In: Antonio Zampolli (ed., 1984): Studies in Honour of Roberto Busa. Linguistica
Computazionale, Vol. IV und V., S. 343-359

MACHINE-AIDED INDEXING OF TEXT CORPORA*

HARALD H. ZIMMERMANN

When the first volumes of the computer-aided indices of German literature[1] appeared in the sixties, linguistic dataprocessing with the help of the computer was still in its infancy. The same could be said (and it holds true to some extent even today) regarding the preparation of more or less representative text corpora[2], such as the LIMAS corpus[3], the news corpus of Lund[4] and various corpora of the Institute for the German Language[5].

What played a major role in the case of these initial attempts at computer-aided text processing were the basic functions of EDP devices pertaining to computation and collation, namely the capacity to sort words, or rather, word forms, (for instance in alphabetical order, spelt forward and backward) to order them according to their frequency (frequency dictionary); and sometimes also the formal function of indicating the stored data in a given context (line or sentence). More often than not such results of text analysis by computer were viewed as end products, and were presented in this raw form to potential users, especially to philologists and linguists, and this is still being done today.

It is, of course, also true of text dictionaries that one should not put everything in a corpus-oriented dictionary of this type that can be accomplished with its help. It is, moreover, true that one cannot foresee all the questions which may be raised in connection with an element or portion of the text.

On the other hand, it cannot be considered as an achievement to have discovered the computer as a substitute for a card-index box: Kaeding[6], for example, has processed running texts of nearly 11 million words without the aid of a computer. This material, even today, is useful for many complementary (intellectual as well as machine-aided) evaluations. Reference may be made in this connection to the linguistic statistics of Meier[7], and the investigations by the Goethe Institute which make use of the Kaeding material[8].

The demands on the computer should, therefore, definitely be somewhat higher in connection with collections of linguistic data. This can especially be done within the field of morphosyntax. Therefore, as early as 1970, in a case of text processing of this kind - namely the indexing of the works of the Austrian poet Georg Trakl - the expectations from a computer-aided product were higher than was usual for literary indices[9]. In order to provide the possibility of better use, especially by philologists, the following tasks should be accomplished:

1. Morphological «lemmatization» of the word-forms, i.e. their reduction to basic forms
2. Decomposition of compound word-forms into meaningful segments
3. Determination of derivations (especially suffixes)
4. Classification of documents according to syntactic categories (word-classes)
5. Semantic disambiguation

1. Lemmatization

By lemmatization in the broadest sense one generally understands the categorisation of word-forms according to key-words which represent them respectively, and which can morphologically, syntactically and semantically be traced back to the same characteristics, for instance one stem morpheme, one part of speech and similar characteristics of meaning[10]. This definition, if used strictly, would lead in some cases to problems (e.g. in the case of WAR and BIST, morphologically diverging forms of the German verb SEIN): on the other hand, however, it has not been possible till now to find semantic characteristics for sufficient differentiation in meaning which would stand up to every test. Finally, even in the syntactical field, the differentiation of parts of speech is not always to be determined empirically (for instance by means of distribution analyses); on the other hand, clumbing together or establishment of relations between word-forms over and above a particular category could be useful (e.g. SINGEN-GESANG; LIEBE-LIEBLING/ LIEB). The differentiation or clumbing together therefore always represents one of many possible ways of viewing the basic material in a particular form (for example, a «lemmatized index» in book-form), whereby the aspect of suitability of use and usefulness receives prime consideration.

2. Decomposition of Compound Words

In the German language compound words appear very frequently in texts. It is just a question of particular writing habits, namely that of writing in one word, as the alternative forms of the so-called «instant compounds» show (AMERIKAREISE - REISE NACH AMERIKA, MINISTERBESUCH - BESUCH DES MINISTERS). However, various possibilities of combining words within the context should not be overlooked. On the other hand, expressions comprising several words, such as JURISTISCHE PERSON which can be seen as one unit of meaning, show that writing together alone is not an indication for a change of meaning, and vice versa.

Therefore, it is useful in the interests of the user to decompose the compounds into their constituent elements, and to make use of them in an index or a register in an appropriate form of representation, at least in all those cases in which the elements of a compound are to some extent self-contained; or, perhaps, already in the case of rather formal decomposition, as in several cases, it is difficult to make a strict distinction.

3. Derivations

In many languages there are certain word derivations (mostly morphologically characterized) which make it seem a good idea to place the corresponding derivation variants together in relation to a core element. By the same token one can link the derivation element itself (above all suffixes) with the basis element, in order to facilitate assessments on frequency, category, etc., in a general linguistic corpus or in an author-related text.

The aims mentioned under *(2)* and *(3)* led to the fact that, at the initial stages of corpus research, simple wordform lists were prepared in which the words were spelt backwards. These lists were capable of satisfying almost all queries in this regard.

4. Wordclass Labelling

The labelling of word-class entries in a text collection is on the one hand an important instrument for separating the rather *functional* elements in a language (functional words, particles - such as conjunctions, articles, prepositions) from the elements which convey meaning (nouns, adjectives, verbs, adverbs). This makes it possible, for example, to restrict oneself to these meaning-conveying words when dealing with line or sentence concordances, since these words as a rule will be the centre of further research.

At the same time, a more clearly differentiated labelling of word-classes in several cases allows a (partial) semantic disambiguation of word-forms, as far as different word-classes are concerned (e.g. KREUZE (verb) derived from KREUZEN as against KREUZE (noun) derived from KREUZ; WAGEN (verb) as opposed to WAGEN (noun); LAUTE (adjective) derived from LAUT as opposed to LAUTE (verb) from LAUTEN; LAUTEN or LAUTE (noun) from (die)

LAUTE/ (der) LAUT...).

Finally, the labelling of word-classes can form the basis for far-reaching evaluation of text material, such as for the frequency of use of the definite article as opposed to the indefinite, for the structure of noun groups used, etc.

5. Semantic Disambiguation

In the case of heterogeneous texts with different themes and also of larger text collections, a distinction between the meanings of text-words where ambiguities occur is sensible for reasons of practicality.

This leads to two fundamental problems:

- How far must the meanings be differentiated? In several cases there occurs an additional partial problem, as besides (completely?) different meanings, semantic family and semantic hierarchies can also be established. Thus text collections with diverging semantic differentiation can no longer be comparable. In such cases, therefore, at least a general standard (or simply a point of reference), for example a reference to a particular lexicon, should be given, which permits a user to comprehend the differentiation. Further, a formal access via an undifferentiated word should be made possible, as can be found in traditional lexica.
- How far is it possible to find out the «correct meaning» in any particular case? Very often the context (for instance of a newspaper article or even of a poem) is not enough, and one inclines in such a situation, even if in rare cases, to interpret (which should actually be left to the user of a product). Therefore, if a semantic disambiguation is undertaken, the doubtful cases must be made recognizable as being such.

In the following it will be undertaken to demonstrate, with the help of the lemmatized index for the complete works of the Austrian poet Georg Trakl, as to how one can achieve appreciable results economically with the most basic use of a computer.

In the preparation of the lemmatized index for the works of Georg Trakl[11], the computer was as usual made use of purely as sorting and storing instrument, as, at that time, further developed computer functions were not available[12].

1) First of all the complete text was put on 5-channel punched tapes, whereby every page of the

critical edition[13] was marked accordingly, and the text was structured according to the line numbers as given by the publisher. With the help of this, later on references to page and line numbers were produced in the index.

2) Only in the case of substantives was the word-class explicitly marked by means of a capitalization characteristic. When an ambiguity cropped up in the case of other word-classes, a differentiating mark was likewise used.

3) Thereafter an alphabetically sorted word-form list was produced by the machine, and punched onto punched cards. The references were output onto reference cards which were corrected with the word-form (punched) cards (= word cards) by means of an identification number.

4) By means of a sorting machine, the word cards were separated mechanically from the reference cards, and the substantives from the remaining word-forms for which a corresponding marking was already there.

5) The word-cards were subsequently printed with the help of an automatic inscriber. With this the contents were made legible. These word-form cards, excepting, of course, the substantives, were divided intellectually into 3 groups: Particles («P», i.e. functional words such as DER, EIN, ODER...), inflected and uninflected adjectives («A») and verbs («V», including the participles). Subsequently, these three piles of cards were automatically given a corresponding word-class marking by means of a card-puncher.

6) The morphological lemmatization (basic form assignment) was to be undertaken in relation to word-class (only the particles remained as word-forms). For this purpose, all the word-forms in which the word and the basic form were identical were taken out manually and subsequently put together automatically in the place already left for this purpose. Others were put together with their basic forms by hand (this had to be done in approximately 2,000 different cases).

7) The basic forms (lemmata) thus received were then sorted by the computer in backward spelling order. Components of compounds and derivation elements for the relevant entries which were detected by simple comparison on specific reference element cards were listed intellectually; the same was true of suffixes.

8) The initial data were then mixed by means of a sorting machine in the order: basic form - word-form - reference instances; subsequently the reference cards were intellectually / manually ordered.

9) Thereafter, the print format for the index was established with the help of a controlling and paging program. Simultaneously, the control symbols were produced, which made it possible to

transfer the output, which was stored on magnetic tape, onto a film by photosetting, for production in book form (cf. Figure 1).

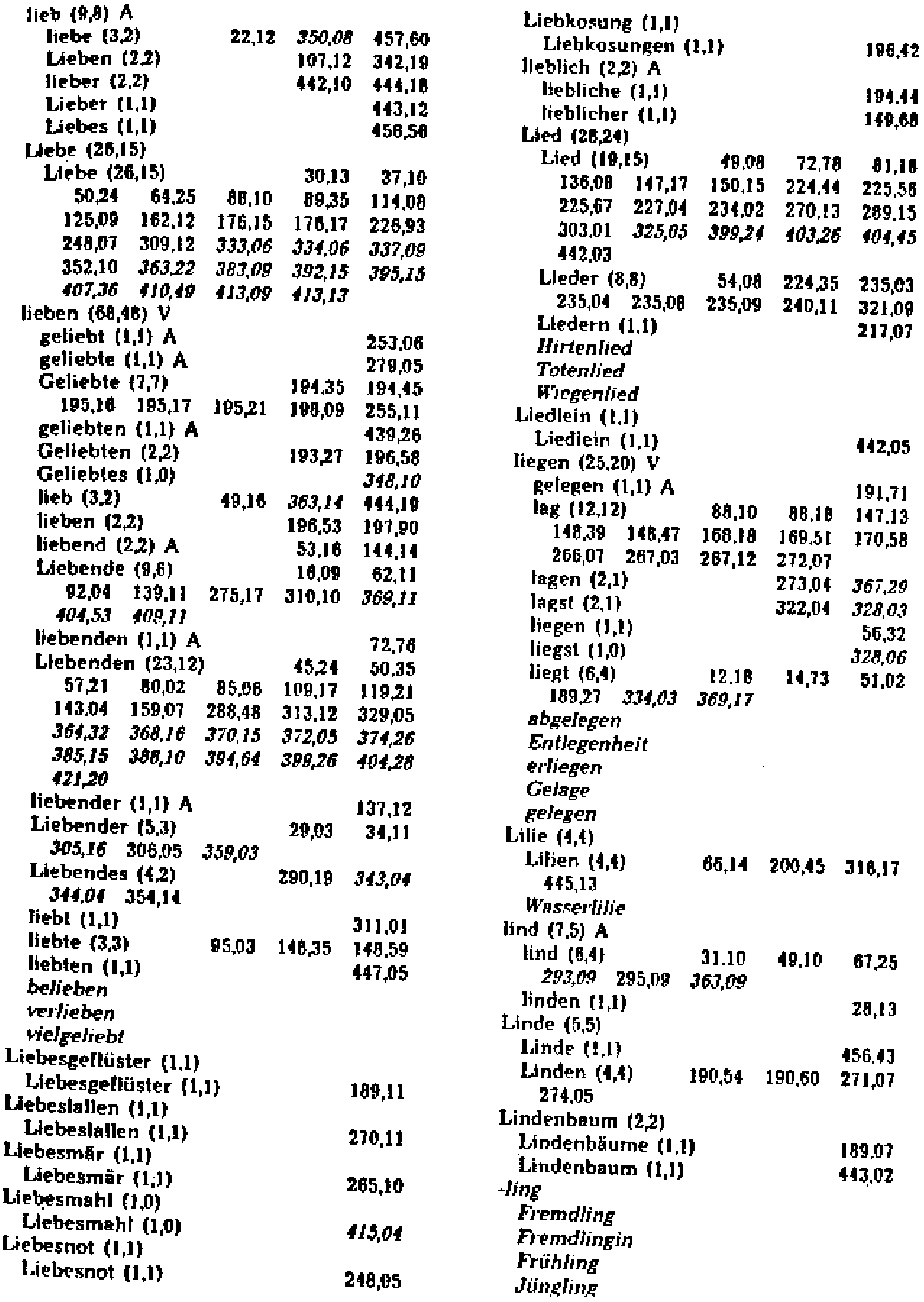
10) With the help of a further mechanical sorting program, a frequency dictionary for the entire data was produced, on the basis of the basic forms. The consideration of the word-classes in printed form showed that the «world» of Georg Trakl and his epoque, as symbolically expressed in the most frequent adjectives, was: dunkel (dark), blau (blue), schwarz (black), leise (low-voiced), still (quiet) .... (cf. Figure 2).

In order to complete phases (3) to (10), about one man-month was necessary. With this it became apparent that an appropriate working concept can lead to economically acceptable, and at the same time appreciable results, even with very limited resources.

Since the completion of this work more than ten years have passed. The unhandy punched tapes and punched cards with their limited stock of characters have in the meantime been replaced by terminals with the capacity for capital and small letters. In some research institutes, for instance at the University of Saarland, there are electronic procedures applicable at least in model form, which make it possible to use the computer not as a sorting instrument, but rather to use its «intelligence» for solving the above-mentioned problems. Such an «intelligent» system has been developed in Saarbrücken in the form of the «Saarbrücken method for automatic text analysis» which can produce from any kind of German language text morphologically, syntactically and semantically differentiated words (lemmata)[14] .

As early as in the sixties several research programs were already being pursued under the supervision of Hans Eggers[15].

In the meantime a practicable version of this system has been developed, and has been used for investigations in several cases.



lieb (9,8) A
liebe (3,2) 22,12 *350,08* 457,60
Lieben (2,2) 107,12 342,19
lieber (2,2) 442,10 444,18
Lieber (1,1) 443,12
Liebes (1,1) 456,58
Liebe (26,15)
Liebe (26,15) 30,13 37,10
50,24 64,25 88,10 89,35 114,08
125,09 162,12 176,15 176,17 226,93
248,07 309,12 *333,06* *334,06* *337,09*
352,10 *363,22* *383,09* *392,15* *395,15*
407,36 *410,49* *413,09* *413,13*
lieben (68,46) V
geliebt (1,1) A 253,06
geliebte (1,1) A 279,05
Geliebte (7,7) 194,35 194,45
195,16 195,17 195,21 198,09 255,11
geliebten (1,1) A 439,26
Geliebten (2,2) 193,27 196,58
Geliebtes (1,0) *348,10*
lieb (3,2) 49,16 *363,14* 444,19
lieben (2,2) 196,53 197,90
liebend (2,2) A 53,16 144,14
Liebende (9,6) 16,09 62,11
92,04 139,11 275,17 310,10 *369,11*
404,53 *409,11*
liebenden (1,1) A 72,76
Liebenden (23,12) 45,24 50,35
57,21 80,02 85,06 109,17 119,21
143,04 159,07 288,48 313,12 329,05
364,32 *368,16* *370,15* *372,05* *374,26*
385,15 *388,10* *394,64* *399,26* *404,28*
421,20
liebender (1,1) A 137,12
Liebender (5,3) 29,03 34,11
305,16 306,05 *359,03*
Liebendes (4,2) 290,19 *343,04*
344,04 354,14
liebt (1,1) 311,01
liebte (3,3) 95,03 148,35 148,59
liebten (1,1) 447,05
belieben
verlieben
vielgeliebt
Liebesgeflüster (1,1)
Liebesgeflüster (1,1) 189,11
Liebeslallen (1,1)
Liebeslallen (1,1) 270,11
Liebesmär (1,1)
Liebesmär (1,1) 265,10
Liebesmahl (1,0)
Liebesmahl (1,0) *415,04*
Liebesnot (1,1)
Liebesnot (1,1) 248,05

Liebkosung (1,1)
Liebkosungen (1,1) 196,42
lieblich (2,2) A
liebliche (1,1) 194,44
lieblicher (1,1) 149,68
Lied (28,24)
Lied (19,15) 49,08 72,78 81,18
136,08 147,17 150,15 224,44 225,56
225,67 227,04 234,02 270,13 289,15
303,01 *325,05* *399,24* *403,26* *404,45*
442,03
Lieder (8,8) 54,08 224,35 235,03
235,04 235,08 235,09 240,11 321,09
Liedern (1,1) 217,07
Hirtenlied
Totenlied
Wiegenlied
Liedlein (1,1)
Liedlein (1,1) 442,05
liegen (25,20) V
gelegen (1,1) A 191,71
lag (12,12) 88,10 88,18 147,13
148,39 148,47 168,18 169,51 170,58
266,07 267,03 267,12 272,07
lagen (2,1) 273,04 *367,29*
lagst (2,1) 322,04 *328,03*
liegen (1,1) 56,32
liegst (1,0) *328,06*
liegt (6,4) 12,18 14,73 51,02
189,27 *334,03* *369,17*
abgelegen
Entlegenheit
erliegen
Gelage
gelegen
Lilie (4,4)
Lilien (4,4) 66,14 200,45 316,17
445,13
Wasserlilie
lind (7,5) A
lind (6,4) 31,10 49,10 67,25
293,09 295,09 *363,09*
linden (1,1) 28,13
Linde (5,5)
Linde (1,1) 456,43
Linden (4,4) 190,54 190,60 271,07
274,05
Lindenbaum (2,2)
Lindenbäume (1,1) 189,07
Lindenbaum (1,1) 443,02
-ling
Fremdling
Fremdlingin
Frühling
Jüngling

Figure 1 - Alphabetical Index for Georg Trakl's poetry.
Example page 90.

Nr.	Rang	Häufigkeit		Wortkl.	Lemma
		rel.	abs.		
1	1	3,789	1240	P	und
2	2	3,673	1202	P	in
3	3	3,493	1143	P	die
4	4	3,056	1000	P	der
5	5	2,698	883	P	ein
6	6	1,295	424	P	das
7	7	1,248	408	V	sein
8	8	1,243	407	P	an
9	9	1,234	404	P	des
10	10	1,057	346	P	den
11	11	0,950	311	P	von
12	12	0,812	266	P	ich
13	13	0,696	228	A	dunkel
14	14	0,693	227	P	sich
15	15	0,684	224	P	es
16	16	0,678	222	P	ihr
17	17	0,620	203	P	auf
18	18	0,602	197	P	o
19	19	0,574	188	P	zu
20	20	0,540	177	P	aus
21			177	P	du
22	21	0,528	173	P	sie
23	22	0,516	169	P	sein
24	23	0,504	165	P	mein
25	24	0,501	164	P	mit
26	25	0,495	162	S	Nacht
27	26	0,492	161	A	blau
28	27	0,479	157	A	schwarz
29	28	0,476	156	P	wie
30	29	0,440	144	P	dem
31	30	0,430	141	P	da
32			141	P	über
33	31	0,424	139	A	leise
34	32	0,421	138	S	Schatten
35	33	0,382	125	P	dein
36	34	0,369	121	P	durch
37	35	0,330	108	V	gehen
38	36	0,317	104	P	mich
39			104	V	schweigen
40	37	0,314	103	P	er
41	38	0,305	100	V	sehen
42	39	0,293	96	S	Abend
43	40	0,290	95	A	still
44	41	0,275	90	P	vor
45	42	0,271	89	A	all
46			89	A	gulden
47	43	0,262	86	A	weiss
48	44	0,250	82	S	Auge
49			82	A	rot
50			82	A	sanft
51	45	0,244	80	P	dies
52	46	0,241	79	P	mir
59	49	0,229	75	S	Hand
60			75	S	Herz
61	50	0,223	73	P	nicht
62	51	0,213	70	S	Gott
63	52	0,210	69	A	purpurn
64	53	0,207	68	A	grün
65			68	V	werden
66	54	0,204	67	A	braun
67			67	V	verfallen
68			67	P	wenn
69	55	0,195	64	V	singen
70	56	0,192	63	P	nach
71	57	0,186	61	A	silbern
72			61	S	Stirn
73	58	0,183	60	S	Antlitz
74			60	S	Baum
75			60	S	Blut
76			60	S	Garten
77	59	0,180	59	S	Stille
78			59	V	tönen
79			59	S	Wein
80			59	P	wir
81	60	0,177	58	V	sinken
82			58	V	treten
83	61	0,171	56	A	kühl
84	62	0,168	55	A	einsam
85			55	V	fallen
86			55	A	wild
87	63	0,165	54	A	tot
88	64	0,161	53	S	Fenster
89			53	P	noch
90			53	V	stehen
91			53	P	voll
92			53	S	Wind
93	65	0,158	52	P	um
94			52	S	Wolke
95	66	0,155	51	P	dich
96			51	V	kommen
97	67	0,152	50	V	dämmern
98			50	P	jen
99			50	S	Leben
100	68	0,149	49	P	aber
101			49	P	als
102			49	S	Haupt
103			49	A	tief
104	69	0,146	48	S	Engel
105			48	A	lang
106			48	V	lieben
107			48	V	sterben
108			48	P	unser
109	70	0,143	47	A	fern
110			47	A	schön

Figure 2 - Frequency Index for Georg Trakl's poetry.
Example page 169.

However, it was not considered sensible to test the system on the «extreme case» of belles-lettres or poetry. It appeared more feasible to use texts written in everyday prose. The work with these texts can also be used for other kinds of text and can be projected onto a more practical use than would be the case in processing a poetical work.

Such a practical use is apparent in the field of specialized information and documentation. The main objective hereby is to index the relevant documents (for instance newspaper articles, texts of judgements, regulations, minutes, patent descriptions, etc.) to such an extent as to be able to retrieve them with the help of the words which appeared in the text. In other words, this can be termed as a type of «automatic indexing».

As it is apparent that the stored data will be quite voluminous, and in certain cases a data-bank or information bank would also contain documents of a heterogeneous nature, the problem of semantic differentiation does gain in importance. (In processing the works of Georg Trakl, this could safely be left aside).

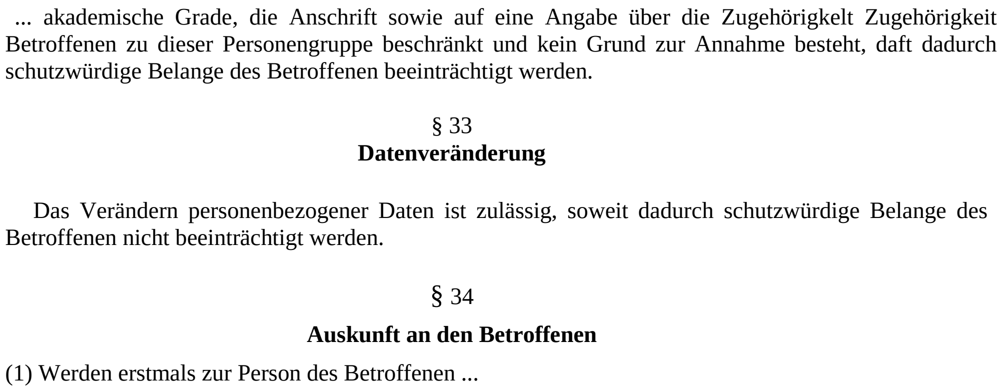
The automatic indexing system which has been developed in Saarbrücken on the basis of the experience gained from the above project comprises all the above-mentioned aspects, i.e.:

- Morphosyntactic lemmatization
- Decomposition
- Derivation
- Categorization according to word-class
- Semantic disambiguation

In certain aspects the system surpasses the above-mentioned framework of questions. For instance, even multiword expressions (such as JURISTISCHE PERSON) can be identified. In addition, between the disambiguated or lexically unambiguous elements a series of semantic relations (synonyms, generic and generated terms, etc.) is established, which is very useful for the process of retrieval. Finally, the syntactic relations which appear in the text (e.g. adjective-substantive, substantive and coordinated substantive) are made available in direct relation (as so-called complex descriptors) for retrieval.

In the following, the basic procedural steps of the system are described briefly[16]:

1) As the first step, the input text is split into individual words which are then «looked up» in a morphosyntactical dictionary via a standardized input interface. Every word-form is then supplied with various pieces of information necessary for further syntactic analysis, and at the same time the basic form is also determined, wherever possible. In the case of those word compounds and derivations which are not found in the general morphosyntactical lexicon, a decomposition or derivation analysis is carried out. In this way, the orthographic mistakes can also be detected at the same time. As our morphosyntactic lexicon for the German language in the meantime has over 142,000 entries of basic forms, the probability of a non-identified token being misspelt is fairly large (cf. Figures 3a and 3b).



... akademische Grade, die Anschrift sowie auf eine Angabe über die Zugehörigkelt Zugehörigkeit Betroffenen zu dieser Personengruppe beschränkt und kein Grund zur Annahme besteht, daft dadurch schutzwürdige Belange des Betroffenen beeinträchtigt werden.

§ 33

Datenveränderung

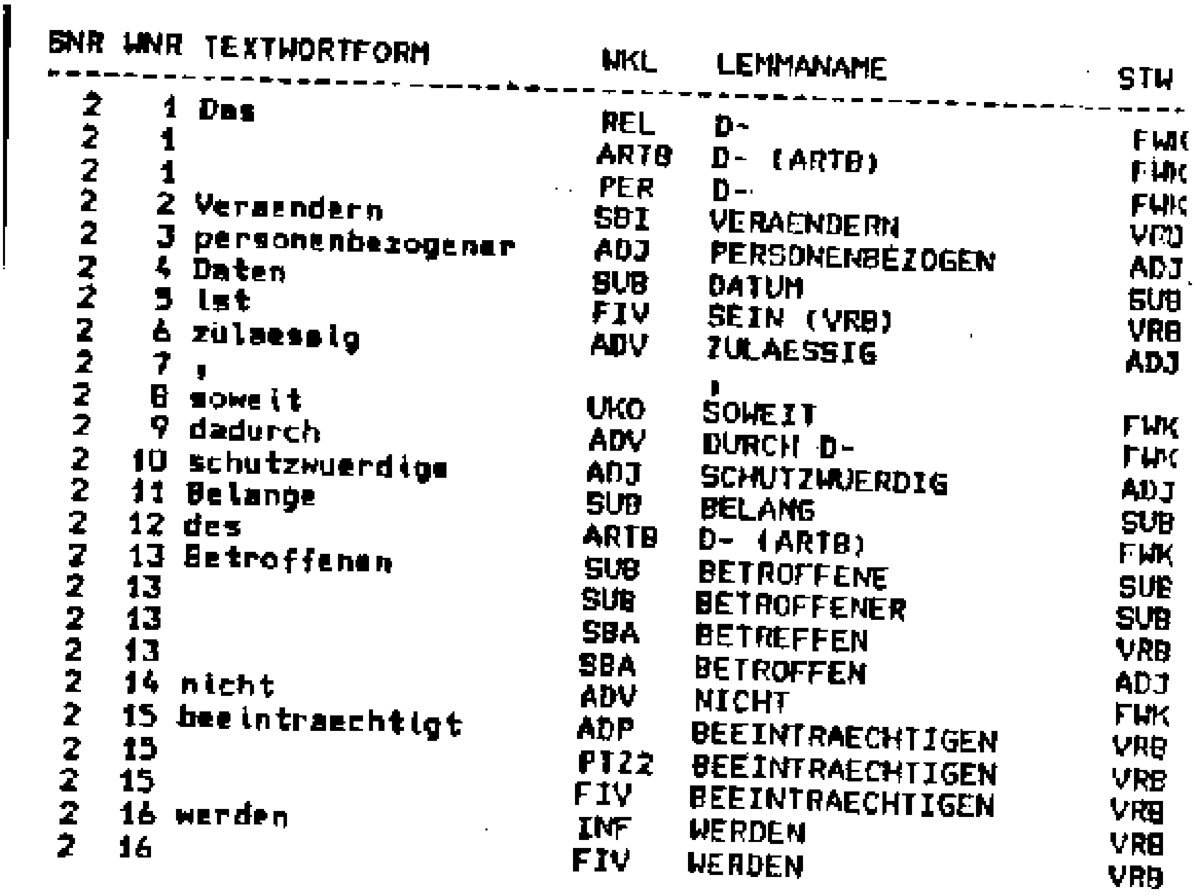
Das Verändern personenbezogener Daten ist zulässig, soweit dadurch schutzwürdige Belange des Betroffenen nicht beeinträchtigt werden.

§ 34

Auskunft an den Betroffenen

(1) Werden erstmals zur Person des Betroffenen ...

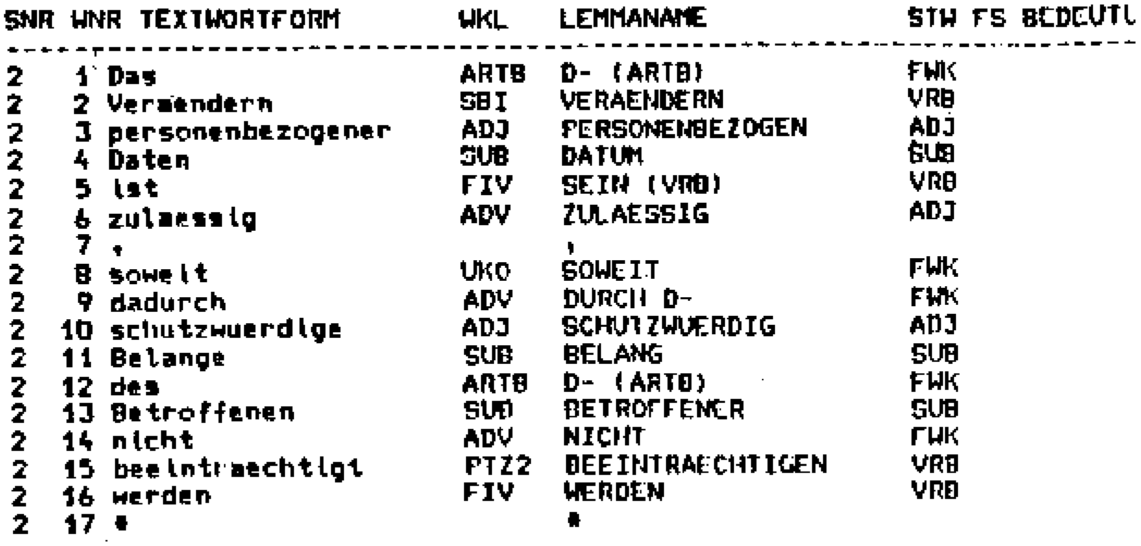
Figure 3a - Example of reduction of text for federal data protection law (BDSE$33).



BNR	WNR	TEXTWORTFORM	WKL	LEMMANAME	STW
2	1	Das	REL	D-	FWK
2	1		ARTB	D- (ARTB)	FWK
2	1		PER	D-	FWK
2	2	Veraendern	SBI	VERAENDERN	VRB
2	3	personenbezogener	ADJ	PERSONENBEZOGEN	ADJ
2	4	Daten	SUB	DATUM	SUB
2	5	ist	FIV	SEIN (VRB)	VRB
2	6	zülaessig	ADV	ZULAESSIG	ADJ
2	7	,		,	
2	8	soweit	UKO	SOWEIT	FWK
2	9	dadurch	ADV	DURCH D-	FWK
2	10	schutzwuerdige	ADJ	SCHUTZWUERDIG	ADJ
2	11	Belange	SUB	BELANG	SUB
2	12	des	ARTB	D- (ARTB)	FWK
7	13	Betroffenen	SUB	BETROFFENE	SUB
2	13		SUB	BETROFFENER	SUB
2	13		SBA	BETREFFEN	VRB
2	13		SBA	BETROFFEN	ADJ
2	14	nicht	ADV	NICHT	FWK
2	15	beeintraechtigt	ADP	BEEINTRAECHTIGEN	VRB
2	15		PT22	BEEINTRAECHTIGEN	VRB
2	15		FIV	BEEINTRAECHTIGEN	VRB
2	16	werden	INF	WERDEN	VRB
2	16		FIV	WERDEN	VRB

Figure 3b – Result of morpho-syntactic analysis.

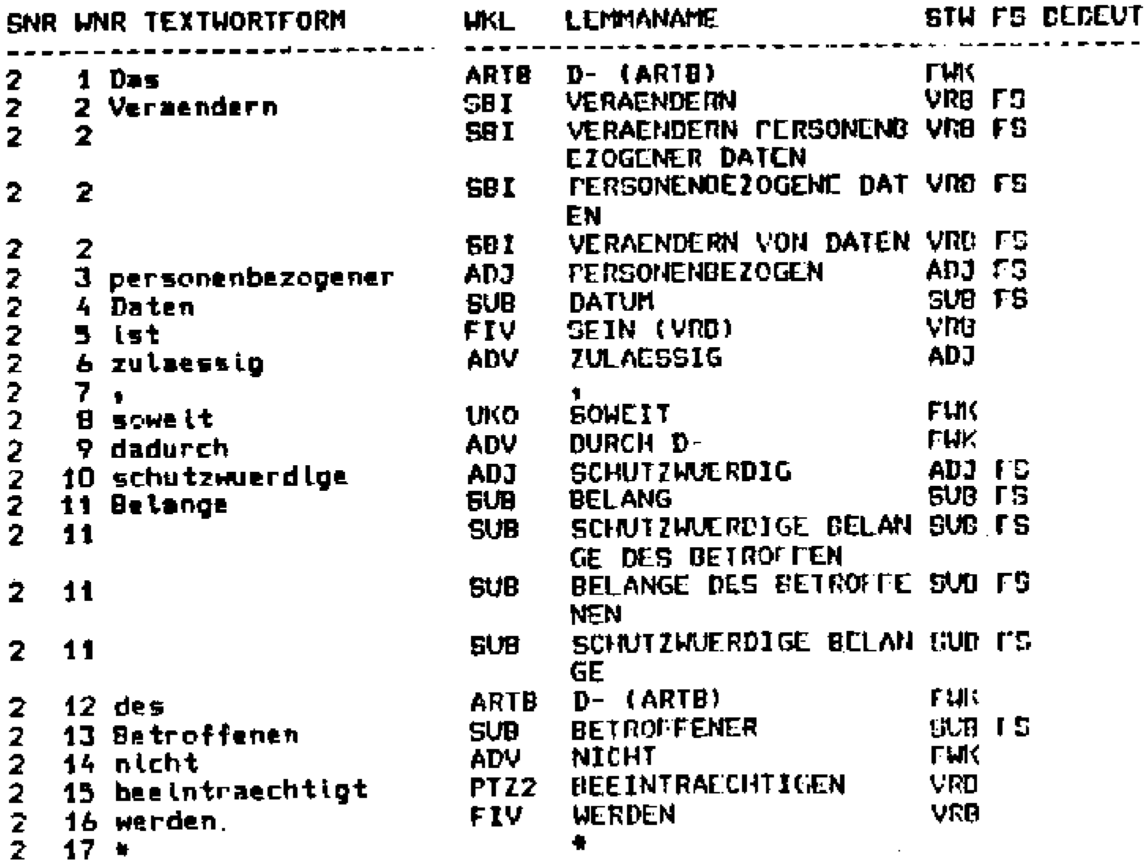
2) The dictionary check, as elaborated above, produces the potential word-classes and the basic forms. Thereafter, the various relevant functions are established on the basis of a sentence or a context-oriented mechanical syntactic analysis. This part of the procedure fulfils the function of determining word-classes and also that of lemmatization, in so far as the potential categories from the lexicon are reduced to the actual ones in the context (cf. Figure 4).



```
SNR WNR TEXTWORTFORM        WKL   LEMMANAME            STW FS BEDEUTU
---------------------------------------------------------------------
2    1 Das                  ARTB  D- (ARTB)            FWK
2    2 Veraendern           SBI   VERAENDERN           VRB
2    3 personenbezogener    ADJ   PERSONENBEZOGEN      ADJ
2    4 Daten                SUB   DATUM                SUB
2    5 ist                  FIV   SEIN (VRB)           VRB
2    6 zulaessig            ADV   ZULAESSIG            ADJ
2    7 ,                          ,
2    8 soweit               UKO   SOWEIT               FWK
2    9 dadurch              ADV   DURCH D-             FWK
2   10 schutzwuerdige       ADJ   SCHUTZWUERDIG        ADJ
2   11 Belange              SUB   BELANG               SUB
2   12 des                  ARTB  D- (ARTB)            FWK
2   13 Betroffenen          SUB   BETROFFENER          SUB
2   14 nicht                ADV   NICHT                FWK
2   15 beeintraechtigt      PTZ2  BEEINTRAECHTIGEN     VRB
2   16 werden               FIV   WERDEN               VRB
2   17 .                          .
```

Figure 4 – Determination of word-class by syntactic analysis.

3) The next step is to identify multi-word expressions and to undertake semantic disambiguation as far as possible on the basis of the sentence-related context. This is done with the help of a semantic lexicon, which, inter alia, contains rules for determining multi-word or inflected expressions as well as characteristics and rules for semantic disambiguation (cf. figure 5).

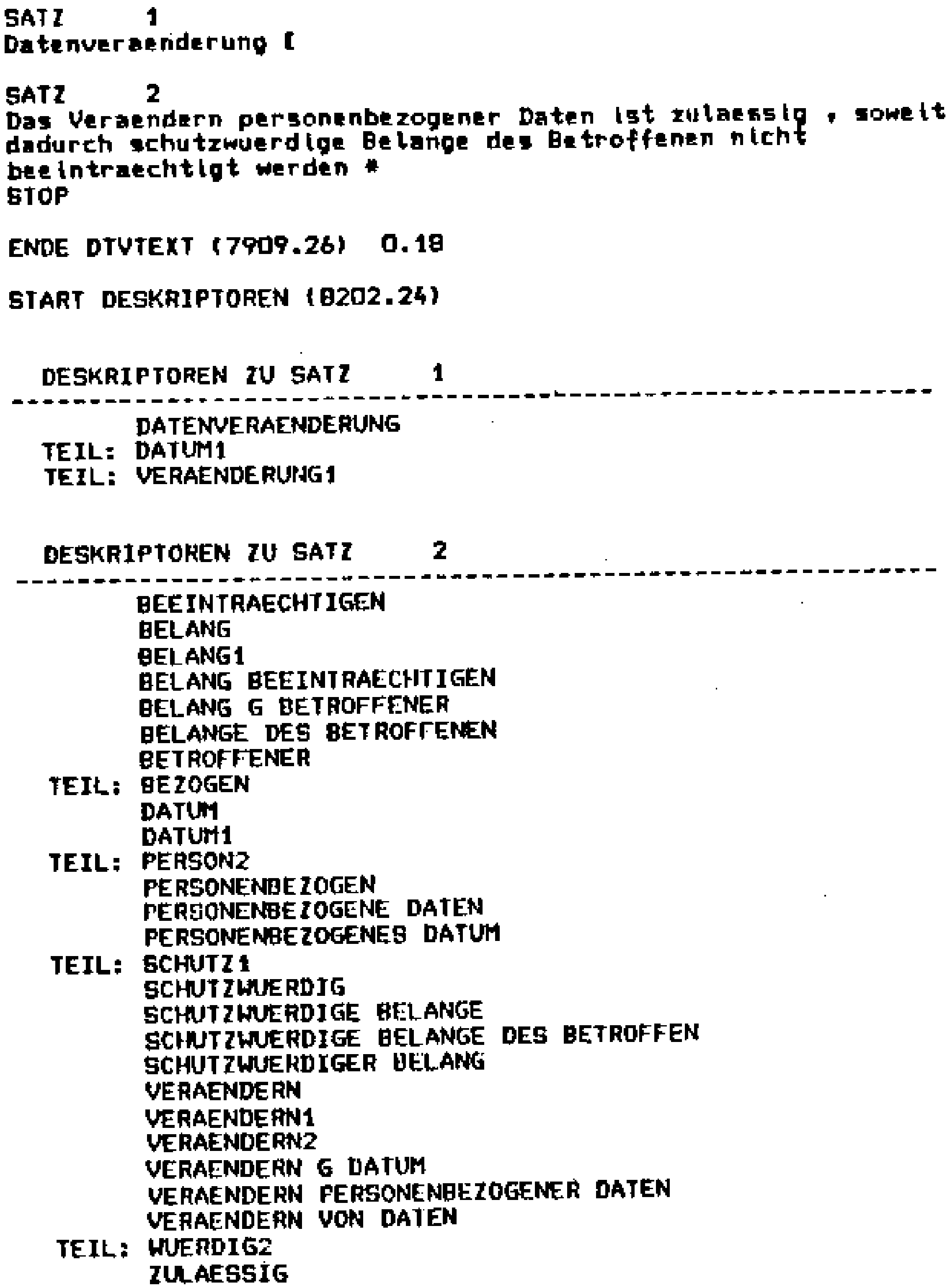


```
SNR WNR TEXTWORTFORM          WKL   LEMMANAME              STW FS BEDEUT
----------------------------  -------------------------- ----------------
2    1 Das                    ARTB  D- (ARTB)              FWK
2    2 Veraendern             SBI   VERAENDERN             VRB FS
2    2                        SBI   VERAENDERN PERSONENB VRB FS
                                    EZOGENER DATEN
2    2                        SBI   PERSONENBEZOGENE DAT VRB FS
                                    EN
2    2                        SBI   VERAENDERN VON DATEN VRB FS
2    3 personenbezogener      ADJ   PERSONENBEZOGEN        ADJ FS
2    4 Daten                  SUB   DATUM                  SUB FS
2    5 ist                    FIV   SEIN (VRB)             VRB
2    6 zulaessig              ADV   ZULAESSIG              ADJ
2    7 ,                            ,
2    8 soweit                 UKO   SOWEIT                 FWK
2    9 dadurch                ADV   DURCH D-               FWK
2   10 schutzwuerdige         ADJ   SCHUTZWUERDIG          ADJ FS
2   11 Belange                SUB   BELANG                 SUB FS
2   11                        SUB   SCHUTZWUERDIGE BELAN SUB FS
                                    GE DES BETROFFEN
2   11                        SUB   BELANGE DES BETROFFE SUB FS
                                    NEN
2   11                        SUB   SCHUTZWUERDIGE BELAN SUB FS
                                    GE
2   12 des                    ARTB  D- (ARTB)              FWK
2   13 Betroffenen            SUB   BETROFFENER            SUB FS
2   14 nicht                  ADV   NICHT                  FWK
2   15 beeintraechtigt        PTZ2  BEEINTRAECHTIGEN       VRB
2   16 werden.                FIV   WERDEN                 VRB
2   17 *                            *
```

Figure 5 – Semantic analysis: disambiguation and determination of multi-word expressions.

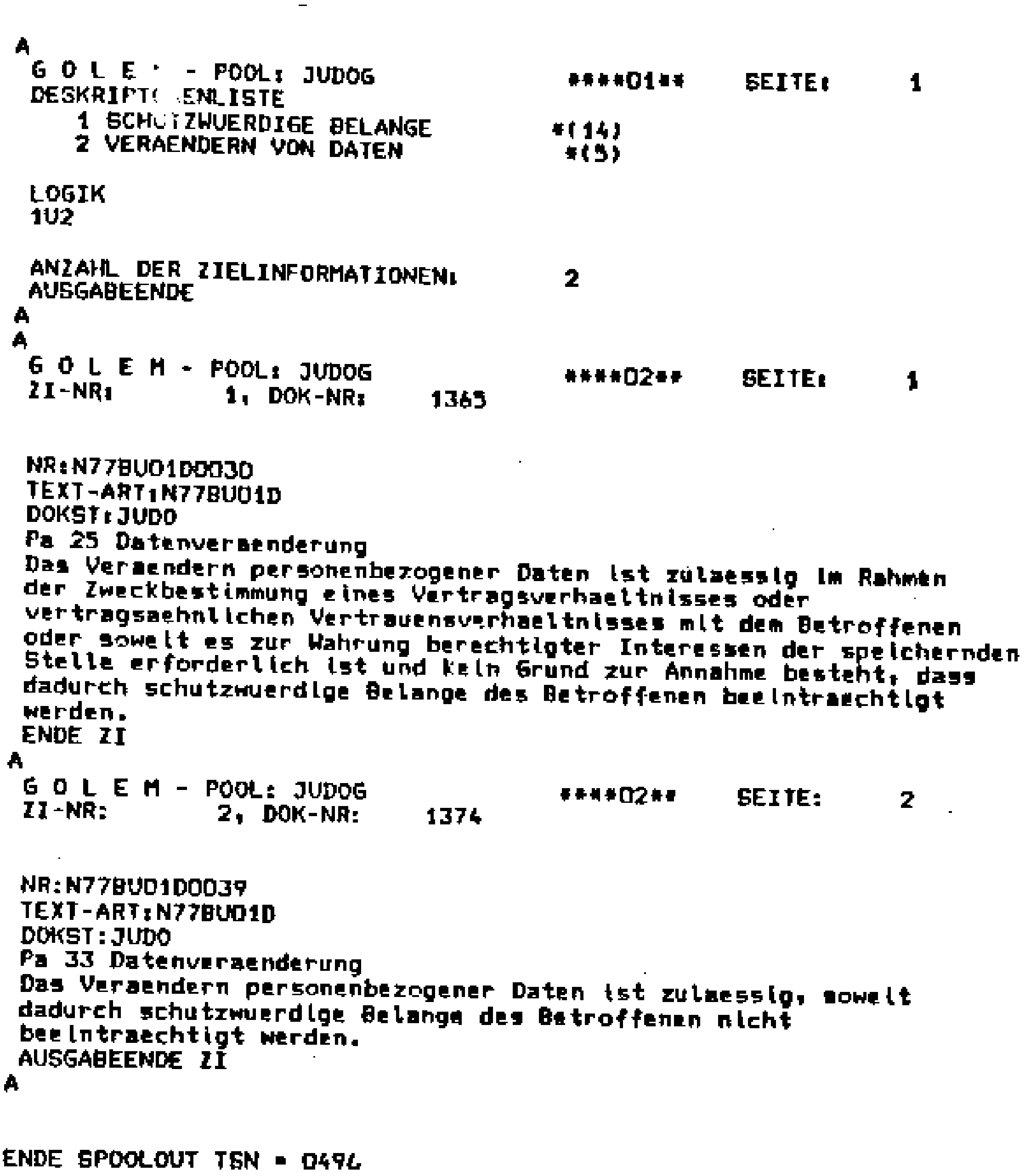
4) In the next step, it is attempted to remove all remaining ambiguities on the basis of statistical data (probability of the use of a word in a particular meaning in a specialized subject area), whereby the semantic relations as given in a technical lexicon are made use of, going beyond the sentence-based context.

5) The last step is a procedure which processes the information in such a way as to make it possible to produce descriptors which can then be stored in a data bank or put in the form of entries which can be put in a register (cf. Figure 6 regarding the generation of descriptors; Figure 7 shows an example of access to a data bank).



```
SATZ      1
Datenveraenderung [

SATZ      2
Das Veraendern personenbezogener Daten ist zulaessig , soweit
dadurch schutzwuerdige Belange des Betroffenen nicht
beeintraechtigt werden *
STOP

ENDE DTVTEXT (7909.26)  0.18

START DESKRIPTOREN (8202.24)


  DESKRIPTOREN ZU SATZ      1
---------------------------------------------------------------
         DATENVERAENDERUNG
  TEIL:  DATUM1
  TEIL:  VERAENDERUNG1


  DESKRIPTOREN ZU SATZ      2
---------------------------------------------------------------
         BEEINTRAECHTIGEN
         BELANG
         BELANG1
         BELANG BEEINTRAECHTIGEN
         BELANG G BETROFFENER
         BELANGE DES BETROFFENEN
         BETROFFENER
  TEIL:  BEZOGEN
         DATUM
         DATUM1
  TEIL:  PERSON2
         PERSONENBEZOGEN
         PERSONENBEZOGENE DATEN
         PERSONENBEZOGENES DATUM
  TEIL:  SCHUTZ1
         SCHUTZWUERDIG
         SCHUTZWUERDIGE BELANGE
         SCHUTZWUERDIGE BELANGE DES BETROFFEN
         SCHUTZWUERDIGER BELANG
         VERAENDERN
         VERAENDERN1
         VERAENDERN2
         VERAENDERN G DATUM
         VERAENDERN PERSONENBEZOGENER DATEN
         VERAENDERN VON DATEN
  TEIL:  WUERDIG2
         ZULAESSIG
```

Figure 6 – List of descriptors for the example text.



```
A
 G O L E M - POOL: JUDOG                     ****01**    SEITE:     1
 DESKRIPTORENLISTE
    1 SCHUTZWUERDIGE BELANGE               *(14)
    2 VERAENDERN VON DATEN                  *(5)

 LOGIK
 1U2

 ANZAHL DER ZIELINFORMATIONEN:             2
 AUSGABEENDE
A
A
 G O L E M - POOL: JUDOG                     ****02**    SEITE:     1
 ZI-NR:         1, DOK-NR:     1365


 NR:N77BU01D0030
 TEXT-ART:N77BU01D
 DOKST:JUDO
 Pa 25 Datenveraenderung
 Das Veraendern personenbezogener Daten ist zulaessig im Rahmen
 der Zweckbestimmung eines Vertragsverhaeltnisses oder
 vertragsaehnlichen Vertrauensverhaeltnisses mit dem Betroffenen
 oder soweit es zur Wahrung berechtigter Interessen der speichernden
 Stelle erforderlich ist und kein Grund zur Annahme besteht, dass
 dadurch schutzwuerdige Belange des Betroffenen beeintraechtigt
 werden.
 ENDE ZI
A
 G O L E M - POOL: JUDOG                     ****02**    SEITE:     2
 ZI-NR:         2, DOK-NR:     1374


 NR:N77BUD1D0039
 TEXT-ART:N77BUD1D
 DOKST:JUDO
 Pa 33 Datenveraenderung
 Das Veraendern personenbezogener Daten ist zulaessig, soweit
 dadurch schutzwuerdige Belange des Betroffenen nicht
 beeintraechtigt werden.
 AUSGABEENDE ZI
A


ENDE SPOOLOUT TSN = 0496
```

Figure 7 – Example for a search in the databank.

The system was developed as a model. In a laboratory application a running text of a little more than 100,000 words from the area of data production was processed, and it could well be established that automatic indexation of text is practicable. During further intensive research at the University of Saarland the system was modified and optimized[17]. The system has been tested in several pilot applications. On the whole, running texts of more than 3 million words from various specialized subject areas (especially from the German Patent Bureau) have been successfully processed and indexed.

Viewing all this, one tends to realize that simple procedures like the generation of word form indices, corresponding frequency lists of KWIC/KWOC concordances seem to have only historical value. The methodology for processing text corpora has entered a phase which could be termed as the «second generation» of text corpora indexation. Any work on text corpora, therefore, should be seen in light of the fact that the «instrument» computer has to be optimized and made more effective by developing appropriate lexica and rules for data-processing to such an extent, that it could easily make use of modern indexing systems as described above, and barring, of course, the cases in which an immediate ad hoc processing of the texts becomes essential.

« »

*. Revised version of a lecture delivered at the symposium “Computer corpus of the Serbo-Croatian language”, Belgrade, 14-18 December, 1981.

Anmerkungen:

1. *Serie «Indices zur deutschen Literatur»,* edited by H. SCHWERTE and H. SCHANZE, Athenaeum, Frankfurt.

2. For the concept of corpus, cf. various contributions in the volume *Empirische Textwissenschaft. und Auswertung von Text-Corpora. Edited by H. BERGENHOLTZ, B. SCHAEDER, Konigstein/Ts., 1979, also including the article by B. RIEGER (pp. 52-70).*

3. Microfiche edition: MCS-Verlag, Nürnberg, 1979. Series: *Regensburger Materialien auf Microfiche (RMM*). For the structure of the LIMAS corpus, cf. R. GLAS: Das LIMAS- Korpus, ein Textkorpus für die deutsche Gegenwartssprache. *Linguistische Berichte* 40 (1975), pp. 63-66.

4. Cf. I. ROSENGREN, Ein Frequenzwörterbuch der modernen Zeitungssprache - wie und wozu? *Beiträge zur Linguistik und Informationsverarbeitung* 14 (1968), pp. 7-21.

5. For the corpora of the Institute of the German Language, cf. B. SCHAEDER: Das Bonner Zeitungskorpus: Eine maschinelle Dokumentation von Tageszeitungen der BRD und der DDR. Mimeo, Bonn 1978, and U. ENGEL: Das Mannheimer Corpus, *Forschungsberichte des Instituts für Deutsche Sprache 2*, Mannheim, 1969, pp. 75-84.

6. F.W. KAEDING (ed.), *Häufigkeitswörterbuch der Deutschen Sprache. Festgestellt durch einen Arbeitsausschuss der deutschen Stenographiesysteme.* Steglitz bei Berlin, 1898. A detailed description can be found in: W.B. ORTMANN (ed.), *Hochfrequente deutsche Wortformen I,* Munich, 1975, pp. 5-26.

7. H. MEIER, *Deutsche Sprachstatistik.* Hildesheim 1964. Investigations by this «idealistic lonewolf» (according to ORTMANN, p. 27) are based on the Kaeding material, whereby the Kaeding material was processed and modified in over 40 years of work done in his free-time.

[8] Machine-aided evaluation of the Kaeding material by the section for scientific didactics of the Goethe Institute by W. D. ORTMANN has since led to a series of publications. These are of use above all in phonological studies.

[9] W. KLEIN, H. ZIMMERMANN: *Index zu Georg Trakl. Dichtungen.* Frankfurt, 1971.

[10] For a formal definition of lemma» in this extended sense, cf. R. DIETRICH: Automatische Textwörterbücher. Studien zur maschinellen Lemmatisierung verbaler Wortformen des Deutschen. In: H.E. BREKLE et al. (eds.), *Linguistische Arbeiten* 2, Tübingen 1973, esp. pp. 1f. Dietrich relies on the definition by H. D. MAAS: Homographie und maschinelle Sprachübersetzung, in *Linguistische Arbeiten des Germanistischen Instituts und des Instituts für Angewandte Mathematik der Universität des Saarlandes,* No. 8, Saarbrücken, 1969.

[11] Cf. footnote 9. Technical instruments used were: Computer of the type Philips Electrologica X1; a CDC2200 computer and punched card sorting machines.

[12] As this method, in spite of the fact that certain changes have taken place in hardware technology in the meantime, still appears to be economical, inter alia for languages in which no procedures for automatic analysis exist, it is described in short.

[13] Georg Trakl. *Dichtungen und Briefe. Historical-critical edition.* Ed. W. KILLY and H. SZKLENAR. Salzburg, 1969 (2 volumes). The index does not include the letters. This was due to the counting of frequency. Cf. also the preface to the index.

[14] For a description of the method, cf.: SALEM *Ein Verfahren zur automatischen Lemmatisierung deutscher Texte.* Ed.: Sonderforschungsbereich 100 «Elektronische Sprachforschung», Projektbereich A. Tübingen 1980.

[15] The most important result of this early research is: H. EGGERS et al., *Elektronische Syntaxanalyse der deutschen Gegenwartssprache,* Tübingen, 1969.

[16] The method is described in: H.H. ZiMMERMANN: Ansätze einer realistischen automatischen Indexierung unter Verwendung linguistischer Verfahren. In: R. KUHLEN (ed.), *Datenbasen, Datenbanken, Netzwerk.* Vol. 1, München, 1979, pp. 311-338. Also: *Bürgernahe Informationsvermittlung am Beispiel des Modellsystems «Juristische Dokumentanalyse im Bereich Datenschutz» (JUDO-DS),* edited by the same. In: Österreichische Gesellschaft für Informatik (ed.): Informationssysteme für die 80-er Jahre (Fachtagung 1980), vol. 1, Linz 1980, pp. 143-168.

[17] For a detailed description, cf. H.H. ZIMMERMANN, E. KROUPA, G. KEIL et al., *CTX - Ein Verfahren zur Computer-gestützten Texterschließung.*