In: Linguistische Studien IV. Festgabe für Paul Grebe zum 65. Geburtstag, Teil 2. Sprache der Gegenwart. 1973. Schriften des Instituts für deutsche Sprache, S. 238-261

HEINZ JOSEF WEBER - HARALD H. ZIMMERMANN

ZUR VERWERTBARKEIT DER GROSSSCHREIBUNG BEI DER AUTOMATISCHEN REDUKTION SYNTAKTISCHER WORTFORMEN-MEHRDEUTIGKEITEN IM DEUTSCHEN

0. Vorbemerkungen

0.1 Bei der maschinellen Bearbeitung sprachlicher Äußerungen ist man auf computerzugängliche (d.h. auf Lochkarten, Lochstreifen aufgenommene) Daten angewiesen. Das Sprachmaterial kann dabei einmal so kodiert sein, wie es der Bearbeiter vorgefunden hat; dieses Verfahren bietet sich z.B. an, wenn aus Setzereien Lochstreifen zur Verfügung stehen, die nicht mehr manuell aufzubereiten sind.[1] Bei der Texterstellung lässt sich das Material aber meist in beliebiger Weise verändern bzw. normieren: So kann der Punkt am Satzende von einem (Abkürzungs-)Punkt im Satzinnern unterschieden werden, oder die Wortformen bestimmter Wortklassen können durch besondere Kennzeichnung von denen anderer Wortklassen abgehoben werden. Die Wortformen der Klasse Substantiv lassen sich z.B. durch große Anfangsbuchstaben von den Formen anderer Wortklassen trennen, wie dies bei der entsprechenden Rechtschreibregelung des DUDEN geschieht.

0.2 Von derartigen Normierungen darf man gewisse Erleichterungen bei der späteren Bearbeitung erwarten. Die Saarbrücker Arbeitsgruppe für Elektronische Sprachforschung hingegen hat bei der Erstellung ihrer Texte die DUDEN-Regelung nicht übernommen, soweit diese die Unterscheidung zwischen groß- und kleingeschriebenen Wortformen bzw. Wortklassen fordert. Wir haben also auf die übliche Großschreibung am Satzanfang und auf die Großschreibung der Substantive (bzw. auf eine entsprechende Normierung) verzichtet. Dabei hat man uns gelegentlich vorgeworfen, ein wichtiges schriftsprachliches Hilfsmittel zur Vermeidung oder Verminderung von Mehrdeutigkeiten zu vernachlässigen.

Die Frage, inwieweit es notwendig ist, bei einer maschinellen syntaktischen Analyse deutscher (Schrift-) Sprache auf die Kennzeichnung der Wortformen durch Groß-Kleinschreibung zurückzugreifen, soll im folgenden eingehender untersucht werden.

0.3 Wir sind uns bewusst, dass diese Frage nur ein Randproblem in der maschinellen Sprachbearbeitung darstellen kann. Darüber hinaus dürfte der Untersuchung jedoch eine allgemeinere Bedeutung zukommen, da dieser Komplex während der Diskussion um eine Reform der deutschen Rechtschreibung oft im Mittelpunkt gestanden hat.

1. Syntaktische Wortformen-Mehrdeutigkeit

1.1 Bei maschineller Sprachbearbeitung - ob im Bereich der automatischen Wörterbucherstellung, der syntaktischen Analyse oder der Übersetzung (um nur einige zu nennen) - muss in irgendeiner Weise das Problem der Mehrdeutigkeit sprachlicher Zeichen bewältigt werden. Diese Mehrdeutigkeit stellt sich auf jeder Analyseebene verschieden dar, sei es als inhaltliche oder syn-

taktische Mehrdeutigkeit von Wortformen (*die Dame des Hauses - die Dame des Schachspiels; billige Äpfel - ich billige es*) von Gruppen (*die schöne Frau* kann als Nominalgruppe sowohl im Nominativ als auch im Akkusativ fungieren) oder von Sätzen (*die Frau sieht das Kind* - beide Nominalgruppen können Subjekt oder Objekt des Satzes sein).[2]

1.2 Bei der Analyse sind solche Mehrdeutigkeiten auf die aktuelle(n) Version(en) zu reduzieren. Gültigkeit und Wirksamkeit eines Analysealgorithmus können weitgehend an der Bewältigung von Mehrdeutigkeiten gemessen werden. Für unsere Fragestellung ist die Behandlung syntaktischer Mehrdeutigkeiten auf Wortformenebene die Voraussetzung, um Angaben machen zu können über

a. das Ausmaß solcher Mehrdeutigkeiten in unseren Texten,
b. die Möglichkeiten, durch Großschreibung bestimmter Wortformen solche Mehrdeutigkeiten zu modifizieren,
b. Auswirkungen einer Verwendung oder Nichtverwendung von Großschreibung auf unser Analyseverfahren und die erzielten Lösungsquoten bei der Disambiguierung von Wortformenmehrdeutigkeiten.

1.3 Die nun folgenden Untersuchungen setzen eine gewisse Vertrautheit mit dem in Saarbrücken entwickelten Analyseverfahren voraus;

Platzgründe verbieten jedoch die Anführung aller theoretischen und technischen Implikationen. Wir verweisen daher auf den 1969 erschienenen Bericht "Elektronische Syntaxanalyse der deutschen Gegenwartssprache" (2).

2. Wortformen-Mehrdeutigkeit und Großschreibung

2.1 Wir begreifen im Folgenden die Texte unserer Corpora als eine Menge W von Wortformen (types). Diese Wortformen können ihrerseits Elemente bestimmter Teilmengen von W sein. Wir nennen diese Teilmengen "Wortklassen". Die Menge W der types kann beschrieben werden als Vereinigungsmenge dieser Wortklassen

W = ADJ v ADV v DEM v FRA ... v ZUA v ZUI

Nun sind diese Teilmengen im Deutschen nicht disjunkt, d.h.: Es gibt types, die Element mehrerer Wortklassen sind. Solche Wortformen seien "Homographen" [3]. Die Wortklassen, denen ein Homograph angehört, nennen wir hier "Wortklassenkombination" (WK-Kombination). Alle Elemente, die der gleichen Wortklassen-Kombination zugeordnet werden können, bilden eine "Homographenklasse" (HO-Klasse). Beispiel: Die Wortformen als, bis, seit gehören sowohl der Wortklasse KON als auch der Wortklasse PRP an. Sie bilden - evtl. mit anderen Wortformen zusammen - eine HO-Klasse (oder eine Durchschnittsmenge KON ^ PRP), die etwa so beschrieben werden kann:

KON ^ PRP = {als, bis, seit, ... }

2.2 Geht man davon aus, dass Wortformen nicht von vornherein einer bestimmten Wortklasse zugeordnet werden können, wird man zur Einsetzung von Kategorien kommen, in denen Wortformen zusammengefasst sind, die mehreren und dabei den gleichen Wortklassen angehören.

2.3 Ein wichtiger Faktor, der Zahl und Umfang dieser Kategorien beeinflusst, ist das zugrundegelegte Wortklassensystem. So hat z.B. AGRICOLA (1) die bestehende Großschreibungsregelung bei der Konstituierung seiner Mehrdeutigkeitstypen akzeptiert: Beispiele sind Schrauben (nach Wortklasse eindeutig, flexivisch mehrdeutig), Arme (wortklassenmehrdeutig, Mask. Pl. und Adj.-Subst., flexivisch mehrdeutig), zwecks (in jeder Hinsicht eindeutig), die, durch Groß- oder Kleinschreibung schon voranalysiert, einem fiktiven Textzusammenhang entnommen sind.[4]

3. Auswirkungen einer Verwendung der Großschreibung auf Homographen-System und Klassifikation des Textmaterials

3.1 Die Auswirkungen der Großschreibung von Substantiven sollen nun zuerst im Hinblick auf Zahl und Umfang der WK-Kombinationen im syntaktischen Wortformenbuch und zum andern in Bezug auf die Aktualisierung von Wortklassenmehrdeutigkeiten im Saarbrücker Textmaterial beschrieben werden.

3.2 Zunächst eine Übersicht der von uns im Deutschen - unter den genannten Voraussetzungen - festgestellten Wortklassen-Mehrdeutigkeiten. Zu den einzelnen HO-Klassen werden jeweils Wortformenbeispiele angeführt. Weiter wird die Anzahl der types dieser HO-Klassen im syntaktischen Wortformenbuch mitgeteilt. Zusätzlich ist angegeben, wie hoch sich der Textanteil der homographen types im RDE-Corpus beläuft. Der Textanteil wurde aufgrund eines Samples von insgesamt 1.475 RDE-Sätzen (mit 27.220 tokens) ermittelt.[5] Weitergehende quantitative Angaben (siehe Abschnitt 3.6) wurden durch Zählungen an 500 zusätzlichen RDE-Sätzen gewonnen.[6]

Nachstehende Tabelle ist in zwei Hauptgruppen unterteilt: In der ersten sind die WK-Kombinationen zusammengefasst, in denen solche Wortklassen vertreten sind, welche nach herkömmlicher Regel Großschreibung verlangen (Gruppe I). Diese Gruppe wird wiederum untergliedert: In I a werden die Wortklassen-Kombinationen behandelt, bei denen Großschreibung (der Klassen SUB, SIN, SAD) homographie-v e r m e i d e n d wäre. In I b stehen die WK-Kombinationen, bei denen Großschreibung lediglich homographie- v e r m i n d e r n d wirken kann. In Gruppe II sind dann jene WK-Kombinationen zusammengefasst, deren Wortklassen keinerlei Großschreibung verlangen.[7]

3.3 Tabelle 1

WK-Kombinationen	Wortformenbeispiele	Anzahl der types[8]	Textanteil[9]
Gruppe I : Großschreibung beteiligt			
Gruppe I a : Großschreibung homographie-vermeidend			
1 SUB^VRB	= *(angriff, berufe, toene,...)*	549	2,6
2 SUB^VZS	= *{kopf, preis, zustande,...}*	40	0,1
3 SUB^PRP	= *{angesichts, gen, kraft,...}*	*17*	-
4 SUB^KON	= *{ehe, falls, plus}*	*3*	-
5 SUB^IND	= *{all}*	*1*	-
6 SUB^POP	= *{willen}*	*1*	-
7 SIN^INF	= *{alleinsein, dasein.... }*	*8*	-
8 SAD^ADJ	= *{bester, naechste, tote,...}*	6.223	6,6%
Gruppe 1 b : Großschreibung homographie-vermindernd			
9 SUB^SAD^ADJ[10]	= *{geraden, oedem, doktrinaere,...}*	461	*0,7*
10 SIN^INF^VRB	= *{bestehen, zusammenleben,...}*	1.650	4,0
11 SUB^SIN^INF^VRB	= *{graben, stellen, toenen.... }*	402	1,2

12	SIN^INF^VRB^PTZ^ADV	= *{erwachsen, missfallen,...}*	53	0,2
13	SUB^SIN^INF^VRB^PTZ^ADV	= *{berufen, erlassen, gefallen,...}*	16	0,1
14	SIN^INF^VRB^ADV^VZS	= *{albern, beschaffen, verlegen,...}*	*8*	-
15	SUB^SIN^INF^VRB^ADV^VZS	= *{bescheiden, buchen, eichen,...}*	*8*	-
16	SIN^SAD^INF^ADJ^VRB	= *{bessern, billigen, vornehmen,...}*	94	0,3
17	SUB^SIN^SAD^INF^ADJ^VRB	= *(bleichen, werten, wundern,...)*	33	0,1
18	SIN^SAD^INF^ADJ^VRB^IND	= *{einigen}*	*1*	
19	SAD^ADJ^VRB	= *{beachtete, billige, vornehme,...}*	742	1,0%
20	SUB^SAD^ADJ^VRB	= *{bleiche, liebe, werte,...}*	51	0,1%
21	SAD^ADJ^ADV^VZS	= *{schoener, solide, weiter,...}*	934	1,2
22	SUB^SAD^ADJ^ADV^VZS	= *{feige, gerade, wunder,...}*	42	0,1
23	SAD^ADJ^VRB^ADV^VZS	= *{lange, rege, truebe}*	*3*	-
24	SUB^SAD^ADJ^VRB^ADV^VZS	= *{bange, irre, weise,...}*	7	-
25	SUB^VRB^PTZ^ADV	= *{betrieben, genossen, umrissen,...}*	*11*	-
26	SUB^VRB^ADV^VZS	= *{schade}*	*1*	-
27	SUB^VRB^VZS	= *{abstand, mass, schritt,...}*	*6*	-
28	SUB^PTZ^ADV	= *{angeboten, bedacht, gefahren)*	*6*	-
29	SUB^ADV^VZS	= *{arm, doktrinaer, scheu,...}*	165	*0,8%*
30	SUB^ADV^VZS^POP	= *{hinweg, weg}*	*2*	-
31	SUB^ADV^VZS^PRP	= *{abseits, bar, laut,...}*	*10*	-
32	SUB^VZS^PRP	= *{dank, trotz}*	*2*	-
33	SUB^VZS^PRP^KON	= *{statt}*	*1*	-
34	SUB^POP^PRP	= *(wegen)*	*1*	-
35	SIN^INF^VRB^IND	= *{einen}*	*1*	0,3
36	SIN^INF^VRB^KON	= *(sondern)*	*1*	*0,1*
37	SIN^INF^VRB^POS	= *{meinen}*	*1*	-
38	SIN^INF^VRB^IZU	= *{hinzukommen, hinzulegen,...}*	*5*	-
39	SIN^INF^POS	= *{sein}*	1	0,3
40	SAD^ADJ^VRB^IND	= *{einige}*	*1*	-
41	SAD^ADJ^VRB^PTZ^ADV	= *{erblichen}*	*1*	-
42	SAD^ADJ^VRB^ADV^VZS^PRP	= *{nahe}*	*1*	-
43	SAD^ADJ^ADV^VZS^IND	= *{einiger, lauter}*	*2*	-
44	SAD^ADJ^ADV^VZS^PRP	= *{voller}*	*1*	-
45	SAD^ADJ^POP	= *{halber}*	*1*	-

Gruppe II : Großschreibung nicht beteiligt

46	VRB^PTZ^ADV	= *(beachtet, ediert, interviewt,...)*	*801*	1,6
47	VRB^ADV^VZS	= *{aeusserst, brach, traut,...}*	15	-
48	VRB^PTZ^ADV^VZS	= *{verloren}*	*1*	-
49	VRB^IND	= *{eine}*	1	0,7
50	VRB^PRP	= *{anstelle, vermoege,...}*	*3*	-
51	VRB^POS	= *(meine)*	*1*	-
52	PTZ^ADV	= *{ausgesprochen, gepflegt .... }*	1.145	1,0%
53	PTZ^ADV^VZS	= *{gefangen}*	*1*	-
54	ADV^VZS	= *{beisammen, dabei, wieder,...}*	145	1,0
55	ADV^VZS^POP	= *(her, hinaus, zusammen,...}*	19	0,1
56	ADV^VZS^PRP	= *{ausserhalb, fern,...}*	37	0,1
57	ADV^VZS^POP^PRP	= *[entsprechend, gemaess.... }*	*2*	-
58	VZS^POP^PRP	= *{an, auf entlang,...}*	10	3,2
59	ADV^VZS^POP^PRP^ZUA^ZUI	= {zu}	1	2,0
60	ADV^VZS^IND	= *[genug, mehr, weniger,...}*	12	0,4
61	VZS^IND	= *{ein)*	1	0,8%
62	VZS^PRP	= *{bei, unter, vor, wider)*	4	0,5
63	KON^ADV	= *{aber, auch, damit,...)*	20	2,6%
64	KON^VZS	= *(bevor)*	*1*	-
65	KON^PRP	= *{als, bis, ohne, seit}*	*4*	*1,2*
66	KON^ADV^VZS	= *{allein, bloss, da, dagegen}*	4	0,1
67	KON^ADV^VZS^PRP	= *(waehrend)*	*1*	-
68	KON^ADV^VZS^POP^PRP	= *[um}*	1	0,3
69	REL^FRA	= *{wann, wer, wieviel,...}*	30	0,5
70	REL^FRA^PRP^ADV	= *(wie)*	1	0,4
71	REL^DEM	= *{das, den, die,...}*	9	11,7
72	PER^POS	= *{ihr, ihrer, meiner,...}*	12	0,3

Zusammenfassung:

Zahl der HO-Klassen	Zahl der an den WK-Kombinationen beteiligten WK-Alternativen	Anzahl der types	Textanteil
Gruppe la (Nr. 1 bis 8):			
8	16	7.069	9,6%
Gruppe Ib (Nr. 9 bis 45):			
37	158	4.726	11,1
Gruppe II (Nr. 46 bis 72):			
27	77	2.277	28,7%
Insgesamt:			
72	251	14.072	49,4%[11]

3.4 Vergleicht man die Ergebnisse in den einzelnen Gruppierungen miteinander, stellt man fest, daß in Gruppe I die Zahl der types - gegenüber Gruppe II - weitaus höher ausfällt. Die Homographen der Gruppe I erreichen dagegen nicht den Textanteil der Gruppe II. Dies hängt in erster Linie damit zusammen, daß in Gruppe II die Wortklassen DEM, REL, PRP und VZS häufiger als in Gruppe I vertreten sind: Wortklassen, die sich einerseits durch ein begrenztes Repertoire an types, andrerseits durch relativ hohen Textanteil auszeichnen (vgl. allein Nr. 71: 9 types - 11,7 % Textanteil).

3.5 Da die einfache Nennung des Textanteils einer HO-Klasse nichts über die Teilhabe der einzelnen an der Mehrdeutigkeit beteiligten Wortklassenalternativen aussagt, sollen im Folgenden die in Gruppe I am häufigsten vertretenen HO-Klassen diesbezüglich eingehender untersucht werden. Dabei wird auf Werte zurückgegriffen, die an den in Abschnitt 4.2 erwähnten 500 RDE-Sätzen (8.000 tokens) ermittelt worden sind.

3.6 Tabelle 2

WK-Kombinationen	Belege f. die einzelnen WK-Kombinationen	Teilhabe von SUB,SIN,SAD absolut	prozentual	Textanteil v. SUB,SAD,SIN
Gruppe 1 a:				
1 SUB^VRB	144	110	78%	1,9%[12]
8 SAD^ADJ	566	50	9 %	0,7
Gruppe 1 b:				
9 SUB^SAD^ADJ	29	8	27%	0,2%
10 SIN^INF^VRB	285	27	10%	0,4
11 SUB^SIN^INF^VRB	78	44	56%	0,7
12 SIN^INF^VRB^PTZ^ADV	17	1	7 %	0,01%
13 SUB^SIN^INF^VRB^PTZ^ADV	3	1	33 %	0,03%
16 SIN^SAD^INF^ADJ^VRB	34	2	7 %	0,05%
17 SUB^SIN^SAD^INF^ADJ^VRB	4	2	50%	0,05%
19 SAD^ADJ^VRB	62	4	7 %	0,08%

20 SUB^SAD^ADJ^VRB	10	8	80%	0,08%
21 SAD^ADJ^ADV^VZS	55	3	5 %	0,06%
29 SUB^ADV^VZS	46	24	55 %	0,5%
35 SIN^INF^VRB^IND	12	-	-	-
36 SIN^INF^VRB^KON	16	-	-	-
39 SIN^INF^POS	23	2	10 %	0,03%

3.7 Die Verwendung der Großschreibung als wortklassenunterscheidendes Merkmal hat für das in Abschnitt 3.2 behandelte Homographensystem folgende Auswirkungen: Die in Tabelle 1 einheitlich kleingeschriebenen types der Gruppe I werden in jeweils zwei Varianten aufgeteilt, in eine klein- und großgeschriebene: *liebe* z.B. wird zu *liebe* - *Liebe*. Dieser Aufteilung der types in zwei Schreibvarianten entspricht eine Aufteilung der vorher zutreffenden WK-Kombinationen, wobei jeweils eine die Klassenzugehörigkeit der kleingeschriebenen Variante und die andere die der großgeschriebenen bezeichnet: *liebe* = SUB^SAD^ADJ^VRB wird aufgeteilt in *liebe* = ADJ^VRB und *Liebe* = SUB^SAD. Dabei entfällt Gruppe 1 a (Nr. 1 bis 8) vollständig, da ihre Wortformen bei Groß - Kleinschreibung jeweils nur e i n e r Wortklasse zufallen - also wortklasseneindeutig werden. Die Wortformen der Gruppe I b dagegen bleiben auch bei Groß-Kleinschreibung homograph, lediglich die Zahl der HO-Klassen, der WK-Alternativen und der Textanteil der HO-Klassen werden reduziert.

3.8 Nachfolgende Aufstellung reflektiert die Auswirkungen der Groß-Kleinschreibung auf das in Tabelle 1 dargestellte Homographensystem. Der besseren Vergleichsmöglichkeit wegen lehnen wir uns weitgehend an den Aufbau von Tabelle 1 an. Die kleingeschriebenen Varianten der Gruppe I b und die types der Gruppe II (von Tabelle 1) werden jetzt zu e i n e r Gruppe K zusammengefasst. Die großgeschriebenen Varianten der Gruppe I b bilden die Gruppe G.

Der Verlust an Textanteil, der mit dem durch Großschreibung bedingten Wegfall von Wortklassenalternativen einhergeht, ist aufgrund der in Tabelle 2 ermittelten Teilhabe am Textanteil bestimmt worden. Den neu entstandenen oder verbliebenen HO-Klassen stellen wir die entsprechenden HO-Klassen-Nummern aus Tabelle 1 voran. Werden mehrere HO-Klassen aus Tabelle 2 zu e i n e r (in Tabelle 3) zusammengefasst, sind auch die jeweiligen Werte für die Anzahl der types und den Textanteil addiert worden.

3.9 Tabelle 3:

WK-Kombinationen	Wortformenbeispiele	Zahl der types	Textanteil
Gruppe K :			
10/11 INF^VRB	= *{bestehen, graben,...}*	2.052	4,1 % [14]
12/13 INF^VRB^PTZ^ADV	= *{erwachsen, berufen,...}*	69	0,3 %
14/15 INF^VRB^ADV^VZS	= *{albern, bescheiden,...}*	16	-
16/17 INF^VRB^ADJ	= *{bessern, bleichen,...}*	127	0,3
18 INF^VRB^ADJ^IND	= *{einigen}*	*1*	-
19/20 ADJ^VRB	= *{beachtete, bleiche,...}*	793	1,0
21/22 ADJ^ADV^VZS	= *{schoener, feige,...}*	976	1,2
23/24 ADJ^VRB^ADV^VZS	= *[lange, bange,...}*	*10*	-
27 VRB^VZS	= *{abstand, .. }*	*6*	-
33 VZS^PRP^KON	= *{statt}*	*1*	-
34 POP^PRP	= *{wegen}*	*1*	-
35 INF^VRB^IND	= *{einen}*	*1*	0,3
36 INF^VRB^KON	= *(sondern)*	*1*	*0,1*
37 INF^VRB^POS	= *{meinen}*	*1*	-
38 INF^VRB^IZU	= *{hinzukommen,...}*	*5*	-

39	INF^POS	= *{sein}*	1	0,3
40	ADJ^VRB^IND	= *{einige)*	1	-
41	ADJ^VRB^PTZ^ADV	= *{erblichen}*	*1*	-
42	ADJ^VRB^ADV^VZS^PRP	= *{nahe}*	*1*	-
43	ADJ^ADV^VZS^IND	= *{einiger, lauter)*	*2*	-
44	ADJ^ADV^VZS^PRP	= *{voller}*	*1*	-
45	ADJ^POP	= *{halber}*	*1*	-
46/25	VRB^PTZ^ADV	= *{betrieben, beachtet,...}*	*812*	*1,6*
47/26	VRB^IADV^VZS	= *{schade, aeusserst,...}*	*16*	-
48	VRB^PTZ^ADV^VZS	= *{verloren}*	*1*	-
49	VRB^IND	= *{eine}*	*1*	*0,7*
50	VRB^PRP	= *{anstelle,...}*	*3*	-
51	VRB^POS	= *{meine)*	*1*	-
52/28	PTZ^ADV	= *{angeboten, ausgesprochen,...}*	*1.151*	*1,0*
53	PTZ^ADV^VZS	= *{gefangen}*	*1*	-
54/29	ADV^VZS	= *{arm, beisammen,..)*	*310*	*1,3*
55/30	ADV^VZS^POP	= *{hinweg, weg, her,...}*	*21*	*0,1*
56/31	ADV^VZS^PRP	= *{abseits, aufwaerts,...)*	*47*	*0,1*
57	ADV^VZS^POP^PRP	= *{entsprechend, gemaess}*	*2*	-
58	VZS^POP^PRP	= *{an,...}*	*10*	*3,2*
59	ADV^VZS^POP^PRP^ZUA ^ZUI	= {zu}	1	2,0
60	ADV^VZS^IND	= *{genug,...}*	12	0,4
61	VZS^IND	= *{ein}*	*1*	*0,8%*
62/32	VZS^PRP	= *[dank, bei,...}*	6	0,5
63	KON^ADV	= *[aber,...}*	20	2,6
64	KON^VZS	= *{bevor}*	*1*	-
65	KON^PRP	= *{als,...}*	*4*	*1,2*
66	*KON^ADV^VZS*	= *{allein,...}*	*4*	*0,1*
67	*KON^ADV^VZS^PRP*	= *{waehrend)*	*1*	-
68	*KON^ADV^VZS^POP^PRP*	= *{um}*	*1*	*0,3*
69	*REL^FRA*	= *{wann,.. }*	*30*	*0,5*
70	*REL^FRA^PRP^ADV*	= *[wie}*	*1*	*0,4*
71	*REL^DEM*	= *{das,...}*	9	11,7
72	*PER^POS*	= *{ihr,...}*	12	0,3

Gruppe G

9/20/22/24	SUB^SAD	= *(Geraden, Bleiche, Feige, Bange,...)*	561	0,2%
11/13/15	SUB^SIN	= *{Graben, Berufen, Bescheiden,...}*	*426*	0,7%
16/18	SAD^SIN	= *[Bessern, Einigen,...}*	*95*	-
17	SUB^SIN^SAD	= *{Bleichen, Werten, Wundern,...}*	33	-

Zusammenfassung:

Zahl der HO-Klassen	Zahl der an den WK-Kombinationen beteiligten WK-Alternativen	Anzahl der types	Textanteil
Gruppe K			
49	146	6.542	37,5
Gruppe G			
4	9	1.115	1,0% [15]
Insgesamt:			
53	155	7.657	38,5

3.10 Fassen wir die bisherigen Ergebnisse zusammen:

Die Verwendung der Großschreibung als wortklassenunterscheidendes Merkmal verursacht für das in Tabelle 2 dargestellte Homographensystem,

1. eine Verringerung der WK-Kombinationen von 72 auf 53 [16]
2. eine Verringerung der WK-Alternativen in diesen Kombinationen von 251 auf 155.[16]

Das Verhältnis zwischen den WK-Alternativen und den WK-Kombinationen verschiebt sich dadurch von 3,5 : 1 (251: 72) auf 3,0 : 1 (155 : 53).[16]

Im untersuchten T e x t m a t e r i a l ergibt die Verwendung der Großschreibung

1. ein Abnehmen des HO-Textanteils von 49,4 %[16] auf 38,5 %,
2. eine Verminderung der homographen types [17] von 14.072 [16] auf 7.657.

Die Verwendung der Großschreibung würde demnach eine Verringerung der Zahlenwerte um ca. ein Drittel (bezogen auf das H o m o g r a p h e n - s y s t e m) und um etwa ein Viertel (hinsichtlich der Mehrdeutigkeiten im untersuchten T e x t m a t e r i a 1) ergeben. Diese Auswirkungen können unter Umständen die Erstellung eines Reduktionsalgorithmus erheblich erleichtern bzw. dessen Erfolgsquoten bei der Reduktion von Mehrdeutigkeiten erhöhen.

4. Auswirkungen einer Verwendung der Großschreibung auf den Umfang des Reduktions-Algorithmus und die erzielten Lösungsquoten bei der Disambiguierung.

4.1 Als Maßstab für die Auswirkungen einer Verwendung der Großschreibung auf das Saarbrücker Verfahren soll die Anzahl der zur Homographenlösung vorgesehenen Programminstruktionen gelten. Mit diesem Kriterium lassen sich der durch unsere Schreibregelung zusätzlich erforderliche Programmieraufwand und der entsprechend dazu vorgesehene Maschinenspeicherraum am zuverlässigsten quantifizieren. Ein Nachteil dieses Kriteriums liegt allerdings darin, dass aus der Anzahl der Instruktionen deren Inhalt nicht ersichtlich ist. Die Komplexität eines Reduktionsvorgangs wird damit also nur sehr oberflächlich wiedergegeben. Da diese Einschränkung jedoch auf alle Teile des Reduktionsalgorithmus gleichermaßen zutrifft, erscheint es uns sinnvoll, die Anzahl der Programmierinstruktionen - neben anderen Daten - als Maßstab zur Beurteilung des Komplexitätsgrades von Homographenlösungen heranzuziehen.

4.2 Auszählungen haben ergeben, dass eine Verwendung der Großschreibung den Reduktionsalgorithmus in seiner bestehenden Konzeption um etwa 680 Programminstruktionen verringern würde (aus Gründen der Platzersparnis haben wir diesbezüglich auf eine genauere Darstellung verzichtet). Der zur Reduktion der übrigen Wortklassenmehrdeutigkeiten erforderliche - noch verbleibende - Algorithmus bestünde dann aus ca. 3.800 Instruktionen. Die durch Verwendung der Großschreibung erzielte Einsparung betrüge demnach etwa ein Sechstel.

4.3 Die Auswirkungen der Großschreibung auf die Höhe der Lösungsquoten können - in stärkerem Maße als in den vorangegangenen Untersuchungen - nur annäherungsweise angegeben werden, da Veränderungen der Lösungsquoten nur im Hinblick auf die Veränderung jeweils einer Bedingung (z.B. Rückgang des Homographen-Textanteils, Verminderung der WK-Kombinationen, Verringerung der WK-Alternativen in diesen WK-Kombinationen,...) errechnet werden können.

Zum Zeitpunkt der Stichprobe, auf deren Ergebnisse im folgenden zurückgegriffen wird, haben sich zudem nicht alle Programmteile des Reduktionsalgorithmus im gleichen Zustand befunden - dies gilt sowohl für die linguistische Basis als auch für die technische Ausführung. Einer Argumentation mit Lösungsquoten allein kann daher nur bedingt Beweiskraft zugestanden werden. Deshalb müssen noch andere Daten (vor allem die zur Lösung benötigte Anzahl Programminstruktionen) herangezogen werden. Wenn sich z.B. zeigen sollte, dass gewisse WKKombinationen trotz relativ hohen Programmieraufwandes niedrigere Lösungsquoten erreichen als andere WK-Kombinationen, die mit wenigen Instruktionen höhere Lösungsquoten aufweisen, kann dies als Indiz für unterschiedliche Lösungsschwierigkeit gewertet werden (wenn wir davon absehen, dass dies ebenso auf eine gewisse "Umständlichkeit" bzw. Ineffizienz beim Programmieren hinweisen könnte).

4.4 Für die Lösung der Mehrdeutigkeiten in Gruppe I a (Tabelle 1) sind insgesamt 292 Programminstruktionen vorgesehen. Damit wird eine durchschnittliche Lösungsquote[18] von 94 % erreicht. Die Mehrdeutigkeiten der Gruppe I b beanspruchen zur Lösung 1.724 Instruktionen; 390 werden zur Ermittlung der WK-Alternativen SUB (+SIN) benötigt. Die durchschnittliche Lösungsquote beträgt hier 88 %. In Gruppe II machen die vorhandenen Programmteile ca. 2.400 Programmanweisungen aus; die durchschnittliche Lösungsquote liegt bei 95 %.

4.5 Ein Vergleich zwischen den durchschnittlichen Lösungsquoten in den drei Gruppen zeigt, dass sich die Werte von Gruppe I a und II etwas über dem Gesamtdurchschnitt von 93,5 % bewegen, während Gruppe I b mit 88 % deutlich unter diesem Durchschnitt liegt. Es ist anzunehmen, dass zwischen der Anzahl der WK-Alternativen (die ja in Gruppe I b ziemlich hoch liegt) und der Höhe der Lösungsquote für eine WK-Kombination eine Abhängigkeit besteht.[19] Mit der durch Großschreibung verursachten Verringerung der WK-Alternativen pro HO-Klasse (bzw. WK-Kombination) müsste demnach eine entsprechende Verbesserung der Lösungsquoten einhergehen. Weitere Untersuchungen haben indes ergeben, dass eine Korrelation zwischen der Zahl der WK-Alternativen pro Kombination und der Höhe der Lösungsquote nicht immer festzustellen ist: So gelingt die Reduktion der Mehrdeutigkeit Nr. 59 (6 WK-Alternativen) mit 230 Programminstruktionen zu 96 %, während sich die Reduktion von Nr. 10/11 (3 WK-Alternativen) mit 722 Programmanweisungen zu nur 86 % bewerkstelligen lässt. Mit der Anzahl der WK-Alternativen pro Kombination ist also der Schwierigkeitsgrad für die Auflösung einer WK-Mehrdeutigkeit nicht in jedem Fall ausreichend angegeben. Wahrscheinlicher ist, dass die Höhe der Lösungsquote von der Wortklassenzugehörigkeit der an der Kombination beteiligten Alternativen abhängt. Wir haben damit ein Resultat vorweggenommen, das im Folgenden (Tabelle 4) belegt werden soll. Dabei stellen wir verschiedene WK-Kombinationen mit einigen inzwischen (vgl. Tabelle 1,2,3) bewerteten Daten zusammen. Die betreffenden WK-Kombinationen machen den weitaus größten Teil der in Tabelle 1 vorgelegten Wortklassenmehrdeutigkeiten aus.

4.6 Tabelle 4

WK-Kombinationen	types	Textanteil	Progr. Instruktionen	Lösungsquoten
1 (2)	549	2,6%	140	92%
9 (2)	461	0,7%	125	93%
10/11[20] (3)	2.052	5,2%	722	86%
12 (5)	53	0,2%	1.253	85%

13	(5)	16	0,1%	1.257	77%
16/19	(4)	836	1,3%	896	85%
17/20	(4)	84	0,2%	916	75%
21	(3)	934	1,2%	102	92%
29	(3)	165	0,8%	157	91%
35	(4)	1	0,3%	898	100%
36	(4)	1	0,1%	737	93%
39	(3)	1	0,3%	143	94%
46	(3)	801	1,6%	784	87%
49	(2)	1	0,7%	174	94%
52	(2)	1.145	1,0%	535	94%
54	(2)	145	1,0%	94	93%
58	(3)	10	3,2%	46	97%
59	(6)	1	2,0%	230	96%
61	(2)	1	0,8%	75	96%
63	(2)	20	2,6%	55	96%
65	(2)	4	1,2%	258	92%
71	(2)	9	11,7%	273	99%

5. Zusammenfassung

5.1 Die Verwendung der Großschreibung verringert das in Tabelle 1 dargelegte Homographensystem um etwa ein Drittel seines bisherigen Umfangs. Die Abnahme des HO-Textanteils ist demgegenüber geringer (ca. ein Viertel). Der bestehende Reduktionsalgorithmus wird um ca. ein Sechstel seines jetzigen Umfangs verkleinert: Die Großschreibung erbringt also in erster Linie klassifikatorische Erleichterungen - die Bearbeitung von WK-Mehrdeutigkeiten in aktualisierten Texten wird nicht in gleichem Maße begünstigt.

5.2 Zwar ist die Bedeutung der Großschreibung bei der Vermeidung oder Verminderung syntaktischer Mehrdeutigkeiten nicht zu übersehen: Die Analyse (auf Wortformenebene) wird durch sie spürbar erleichtert. Überschätzen darf man indes die Hilfe der Großschreibung nicht; viele schwierig lösbaren Mehrdeutigkeiten bleiben trotz Großschreibung bestehen; auf andere Typen syntaktischer Mehrdeutigkeit, die eine syntaktische Analyse häufig mehr als WK-Mehrdeutigkeiten erschweren, hat Großschreibung ebenfalls keinen direkten Einfluss. Die WK-Kombinationen, die durch Großschreibung beseitigt werden können, sind meist mit geringem Programmieraufwand und relativ häufig korrekt zu lösen.

Anmerkungen

1 Siehe (4), S. 13.

2 Vgl. dazu (1), S. 26 f. und S. 137 ff. Es ist nicht beabsichtigt, alle Typen von Mehrdeutigkeit aufzuzählen oder sie zu ordnen. Zum Problem der Mehrdeutigkeit auf Wortformenebene vgl. weiter: (1), S. 144; (2), Kapitel 6.1; (3), S. 2 ff.

3 Da das Saarbrücker Wortformenbuch keine lexikalisch-semantischen Bedeutungen verzeichnet, erübrigt sich hier die gelegentlich vorgenommene Unterscheidung zwischen "Homonymie" bzw. "Homographie" und "Polysemie". Maßgebend für die Konstituierung des Begriffs "Homograph" sind Wortklassenzugehörigkeit und Graphemfolge.

4 Vgl. (1), S. 15, S. 17, S. 26 und 27.

5 Auswahl und Zusammensetzung dieses Testmaterials siehe in (2), Kapitel 6.1.4.

6 Die 500 zusätzlichen RDE-Sätze dienen auch als Kontrollgruppe für die Ergebnisse des ersten Samples (vgl. dazu Tabelle 2). Es handelt sich hier um die RDE-Sätze 3.000 bis 3.500 (zu je 16 tokens).

7 Die generelle Möglichkeit der Substantivierung - bzw. Großschreibung - lassen wir - bis auf die von INF(SIN) und ADJ(SAD) - außer acht. Wendungen wie *das Ob und Wie, Wenn und Aber, Auf und Ab* werden demnach nicht berücksichtigt. Ebenfalls vernachlässigt wird die Großschreibung am Satzanfang bzw. nach Doppelpunkt oder Semikolon. Zählungen haben uns darin bestärkt, diese Einschränkungen vorzunehmen. Mit den Klassen SUB, SIN und SAD erfassen wir etwa 80% aller Fälle von Großschreibung in unseren Texten (einschließlich der Großschreibung am Satzanfang).

8 Die Angaben für die Anzahl der types gelten für das Saarbrücker syntaktische Wortformenbuch. Dieses bietet jedoch nur eine Auswahl des deutschen Wortformenbestandes. Mit Sicherheit sind daher einige Zahlenangaben in dieser Rubrik - vergleicht man sie etwa mit Zahlenwerten, die anhand zusätzlicher Lexika (z.B. DUDEN, Bd. 1; MACKENSEN) gewonnen sind - zu niedrig ausfallen. Die betrifft vor allem die HO-Klassen Nr. 1, 8, 10, 11, 19, 21, 46 und 52. Für die HO-Klassen Nr. 1 z.B. können - nach unseren Schätzungen - ca. 1.500 types berechnet werden.

9 Werte unter 0,1 % sind in dieser Tabelle nicht aufgeführt, werden jedoch in der Zusammenstellung im Anschluss an die Tabelle und bei allen folgenden Auszählungen mitberücksichtigt.

10 Großgeschriebene Wortklassen sind fett gedruckt.

11 In (2), Kapitel 6.1.4 wurde ein Homographentextanteil von ca. 43 % angegeben. Die Diskrepanz zwischen diesen beiden Werten (49,4 % und 43 %) ist dadurch zu erklären, dass in (2) auf die Mehrdeutigkeit Nr. 8 (SAD^ADJ) verzichtet worden ist. Vgl. dazu Tabelle 2, Anm. 13. Nach Abzug des Textanteils von Nr. 8 ergibt sich für obige Zusammenfassung ein Homographentextanteil von 42,8 % (49,4 % minus 6,6%), der mit dem in (2) angegebenen Wert vergleichbar ist.

12 Der Textanteil der Klassen SUB, SIN, SAD in einer WK-Kombination ist anhand der Teilhabe dieser Klassen an der WK-Kombination errechnet worden. Als Grundwert gilt dabei die Angabe für den Textanteil in Tabelle 1. Die Teilhabe an der WK-Kombination ist - wie schon gesagt - an 500 RDE-Sätzen bestimmt worden. Im folgenden ein Beispiel für die Errechnung des Textanteils von SUB, SIN und SAD: Der Textanteil der HO-Klasse Nr. 1 (SUB^VRB) beträgt laut Tabelle 2 um 2,6 %. Die Teilhabe von SUB, SIN und SAD an der WK-Kombination beträgt - laut Tabelle 3 - um 78 %. Der Textanteil von SUB, SIN und SAD ergibt dann (78 % von 2,6 % Textanteil) ca. 1,9 %.

13 Von der generellen Möglichkeit der Substantivierung des Adjektivs (SAD) - vgl. dazu Tabelle 1, Anm. 7 und Tabelle 2, Anm. 12 - wird im untersuchten Textmaterial verhältnismäßig

wenig Gebrauch gemacht (9 % bei Nr. 8; 7 % bei Nr. 19). Dieser Umstand und die Überlegung, dass eine Unterscheidung zwischen ADJ und SAD (zwischen *der naechste* und *der Naechste* z.B.) syntaktisch nicht unbedingt erforderlich ist - im Gegensatz zur Unterscheidung zwischen INF und SIN -, haben uns bewogen, für unser Verfahren auf die Wortklasse SAD - und damit auch auf die Homographie ADJ/SAD - zu verzichten. Diese Untersuchung hat die generelle Möglichkeit der Substantivierung von ADJ jedoch berücksichtigt.

14 Die Einbuße an Textanteil, die durch den Wegfall der WK-Alternative SUB, SIN und SAD verursacht würde, betrüge z.B. in Klasse 10 ca. 0,4 %: Der Textanteil dieser HO-Klasse beträgt - nach Tabelle 1 - 4,0 %, die Teilhabe der Wortklasse SIN daran macht etwa 10 % aus (siehe Tabelle 2): das ergibt einen Textanteil von 0,4 %. Die Werte für die übrigen HO-Klassen sind auf gleiche Weise errechnet.

15 Laut Tabelle 2 beträgt der Textanteil der Klassen SUB, SIN, SAD in den HO-Klassen 9, 11 und 16 etwa 0,2 %, 0,7 % und 0,05 %. Für Gruppe G kann also ein Textanteil von ca. 1,0 % veranschlagt werden.

16 Da in unserem Verfahren - im Gegensatz zu dieser Untersuchung - auf die WK-Alternativen SAD und SIN und damit auf die Mehrdeutigkeiten SUB^SAD^SIN... und SAD^ADJ ... verzichtet worden ist, ergeben sich hierfür auch andere Werte:

Die WK-Kombinationen werden von 64 auf 49 vermindert, die WK-Alternativen in diesen Kombinationen von 203 auf 146. Das Verhältnis zwischen den WK-Alternativen und den WK-Kombinationen beträgt dann noch (statt vorher 3,2:1) 2,9:1.

Im untersuchten Textmaterial ergeben sich folgende Veränderungen: Der HO-Textanteil von 42,8 % (vgl. Anm. 11) wird auf 38,5 % reduziert, die homographen types von 7.622 auf 6.542.

17 Dies heißt natürlich nicht, dass sich die Anzahl der Wörterbucheinträge insgesamt vermehrt - im Gegenteil: durch Zuordnung einer Wortform in eine HO-Klasse werden ja mehrfache Wörterbucheinträge überflüssig -, sondern dass eine Umschichtung des Wortformenbestandes vorgenommen wird.

18 Bei der Errechnung der durchschnittlichen Lösungsquoten wurden jeweils Lösungsquote und Textanteil der einzelnen HO-Klassen aufeinander bezogen.

19 Das in den drei Gruppierungen unterschiedliche Verhältnis zwischen der Zahl der WK-Alternativen (WKA) und der WK-Kombinationen (WKK) bestätigt dies offenbar:

Gruppe 1 a: 16 WKA : 8 WKK - 2,0; 1 - Lösungsquote 94 %.
Gruppe 1 b: 158 WKA : 37 WKK - 4,3 : 1 - Lösungsquote 88 %.
Gruppe II: 77 WKA : 27 WKK - 2,9 : 1 - Lösungsquote 95 %.

20 Wenn WK-Kombinationen im Reduktionsalgorithmus den gleichen Programmteil durchlaufen, werden sie hier zusammengefasst. Vgl. auch Anm. 16.

Literatur

(1) AGRICOLA, Erhard
Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und Englischen. In: Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung, Nr. 12 (1968).

(2) EGGERS, Hans und Mitarbeiter
Elektronische Syntaxanalyse der deutschen Gegenwartssprache. Tübingen 1969.

(3) MAAS, Heinz Dieter
Homographie und maschinelle Sprachübersetzung. In: Linguistische Arbeiten 8 (1969), Saarbrücken.

(4) ROSENGREN, Inger
Ein Frequenzwörterbuch der modernen Zeitungssprache - wie und wozu? In: Beiträge zur Linguistik und Informationsverarbeitung 14 (1968), S. 7-21.

(5) TRNKA, Bohumil
Bemerkungen zur Homonymie.
In: Travaux du Cercle Linguistique de Prague 4 (1931), S. 152 - 156.