(9 samples)						
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
end	(the last 9 F-classes)						
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
sums			551 116	1 086 559	83.595012 %	1.210778 %	

15.4.4 Correlation of type and token frequency



15.4.5 Semantic significance

The higher the freqency, the lower the semantic significance.
Examples: the, of, and, to, a, in, that, was

The lower the freqency, the higher the semantic significance.
Examples: audiophile, butternut, customhouse, dustheap

15.4.6 Hapaxlegomena

Word forms in a corpus which occur only once.

15.4.7 Zipf's law

frequency · rank = constant

15.4.8 Illustration of Zipf's law

word form	rank	·	frequency	=	constant
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 Statistical tagging

15.5.1 Top of Brown corpus frequency list

```
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
```

The entry `9543-15-428 HE`, for example, indicates that the word form `HE` occurs 9 543 times in the Brown corpus, in all 15 genres, and in 428 of the 500 sample texts.

15.5.2 Statistical tagging

is based on categorizing by hand – or half automatically with careful post-editing – a small part of the corpus, called the *core corpus*. The categories used for the classification are called *tags* or *labels*. After hand-tagging the core corpus, the probabilities of the transitions from one word form to the next are computed by means of *Hidden Markov Models* (HMMs).

15.5.3 Subset of the *basic (C5) tagset*

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 Sample from the alphabetical word form list of the BNC

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

Each entry consists (i) of a number detailing the frequency of the tagged word form in the whole corpus, (ii) the surface of the word form, (iii) the label, and (iv) the number of texts in which the word form was found under the assigned label.

15.5.5 Error rates in statistical tagging

The error rate of CLAWS4 is quoted by Leech 1995 at 1.7%, which may seem very good. However, given that the last 1.2% of the low frequency tokens requires 83.6% of the types (cf. 15.4.4), an error rate of 1.7% may also represent a very bad result – namely that about 90% of the types are not analyzed or not analyzed correctly. This conclusion is born out by a closer inspection of sample 15.5.4.

15.5.6 Weaknesses of statistical tagging

1. The categorization is too unreliable to support rule-based syntactic parsing.
2. Word forms can be neither reduced to their base forms (lemmatization) nor segmented into their allomorphs or morphemes.
3. The overall frequency distribution analysis of a corpus is distorted by an artificial inflation of types (e.g., 37.5% in the BNC).
4. Even if the tagger is successfully improved as a whole, its results can never be more than probabilistically-based conjectures.