Für den Minimalwert wurden x_{min} = 50 Token erfasst. Die Spannweite der insgesamt genutzten Token betrug *R* = 1 526. Der Durchschnitt lag bei *M* = 624.72 Token und der Median lag bei x_{med} = 515.5 Token (Abb. 21).

Lemma-Types

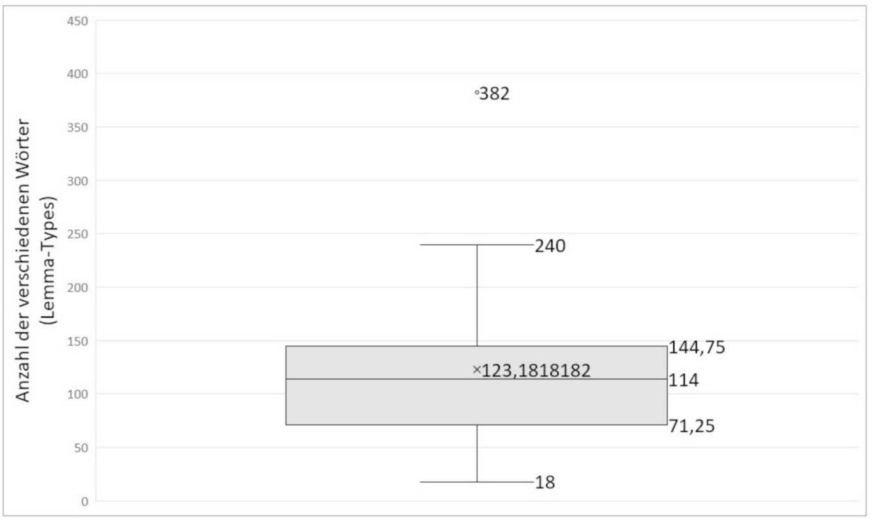

Abb. 22: Verteilung der erfassten Lemma-Types im Primärkorpus (N = 22)

Die Hälfte der Untersuchungsgruppe nutzte zwischen dem Interquartilsabstand (IQA) IQA_1 = 71.25 und IQA_3 = 144.75 Lemma-Types. Der Maximalwert lag bei x_{max} = 240 Lemma-Types. Für den Minimalwert wurden x_{min} = 18 Lemma-Types erfasst. Die Spannweite der Lemma-Types betrug *R* = 222. Der Durchschnitt lag bei *M* = 123.18 Lemma-Types und der Median lag bei x_{med} = 114 Lemma-Types. Der Wert 382 Lemma-Types wurde gemäß der Beschreibung von Boxplots als Ausreißer gewertet. Vermutlich lagen die sprachlich-kommunikativen Fähigkeiten des Schülers/der Schülerin über denen der Untersuchungsgruppe (Abb. 22).

12.1 Kernvokabular

Einzelauswertung Primärkorpus

Das Kernvokabular wurde pro Fall nach der 80 %-Marke von Boenisch (2014b) berechnet. Bei der Betrachtung der 80 %-Marke in Relation zur Häufigkeit pro Lemma-Type konnte eine verhältnismäßig geringe Anzahl an Mindestnennungen (*H*) identifiziert werden. Mithilfe der Analyse der Lemma-Types Beziehungen zeigte sich, dass bei über der Hälfte

der Untersuchungsgruppe ($n = 12$) mehr als ein Drittel der Lemma-Types (35 %) aus dem Gesamtwortschatz innerhalb der 80 %-Marke lagen und nur ein- bis dreimal ($H_{min} \leq 3$) gebraucht wurden (Tab. 36).

Tab. 36: Primärkorpus: Auswertung der 80 %-Marke pro Fall in Relation zu Token, Häufigkeit (H), Lemma-Types, Rang (R)

	80 %-Marke				
Fall	**Token**	**H**	**Lemma-Types**	**%-Anteil**	**R**
S002	870 (872)	6 (95)	37 (121)	30.58 %	34 (97)
S005	203 (204)	3 (31)	22 (59)	37.29 %	19 (32)
S006	139 (139)	2 (50)	16 (39)	41.03	16 (29)
S008	823 (822)	6 (85)	35 (133)	26.32	35 (94)
S011	314 (316)	4 (59)	21 (83)	25.30	18 (33)
S012	1 261 (1 259)	6 (182)	48 (204)	23.53	45 (122)
S013	129 (130)	2 (18)	22 (47)	46.81	21 (29)
S027	1 082 (1 083)	4 (97)	74 (240)	30.83	59 (151)
S036	286 (286)	2 (24)	61 (131)	46.56	39 (63)
S038	896 (894)	8 (102)	29 (128)	22.66	28 (75)
S039	774 (774)	3 (60)	67 (200)	33.50	56 (119)
S042	558 (558)	3 (68)	55 (160)	34.38	43 (86)
S053	511 (511)	4 (78)	36 (107)	33.64	36 (73)
S054	231 (232)	2 (25)	30 (78)	38.46	28 (40)
S084	216 (217)	2 (45)	25 (69)	36.33	25 (35)
S085	824 (824)	1 (146)	176 (382)	46.07	136 (136)
S086	178 (178)	1 (22)	40 (85)	47.06	40 (40)
S094	555 (554)	3 (56)	51 (145)	35.17	44 (87)
S099	40 (40)	2 (8)	12 (18)	67.67	4 (17)
S100	42 (42)	1 (8)	23 (34)	67.65	12 (12)
S116	808 (809)	4 (118)	39 (144)	27.08	35 (95)
S125	254 (255)	2 (30)	50 (104)	48.08	38 (59)

Anmerkungen: Token: Wenn der tatsächliche Wert (in Klammern) beispielsweise 2 Token vor oder nach dem errechneten Wert (ohne Klammer) lag, wurde der höhere Wert für die Bestimmung der 80 %-Marke sowie der Vergabe des Rangplatzes gewählt.
Häufigkeit: In der Klammer ist die maximale Häufigkeit (Nennungen) eines Wortes aufgeführt.
Lemma-Types: In der Klammer sind die Lemma-Types gesamt aufgeführt, sodass die 80 %-Marke in Relation betrachtet werden kann.
Rang: In der Klammer ist der höchste Rang (R_{max}) aufgeführt.

Gesamtliste Primärkorpus

Die erhobenen Einzellisten ($N = 22$) wurden zu einer Gesamtliste zusammengefügt. Insgesamt zählten *13 734 Token* und *877 Lemma-Types* zur Gesamtliste ($R_{max} = 562$).

Anhand der Verteilung der absoluten Häufigkeiten im Liniendiagramm ließ sich erkennen, dass eine begrenzte Anzahl an Lemma-Types ($n = 103$) aus dem Gesamtwortschatz hoch frequent genutzt wurde (Abb. 23). Die Häufigkeit des Token *ich* (Rang 1) lag bei $H = 985$ und von *Bruder* ($R = 140$) bei $H = 13$.

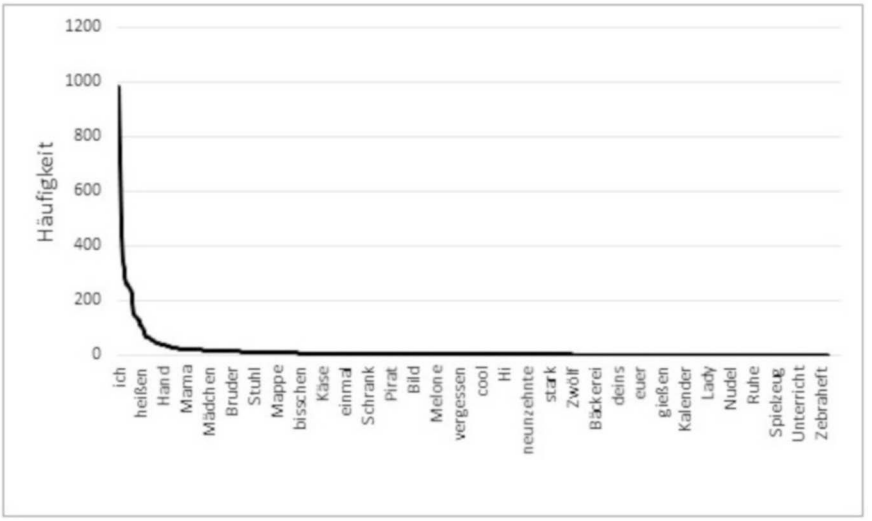

Abb. 23: Absolute Häufigkeiten Gesamtliste Primärkorpus ($N = 22$, $\Sigma = 13\,734$ Token)
Anmerkungen: Ränge (R): ich $R = 1$; heißen $R = 29$; Hand; $R = 56$; Mama $R = 80$; Mädchen $R = 108$; Bruder $R = 140$; Stuhl $R = 161$; Mappe R = 185; bisschen $R = 223$

Von dieser Gesamtliste wurde das *Kernvokabular* in Anlehnung an die Operationalisierung von Boenisch (2014b) berechnet. Nach der 80 %-Marke zählten *10 987 Token* sowie *112 Lemma-Types* zum Kernvokabular. Am häufigsten wurde das Wort *ich* ($H = 985$) verwendet. Am wenigsten wurde das Wort *Schere* ($H = 17$) genutzt (Tab. 37).

Tab. 37: Kernvokabular 80 %-Marke Primärkorpus

Kernvokabular 80 %-Marke ($N = 22$)	
Token-80 %	10 987 (10 992)
Token-Gesamt	13 734
H_{max}	985
H_{min}	17

Kernvokabular 80 %-Marke (*N* = 22)	
R-80 %	108 (562)
Lemma-Types-80 %	112 (877)

Da insgesamt sieben Lemma-Types auf Rang 108 mit *H* = 17 platziert waren, zur 80 %-Marke aber nur fünf Lemma-Types zählten, mussten zwei weitere Einschlusskriterien zur Auswahl der Kernvokabularwörter formuliert werden. Zum einen wurden die Lemma-Types mit der größten Streuung in der Untersuchungsgruppe berücksichtigt. Daher wurden die Wörter *heute*, *Schere*, *dies* und *Mädchen* im Kernvokabular aufgenommen. Andererseits wurde bei den Lemma-Types mit gleicher Streuungsanzahl (*leise*, *fünf*) nach der Wortart mit der potentiell größeren Verwendungshäufigkeit entschieden. Somit wurde das Wort *leise* anstelle von *fünf* innerhalb der 80 %-Marke eingeschlossen. Das siebte Lemma-Type auf Rang 108 war *links*, wurde nur von einem Kind verwendet und wurde deshalb nicht im Kernvokabular berücksichtigt.

Einen Überblick zum Kernvokabular entsprechend der 80 %-Marke bei Kindern im anfänglichen Erwerb von Deutsch als Zweitsprache liefert die nachfolgende Tabelle 38.

Tab. 38: Kernvokabular entsprechend der 80 %-Marke

Wort	Wortart	*H*	*S*	*R*	Wort	Wortart	*H*	*S*	*R*
ich	Pronom.	*985*	*21*	*1*	Bleistift	Subst.	34	12	60
sein	Verb	*744*	*20*	*2*	jetzt	Adverb	34	11	60
das	Artikel	*696*	*21*	*3*	wippen	Verb	32	5	62
ja	Partikel	*489*	*20*	*4*	Tschüss	Interjek.	31	11	63
nein	Partikel	*435*	*20*	*5*	zwei	Zahlwort	31	10	63
du	Pronom.	*339*	*20*	*6*	kein	Pronom.	30	8	65
die	Artikel	*314*	*19*	*7*	dann	Adverb	29	13	66
nicht	Adverb	*284*	*19*	*8*	Jahr	Subst.	28	8	67
Frau	Subst.	*268*	*15*	*9*	es	Pronom.	27	9	68
der	Artikel	*262*	*18*	*10*	bei	Präposi.	26	6	70
haben	Verb	*261*	*18*	*11*	Nummer	Subst.	26	5	70
okay	Adverb	*256*	*13*	*12*	malen	Verb	25	7	72
was	Pronom.	*251*	*20*	*13*	klein	Adjektiv	24	11	73
gucken	Verb	*251*	*17*	*13*	in	Präposi.	24	11	73
so	Adverb	*237*	*17*	*15*	Stift	Subst.	23	12	75
hallo	Interjek.	*236*	*20*	*16*	schon	Adverb	23	6	75
mein	Pronom.	*227*	*15*	*17*	warum	Pronom.	23	6	75

Wort	Wortart	*H*	*S*	*R*
hier	Adverb	*149*	*18*	*19*
machen	Verb	*147*	*16*	*20*
und	Konjunk.	*145*	*17*	*21*
auch	Adverb	*137*	*17*	*22*
mal	Adverb	*135*	*13*	*23*
wie	Pronom.	*130*	*18*	*24*
da	Adverb	*126*	*16*	*25*
warten	Verb	*126*	*15*	*25*
gut	Adjektiv	*111*	*17*	*27*
können	Verb	*110*	*14*	*28*
heißen	Verb	*103*	*17*	*29*
alle	Adverb	*101*	*15*	*30*
kommen	Verb	*100*	*16*	*31*
gehen	Verb	*87*	*15*	*32*
wollen	Verb	*77*	*10*	*33*
fertig	Adjektiv	*69*	*14*	*34*
bitte	Partikel	*68*	*10*	*35*
zu	Adverb	*68*	*9*	*35*
hey	Interjek.	*66*	*6*	*37*
spielen	Verb	*63*	*14*	*38*
wissen	Verb	*62*	*9*	*39*
sagen	Verb	*50*	*14*	*45*
alt	Adjektiv	*47*	*8*	*46*
doch	Adverb	*46*	*16*	*47*
Pause	Subst.	*45*	*9*	*48*
noch	Adverb	*43*	*12*	*49*
oder	Konjunk.	*42*	*8*	*50*
dein	Pronom.	*41*	*11*	*51*
mir	Pronom.	*40*	*10*	*52*
rot	Adjektiv	*39*	*8*	*53*
geben	Verb	*37*	*10*	*54*
helfen	Verb	*37*	*8*	*54*
Hand	Subst.	*36*	*12*	*56*
müssen	Verb	*36*	*9*	*56*
dürfen	Verb	*36*	*8*	*56*
mit	Präposi.	*35*	*10*	*59*

Wort	Wortart	*H*	*S*	*R*
bis	Adverb	23	4	75
Mama	Subst.	22	10	80
Morgen	Subst.	22	9	80
Kleber	Subst.	22	8	80
Apfel	Subst.	22	7	80
gelb	Adjektiv	22	6	80
los	Adverb	22	6	80
blau	Adjektiv	22	4	80
schwarz	Adjektiv	22	3	80
rechnen	Verb	22	2	80
nehmen	Verb	21	7	89
Auto	Subst.	21	3	89
Drache	Subst.	21	1	89
von	Präposi.	20	10	92
sechs	Zahlwort	20	9	92
Klasse	Subst.	20	8	92
essen	Verb	20	8	92
nur	Adverb	20	6	92
mich	Pronom.	20	5	92
danke	Partikel	19	9	98
auf	Adverb	19	9	98
Wasser	Subst.	19	8	98
Stopp	Interjek.	19	8	98
nichts	Pronom.	19	7	98
Mikrofon	Subst.	19	5	98
Fahrrad	Subst.	19	3	98
Tafel	Subst.	18	9	105
lassen	Verb	18	7	105
Fahrrad	Subst.	19	3	98
Buch	Subst.	18	6	105
heute	Adverb	17	10	108
Schere	Subst.	17	9	108
dies	Pronom.	17	7	108
Mädchen	Subst.	17	7	108
leise	Adjektiv	17	5	108

Anmerkungen: *H* = Häufigkeit; *S* = Streuung: (Wie viele Kinder haben das Wort genutzt?); *R* = Rang

Es zeigte sich, dass der Abstand zwischen den Rängen nach der 80 %-Marke deutlich anstieg ($M = 35$ Ränge, Ausnahme: $R = 123$, $R = 152$, $R = 161$), da sich immer mehr verschiedene Wörter die gleiche Anzahl an Häufigkeiten teilten (Tab. 39)

Tab. 39: Rangfolge nach der 80%-Marke (R = Rang; H = Häufigkeit)

R	*H*	Lemma-Types
108	17	2
115	16	8
123	15	7
130	14	10
140	13	12
152	12	9
161	11	11
172	10	13
185	9	20

R	*H*	Lemma-Types
205	8	18
223	7	20
243	6	31
274	5	42
316	4	39
354	3	74
429	2	133
562	1	316

Nach der 80 %-Marke (TOP 108) nahm die Häufigkeitsverteilung pro Wort kontinuierlich ab. Insbesondere nach den 50 am häufigsten genutzten Lemma-Types (TOP 50) ließ sich eine abrupte Absenkung in der Häufigkeitsverteilung erkennen (Abb. 24).

Die Spannweite über die Streuung (*Wie viele Kinder benutzten das Lemma-Type?*) betrug $R = 21$ ($x_{max} = 22 - x_{min} = 1$). Konkret ließ sich über die Spannweite ableiten, dass das Kernvokabular innerhalb der 80 %-Marke von $n = 1$ bis $n = 22$ Fälle genutzt wurde.

Vor diesem Hintergrund wurde das Kernvokabular unter Beachtung des Streuungswertes ≥50 % ($n = 11$) analysiert und auf den Datensatz der 80 %-Marke angewendet. Zu dem Kernvokabular nach der 80 %-Marke plus dem Streuungswert ≥50 % zählten 9 088 Token. Die maximale absolute Häufigkeit lag bei $H = 985$ (*ich*). Die geringste absolute Häufigkeit betrug $H = 23$ (*Stift*). Insgesamt zählten *48 Lemma-Types* zum Kernvokabular ($Liste_{Kv}$reduziert) (Tab. 40). Der Anteil am Gesamtkorpus lag bei 66.17 %.

Unter Verwendung des Streuungskriterium betrug die Differenz zwischen den erhobenen Kernvokabularlisten -64 Lemma-Types. Die Differenz zwischen der kleinsten Anzahl absoluter Nennungen lag bei -6. Anhand Tabelle 41 wird deutlich, welche Wörter zum Kernvokabular unter zusätzlicher Beachtung des Streuungskriteriums gehörten (reduzierte Kernvokabularliste). Erst ab Rang 34 traten Unterschiede innerhalb der Rangverteilung zwischen den Kernvokabularlisten mit und ohne Streuungskriterium auf.

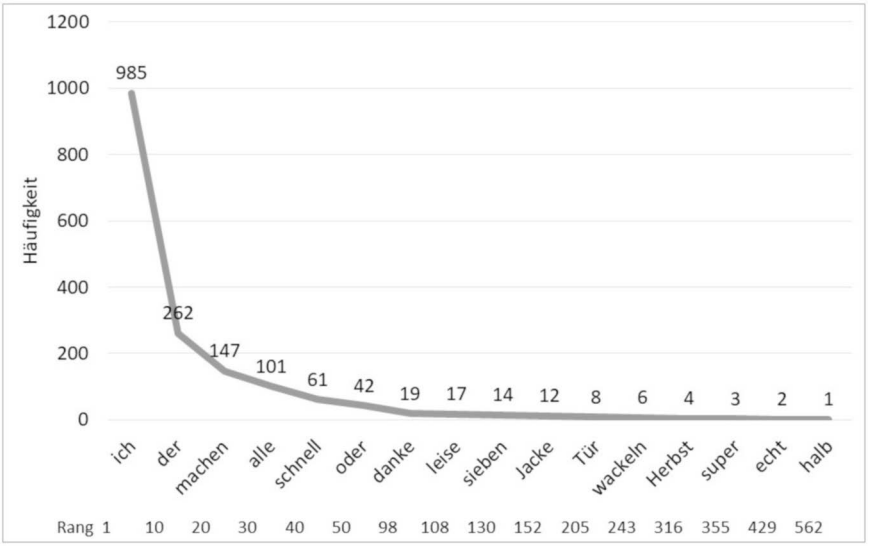

Abb. 24: Häufigkeitsverteilung innerhalb und außerhalb der 80 %-Marke (N = 22, Σ = 13 734 Token)

Tab. 40: Reduzierte Kernvokabularliste nach Anwendung des Streuungskriteriums ≥50 %

Reduzierte Kernvokabularliste	
Token	9 088
H_{max}	985
H_{min}	23
Lemma-Types	48

Anmerkungen: H_{max} = maximale Häufigkeit, H_{min} = minimale Häufigkeit

Tab. 41: Reduzierte Kernvokabularliste, Vergleich der Ränge

Wort (WA)	***H (S)***	***R* (80 %-Marke)**	***R* (≥50 %)**
ich (Pron.)	985 (21)	1	1
sein (Verb)	744 (20)	2	2

Wort (WA)	***H (S)***	***R* (80 %-Marke)**	***R* (≥50 %)**
da (Adverb)	126 (16)	25	25
warten (Verb)	126 (15)	25	25

Wort (WA)	*H (S)*	*R* (80 %-Marke)	*R* (≥50 %)
das (Artikel)	696 (21)	3	3
ja (Partikel)	489 (20)	4	4
nein (Partikel)	435 (20)	5	5
du (Pron.)	339 (20)	6	6
die (Artikel)	314 (19)	7	7
nicht (Adverb)	284 (19)	8	8
Frau (Subst.)	268 (15)	9	9
der (Artikel)	262 (18)	10	10
haben (Verb)	261 (18)	11	11
okay (Adverb)	256 (13)	12	12
was (Pron.)	251 (20)	13	13
gucken (Verb)	251 (17)	13	13
so (Adverb)	237 (17)	15	15
hallo (Interjek.)	236 (20)	16	16
mein (Pron.)	227 (15)	17	17
ein (Artikel)	190 (18)	18	18
hier (Adverb)	149 (18)	19	19
machen (Verb)	147 (16)	20	20
und (Konj.)	145 (17)	21	21
auch (Adverb)	137 (17)	22	22

Wort (WA)	*H (S)*	*R* (80 %-Marke)	*R* (≥50 %)
gut (Adjek.)	111 (17)	27	27
können (Verb)	110 (14)	28	28
heißen (Verb)	103 (17)	29	29
alle (Adverb)	101 (15)	30	30
kommen (Verb)	100 (16)	31	31
gehen (Verb)	87 (15)	32	32
fertig (Adjek.)	69 (14)	34	33
spielen (Verb)	63 (14)	38	34
wo (Pron.)	56 (11)	42	35
aber (Partikel)	53 (15)	44	36
sagen (Verb)	50 (14)	45	37
doch (Adverb)	46 (16)	47	38
noch (Adverb)	43 (12)	49	39
sein (Pron.)	41 (11)	51	40
Hand (Subst.)	36 (12)	56	41
Bleistift (Subst.)	34 (12)	60	42
jetzt (Adver)b)	34 (11)	60	42
tschüss (Interjek.)	31 (11)	63	44
dann (Adverb)	29 (13)	66	45
klein (Adjek.)	24 (11)	73	46

Wort (WA)	*H (S)*	*R* (80 %-Marke)	*R* (≥50 %)
mal (Adverb)	135 (13)	23	23
wie (Pron.)	130 (18)	24	24

Wort (WA)	*H (S)*	*R* (80 %-Marke)	*R* (≥50 %)
in (Präpo.)	24 (11)	73	46
Stift (Subst.)	23 (12)	75	48

Anmerkungen: WA = Pronomen; *H* = Häufigkeit; *S* = Streuung; *R* = Rang

Zusammensetzung des Kernvokabulars

Um einen Überblick über die Wortartenverteilung im Gesamtkorpus zu gewinnen, wurde die Gesamtliste (*N* = 22) in Excel nach Wortarten analysiert. Die absolute Häufigkeit pro Wortart wurde deskriptiv berechnet. Die Rangfolge der *fünf häufigsten Wortarten* wurde in der Ergebnisbeschreibung berücksichtigt. Eine detailliertere Verteilung aller erfassten Wortarten pro Korpus ist den jeweiligen Abbildungen zu entnehmen.

Die am häufigsten verwendeten Wortarten (gemessen an den Token) im Gesamtkorpus waren Verben (Vollverben, Hilfs- und Modalverben) (2 930 ≈ 21.33 %), Pronomen (2 424 ≈ 17.65 %), Substantive (2 158 ≈ 15.71 %), Adverbien (2 003 ≈ 14.58 %) und Artikel (1 485 ≈ 10.81 %) (Abb. 25).

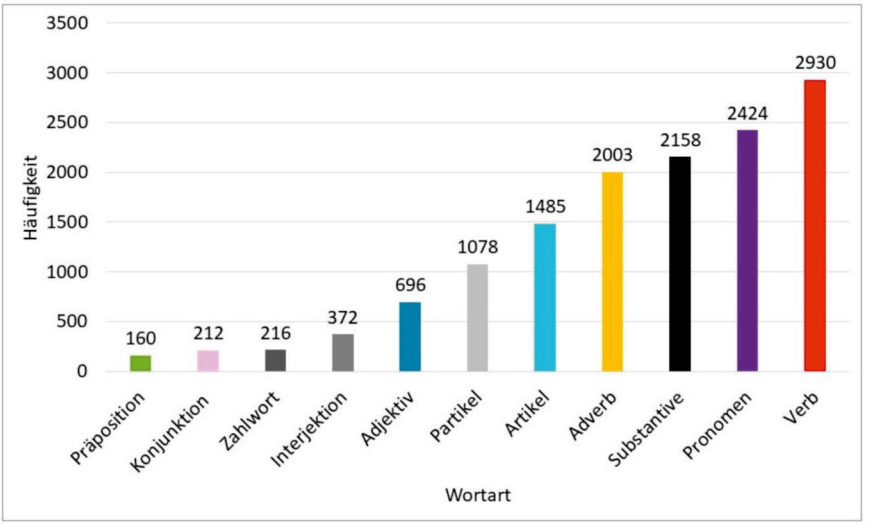

Abb. 25: Wortartenverteilung Gesamtliste im Primärkorpus (N = 22, Σ = 13 734 Token)

Neben der Betrachtung der absoluten Häufigkeiten pro Wortart wurde zusätzlich die Anzahl der unterschiedlichen Wörter innerhalb einer Wortart berechnet. Wenngleich Verben (Vollverben, Hilfs- und Modalverben) und Pronomen am häufigsten genutzt wurden, machten die Substantive (555 Lemma-Types) den größten Anteil am Gesamtwortschatz aus. Danach folgten die Verben (Vollverben, Hilfs- und Modalverben 167 Lemma-Types), die Adverbien (78 Lemma-Types) sowie die Adjektive (72 Lemma-Types) (Abb. 26). Die Substantive und Verben (Vollverben, Hilfs- und Modalverben) machten einen Anteil von 82.33 % an den insgesamt verschiedenen verwendeten Wörtern aus.

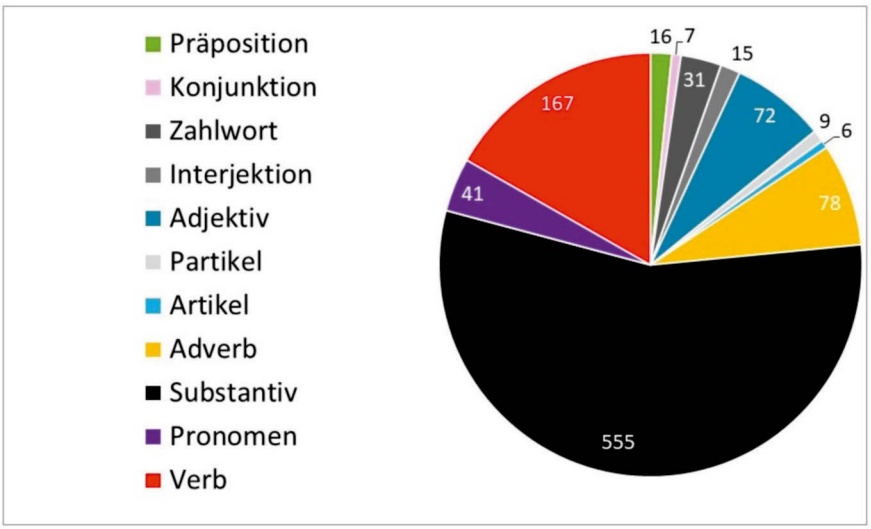

Abb. 26: Wortartenverteilung (Lemma-Types) Gesamtliste im Primärkorpus (N = 22, Σ = 13 734 Token)

Daran anschließend wurde die Wortartenverteilung in Bezug auf das Kernvokabular herausgearbeitet. Es zeigte sich, dass innerhalb der 80 %-Marke die Verben (Vollverben, Hilfs- und Modalverben) (2 465 ≈ 22.43 %), Pronomen (2 259 ≈ 20.55 %), Adverbien (1 769 ≈ 16.09 %), Artikel (1 462 ≈ 13.30 %) sowie Partikel (1 064 ≈ 9.68 %) am meisten verwendet wurden (Abb. 27).

Im Unterschied zur Gesamtliste wurde deutlich, dass im Kernvokabular die Substantive deutlich seltener verwendet wurden (Gesamtliste: 2 138 ≈ 15.57 % vs. Kernvokabularliste: 817 ≈ 7.43 %). Dennoch machten im Kernvokabular die Substantive (23 Lemma-Types) gemeinsam mit den Verben (Vollverben, Hilfs- und Modalverben (23 Lemma-Types) den größten Anteil unterschiedlicher Wörter aus. Danach folgten Adverbien (19 Lemma-Types), Pronomen (15 Lemma-Types) und Adjektive (10 Lemma-Types).

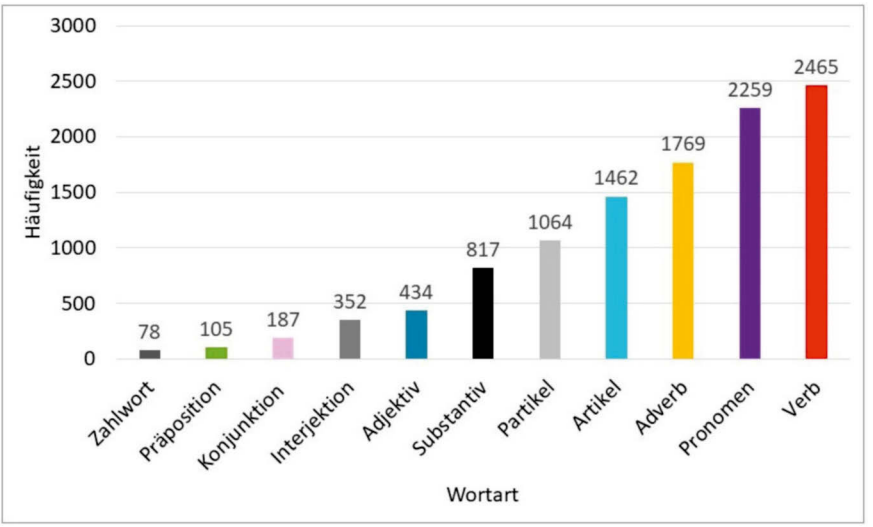

Abb. 27: Wortartenverteilung Kernvokabular 80 %-Marke im Primärkorpus (N = 22, Σ = 13 734 Token)

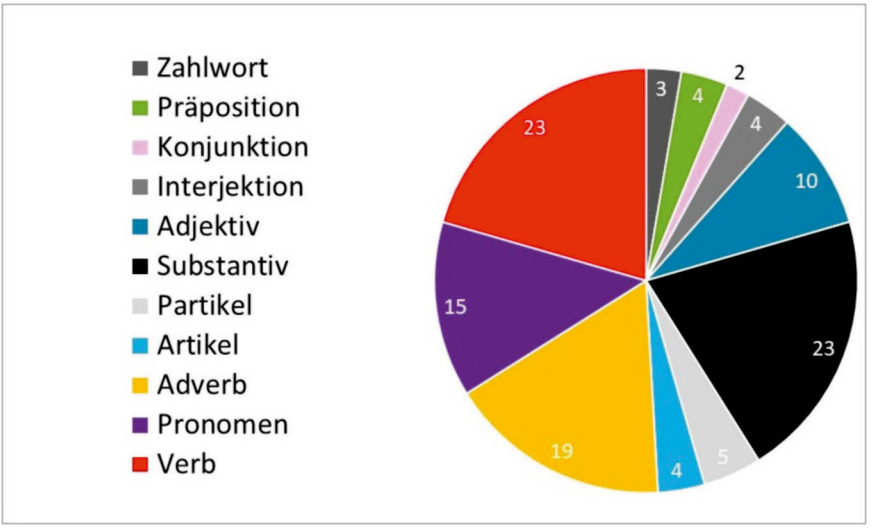

Abb. 28: Anzahl Lemma-Types Kernvokabular 80 %-Marke pro Wortart im Primärkorpus (N = 22, Σ = 10 992 Token)

Die Wortartenverteilung im Kernvokabular unter Hinzunahme des Streuungskriteriums ergab einen ähnlichen Trend, wobei die Artikel und Pronomen noch stärker anteilig gebraucht wurden: Verben (Vollverben, Hilfs- und Modalverben) (2 042 ≈ 22.27 %),

Pronomen (2 029 ≈ 22.33 %), Adverbien (1 577 ≈ 17.35 %), Artikel (1 462 ≈ 16.1 %), Partikel (977 ≈ 10.75 %) (Abb.29).

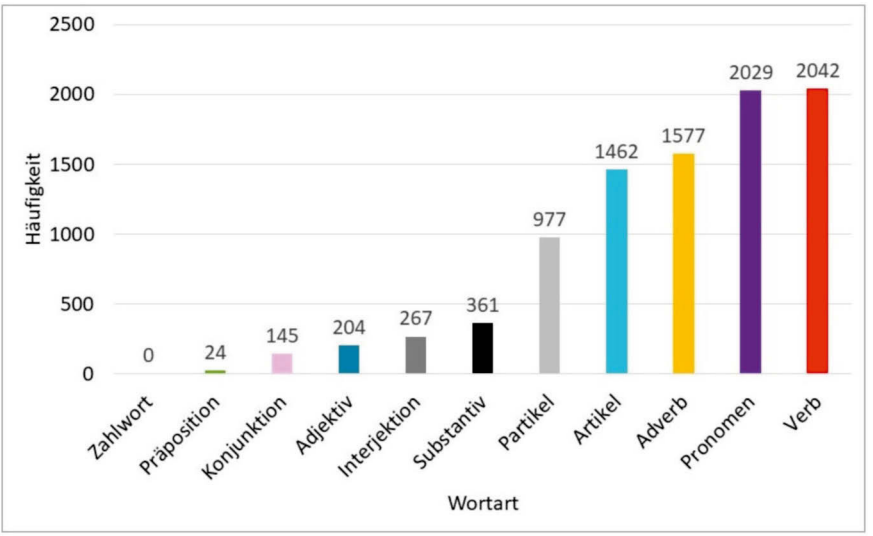

Abb. 29: Wortartenverteilung reduzierte Kernvokabularliste im Primärkorpus (N = 22, Σ = 9 088)

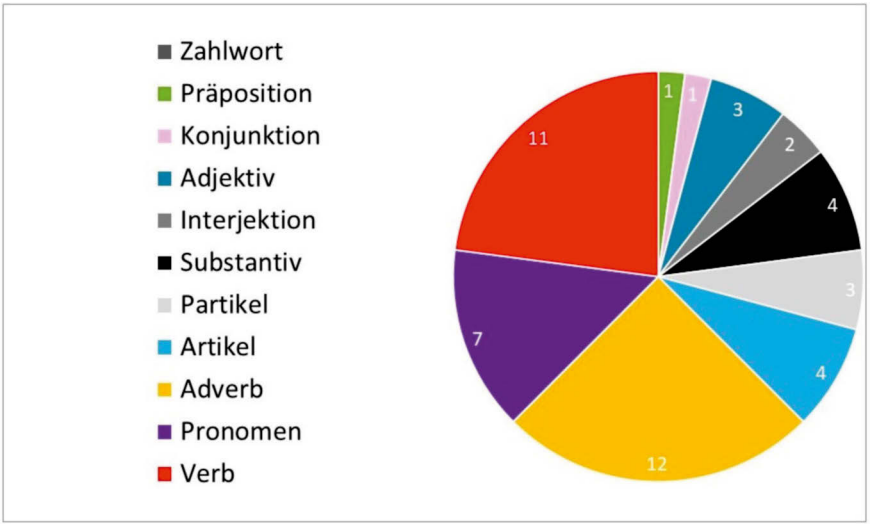

Abb. 30: Anzahl Lemma-Types reduzierte Kernvokabularliste pro Wortart im Primärkorpus (N = 22, Σ = 9 088 Token)

Bei Betrachtung der Anzahl verschieden genutzter Wörter pro Wortart wurde deutlich, dass die Adverbien (12 Lemma-Types) die größte Gruppe im Kernvokabular ausmachten. Danach folgten die Verben (Vollverben, Hilfs- und Modalverben) (11 Lemma-Types) sowie die Pronomen (7 Lemma-Types), Artikel (4 Lemma-Types) und Substantive (4 Lemma-Types) (Abb. 30).

Die Analyse der Wortartenverteilung im Randvokabular – also jenseits der 80 %-Marke – ergab folgende Verteilung: Substantive (1 341 ≈ 48.91 %), Verben (Hilfs- und Modalverben, Vollverben) (465 ≈ 16.96 %), Adjektive (262 ≈ 9.56 %), Adverbien (234 ≈ 8.53 %) und Pronomen (165 ≈ 6.02 %). Insgesamt zählten 2 742 Token sowie 765 Lemma-Types zum Randvokabular (Abb. 31).

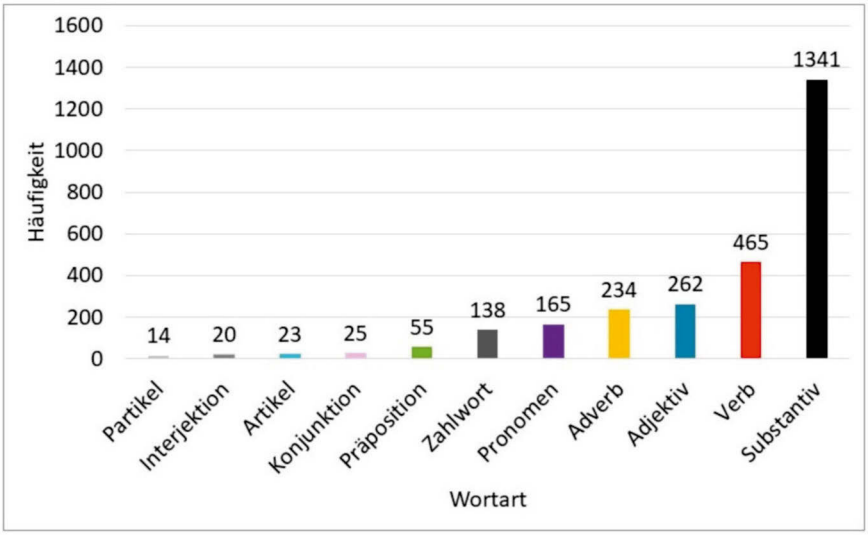

Abb. 31: Wortartenverteilung Randvokabular im Primärkorpus (N = 22, Σ = 2 742 Token)

Knapp die Hälfte aller verwendeten Token im Randvokabular konnte den Substantiven zugeordnet werden. Funktionswörter, wie die Pronomen waren nur vereinzelt vertreten.

Der *Anteil von Funktionswörtern und Inhaltswörtern* wurde in einer weiteren Analyse in Bezug auf die drei Listen *Gesamtliste*, *Kernvokabular 80 %-Marke* und *reduzierte Kernvokabularliste (80 %-Marke + ≥50 %)* beschrieben. Die Zuordnung der Wortarten zu Funktionswörtern und Inhaltswörtern erfolgte angelehnt an internationale Kernvokabularstudien (u.a. Boenisch & Soto, 2015; Hattingh & Tönsing, 2020, Kap. 7.1.1). Demzufolge zählten Substantive, Vollverben, Adjektive, Adverbien und Zahlwörter zu den Inhaltswörtern, wenngleich Hilfs- und Modalverben, Präpositionen, Konjunktionen, Artikel, Partikel, Interjektionen und Pronomen den Funktionswörtern zugeteilt wurden.

Anhand der Gegenüberstellung der absoluten Häufigkeit von Funktions- und Inhaltswörtern in der jeweiligen Korpusliste ließ sich erkennen, dass der absolute Anteil

der Funktionswörter im Vergleich innerhalb der reduzierten Kernvokabularliste am größten war (Abb. 32).

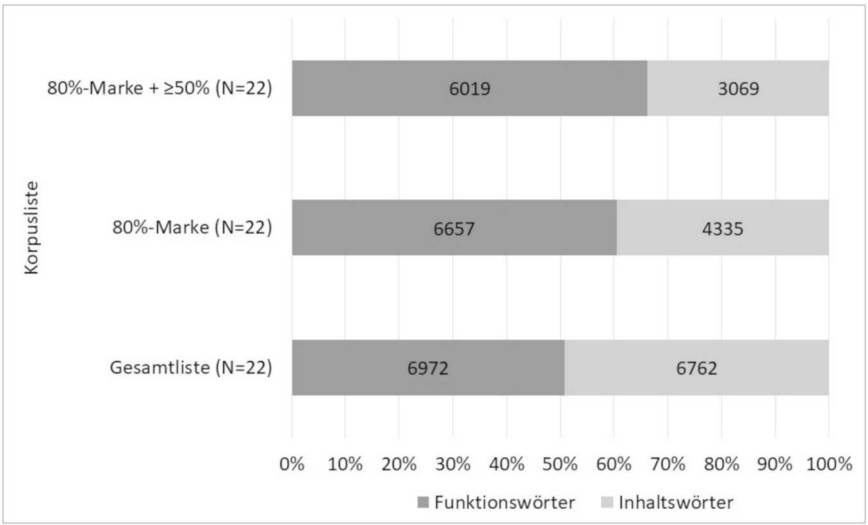

Abb. 32: Vergleich Anteil von Funktions- und Inhaltswörtern im Primärkorpus

Der prozentuale Anteil der Funktionswörter in der reduzierten Kernvokabularliste betrug ≈ 66.23 %, in der Kernvokabularliste ≈ 60.56 % und in der Gesamtliste ≈ 50.76 %. Bei den Inhaltswörtern lag der prozentuale Anteil in der reduzierten Kernvokabularliste bei ≈ 33.77 %, in der Kernvkabularliste bei ≈ 39.44 % und in der Gesamtliste bei ≈ 49.24 % (Tab. 43).

Tab. 42: Vergleich Prozentualer Anteil von Funktions- und Inhaltswörtern im Primärkorpus

	Liste$_{gesamt}$	**Liste$_{80\,\%\text{-Marke}}$**	**Liste$_{80\,\%\text{-Marke}+\geq 50\,\%}$**
Funktionswörter	50.76 %	60.56 %	66.23 %
Inhaltswörter	49.24 %	39.44 %	33.77 %

In den beiden Kernvokabularlisten machten die Funktionswörter über die Hälfte des Korpus aus. Der relative Anteil am Korpus stieg an, umso kleiner der untersuchte Korpus war. In der Gesamtliste nahmen die Funktionswörter die Hälfte des Korpus ein.

Auffällig hoch war die Verwendungshäufigkeit und der prozentuale Anteil der Hilfs- und Modalverben innerhalb der Funktionswörter[1] (Tab. 43).

Tab. 43: Vergleich Nutzungshäufigkeit Hilfs- und Modalverben im Primärkorpus (möchten als mögen ausgewertet)

Funktionswörter	**Häufigkeit (H)**		
Hilfs- und Modalverb (S)	$Liste_{gesamt}$	$Liste_{80\,\%\text{-}Marke}$	$Liste_{80\,\%\text{-}Marke\,+\,\geq 50\,\%}$
sein (20)	744	744	744
haben (18)	261	261	261
werden (3)	4	N.V.	N.V.
können (14)	110	110	110
sollen (7)	13	N.V.	N.V.
müssen (9)	36	36	N.V.
wollen (10)	77	77	N.V.
dürfen (8)	36	36	N.V.
möchten (4)	16	N.V.	N.V.
Summe	1 297	1 264	1 151
Anteil (%) Funktionswörter	18.60	18.99	19.12
Anteil (%) Korpusliste	9.44	11.50	12.67

In allen drei Listen lag der relative Anteil der Hilfs- und Modalverben innerhalb der Funktionswörter bei ca. 19 %. Die Hilfsverben *sein* und *haben* sowie das Modalverb *können* tauchten auf allen Korpuslisten auf. Das Hilfsverb *werden* konnte nicht im Kernvokabular identifiziert werden. Von den Modalverben kam lediglich *sollen* nicht im Kernvokabular vor. Die Modalverben *müssen* (10 Kinder), *wollen* (9 Kinder) und *dürfen* (8 Kinder) waren aufgrund der geringen Streuung nicht auf der reduzierten Kernvokabularliste enthalten. Der relative Anteil in der jeweiligen Korpusliste lag zwischen 9.44 % und 12.67 %. Umso kleiner die Korpusliste, desto höher war der Anteil am untersuchten Korpus.

Vergleichsanalyse

Beschreibung Referenzkorpus, Vergleich der Gesamtlisten

Der Referenzdatensatz bestand aus $N = 28$ Schüler:innen (weiblich $n = 10$; männlich $n = 18$). Die Sprachaufnahmen wurden in der 2. Klasse ($n = 13$) und 4. Klasse ($n = 15$) der Grundschule erhoben (Boenisch, 2013, S. 20). Die Dauer der Sprachaufnahmen betrug ca. Σ = 72:15:00 Std (4 335 min). Die durchschnittliche Länge der Sprachaufnahme pro Schüler:in

1 Hilfs- und Modalverben können auch als Vollverben verwendet werden (Fabricius-Hansen, 2009, S. 416). In der Auswertung wurde diese Unterscheidung nicht explizit vorgenommen, da die Abgrenzung nicht immer eindeutig getroffen werden kann (Fabricius-Hansen, 2009, S. 416).

lag bei 155 min (ca. 3 % über dem Primärkorpus). Die Datenerhebung fand am 17.05.2010 und 24.02.2011 statt. Angaben zu den formalsprachlichen Aspekten im Vergleich zum *Primärkorpus Gesamtliste* ($N = 22$) sind Tabelle 44 zu entnehmen. Auf der Ebene der Token und Lemma-Types zeigte sich, dass der Referenzkorpus ca. dreimal so groß war wie die Primärkorpus (Tab. 44).

Tab. 44: Datensätze Primärkorpus ($N = 22$) und Referenzkorpus ($N = 28$) im Vergleich

	Primärkorpus ($N = 22$)	**Referenzkorpus ($N = 28$)**
Token	13 734	48 961
H_{max}	985	2 664
H_{min}	1	1
Lemma-Types	877	2 730
$Rang_{max}$	562	1 577

Die fünf am häufigsten verwendeten Wortarten im Referenzkorpus ($N = 28$) waren Pronomen (11 930 ≈ 24.4 %), Verben (Hilfs- und Modalverben, Vollverben) (11 263 ≈ 23 %), Substantive (5 412 ≈ 11.1 %), Adverbien (5 238 ≈ 10.7 %) sowie Partikel (4 206 ≈ 8.6 %) (Abb. 33).

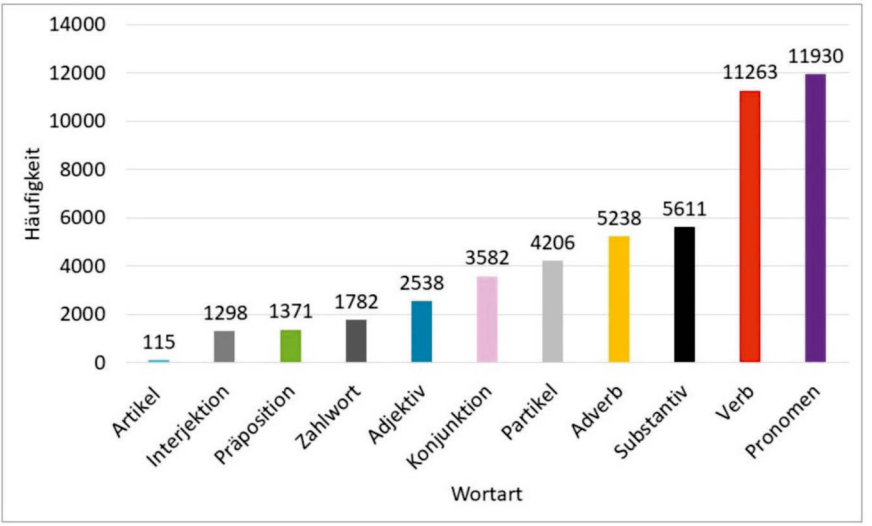

Abb. 33: Wortartenverteilung im Referenzkorpus ($N = 28$, $\Sigma = 48\,961$ Token)

Der hohe Anteil an Pronomen ist mit der Wortartenannotation zu erklären. Die Lemma-Types *der*, *die*, *das* wurden im Referenzkorpus als Pronomen kodiert und nicht als Artikel so wie im Primärdatensatz.

Die größte Anzahl unterschiedlicher Wörter innerhalb einer Wortart bildeten die Substantive (1 210 Lemma-Types). Danach folgten die Verben (Vollverben, Hilfs- und Modalverben) (728 Lemma-Types), Adjektive (279 Lemma-Types), Adverbien (169 Lemma-Types) und die Zahlwörter (84 Lemma-Types). Die Substantive und Verben (Vollverben, Hilfs- und Modalverben) machten einen Anteil von 72.64 % an den insgesamt verschiedenen verwendeten Wörtern aus (Abb. 34).

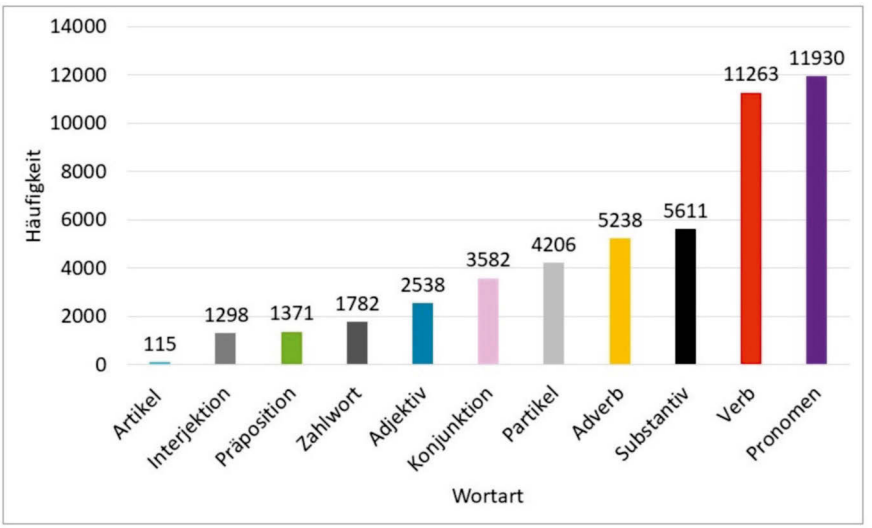

Abb. 34: Anzahl Lemma-Types pro Wortart im Referenzkorpus(N = 28, Σ = 48 961 Token)

Vergleichsanalyse: Umfang des Kernvokabulars

Anhand der Analyse der absoluten Häufigkeiten wurde deutlich, dass eine vergleichsweise geringe Anzahl an Wörtern hochfrequent genutzt wurde (Abb. 35).

Dieser Wortschatz spiegelte das *Kernvokabular entsprechend der 80 %-Marke* wider (39 169 Token). Dazu zählten *210 Lemma-Types* von *ich* (*H* = 2 664) bis *helfen* (*H* = 28). Alle Lemma-Types ab *hinten* zählten zum Randvokabular. Die 80 %-Marke war im Referenzkorpus um 98 Lemma-Types größer als im Primärkorpus (Tab. 45).

Tab. 45: 80 %-Marke im Vergleich zwischen Primärkorpus (N = 22) und Referenzkorpus (N = 28)

	80 %-Marke (*N* = 22)	**80 %-Marke (*N* = 28)**
$Token_{80\,\%}$	10 987 (10 992)	39 169 (39 165)
$Token_{Gesamt}$	13 734	48 961

	80 %-Marke (*N* = 22)	**80 %-Marke (*N* = 28)**
H_{max}	985	2 664
H_{min}	17	28
$R_{80\,\%}$	108	209
Lemma-Types$_{80\,\%}$	112 (877)	210 (2 730)

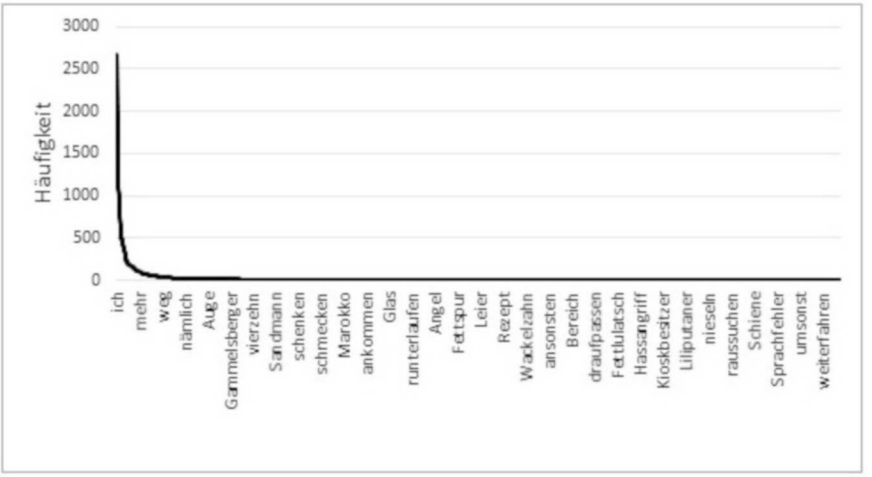

Abb. 35: Absolute Häufigkeiten Gesamtliste Referenzkorpus (N = 28, Σ = 48 961 Token)
Anmerkungen: Ränge (*R*): ich *R* = 1; weg *R* = 171; nämlich *R* = 253; Auge *R* = 342

Vergleichsanalyse: Zusammensetzung des Kernvokabulars

Innerhalb der 80 %-Marke wurden die Pronomen (11 741 ≈ 29.98 %), die Verben (Vollverben, Hilfs- und Modalverben (8 781 ≈ 22.42 %), Adverbien (4 257 ≈ 10.87 %), Partikel (4 088 ≈ 10.44 %) sowie die Konjunktionen (3 519 ≈ 8.99 %) am meisten verwendet (Abb. 36).

Die größte Anzahl unterschiedlicher Wörter im Kernvokabular des Referenzkorpus bildeten die Verben (Hilfs- und Modalverben, Vollverben) (46 Lemma-Types). Die zweitgrößte Gruppe bildeten die Pronomen sowie die Adverbien mit ja 30 Lemma-Types. Danach folgten die Substantive (25 Lemma-Types) und die Adjektive (23 Lemma-Types) (Abb. 37).

Mit der Gegenüberstellung von Primärkorpus (*N* = 22) und Referenzkorpus (*N* = 28) ließ sich in der Wortartenverteilung eine weitestgehende Übereinstimmung im Verlauf der absoluten Häufigkeiten erkennen (Abb. 38).

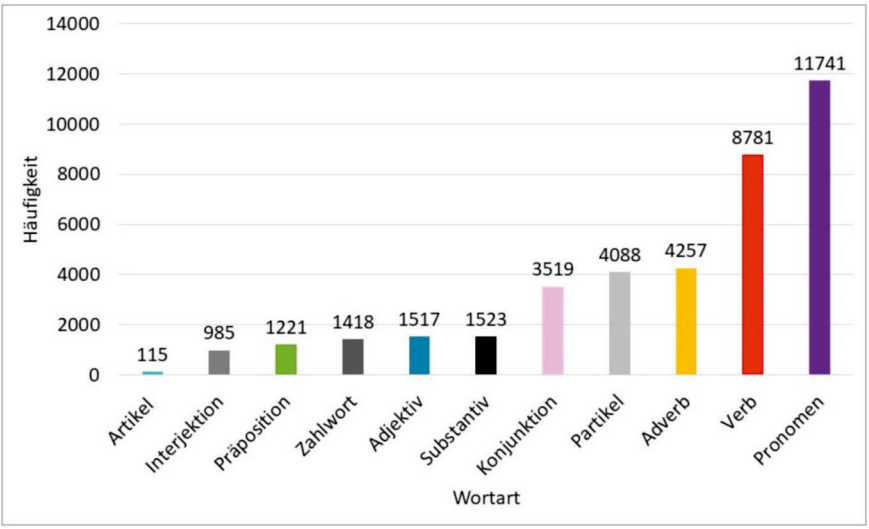

Abb. 36: Wortartenverteilung Kernvokabular 80 %-Marke im Referenzkorpus (N = 28, Σ = 39 165 Token)

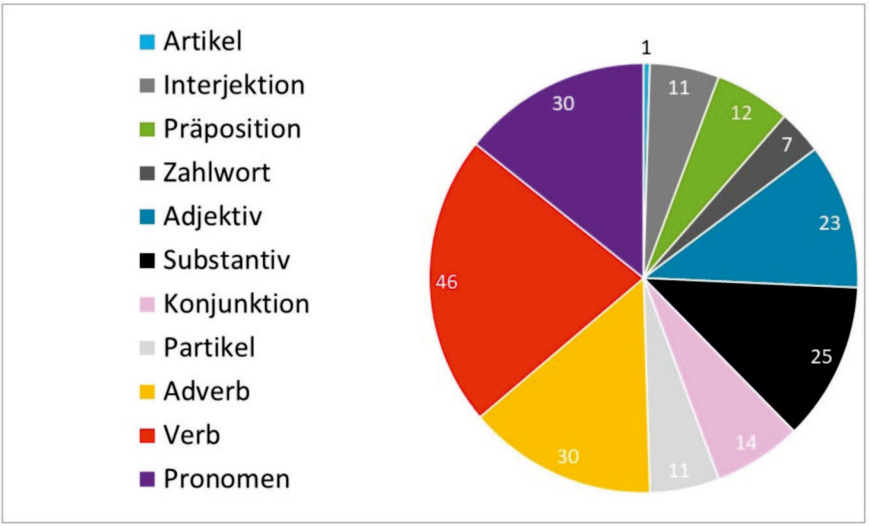

Abb. 37: Anzahl Lemma-Types 80 %-Marke pro Wortart im Referenzkorpus (N = 28, Σ = 39 164)

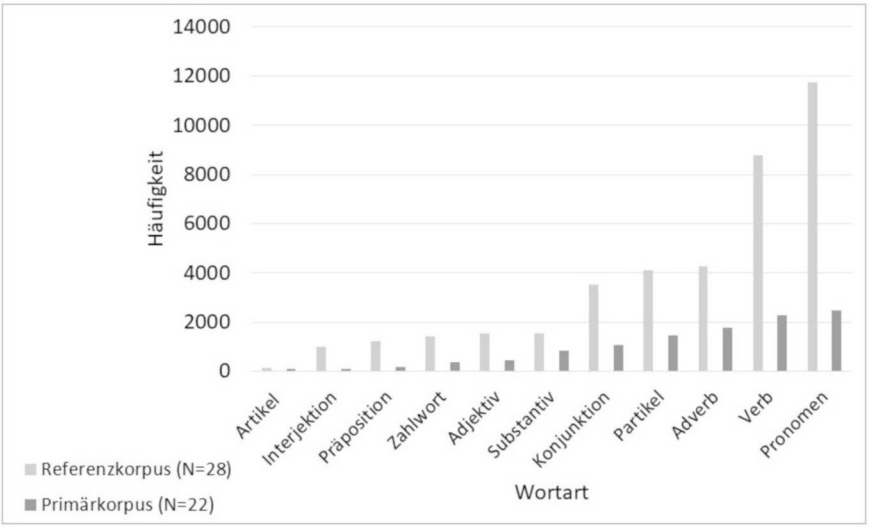

Abb. 38: Vergleichsanalyse Wortartenverteilung im Kernvokabular (80 %-Marke)

Gemeinsamkeiten und Unterschiede (in Tabelle **fett** markiert) in der Wortartenverteilung ließen sich mithilfe der Berechnung der prozentualen Anteile noch deutlicher herausarbeiten (Tab. 46).

*Tab. 46: Vergleichsanalyse über den relativen Anteil der Wortarten innerhalb des Kernvokabulars (80 %-Marke, Unterschiede in **fett** markiert)*

	Primärkorpus (*N* = 22)	Referenzkorpus (*N* = 28)	Differenz
Pronomen	22.43 %	29.98 %	**-7.55 %**
Verb	20.55 %	22.42 %	-1.87 %
Adverb	16.09 %	10.87 %	**5.22 %**
Partikel	9.68 %	10.44 %	-0.76 %
Konjunktion	1.70 %	8.99 %	**-7.29 %**
Substantiv	7.43 %	3.89 %	3.54 %
Adjektiv	3.95 %	3.87 %	0.08 %
Zahlwort	0.71 %	3.62 %	-2.91 %
Präposition	0.96 %	3.12 %	-2.16 %
Interjektion	3.20 %	2.52 %	0.68 %
Artikel	13.30 %	0.29 %	**13.01 %**

Ein ähnlicher Anteil im Gebrauch der Wortarten ließ sich für folgende Wortarten nachweisen: Verben (Hilfs- und Modalverben, Vollverben), Partikel, Substantive, Adjektive, Zahlwörter, Präpositionen, Interjektionen.

Deutlichere Unterschiede ließen sich im relativen Anteil bei den Pronomen, Adverbien, Konjunktionen und den Artikeln erkennen, welche methodisch zu begründen sind. Zum Beispiel wurden *wie* und *wo* aufgrund der Verwendungsweise im Primärkorpus als Adverb kodiert, wohingegen im Referenzkorpus das Lemma-Type als Konjunktion annotiert wurde. Das Wort *gleich* wurde im Primärkorpus als Adjektiv gezählt und im Referenzkorpus als Präposition. Die Wörter *der*, *die*, *das* wurden im Primärdatensatz als Artikel annotiert und im Referenzkorpus als Pronomen.

Insgesamt wurden im Kernvokabular des Referenzkorpus 26 764 *Funktionswörter* und 12 401 *Inhaltswörter* verwendet. Der relative Anteil der Funktionswörter am Kernvokabular lag bei 67.17 %. Die Inhaltswörter hatten einen Anteil von 32.83 %. Der relative Anteil von Funktions- und Inhaltswörtern lag demnach ähnlich hoch, wie im Primärdatensatz (60.56 % Funktionswörter, 39.44 % Inhaltswörter). Der prozentuale Anteil der verwendeten Hilfs- und Modalverben innerhalb des Kernvokabulars betrug 13.06 % (*sein H* = 1 888, *haben H* = 1 641, *werden H* = 182, *können H* = 463, *müssen H* = 324, *wollen H* = 178, *sollen H* = 144, *dürfen H* = 209, *möchten*[2] *H* = 66) und lag etwas höher als im Primärkorpus (+11.56 %). Alle Hilfs- und Modalverben zählten zum Kernvokabular. Der Anteil der Hilfs- und Modalverben am Gesamtwortschatz betrug 10.41 %.

In einer abschließenden Analyse wurden die Listen der *TOP 20*, *TOP 50* und *TOP 100 Wörter* aus dem Primär- und Referenzkorpus gegenübergestellt, qualitativ verglichen sowie die prozentuale Übereinstimmung auf der Ebene der Lemma-Types berechnet. In der Liste der TOP 20 bestand eine Übereinstimmung von 15 Lemma-Types (75 %). Die Übereinstimmung in der Liste TOP 50 betrug 36 Lemma-Types (72 %). In der Liste TOP 100 konnten 60 Lemma-Types (57.69 %) übereinstimmend nachgewiesen werden (Tab. 47).

Tab. 47: Vergleichsanalyse TOP 20, TOP 50, TOP 100: Übereinstimmungen und Differenzen in den Lemma-Types im Primärkorpus und Referenzkorpus

	Übereinstimmungen Lemma-Types		**Differenz Lemma-Types**	
	prozentual	*absolut*	*prozentual*	*absolut*
TOP 20	75 %	15	25 %	4
TOP 50	72 %	36	28 %	14
TOP 100	59.02 %	61	40.98 %	40

Anmerkungen: Differenz Lemma-Types/Spalte absolut: Anzahl Lemma-Types, die nur im Referenzkorpus ($N = 28$) zu finden waren. TOP 100: Die TOP 100 der Kernvokabularliste ($N = 22$) entsprach $R = 98$. Insgesamt konnten zusammen mit der Referenzliste 206 Lemma-Types in der Vergleichsanalyse TOP 100 berücksichtigt werden.

2 als Form von *mögen* ausgewertet

Von der TOP 20-Liste aus dem Primärkorpus konnten die vier fehlenden Lemma-Types (*auch*, *da*, *mal*, *und*) unter den TOP 50 wiedergefunden werden.

Von den 14 abweichenden Lemma-Types in der TOP 50-Liste konnten 10 Lemma-Types in der TOP 100-Liste identifiziert werden (*dann*, *dürfen*, *es*, *in*, *jetzt*, *kein*, *mir*, *mit*, *müssen*, *schon*; fehlend: *den*, *man*, *oh*, *werden*).

Die prozentuale Übereinstimmung zwischen den TOP 100-Listen betrug 59.02 % (61 Lemma-Types). Im Primärkorpus fehlten insgesamt 40 Lemma-Types auf der TOP 100-Liste im Vergleich zum Referenzkorpus. Von diesen tauchten wiederum 35 Lemma-Types in der Gesamtliste auf (fehlend: *ah*, *aufnehmen*, *boah*, *ey*, *oh*) (Tab. 48).

Tab. 48: Vergleichsanalyse TOP 100: fehlende Lemma-Types im Primärkorpus und Position in der Gesamtliste (H = Häufigkeit; R = Rang)

Lemma-Types	*H*	*R*
ach	1	562
Affe	9	185
also	2	429
brauchen	9	185
dass	2	429
dem	12	152
den	11	161
denn	3	355
dich	11	161
dies	17	108
Ding	2	429
dir	15	123
drei	15	123
ganz	6	243
gar	7	223
gerade	1	562
gleich	7	223
hören	11	161

Lemma-Types	*H*	*R*
ihr	16	115
immer	16	115
lang	2	429
man	6	243
mehr	9	185
Minute	13	140
ne	7	223
sehen	4	316
sollen	13	140
Stunde	2	429
uns	9	185
viel	11	161
voll	2	429
weil	3	355
wenn	15	123
werden	4	316
wieder	9	185

12.2 Feste Wortkombinationen

Insgesamt konnten *2 998 Dreiwortkombinationen* (Token) im Primärkorpus identifiziert werden. Von diesen wurden *1976 unterschiedliche Dreiwortkombinationen* (Types-Kombinationen) verwendet. Am häufigsten wurde die Dreiwortkombination *Was ist das* ($H = 72$) verwendet. Der höchste Rang lag bei $R_{max} = 384$. Von den 1 976 unterschiedlichen Dreiwortkombinationen waren 383 Dreiwortkombinationen mit einem Anteil