Quantitative Aspekte der Modalpartikelverwendung. Untersuchungen zum automatisch annotierten Korpus für gesprochenes Deutsch FOLK*

Peter Paschke (Venedig)

Abstract

The topic of this article are quantitative aspects of the use of modal particles (MPs) in corpora of spoken German, i. e. the token rates of individual or all MPs and the ranking of their frequency. Since MPs consistently have heterosemes in other word classes (adverb, focus particle, interjection, etc.), in the past such analyses had to be conducted manually. It was only in 2017 that automatic POS tagging adapted to spoken language was released for the FOLK corpus, enabling an automatic search for MPs. By comparing the automatic counts in FOLK to the frequency data of the manually analysed corpora of Hentschel (1986) and Brünjes (2014) and by checking the POS tagging of samples randomly extracted from FOLK, the paper seeks to answer the question of how reliable the automatically generated MP-data of FOLK are. With regard to the list of lexemes considered in FOLK, errors are essentially limited to the MP *eigentlich* and to some quantitatively marginal cases. The overall frequency of MPs in FOLK (token rate 2.55%) also seems plausible. Major deviations from previous studies arise in the frequencies of some single MPs, of which *auch, mal* and *halt* are analysed in more detail. While the discrepancies for *auch* are due to deficits in POS tagging, for *mal* and *halt* corpus characteristics (discourse types and survey periods) play a major role. When extrapolating the adjusted frequencies found in the random samples to the whole corpus, the MP frequency rankings of FOLK however correlate just as well with those of manual counts (r=0.81/0.82) as the manually determined MP rankings of different corpora do with each other.

1 Einleitung

Modalpartikeln (MPn) sind vor allem ein Phänomen der (konzeptionell) gesprochenen Sprache in der Interaktion, i. e. von Gesprächen, weshalb es sinnvoll ist, sie anhand von geeigneten gesprochensprachlichen Korpora zu analysieren. Korpora erlauben zudem quantitative Aussagen über die Verwendung von MPn, etwa über die Rangfolge ihrer Häufigkeit oder den Anteil aller bzw. einzelner MPn an den laufenden Wörtern eines Korpus (Tokenrate). Da aber MPn systematisch Homonyme bzw. Heteroseme[1] in anderen Wortklassen aufweisen, war es in

* Für wertvolle Anregungen zu einer früheren Version dieses Aufsatzes danke ich den anonymen Gutachter:innen.

[1] Im Anschluss an Brünjes (2014: 18) verwende ich im Folgenden den Terminus „Heterosem".

einschlägigen Untersuchungen (cf. z. B. Hentschel 1986; Hentschel/Keller 2006; Brünjes 2014) stets notwendig, die Belege für Lexeme wie *auch*, *ja*, *mal* etc. einzeln zu prüfen, um die MPn von Verwendungen als Fokuspartikel, Adverb, Responsiv etc. abzugrenzen. Um den Zeitaufwand in einem akzeptablen Rahmen zu halten, mussten jeweils der Umfang des Korpus, die Zahl der Belege und/oder die Liste der untersuchten MPn eingeschränkt werden.

Das ab 2008 aufgebaute Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK) in der Datenbank Gesprochenes Deutsch (DGD) des IDS Mannheim, da sich in den letzten Jahren zu einer Art Referenzkorpus für gesprochenes Deutsch entwickelt hat, verspricht hier Abhilfe. FOLK ist ein Korpus, „das Gesprächsdaten aus unterschiedlichsten Bereichen des gesellschaftlichen Lebens (Arbeit, Freizeit, Bildung, öffentliches Leben, Dienstleistungen etc.) im deutschen Sprachraum beinhaltet“ (FOLK: Beschreibung des Korpus). Die Tonaufnahmen stehen online mit alignierten Annotationen (literarische Umschrift, orthographische Normalisierung, Lemmatisierung) und Recherchewerkzeugen zur freien Verfügung. 2017 wurde ein POS-Tagging freigeschaltet, das dank Anpassung an Phänomene gesprochener Sprache (Interjektionen, Häsitationspartikeln, Rezeptionssignale, MPn etc.)[2] mit ca. 95%iger Genauigkeit die Wortart (POS, *part of speech*) der Lemmata bestimmt (Westpfahl/Schmidt 2016, Westpfahl 2020). Ein Beispiel für die verschiedenen Annotationsebenen inkl. POS-Tagging zeigt Abbildung 1.

Annotationen für FOLK_E_00042_SE_01_T_01_DF_01 / c1077

ID	w6512	w6513	w6514	w6515	w6516	w6517	w6518	w6519	w6520	w6521	w6522	w6523	w6524	w6525	w6526
Transkription	ich	weiß	nich	is	mir	halt	irgendwie	total	peinlich	weil	des	referat	so	schlecht	is
Normalisierung	ich	weiß	nicht	ist	mir	halt	irgendwie	total	peinlich	weil	das	Referat	so	schlecht	ist
Lemma	ich	wissen	nicht	sein	ich	halt	irgendwie	total	peinlich	weil	d	Referat	so	schlecht	sein
POS	PPER	VVFIN	PTKNEG	VAFIN	PRF	PTKMA	ADV	PTKIFG	ADJD	KOUS	ART	NN	PTKIFG	ADJD	VAFIN

Abbildung 1: Annotationen in FOLK (PTKMA = Modal- bzw. Abtönungspartikel); FOLK_E_00042

Das POS-Tagging von FOLK bietet erstmals die Möglichkeit, in einem vergleichsweise großen Korpus gesprochener Sprache (Version 2.16 vom 17.5.2021 umfasst 2.990.421 laufende Wörter) gezielt nach MPn zu suchen und quantitative Analysen vorzunehmen.

Der vorliegende Beitrag kreist vor allem um die Frage, wie zuverlässig die von FOLK ermittelten Daten hinsichtlich Auswahl und Häufigkeit der MPn sind: Einerseits geht es also um die Frage, welche Lexeme überhaupt als MPn klassifiziert werden, andererseits um die mit den Recherchewerkzeugen der DGD ermittelte Häufigkeitshierarchie der MPn sowie ihre auf die Korpusgröße bezogene Frequenz (Tokenrate). Durch den Vergleich mit gängigen MPn-Listen und vorliegenden (manuellen) Auszählungen wird die Zuverlässigkeit der dank FOLK möglichen automatischen Recherche einer Prüfung unterzogen, wobei auffällige Abweichungen im Mittelpunkt des Interesses stehen. Abweichungen in der relativen Häufigkeit einzelner MPn, aber auch hinsichtlich der Tokenrate der MPn insgesamt, können auf verschiedene Faktoren zurückgehen: Neben Defiziten beim automatischen POS-Tagging kommen insbesondere unterschiedliche Abgrenzungen von MP und jeweiligem Heterosem und/oder Unterschiede in

[2] Das erweiterte Tagset trägt die Bezeichnung STTS 2.0 (Stuttgart-Tübingen-TagSet 2.0). Eine Auswahl von Tags findet sich im Anhang.

der Zusammensetzung der verglichenen Korpora (cf. Hentschel/Keller 2006; Silberstein 2021) in Frage.

Die Definition von MPn ist alles andere als eindeutig und unstrittig. Im vorliegenden Beitrag soll der von Brünjes (2014) formulierte Minimalkonsens zugrunde gelegt werden, der vor allem auf syntaktischen Eigenschaften beruht, wie sie bereits Thurmair (1989) systematisch dargestellt hat:

> Als Modalpartikeln werden diejenigen Unflektierbaren bezeichnet, die nicht satzgliedfähig sind, im Mittelfeld auftreten, Heteroseme in anderen Wortklassen haben, zu den Synsemantika gehören, keinen Beitrag zur Proposition liefern, affin zu bestimmten Satzarten/-modi sind und Satz- bzw. Äußerungsskopus haben.
>
> (Brünjes 2014: 18)

Zum Aufbau des Beitrags: Im folgenden Kapitel geht es zunächst (2.1) darum, welche Lexeme in FOLK überhaupt als MPn erfasst sind. Diskutiert werden sowohl fragliche Kandidaten (2.1.1) wie *ausgerechnet*, *glatt* oder *echt* als auch nicht erfasste Partikeln (2.1.2) wie *eigentlich*, *etwa* und *vielleicht*. In Abschnitt 2.2 steht dann die Häufigkeit des Vorkommens der (unstrittigen) MPn im Mittelpunkt. Durch Vergleich mit den (manuell durchgeführten) quantitativen Analysen von Hentschel (1986) und Brünjes (2014) wird geprüft, ob das automatische POS-Tagging von FOLK plausible Häufigkeitshierarchien und Frequenzangaben (Tokenraten) für MPn hervorbringt. Kapitel 3 ist einigen auffälligen Abweichungen, nämlich bei den MPn *auch* (3.1), *mal* (3.2) und *halt* (3.3) gewidmet. Die automatisch ermittelte Häufigkeit dieser MPn weicht in eklatanter Weise von einem (*mal*) bzw. von beiden Vergleichskorpora (*auch*, *halt*) ab. Hier wird jeweils erörtert, ob die Abweichungen auf unterschiedliche Definitionen der betroffenen MPn, auf Defizite des POS-Taggings oder auf bestimmte Korpusmerkmale zurückgeführt werden können. Als problematisch erweist sich die automatische Annotation besonders bei *auch*; im Übrigen korrelieren die MP-Häufigkeiten von FOLK recht gut mit manuell ausgezählten Korpora (3.4). Im abschließenden Resümee (Kapitel 4) werden die wesentlichen Erkenntnisse rekapituliert und Desiderata für weitergehende Studien formuliert.

2 Modalpartikeln (MPn) in FOLK im Vergleich

Eine DGD-Wortlisten-Recherche in FOLK (Version 2.16 vom 17.05.2021) mit Hilfe des POS-Tags „PTKMA“ (Modal- bzw. Abtönungspartikel) ergibt 76.331 Treffer. Bei einem Gesamtumfang des Korpus von 2.990.421 laufenden Wörtern (Token) entspricht das einer Tokenrate von 76.331/2.990.421 = 2,55%, i. e. ungefähr jedes 39. Wort ist in FOLK als PTKMA annotiert. Die Verteilung der Treffer auf die Lemma-Types geht aus Tabelle 1 hervor.

MP-Lemma	Treffer
ja	23.882
mal	10.583
halt	9.694
doch	7.070
schon	6.154
einfach	5.352
denn	4.876
aber	4.384
eben	3.038
wohl	479
überhaupt	308
bloß	139

MP-Lemma	Treffer
ruhig	114
echt	110
wirklich	43
auch	29
nur	23
glatt	21
ausgerechnet	12
fei	10
nun	5
zu	3
einmal	2
Summe	**76.331**

Tabelle 1: Häufigkeit der MP-Lemmata im FOLK-Korpus

Die Angaben in Tabelle 1 werden im Folgenden unter dem Gesichtspunkt der Auswahl (2.1) und der relativen Häufigkeit (2.2) mit Angaben aus anderen Quellen verglichen.

2.1 Auswahl der MP-Lexeme

Darüber, welche Partikeln zur Klasse der MPn zu rechnen sind, besteht zwar keine Einigkeit (cf. Hentschel 2013: 64), aber ein gewisser Konsens ist dennoch festzustellen. Schoonjans (2018: 24) bietet eine Übersicht der in der Fachliteratur gewöhnlich als MPn gelisteten Lexeme, die auf der Auswertung von 14 einschlägigen Publikationen aus den Jahren 1977–2010 beruht. MPn, die nur von einer Minderheit (3–7 Autor:innen) aufgeführt werden, sind in Tabelle 2 in Klammern gesetzt; die übrigen finden sich in mindestens 12 der ausgewerteten Publikationen (cf. die ähnliche Liste bei Hentschel/Keller 2006: 75)[3].

aber	*eben*	(*gleich*)	(*noch*)	*vielleicht*
auch	(*eh*)	*halt*	*nur*	*wohl*
bloß	*eigentlich*	*ja*	*ruhig*	
denn	*einfach*	*mal*	*schon*	
doch	*etwa*	(*nicht*)	(*sowieso*)	

Tabelle 2: In der Fachliteratur (1977–2010) allgemein anerkannte MPn nach Schoonjans (2018: 24)

Beim Abgleich von Schoonjans' (2018) Liste mit dem FOLK-Recherche-Ergebnis fallen zwei Dinge ins Auge:

1. Die in FOLK ermittelten Partikeln *ausgerechnet*, *echt*, *einmal*, *fei*, *glatt*, *nun*, *überhaupt*, *wirklich* und *zu* werden gewöhnlich nicht als MPn anerkannt (siehe 2.1.1 Zweifelhafte MPn).
2. Umgekehrt fehlen in FOLK die von einer Mehrheit anerkannten MPn *eigentlich*, *etwa* und *vielleicht* (siehe 2.1.2 Nicht erfasste MPn).

[3] Von den 17 mehrheitlich anerkannten MPn fehlt dort nur *auch*; zusätzlich wird *nun mal* genannt.

2.1.1 Zweifelhafte MPn

Um zu verstehen, wie es möglich ist, dass in FOLK einige eher zweifelhafte Kandidaten als PTKMA getaggt wurden, ist zu beachten, dass zwar für „alle mehr oder weniger geschlossene[n] Wortartenklassen“ (Westpfahl 2020: 325), nicht jedoch für die MPn Wortformlisten entwickelt wurden (cf. ibd. und Westpfahl/Schmidt 2016: 1496)[4]. Die Entscheidungen des POS-Taggers beruhen bei den MPn also allein auf dem Training am manuell annotierten (bzw. korrigierten) „Goldstandard“-Korpus mit ca. 100.000 Token (cf. Westpfahl 2020: 257). Die Annotations-Richtlinie für STTS 2.0 (cf. Westpfahl et al. 2017), die der Goldstandard-Annotation zugrunde liegt, enthält eine Reihe von Entscheidungshilfen für die Abgrenzung von MPn und Heterosemen, aber ihrerseits keine Liste möglicher MP-Kandidaten. Grundlegend für die Bestimmung von MPn ist gemäß dieser Richtlinie das distributionelle Kriterium, demzufolge MPn (anders als Adverbien oder Fokuspartikeln) nur im Mittelfeld auftreten können (cf. Westpfahl et al. 2017: 21). Ob die Anwendung dieses Kriteriums auf das Goldstandard-Korpus geeignet ist, ein korrektes Training des POS-Taggers und damit eine zutreffende automatische Annotation von FOLK zu gewährleisten, soll ein Blick auf einige zweifelhafte MP-Kandidaten klären helfen.

	Sprechereignis	Sprecher	Treffer
1	FOLK_00011_01	NK	warum muss ich **ausgerechnet** da
2	FOLK_00021_01	NI	musste **ausgerechnet** das dein dritter spieler werden
3	FOLK_00026_01	AW	muss des doch üben abber dass des **ausgerechnet** bei ner klassenfahrt üben muss mit diesem arsch
4	FOLK_00060_01	KR	und äh äh wenn **ausgerechnet** äh
5	FOLK_00070_01	MO	hatte **ausgerechnet** dass für einen projektabbruch rund eins komma fünf milliar…
6	FOLK_00129_01	MF	mein warum ausgerechnet in warum ist das **ausgerechnet** hier so
7	FOLK_00132_01	KA	nee er hat s nich geschafft passau **ausgerechnet** passau fehlt
8	FOLK_00163_01	CD	etwas weil äh der eberhard hat jetzt **ausgerechnet** meinen computer ausenandergebaut
9	FOLK_00204_01	EW	des sagst **ausgerechnet** du
10	FOLK_00210_01	EJ	die gegen oligarchen gegen korruption aufgetreten hat **ausgerechnet** im osten oligarchen einsetzt als ihre ve vertreter
11	FOLK_00267_01	LV	wieso sacht er wir haben das doch **ausgerechnet** dafür müssen sie doch sechsenhalb oder sieben stunden a…
12	FOLK_00374_01	TH	wieso denn ausgerechnet kommst du jetzt **ausgerechnet** darauf

Ergebnisse 1 bis 12 von 12 (12 / 0 aus-/abgewählt) Seite 1 von 1

Abbildung 2: Treffer in FOLK für das PTKMA-Lemma *ausgerechnet*

Das Lemma *ausgerechnet* ist in FOLK 15-mal als adverbiales oder prädikatives Adjektiv (ADJD) und 12-mal als MP (PTKMA) annotiert, aber nie als Intensitäts-, Fokus- bzw. Gradpartikel (PTKIFG). Ein Blick auf die KWIC-Ausgabe in Abbildung 2 zeigt jedoch, dass es sich bei den mutmaßlichen MPn meist um Fokuspartikeln handelt, während in zwei Fällen (Belege 5 und 11) das Partizip Perfekt des Vollverbs *ausrechnen* vorliegt (das hier fälschlich als Lemma *ausgerechnet* annotiert wurde). Jedenfalls ist kein Beleg zu erkennen, bei dem *ausgerechnet* auf die Mittelfeldstellung beschränkt wäre, so dass eine Klassifizierung als MP in Frage käme. Die 15 ADJD-Treffer entpuppen sich ebenfalls als Fokuspartikeln (10 Belege) bzw. als Partizipien (5 Belege). Wie lässt sich erklären, dass alle 27 Treffer für das Lemma *ausgerechnet* inkorrekt getaggt wurden? Eine Durchsicht des Goldstandard-Korpus[5] zeigt, dass dieses gar keine Vorkommen von *ausgerechnet* enthält, i. e. es fehlte offensichtlich eine adäquate Trainingsbasis für den POS-Tagger.

[4] Bestätigt durch persönliche Mitteilung von Henrike Helmer (AGD/DGD-Support des IDS) vom 20.10.2021.

[5] Leider kann der Goldstandard nicht als Subkorpus bzw. virtuelles Korpus innerhalb der DGD abgefragt werden. Man kann ihn in der DGD aber mit Transkripten und Annotationen herunterladen (Menüpunkt Download → Korpusauswahl: POS Goldstandard FOLK) und dann mit geeigneter Software analysieren.

Ein positives Gegenbeispiel liefert das Lemma *glatt*: Es wird in FOLK 21mal als PTKMA ausgewiesen und 14 mal als ADJD. Die ADJD-Belege sind durchwegs korrekt getaggt, aber auch knapp die Hälfte der MP-Vorkommen hält einer Überprüfung stand, wie Abbildung 3 bestätigt. In der einschlägigen Literatur (cf. Heinrich 2007: 167–176; Autenrieth 2002) ist *glatt* ebenfalls als MP-Kandidat bekannt. Im Goldstandard-Korpus tritt die transkribierte Wortform *glatt* lediglich einmal auf, wobei sie korrekt als MP annotiert ist (das Vorkommen ist identisch mit Beleg 2 in Abbildung 3).

	Sprechereignis	Sprecher	Treffer		
1	FOLK_00021_01	XM1	der brauch	glatt	n tor oder
2	FOLK_00022_01	AW	würd isch	glatt	machen
3	FOLK_00026_01	AW	die behaupten doch	glatt	der hatte früher a de es
4	FOLK_00266_01	JS	mehr hier auftauchen überleg ich mir doch	glatt	meine zelte in deutschland abzubrechen und hierher nach m…
5	FOLK_00267_01	AR	nein also dann muss man sie ma	glatt	vorn kopp sagn willste überhaupt noch
6	FOLK_00287_01	XW	für fümundsiebzig öh pfund würd ich dat	glatt	kaufen
7	FOLK_00311_01	TO	aber wie heißt jetzt fällt mir doch	glatt	dieser na sach ma der architekt
8	FOLK_00313_01	UR	sieht man	glatt	
9	FOLK_00313_01	BB	ah ha ha ha sieht man	glatt	klar sehr gut

Ergebnisse 1 bis 9 von 9 (9 / 0 aus-/abgewählt) Seite 1 von 1

Abbildung 3: Neun korrekte von insgesamt 21 Treffern in FOLK für das PTKMA-Lemma *glatt*

Auch einige weitere Lexeme, die in Schoonjans' (2018: 24) Literatur-Übersicht fehlen, kommen als MP-Kandidaten in Betracht: Von den 1.604 FOLK-Belegen des Lemmas *überhaupt* sind 308 als PTKMA getaggt (z. B. *warum muss ich überhaupt schlafn*, E00014, T01, B0092[6]; *weiß jemand zufälligerweise wer des is überhaupt*, E00004, T02, B0226). Obschon oft nicht berücksichtigt (cf. Thurmair 1989: 9, 27f.; Kwon 2005; Engel 1991; Brünjes 2014), ist *überhaupt* andernorts durchaus als MP (cf. Weinrich 1993: 852f.) oder zumindest als zum „Randbereich" (Zifonun/Hoffmann/Strecker 1997: 1029) gehörig anerkannt.

Auch die Lexeme *echt* (110 Belege) und *wirklich* (43 Belege), die in FOLK – abweichend von der Literatur[7] – häufig als MP getaggt sind, scheinen teilweise aufs Mittelfeld beschränkt zu sein (*der krankenwagen der is echt schon sehr alt bestimmt*, E00175, T02, B0002; *die wehren sich wirklich dagegen*, E00185, T01, B0356)[8] und verdienen zumindest eine genauere Analyse. Gleiches gilt für die Partikel *nun*: Zwar können die 5 angeblichen MP-Vorkommen in FOLK kaum überzeugen, aber das zweigliedrige *nun mal* (mit gleicher Bedeutung wie die MPn *halt* und *eben*), das im Goldstandard-Subkorpus drei Mal mit der doppelten Annotation PTKMA versehen wurde (z. B. in *die schwaben sind nun mal damals die peripherien*, E00059, T01, B0545)[9] ist ein überzeugender MP-Kandidat. Bei den beiden in FOLK als PTKMA getaggten

[6] Hier und im Folgenden wird eine abgekürzte Schreibweise verwendet: E00014, T01, B0092 steht für: FOLK_E_00014_SE_01_T_01, 0092. Darin ist E_00014 die Ereignis-ID, SE_01 die Sprechereignis-Nummer (stets =01 in den zitierten Belegen), T_01 die Transkriptnummer, während nach dem Komma die Nummer des Beitrags angegeben ist. Der hier zitierte Beleg kann in FOLK aufgefunden werden, indem man in den DGD-Metadaten die DGD-Kennung FOLK_E_00014 sucht und vom Ergebnis ausgehend das Transkript 01 aufruft.

[7] Autenrieth 2002 analysiert immerhin die untergegangene Form *e(cher)t*.

[8] Jedenfalls scheint *wirklich* nicht allein das Vorfeld besetzen zu können. Eine Recherche in FOLK (Lemma *wirklich*, Kontext 1 Token rechts: VAFIN) hat bei L1-Sprecher:innen keine alleinige Vorfeldbesetzung belegen können. Eines der wenigen L2-Beispiele liefert E00160, T02, B0395: *wirklisch hast du gesagt äh hast du keine parkplatz gefunden*.

[9] Beim automatischen POS-Tagging im FOLK-(Gesamt)Korpus dagegen ist *nun* als ADV annotiert.

Vorkommen des Lemmas *einmal* handelt es sich hingegen um dialektale Varianten (*wart amal*; *sog emal*), die vermutlich (wie in zwei anderen Fällen geschehen) besser als *mal* lemmatisiert worden wären. Die ebenfalls dialektale MP *fei* (10 Belege in FOLK, z. B. *des is fei komisch heid*, E00319, T01, B2921) kommt als eigene, nicht standardsprachliche MP in Betracht (vgl. das Lemma „fei“ im Onlinewörterbuch DWDS). Bei den drei Belegen für *zu* als MP handelt es sich um Fehlklassifizierungen von Präposition und Gradpartikel.

Zusammenfassend lässt sich sagen, dass die in FOLK (siehe Tabelle 1) zu Tage tretenden Abweichungen vom gängigen Kanon anerkannter MPn (siehe Tabelle 2) nur zum kleineren Teil auf offensichtlich problematischen POS-Annotationen (*ausgerechnet*, *einmal*, *zu*) beruhen, während in den meisten Fällen (*glatt*, *überhaupt*, *echt*, *wirklich*, *nun* bzw. *nun mal*, *fei*) eine MP-Funktion zumindest nicht ausgeschlossen werden kann. Dass dem POS-Tagging der MPn keine geschlossene Wortliste zugrunde liegt, erscheint vor diesem Hintergrund nicht als Nachteil, sondern als Entscheidung, die es ermöglicht hat, neue MP-Kandidaten zu ermitteln.

2.1.2 Nicht erfasste MPn

In der quantifizierten MP-Liste von FOLK (Tabelle 1) fehlen die in Tabelle 2 aufgeführten und mehrheitlich anerkannten MPn *eigentlich*, *etwa* und *vielleicht*.

Sämtliche 5509 Vorkommen der nicht flektierten Form *eigentlich* in FOLK sind als Adverb (ADV) getaggt. Selbstverständlich sind darunter auch (vorfeldfähige) adverbiale Verwendungen (z. B. *hm eigentlich will ich im moment informatik studieren*, E00129, T01, B0877), das Problem aber sind modale Verwendungen, die nicht als solche klassifiziert wurden:

(1) E00021, T02, B0242: *is deine familie eigentlich geflüchtet* [Fokusakzent auf *flüch*]
(2) E00021, T15, B1442: *was gibt s eigentlich noch für torhüter so* [Fokusakzent *tor*]
(3) E00026, T03, B0101: *wo is die hin eigentlich* [Fokusakzent auf *hin*]

Die Gründe liegen vermutlich in der (manuellen) Annotation des Goldstandard-Korpus, bei der von 15 Vorkommen der MP *eigentlich* nur eine einzige als PTKMA gekennzeichnet wurde (siehe Tabelle 3).[10]

E00005, T02_DF_01_S_1	**LB**	was will er	**eigentlich**	erreischen (.) was will (.) herr kleink	ADV
E00005, T02_DF_01_S_1	**LB**	on den andern hörn (.) welche ghörn denn	**eigentlich**	zum system	ADV
E00009, T01_DF_01_S_1	**LB**	gut (.) was isch	**eigentlich**	ungewöhnlich wenn man die sekundärspule	ADV
E00016, T01_DF_01_S_1	**CJ**	was spielen die denn	**eigentlich**	indiander und cowboy oder so was	ADV
E00020, T01_DF_01_S_2	**HM**	wieso hat	**eigentlich**	niemand ne gscheite kühltruhe	PTKMA
E00040, T01_DF_01_S_2	**EP**	wen ham die	**eigenlisch**	gholt als ersatz für de ronaldo	ADV
E00040, T01_DF_01_S_2	**EP**	°h (.) was_n jetz	**eigenlisch**	obe bei eusch in der garasch wo du vorh	ADV
E00040, T03_DF_01_S_1	**EP**	was_enn	**eigenlisch**	mi_m linsenreich schon wieder jemand ge	ADV
E00053, T01_DF_01_S_3	**LS**	[dass ich mich … frage] wie dumm die	**eigentlich**	is dass die sich da wirklich au fast so	ADV
E00121, T01_DF_01_S_1	**BB**	kommt da	**eigentlich**	net en minus hin +++ +++	ADV
E00135, T01_DF_01_S_1	**NH5**	warum hab ich	**eigentlich**		ADV
E00136, T01_DF_01_S_1	**XM**	(.) is des	**eigentlich**	in euerm sinne	ADV
E00139, T01_DF_01_S_1	**NH10**	is sie	**eigenlisch**	vaschwitzt oder so	ADV
E00144, T01_DF_01_S_1	**LS**	or dem hintergrund wie handhabsch du des	**eigentlich**	machsch du des da auch dass du da sagsc	ADV
E00185, T02_DF_01_S_1	**MF**	unwahrscheinlich wie viel schüler hat	**eigentlich**	das gymnasium hier	ADV

Tabelle 3: Vorkommen der MP *eigentlich* mit jeweiliger POS-Annotation im Goldstandard-Korpus

[10] Die 135 Vorkommen der transkribierten Formen *eigentlich/eigenlisch* wurden mit EXMARaLDA EXAKT 1.3 aus den Transkriptionen (.fln) ermittelt. Die POS-Tags wurden in den zugehörigen xml-Dateien gesucht.

Es ist anzunehmen, dass der POS-Tagger im Training mit dem so annotierten Goldstandard keine korrekten Parameter für die Identifizierung der MP *eigentlich* ausbilden konnte.

Auch die 246 Vorkommen des Lemmas *etwa* sind durchgängig als ADV getaggt,[11] obwohl modale Verwendungen in FOLK belegt sind, wie die folgenden Beispiele zeigen:

(4) E00243, T01, B0501 *geht s etwa schon weiter*
(5) E00260, T02, B0045 *hast du etwa so was wie n geburtstagstisch vorbereitet*
(6) E00355, T01, B0809 *magst du mira marco mich und maja etwa nich*

Der Fehler beruht hier nicht auf einer inkorrekten Annotation des Trainingskorpus, sondern auf der Tatsache, dass dieses keinerlei Vorkommen der MP *etwa* enthält.

Für das Lexem *vielleicht* enthält FOLK 4403 Belege, die allesamt als Adverb getaggt sind. Darunter sind zweifellos zahlreiche adverbiale Verwendungen (z. B. *vielleicht könn mer des ganze mal von vorne mache*; E00001, T01, B171), aber auch die relative seltene MP *vielleicht* (cf. Brünjes 2014: 165f.) ist belegt, und zwar in Exklamativsätzen:

(7) E00424, T01, B0327 (*ai*) *du bisch vielleicht n clown*
(8) E00329, T03, B1184 *das is vielleicht lecker*
(9) E00430, T02, B0267 […] *das war vielleicht witzig* […]
(10) E00308, T02, B0162 […] *das is vielleicht en mist ah*

Hier drückt die MP *vielleicht* aus, dass der „Inhalt der Äußerung erwartet [wurde], allerdings in einem geringeren Ausmaß“ (Brünjes 2014: 167; cf. auch Thurmair 1989: 192f.). Eine Interpretation als Modalwort zwecks „Einschätzung der Wahrscheinlichkeit eines Sachverhalts“ wäre inkompatibel mit dem Exklamativmodus (cf. Brünjes 2014: 167). Für die Verwendung in rhetorischen Entscheidungsfragen (z. B. *Sollten wir vielleicht tatenlos zusehen*?, Engel 1991: 238), in denen *vielleicht* durch die MP *etwa* ersetzt werden kann (cf. Thurmair 1989: 194; Cognola/Coniglio 2026) konnten bei einer kursorischen Durchsicht keine Beispiele in FOLK gefunden werden. Dagegen gibt es Belege für die MP *vielleicht* in indirekten Aufforderungen (z. B. *kannst du vielleicht hier noch ma nachkucken*, E00329, T01, B1264). Diese aber sind laut Brünjes (2014: 169) prinzipiell ambig und können auch als Modalwort bzw. Adverb interpretiert werden. In dem zitierten Beleg könnte *vielleicht* als Modalwort eine freundliche, unaufdringliche Bitte ausdrücken. Nur bei Signalisierung von Ungeduld erkennt Brünjes *vielleicht* in indirekten Aufforderungen als MP an, z. B. bei einem unfreundlich geäußerten *Würdest du mir vielleicht mal zuhören?*. Die Partikel verweist dann laut Brünjes (2014: 170) auf eine weniger nachdrückliche Variante des Sprechakts, wodurch die tatsächlich geäußerte Aufforderung als stärker markiert wird.

Genau wie in FOLK ist *vielleicht* auch im Goldstandard-Korpus durchgehend als Adverb annotiert. Rhetorische Fragen und Exklamative sind im Trainingskorpus nicht belegt. So ist erklärlich, dass die oben angeführten Belege für Exklamative vom POS-Tagger nicht identifiziert werden konnten. Für Aufforderungen existieren zwar einige Belege (z. B. […] *ruft ihn*

[11] Die sowohl in FOLK (z. B. *etwa zwanzig güterzüge*) wie im Goldstandard (z. B. *etwa fünf millimeter*) zahlreich vorkommende Gradpartikel ist überraschenderweise ebenfalls durchgängig als Adverb (ADV) annotiert.

vielleicht grad nochma an drüben, E00111, T01, B0117), aber diese könnten aufgrund ihrer prinzipiellen Ambiguität als adverbiale Verwendung annotiert worden sein.

Zusammenfassend können wir sagen, dass bei *eigentlich* die problematische (manuelle) Annotation des Goldstandard-Subkorpus aller Voraussicht nach ein angemessenes Training des POS-Taggers und somit eine zuverlässige automatische Annotation des FOLK-Korpus verhindert hat. Bei *etwa* dagegen liegt das lückenhafte Training des POS-Taggers am Mangel von MP-Vorkommen im Goldstandard-Subkorpus. Im Goldstandard-Material fehlen auch Belege für die MP *vielleicht* in Exklamativsätzen, so dass die entsprechenden Belege in FOLK vom POS-Tagger nicht korrekt identifiziert werden konnten. Andere Verwendungen (rhetorische Fragen, indirekte Aufforderungen) sind entweder nicht belegt oder weisen Ambiguitäten auf.

2.2 Relative Häufigkeit der erfassten MPn

Während in Abschnitt 2.1 untersucht wurde, inwieweit die in FOLK ermittelten MPn mit den in der Fachliteratur üblichen Auflistungen übereinstimmen, geht es im Folgenden um die Frage, ob die in FOLK berechneten Häufigkeiten (siehe Tabelle 1) mit Angaben in der einschlägigen Forschung übereinstimmen und welche Gründe es für Abweichungen geben könnte. Als Vergleichsbasis dienen Hentschel (1986) und Brünjes (2014), die einzigen mir bekannten Arbeiten, die korpusbasierte quantitative Untersuchungen zu einer größeren Zahl von MPn präsentieren.[12] Die beiden Korpora und die darin ermittelten MP-Häufigkeiten werden zunächst kurz vorgestellt.

Das Korpus von Hentschel (1986: 240) umfasst 18 Gespräche aus dem Freiburger Korpus (FR, cf. Fuchs/Schenk 1975) sowie 5 Gespräche, die von Studierenden der FU Berlin im Rahmen von Lehrveranstaltungen aufgezeichnet wurden. Laut Hentschel können die Gespräche des Freiburger Korpus (FR) mit Einschränkungen „als eine Zufallsstichprobe aus dem Bereich alltäglicher Kommunikation betrachtet werden“ (ibd.: 239). Die Texte stammen aus dem südwestdeutschen Raum, dialektale Einflüsse im Partikelgebrauch können – so die Autorin (ibd.: 240) – nicht völlig ausgeschlossen werden, auch wenn es dafür keine konkreten Anhaltspunkte gebe. Dagegen waren an den von Studierenden in Berlin aufgezeichneten Gespräche nur Norddeutsche beteiligt. Der Anteil der MPn am „Gesamtwortaufkommen“, i. e. die MP-Tokenrate, liegt insgesamt bei 2,2%, schwankt bei den einzelnen Texten jedoch zwischen 0,91% und 5,13% (cf. Hentschel 1986: 240). Die Autorin weist eine Korrelation zwischen MP-Frequenz und Privatheitsgrad der Gespräche nach (cf. ibd.: 238–246). Da Hentschel (1986: 247) insgesamt 467 Partikelbelege angibt, dürfte ihr Korpus

[12] Thurmair (1989) hat ca. 2.000 Belege aus unterschiedlichen Quellen gesammelt (cf. ibd.: 5), darunter auch einzelne Hörbelege in Form von Gedächtnisprotokollen (cf. ibd.: 6), was quantitative Aussagen nicht erlaubt. Kwon (2005), Coniglio (2011) und Moroni (2010) behandeln Fragen der Frequenz nicht, Schoonjans (2018) nur mit Bezug auf begleitende Gestik/Mimik, Müller (2018) beschränkt auf MP-Kombinationen. Hentschel/Keller (2006) untersuchen ein sehr spezielles Korpus von Interviews mit jungen Müttern und schließen einige MPn aus der Analyse aus (*auch, denn, mal, wohl*). Die 2020 abgeschlossene Dissertation von Dagmar Silberstein lag mir bei Abfassung des Beitrags im Frühjahr 2022 nicht vor. Sie wurde als Silberstein (2024) publiziert.

ca. 21.200 Token umfassen, i. e. ca. jedes 45. Wort ist eine MP. Eine Übersicht über die Häufigkeit der einzelnen MPn in ihrem Korpus gibt Tabelle 4.

Partikel	Einzelvor-kommen	Vorkommen in Kombination	Gesamt	in %
ja	112	11	123	26,3
doch	72	12	84	18,0
mal	66	9	75	16,1
auch	34	8	42	9,0
eben	39	2	41	8,8
denn	23	0	23	4,9
schon	13	4	15	3,2
eigentlich	11	2	12	2,8
einfach	10	2	12	2,6
wohl	8	3	11	2,4
halt	9	0	9	1,9
aber	2	1	3	0,6
bloß	1	0	1	0,2
etwa	0	1	1	0,2
nun nur	1	0	1	0,2
mal	0	1	1	0,2
ruhig	1	0	1	0,2
Partikeln mit abtönungsähnlichen Funktionen:				
überhaupt	5	0	5	1,1
immerhin	2	0	2	0,4
jedenfalls	3	0	3	0,6
allerdings	2	0	2	0,4
SUMME:	**411**	**56**	**467**	**100,1**

Tabelle 4: Tabelle der MP-Häufigkeiten des Hentschel-Korpus (Hentschel 1986: 247)[13]

Brünjes (2014: 2f.) beschreibt ihr erheblich größeres Korpus wie folgt:

> Die Datenbasis bildete das Korpus Gespräche im Fernsehen, das vom Institut für deutsche Sprache (IDS) in Mannheim zur Verfügung gestellt wird. Das Korpus beinhaltet unterschiedliche Fernsehsendungen (Talkshows, Diskussionen, Interviews), die in den Jahren 1989 bis 2003 ausgestrahlt wurden. Die zur Verfügung gestellten 68 Transkripte umfassen insgesamt 740.826 Wörter.[14]

Brünjes (2014: 179f.) ermittelt in dem von ihr untersuchten Korpus von Fernsehsendungen die in Tabelle 5 gezeigten Häufigkeiten von MPn bzw. den entsprechenden Lexemen (inkl. Heteroseme). Bei mehr als 500 Belegen für eine MP ist die Gesamtzahl geschätzt:

[13] Kleinere Unstimmigkeiten in der Tabelle, z. B. bei der Berechnung der Gesamtvorkommen von *schon* und *eigentlich*, wurden unverändert übernommen.

[14] Ein Link zu Informationen über das Korpus Gespräche im Fernsehen (GF) findet sich im Literaturverzeichnis (Korpora und Online-Ressourcen). GF ist nicht Teil der DGD und kann nur über den persönlichen Service des AGD zugänglich gemacht werden, und zwar in einem Umfang von maximal 60 Transkripten.

	Modalpartikelbelege		davon ambig	Gesamtbeleganzahl des Lexems	
aber		27	3		3.403
auch	500/	2.415	-	984/	4.752
bloß		4	-		62
denn	500/	623	-	842/	1.039
doch	500/	1.713	-	524/	1.803
eben		453	-		696
eigentlich		193	-		835
etwa		10	-		131
halt		223	-		258
ja	500/	3.956	-	1.243/	9.890
mal		211	-		2.066
nur		9	-		1.499
schon		249	6		1.274
vielleicht		4	3		734
wohl		17	6		147

Tabelle 5: Überblick über die MP-Frequenzen im Korpus von Brünjes (2014: 179f.)

Die Gesamtzahl der von Brünjes (2014) in ihrem Korpus ermittelten MP-Vorkommen beträgt 10.107 (meine Berechnung). Bei einer Korpusgröße von 740.826 laufenden Wörtern ergibt sich eine Tokenrate von 1,36%, i. e. ca. jedes 73. Wort ist eine MP. In Tabelle 6 werden die MP-Tokenraten der drei hier verglichenen Korpora gegenübergestellt. Auffällig ist die deutlich geringere Tokenrate im Fernsehgespräch-Korpus von Brünjes.

Korpus	Korpusgröße	MP-Gesamt-vorkommen	MP-Tokenrate
Hentschel (1986)	ca. 21.300	467	2,20% (jedes 45. Wort)
Brünjes (2014)	740.826	10.107	1,36% (jedes 73. Wort)
FOLK (2021)	2.990.421	76.331	2,55% (jedes 39. Wort)

Tabelle 6: Vergleich der MP-Tokenraten in FOLK und den beiden Vergleichskorpora

Welche Erklärungen könnte es dafür geben? Zunächst einmal ist denkbar, dass im Brünjes-Fernsehkorpus weniger informelle Gespräche vertreten sind als in den beiden anderen Korpora. Wie erwähnt, hat Hentschel (1986: 238–246) gezeigt, dass die MP-Häufigkeit vom „Privatheitsgrad“ der Gespräche abhängt (cf. auch Thurmair 1989: 4). Es ist nun durchaus vorstellbar, dass Gespräche im Fernsehen vor Publikum (im Sendesaal und zu Hause), in einer für die Gäste wenig vertrauten Umgebung und zwischen Gesprächspartnern, deren Vertrautheitsgrad ebenfalls eher gering ist, anders verlaufen als die (in FOLK vielfach vertretenen) informellen und oftmals sehr assoziativen Gespräche im Familien- und Freundeskreis und somit eine geringere MP-Tokenrate aufweisen (cf. Silberstein 2021: 703). Auch Brünjes selbst erwägt bei der Darstellung einzelner MPn den Einfluss der Merkmale ihres Fernsehgesprächskorpus auf die jeweilige Häufigkeit (cf. Brünjes 2014: 103, 116, 132, 143, 166). Unterschiede hinsichtlich der Korpusmerkmale müssen immer mitbedacht werden, wenn Häufigkeiten von MPn – einzeln oder in ihrer Gesamtheit – miteinander verglichen werden. Eine zweite denkbare Erklärung

dafür, dass die MP-Tokenrate in FOLK fast doppelt so hoch ausfällt wie bei Brünjes (2014), wäre, dass beim automatischen POS-Tagging in FOLK irrtümlich eine Vielzahl von Heterosemen als MPn gewertet wurden. Angesichts der ebenfalls recht hohen Häufigkeit von MPn im manuell ausgewerteten Korpus von Hentschel (1986), vermag eine solche Erklärung die beobachteten Differenzen aber kaum in vollem Umfang zu erklären. Auf jeden Fall sind beide Erklärungsansätze auch bei der Untersuchung der Häufigkeitsunterschiede einzelner MPn im Blick zu behalten. Zwecks besser Vergleichbarkeit beschränke ich mich dabei im Folgenden auf die 17 MPn, die laut Schoonjans (2018: 24) von einer Mehrheit von Autor:innen anerkannt sind. In Tabelle 7 werden die absoluten Häufigkeiten, die prozentualen Anteile an allen (berücksichtigten) MPn und die Tokenraten dieser 17 MPn in den untersuchten drei Korpora gegenübergestellt.

	Zahl der MP-Belege			**Anteil an den MP-Belegen im jeweiligen Korpus**			**MP-Tokenrate**		
	Hentschel (1986)	**Brünjes (2014)**	**FOLK (2021)**	**Hentschel (1986)**	**Brünjes (2014)**	**FOLK (2021)**	**Hentschel (1986)**	**Brünjes (2014)**	**FOLK (2021)**
				=a/453	**=b/10.107**	**=c/75.817**	**=a/21.200**	**=b/740.826**	**=c/2.990.421**
	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**
Aber	3	27	4.384	1%	0%	6%	0,01%	0,00%	0,15%
Auch	42	2.415	29	**9%**	**24%**	**0%**	0,20%	0,33%	0,00%
Bloß	1	4	139	0%	0%	0%	0,00%	0,00%	0,00%
Denn	23	623	4.876	5%	6%	6%	0,11%	0,08%	0,16%
Doch	84	1.713	7.070	19%	17%	9%	0,40%	0,23%	0,24%
Eben	41	453	3.038	9%	4%	4%	0,19%	0,06%	0,10%
Eigentlich	12	193	-	3%	2%	-	0,06%	0,03%	-
Einfach	12	-	5.352	3%	-	7%	0,06%	-	0,18%
Etwa	1	10	-	0%	0%	-	0,00%	0,00%	-
Halt	9	223	9.694	**2%**	**2%**	**13%**	0,04%	0,03%	0,32%
Ja	123	3.956	23.882	27%	39%	31%	0,58%	0,53%	0,80%
Mal	75	211	10.583	**17%**	**2%**	**14%**	0,35%	0,03%	0,35%
Nur	-	9	23	-	0%	0%	-	0,00%	0,00%
Ruhig	1	-	114	0%	-	0%	0,00%	-	0,00%
Schon	15	249	6.154	3%	2%	8%	0,07%	0,03%	0,21%
Vielleicht	-	4	-	-	0%	-	-	0,00%	-
Wohl	11	17	479	2%	0%	1%	0,05%	0,00%	0,02%
Summe	**453**	**10.107**	**75.817**	**100%**	**100%**	**100%**	**2,14%**	**1,36%**	**2,54%**

Tabelle 7: Vergleich der Häufigkeiten, prozentualen Anteile und Tokenraten der MPn[15]

[15] Die MP-Gesamt-Tokenraten in der Tabelle (rechts unten) weichen von den in Tabelle 6 genannten Werten z. T. geringfügig ab, da einige MP aus der Berechnung ausgeschlossen wurden.

Was die Frequenz der einzelnen MPn betrifft, so ist der Tabelle zu entnehmen, dass in allen drei Korpusanalysen die MP *ja* die Rangliste anführt (27%-39%), was sich auch in hohen Tokenraten (0,53–0,80%) niederschlägt. Tabelle 8 gibt Aufschluss darüber, welche fünf MPn in den drei Korpora am häufigsten sind.

Rang	Korpus von Hentschel (1986)	Korpus von Brünjes (2014)	FOLK (Stand 2021)
1	ja 27%	ja 39%	ja 31%
2	doch 19%	auch 24%	mal 14%
3	mal 17%	doch 17%	halt 13%
4	auch 9%	denn 6%	doch 9%
5	eben 9%	eben 4%	schon 8%

Tabelle 8: Rangliste der häufigsten MPn in den angegebenen Korpora

Außer *ja* steht nur *doch* in allen drei Ranglisten auf den ersten fünf Plätzen. *Mal*, *auch* und *eben* sind je zweimal vertreten; *denn*, *halt* und *schon* nur einmal. Die MPn *mal*, *auch* und *halt* sollen im Folgenden genauer untersucht werden, denn wie aus Tabelle 7 ersichtlich, ergeben sich bei diesen drei Partikeln hinsichtlich ihres Anteils am MP-Gesamtvorkommen gravierende Differenzen: bei *auch* fällt FOLK mit einem Anteil von (gerundet) 0% aus der Reihe, während die Anteile bei Hentschel (9%) und Brünjes (24%) sehr viel höher liegen. Bei *halt* ist es genau umgekehrt: in FOLK rangiert diese MP mit 13% an dritter Stelle, in den anderen beiden Korpora mit je 2% in der Schlussgruppe. Die MP *mal* zeigt dagegen vergleichbare Häufigkeiten bei Hentschel (17%, Rang 3) und in FOLK (14%, Rang 2), während das Korpus von Brünjes mit lediglich 2% aus der Reihe fällt. In Abschnitt 3 werden mögliche Gründe für die auffälligen Frequenzunterschiede bei den drei MPn *auch*, *halt*, *mal* erörtert. Tabelle 9 zeigt für Brünjes (2014) und FOLK den MP-Anteil an der Gesamtzahl der Lexeme (inkl. Heteroseme).

	Brünjes (2014)		FOLK (2021)	
	Gesamtzahl der Belege	Anteil der MPn	Gesamtzahl der Belege	Anteil der MPn
auch	4.752	51%	37.661	0%
halt	258	86%	9.780	99%
mal	2.066	10%	21.243	50%

Tabelle 9: Anteil der MPn an der Gesamtzahl der Belege für die Lexeme *auch*, *halt*, *mal* (inkl. Heteroseme)

3 Analyse auffälliger Unterschiede in der Häufigkeit von MPn

In diesem Kapitel sollen exemplarisch die Häufigkeiten der MPn *auch*, *mal* und *halt* genauer analysiert werden, um mögliche Gründe für die (in Abschnitt 2.2 dargestellten) eklatanten Unterschiede zwischen den drei Korpora zu finden. Zu klären ist insbesondere, welche Rolle Fehler beim automatischen POS-Tagging, i. e. bei der Abgrenzung der MPn von ihren jeweiligen Heterosemen spielen bzw. ob umgekehrt eher Korpusmerkmale und/oder abweichende MP-Definitionen als Erklärung heranzuziehen sind. Die Untersuchung konzentriert sich auf die beiden größeren Korpora, i. e. FOLK (Stand 2021) und das Korpus von Brünjes (2014), für die auch Daten zu den Heterosemen vorliegen.

3.1 Die MP *auch*

In FOLK (2021) finden sich 37.661 Belege für das Lemma *auch* (Tokenrate 1,26%), von denen aber nur 29 Treffer (entsprechend 0,1% des Gesamtvorkommens) als MP (i. e. als PTKMA in STTS 2.0) getaggt sind; 99,9% der Belege sind mithin Heteroseme (586 ADV, 37.046 PTKIFG). Für das Fernsehgespräch-Korpus gibt Brünjes (2014) 4.752 Lexem-Treffer an, was einer fast exakt halbierten Tokenrate von 0,64% entspricht. Andererseits jedoch ermittelt sie einen deutlich höheren Anteil der MP *auch*: Mit 2.415 Belegen entfallen 50,8% des Gesamtvorkommens auf die MP, während die Heteroseme mit 2.337 Belegen die andere Hälfte (49,2%) bilden (siehe Abbildung 4).

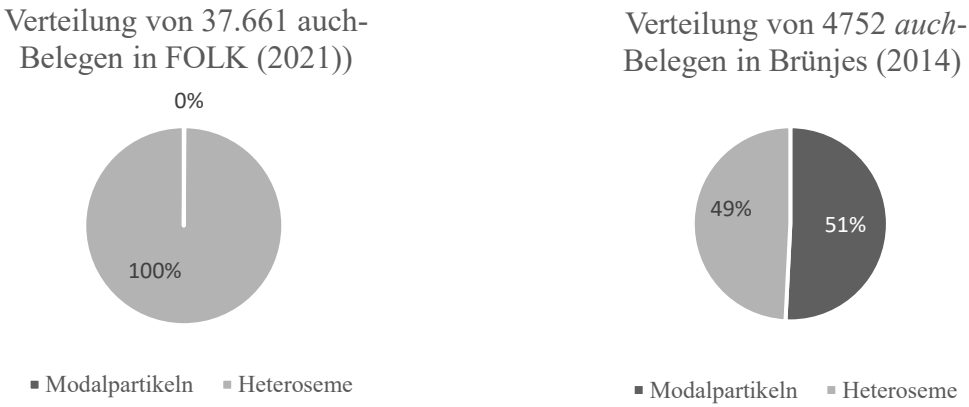

Abbildung 4: Vergleich des Anteils der MP *auch* am Gesamtvorkommen des Lexems

Die fundamental unterschiedlichen Anteile von MP und additiver Fokuspartikel erklären, warum in FOLK die Tokenrate der MP *auch* – trotz eines großen Vorsprungs beim Gesamtvorkommen des Lexems – gegen Null tendiert (0,001%), während sie im Korpus von Brünjes (2014) den (330-mal so hohen) Wert von 0,33% annimmt. Da die MP *auch* im Korpus von Hentschel (1986) eine Tokenrate vergleichbarer Größenordnung (0,20%) erreicht, ist zu vermuten, dass nicht Korpusmerkmale, sondern die definitorische Abgrenzung der MP von ihren Heterosemen und/oder die Anwendung dieses Kriteriums beim automatisierten POS-Tagging für die Häufigkeitsunterschiede verantwortlich sind.

Die MP *auch* ist offenbar nur schwer von der gleichlautenden Fokuspartikel und vom Adverb *auch* zu unterscheiden, denn in den STTS 2.0-Guidelines (cf. Westpfahl et al. 2017: 33f.) ist ihr ein eigener Abschnitt im Kapitel über „Abgrenzungsprobleme“ gewidmet (*mal*, *nur*, *schon* sind weitere hier behandelte MPn, cf. auch Westpfahl 2020: 313, 318). Entsprechend dem distributionellen Kriterium wird *auch* als (Konjunktional-)Adverb (ADV) eingestuft, wenn es allein im Vorfeld steht (*Auch hatte niemand daran gedacht, die Verantwortlichen zu fragen*.). Auch bei Stellung im Mittelfeld wird es als Adverb gewertet, wenn ein „Bezug auf kognitive Verben“ (ibd.: 33) vorliegt: *Ich meine*/*sage*/*denke*/*auch* […]). Aus meiner Sicht gilt dies aber nur bei additiver Bedeutung, i. e. wenn *auch* durch *im Übrigen*, *zudem*, *und* ersetzbar ist, nicht akzentuiert ist und in gleicher Bedeutung nicht gemeinsam mit dem Subjekt ins Vorfeld rücken kann (*Auch ich meine/ sage* […]).

Zur Funktion als Fokuspartikel (PTKIFG) heißt es in den *Guidelines*:

> *Auch* kann als Fokuspartikel auftreten, hierbei drückt *auch* zusammen mit seinem Bezugsausdruck eine Alternative bzw. einen weiteren Faktor des Gesagten aus. Die Bezugsausdrücke können beliebig komplex sein, *auch* steht dabei meistens vor dem Bezugsausdruck, kann aber auch dahinter bzw. in Distanzstellung stehen, jedoch nie alleine im Vorfeld.[16]
>
> (ibd.: 33)

Alle neun Beispiele für Mittelfeldstellung der Fokuspartikel illustrieren die Stellung direkt vor dem (unterstrichenen) Bezugsausdruck (*der Leo hat auch eine Sonnenbrille*). Für die Nach- bzw. Distanzstellung wird kein Beispiel gegeben, obwohl eine entsprechende Interpretation bei manchen Beispielen (*hätte ich auch so ein Tier zum Schmusen gewollt*) durchaus naheläge (*hätte ich AUCH so ein Tier zum Schmusen gewollt*). Deutlich wird, dass ohne einen größeren Kontext (und ohne prosodische Information) der Bezugsausdruck der Fokuspartikel kaum zuverlässig ermittelt werden kann. Für manche Beispiele (*die Folien sind auch wirklich gut*; *ich habe auch gearbeitet heute*) lassen sich sogar Kontexte vorstellen, in denen *auch* als MP fungiert (A: *Ich finde die Folien eigentlich ganz gut.* – B: *Die Folien SIND auch wirklich gut, aber das Vortragstempo war zu schnell*; A: *Du siehst erschöpft aus.* – B: *Ich habe auch geARbeitet heute*.).

Auch als MP ist laut den Guidelines (cf. Westpfahl et al. 2017: 33) ans Mittelfeld gebunden, bildet keine Phrasen und kann nicht erfragt werden. Diese aus dem „Grammatischen Informationssystem“ (grammis) des IDS übernommenen Merkmale[17] taugen aber nur bedingt zur Abgrenzung von Fokuspartikeln (die ebenfalls nicht erfragbar sind und keine Phrasen bilden). Ferner wird darauf hingewiesen, dass die MP *auch* mit anderen MPn wie z. B. *ja* kombiniert werden kann und häufig in Frage- oder Aufforderungssätzen auftritt (Beispiele: *Warum gehst du auch immer so spät ins Bett*; *Hast du auch schon deine Hausaufgaben gemacht*; *Wie auch immer*; *Du musst ja auch nicht immer petzen*). Die Kombinierbarkeit mit der MP *ja* dürfte aber ebenso wenig eine eindeutige Abgrenzung von Adverbien und Fokuspartikeln (z. B. *Er hat ja AUCH keine Zeit*) gewährleisten, so dass zu vermuten ist, dass sich die Annotator:innen vor allem am Satzmodus orientiert haben könnten.

Brünjes (2014) geht auf die Funktion von *auch* als Gradpartikel und Konjunktionaladverb nur sehr kurz ein (ibd.: 95), diskutiert aber ausführlich die Verwendung als MP, und zwar in Assertionen, Direktiva und Erotetika (cf. ibd.: 95–101; Thurmair 1989: 155–160).[18] Anknüpfend an Diewalds (1997) Konzept des „pragmatischen Prätexts“ arbeitet sie an Beispielen aus ihrem Fernsehgesprächskorpus (stets mit Kontextbeschreibung) folgendes gemeinsames Kennzeichen heraus: „*Auch* verweist auf eine Proposition p, die als mit der Situation normalerweise verknüpft, i. e. als Standardannahme in dieser Situation dargestellt wird. Die präsupponierte Proposition ist identisch mit der Proposition der tatsächlichen Äußerung.“ Brünjes‘ ausführliche Beleg-Analysen können hier aus Raumgründen nicht nachvollzogen werden, aber das Grundprinzip sei an dem bereits eingeführten Beispiel (A: *Du siehst erschöpft aus.* – B: *Ich*

[16] Bezüglich der Distanzstellung verweisen die Autor:innen auf den Artikel „Fokuspartikel“ im Grammatischen Informationssystem (grammis) des IDS.

[17] Artikel „Abtönungspartikel“ in der systematischen Grammatik von grammis.

[18] Cf. auch die kontrastiven Studien (Deutsch-Italienisch) von Moroni/Bidese (2021) und Cognola/Moroni/Bidese (2022).

habe auch geARbeitet heute.) illustriert. Die Standardannahme ist in diesem Fall, dass B heute gearbeitet hat, denn Erschöpfung ist oft die Folge von Arbeit (andere mögliche Gründe wären zu wenig Schlaf, sportliche Betätigung etc.). Diese Proposition entspricht der tatsächlichen Äußerung *Ich habe gearbeitet heute*. Mit der Verwendung der MP *auch* signalisiert die Sprecherin, dass das Gearbeitet-Haben in der gegebenen Situation (sichtbare Erschöpfung) eine Standardannahme ist und dass diese auch (!) zutrifft. Resultat ist laut Brünjes (2014: 96) eine „augmentative Relation zwischen präsupponierter Einheit und relevanter Situation".

Vergleicht man die Darstellung bei Westpfahl et al. (2017) mit der bei Brünjes (2014), so wird deutlich, dass die Beispiele in den STTS 2.0-Guidelines aufgrund fehlenden Kontextes oft wenig aussagekräftig und die distributionellen Kriterien nicht immer hilfreich sind, während die ausführliche pragmatische Charakterisierung der Verwendung als MP bei Brünjes (2014) die Berücksichtigung eines größeren Kontextes erfordert, relativ abstrakt und kaum mit distributionellen Merkmalen verknüpft ist, so dass sie sich möglicherweise einer Anwendung in der automatischen Wortartenerkennung entzieht. Vor allem aber könnte die Charakterisierung der MP *auch* qua Satzmodus (Frage- und Aufforderungsätze mit den o. a. Beispielsätzen) in Westpfahl et al. (2017) bei einer darauf basierenden Annotation zur Vernachlässigung von MP-Verwendungen in V2-Assertiva geführt haben. Bedenkt man, dass Brünjes (2014) 96,2% aller 500 MP-Belege in Assertiva gefunden hat und nur 1,2% in Direktiva sowie 2,6% in Erotetika (ibd.: 95f.), dann muss eine Vernachlässigung von Assertiva zwangsläufig zur Ausblendung eines Großteils der MP-Belege von *auch* führen. Unterschiedliche Definitionen der MP-Verwendung von *auch* könnten also möglicherweise die eklatant geringe Tokenrate in FOLK erklären.

Im letzten Schritt soll an Stichproben geprüft werden, wie zuverlässig das POS-Tagging des Lexems *auch* in FOLK ist, und zwar getrennt nach den MP- (29 Belege) und Heterosem-Vorkommen (37.632 Belege). Die händische Überprüfung orientiert sich an der MP-Definition von Brünjes (2014). Aus den 37.632 Heterosem-Vorkommen (37.046 Fokuspartikeln [PTKIFG] und 586 Adverbien [ADV]) wurde zunächst eine Stichprobe von 100 Belegen[19] gezogen, von denen sich 3 als nicht klassifizierbar herausstellten.[20] Von den restlichen 97 waren 74 tatsächlich Heteroseme (27 ADV, 47 PTKIFG), was hinsichtlich der Abgrenzung von der MP einer Zuverlässigkeit von 76,3% entspricht. Als Adverb wurde z. B. das Auftreten in der Äußerung *ähm [und dann] und dann geht auch plötzlich dieser große gelbe pfeil (.) weg °h* (E00358, T02, B0897) eingestuft, weil *auch* (im Sinne von ‚zudem', ‚außerdem') hier ohne Bedeutungsveränderung allein ins Vorfeld rücken kann (*auch geht dann plötzlich…*). Ein Beispiel für die Verwendung als Fokuspartikel ist hingegen: *nö glaub ich auch nicht* (E00182, T02, B0595), denn

[19] Über einen speziellen Button des DGD-Abfragesystems lässt sich aus jedem Abfrageergebnis eine Zufallsstichprobe ziehen, deren Umfang von den Benutzer:innen festgelegt wird. Alle randomisierten Stichproben der vorliegenden Untersuchung wurden mit Hilfe dieser Funktion gezogen. Bei mehr als 10.000 Treffern wird zunächst automatisch eine Zufallsstichprobe von 10.000 Belegen gezogen, aus der sich dann manuell wiederum kleinere Stichproben ziehen lassen.

[20] In E00134, T01, B0645 (*mehrere Kabel als an unser Fernseh dahäm o*), spricht der Fokusakzent auf dem nachgestellten *o* für eine Fokuspartikel (*auch unser Fernseher daheim*), was aber in dem komparativen Kontext keinen Sinn ergibt. In E00357, T01, B1253, liegt ein Redeabbruch vor, der eine zweifelsfreie Bestimmung unmöglich macht. In E00412, T01, B0347 (*das is (.) is is aber auch ++++++ da is jetz nich irgendwie dass ich sage boah*) gibt es eine Lücke (++++++) in der Transkription, welche die Analyse behindert.

(das akzentuierte) *auch* bezieht sich auf das Subjekt und kann gemeinsam mit diesem im Vorfeld platziert werden (*auch ich glaub [das] nicht*).[21] Die übrigen 23 Belege der Stichprobe wurden als MP klassifiziert. Darunter fällt z. B. das zweite *auch* in: *war en fehler (.) ich habe ne falsche entscheidung [getroff]en dazu steh ich auch (.) die verantwortung muss ich ja jetz auch tragen komm ich gar nich drum [rum °h]* (E00173, T02, B0592). Da *auch* nicht akzentuiert ist und mit keinem anderen Element bedeutungsgleich ins Vorfeld rücken kann (z. B. *auch die Verantwortung/ auch ich/ auch jetzt …*), kann die FOLK-Annotation als Fokuspartikel PTKIFG nicht überzeugen. Vielmehr liegt in dieser Assertion eine MP-Verwendung im Sinne von Brünjes (2014) vor: Die Partikel *auch* signalisiert, dass die geäußerte Proposition (‚ich muss die Verantwortung tragen') mit einer in diesem Kontext vom Sprecher vorausgesetzten Präsupposition identisch ist. Einschränkend muss gesagt werden, dass die Abgrenzung von MP und Heterosemen im Fall von *auch* aufwändige Untersuchungen (Berücksichtigung eines größeren Kontextes, mehrfaches Abhören der alignierten Tonaufnahmen) erfordert, ohne dass sich bei Mittelfeldposition und fehlender Akzentuierung von *auch* immer völlige Eindeutigkeit erzielen ließe (cf. Hentschel/Keller 2006: 76f.; Moroni/ Bidese 2021: 196–202). Abschließend ein Blick auf die 29 *auch*-Belege, die in FOLK als MP (PTKMA) getaggt sind. Angesichts der geringen Anzahl konnten sämtliche Vorkommen, ohne Stichproben zu ziehen, manuell überprüft werden. Von 26 klassifizierbaren Vorkommen[22] wurden nur 8 Belege (entsprechend einer Zuverlässigkeit von 30,8%) als MP (PTKMA) bestätigt, während die übrigen als Heteroseme (3 ADV, 15 PTKIFG) eingestuft wurden. Rechnet man die Ergebnisse der Stichprobe auf das Gesamtkorpus hoch, so ergeben sich deutlich gesteigerte Werte für das Vorkommen der MP *auch* (Abbildung 5).

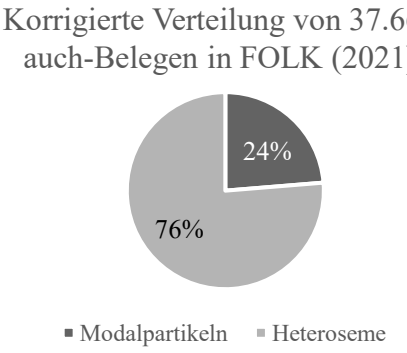

Abbildung 5: Korrigierter Anteil der MP *auch* am Gesamtvorkommen des Lexems in FOLK

Der korrigierte Anteil (23,7%) der MP an allen Vorkommen des Lexems *auch* in FOLK (siehe Abbildung 5) ist zwar noch immer deutlich geringer als im Korpus von Brünjes (2014), die einen Anteil von 50,8% ermittelte (siehe Abbildung 4 rechtes Diagramm), aber da die Tokenrate des Lexems *auch* in FOLK ca. doppelt so hoch ist wie bei Brünjes (2014) (s. oben), kommt es zu einer Angleichung der MP-Tokenraten (siehe Tabelle 10): Während

[21] In FOLK sind beide Belege als PTKIFG annotiert.

[22] In zwei Fällen E00152, T03, B0653; E00370, T02, B0155) wurde die Partikel *auch* von einem L2-Sprecher/einer L2-Sprecherin in unklarer Funktion verwendet; in einem weiteren Fall (B00207, T01, B1423) lag ein Redeabbruch vor.

Brünjes (2014) (aufgrund der Analyse der ersten 984 Vorkommen des Lexems *auch* (cf. Tabelle 5) eine Tokenrate von 0,33% ermittelt, ergibt sich aus der Überprüfung aller 29 MP-Vorkommen sowie einer Stichprobe von 100 Heterosem-Vorkommen in FOLK eine geschätzte Tokenrate von 0,30%.

	Hentschel (1986)	**Brünjes (2014**	**FOLK (2021)**
Korpusumfang	ca. 21.200	740.826	2.990.421
Belege MP *auch*	42	2.415	8.931*
Tokenrate	0,20%	0,33%	0,30%*

Tabelle 10: Vergleich der Tokenraten der MP *auch* (* = neuberechnete, geschätzte Werte)

Die Partikel *auch* bereitet schon bei der manuellen Annotation erhebliche Probleme. Nicht von ungefähr gehörte *auch* (neben *so*) bei der Annotierung des FOLK-Goldstandards zu den Partikeln mit den größten Abweichungen zwischen den beiden Annotatorinnen (Westpfahl 2020: 313). Auch die hier durchgeführte Überprüfung einer Stichprobe von 100 Heterosem-Vorkommen erwies sich als komplexe und zeitraubende Aufgabe. Immerhin ergab sich bei der Anwendung der MP-Definition von Brünjes (2014) auf die FOLK-Stichprobe eine weitgehende Angleichung der Tokenraten. Vor diesem Hintergrund erscheinen die eingangs dargestellten enormen Häufigkeitsunterschiede nicht durch (unterschiedliche) Merkmale der beiden Korpora bedingt zu sein. Die Gründe müssen eher in unterschiedlichen Kriterien zur Abgrenzung der MP *auch* von ihren Heterosemen bzw. in deren Anwendung auf die *auch*-Belege gesucht werden. Zu klären bleibt eine Frage, die hier nicht im Mittelpunkt des Interesses steht, nämlich warum das Fernsehgesprächskorpus von Brünjes (2014) insgesamt nur halb so viele Belege des Lexems *auch* aufweist wie FOLK bzw. warum die *auch*-Heteroseme in FOLK viel häufiger auftreten als im Brünjes-Korpus.

3.2 Die MP *mal*

Mit einem Anteil von 14% am MP-Gesamtvorkommen rangiert *mal* in FOLK auf dem zweiten Rang, nur übertroffen von der MP *ja*, in Hentschels (1986) Korpus nach *ja* und *doch* mit einem Anteil von 17% immerhin auf Rang 3, während sie bei Brünjes (2014) nur 2% aller MPn ausmacht und nicht einmal zu den 5 häufigsten MPn gehört.

Auch was den Anteil der MP am Gesamtvorkommen des Lexems *mal* angeht, sind gravierende Abweichungen zu beobachten. In FOLK gibt es 21.243 Treffer für das Lexem *mal* (Tokenrate 0,71%), die sich zu fast gleichen Teilen auf Adverb (50,2%) und MP (49,8%) verteilen. Im Korpus von Brünjes (2014) hingegen ist nicht nur das Lexem *mal* insgesamt deutlich seltener (2.066 Treffer, Tokenrate 0,28%), sondern auch der Anteil der MP beläuft sich auf ganze 10,2% (siehe Abbildung 6). So kommt es, dass die Tokenrate der MP *mal* bei Brünjes (2014) mit 0,03% nicht einmal ein Zehntel des Werts von FOLK bzw. Hentschel (1986) (beide 0,35%) erreicht.

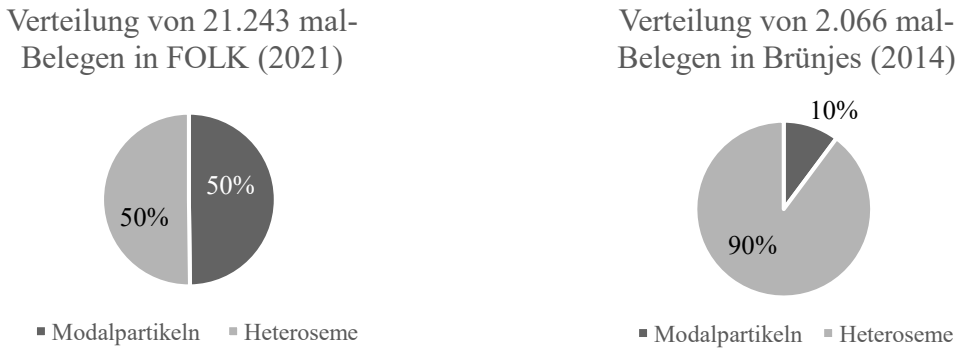

Abbildung 6: Vergleich des Anteils der MP *mal* am Gesamtvorkommen des Lexems

Da sich FOLK nur von einem der beiden manuell ausgewerteten Korpora unterscheidet, i. e. von Brünjes (2014), liegt die Vermutung nahe, dass die genannten Unterschiede in der Häufigkeit nicht auf Probleme beim POS-Tagging zurückzuführen sind, sondern durch unterschiedliche Korpusmerkmale oder MP-Definitionen bedingt sind.

Laut den STTS 2.0-Guidelines für FOLK ist beim Lexem *mal* die Abgrenzung von Adverb und MP problematisch, wenn die MP noch Reste der temporalen Bedeutung in sich trägt, zumal auch das Adverb *mal* (anders als *einmal*) „normalerweise nicht vorfeldfähig ist“ (cf. Westpfahl et al. 2017: 35; cf. auch Thurmair 1989 und 2026). Eine adverbiale Verwendung wird angenommen, wenn *mal* ausdrückt, dass eine Handlung nicht sofort bzw. nicht dauerhaft stattfindet und *mal* durch *irgendwann*, *über kurz oder lang*, *ab und zu* ersetzbar ist. Die MP liegt laut STTS 2.0-Guidelines dagegen vor, wenn die Partikel nicht erfragbar, nicht ins Englische übersetzbar und/oder auf die aktuelle Situation bezogen ist. Speziell hingewiesen wird auf die abmildernde Wirkung in Aufforderungen (*pass mal auf*), beim Aufstellen von Hypothesen (*ich sag mal*, *nehmen wir mal an*) und in Äußerungen, in denen ein weiterer temporaler Ausdruck eine temporale Verwendung von *mal* „unmöglich/unnötig“ macht (z. B. in *dann habt ihr jetzt mal nichts zu tun*, *heute mal nicht*, cf. ibd.: 36). Der Desambiguierung dient ein spezieller Entscheidungsbaum (cf. ibd.: 36) mit diversen Anwendungsbeispielen; hier wird u. a. darauf verwiesen, dass *mal* in Fragesätzen als Adverb anzusehen ist, wenn die Partikel ohne Bedeutungsänderung auch in der Antwort verwendet werden kann. Insgesamt ist die Behandlung von *mal* (z. B. im Vergleich mit der von *auch*) sehr ausführlich (cf. ibd.: 35–37) und bietet einen brauchbaren Leitfaden für die Annotation.

Auch Brünjes (2014: 146) bemerkt, dass die MP z. T. noch Reste der temporalen Bedeutung bewahrt und sich deshalb oft nur schwer vom temporalen Adverb (Bezug auf einen unbestimmten Zeitpunkt in Vergangenheit oder Zukunft) abgrenzen lässt. Die MP kommt laut Brünjes (2014: 147f.) nicht nur in direktiven Sprechakten (41 Belege in ihrem Korpus, Bsp. *Halten Sie mal meine Tasche?*) vor, sondern auch in festen, redeeinleitenden Wendungen wie *hör mal*, *sag mal* (80 Belege) sowie in assertiven V2-Aussagesätzen (84 Belege), dort meist als Teil eines eingeschobenen Heckenausdrucks wie *ich sag mal*, *ich denke mal*. Brünjes (ibd.: 154f.) zeigt, dass herkömmliche temporale Bedeutungsbeschreibungen der MP *mal* (Nicht-Repetitionalität, Nicht-Sofortigkeit, Nicht-Dauerhaftigkeit), aber auch die Annahme einer Abschwächung von Direktiva immer nur auf einen Teil der Belege zutreffen,

also eher kontextbedingt sind. Das gemeinsame Merkmal aller *mal*-Verwendungen besteht laut Brünjes (2014: 148) darin, dass eine negierte Variante des vollzogenen Sprechakts präsupponiert wird, die als Standardverhalten gekennzeichnet wird. „Auf diese Weise markiert der Sprecher den von ihm vollzogenen Sprechakt als außergewöhnlich und grenzt ihn von sonst üblichem Verhalten ab. Die Bedeutung von *mal* ist paraphrasierbar als: *Eigentlich tue (frage/erbitte/...) ich so etwas nicht.*" (ibd.: 148). Insgesamt erscheint Brünjes' Definition der MP *mal* nicht restriktiver als die der STTS 2.0-Guidelines, eventuell mit Ausnahme der für FOLK angenommenen MP-Funktion in Äußerungen mit einem weiteren temporalen Ausdruck. Diese Fälle aber dürften kaum so zahlreich sein, dass sie die oben dargestellten Häufigkeitsunterschiede erklären. Ferner wäre es möglich, dass Brünjes bei der Abgrenzung von Adverb und MP insgesamt etwas restriktiver vorgegangen ist, denn sie schreibt, dass von 2.066 Vorkommen des Lexems *mal* in 211 Belegen „eindeutig eine Modalpartikelfunktion vorliegt" (ibd.: 146). Da Brünjes aber auch Belege einschließt, in denen „Reste der älteren Temporaladverbbedeutung" (ibd.) enthalten sind, ist auch diese Erklärung letztlich wenig plausibel.

Obwohl die Zuverlässigkeit des POS-Taggings in FOLK als Quelle der eklatanten Frequenzunterschiede zwischen FOLK und dem Fernseh-Korpus von Brünjes (2014) nur bedingt in Frage kommt, sollen gleichwohl zwei Zufallsstichproben von FOLK-Belegen des Lexems *mal* (100 MP-Belege, 100 Adverb-Belege) überprüft werden. Von den 100 als PTKMA getaggten Vorkommen sind demnach 79% tatsächlich MPn, 21% hingegen Adverbien. Letzteres trifft z. B. auf die Äußerung *(.) ach so ja ihr wart ja mal miteinander in urlaub* (E00143, T02, B0296) zu, denn hier bezieht sich (das in FOLK als MP annotierte) *mal* eindeutig auf einen Zeitpunkt in der Vergangenheit und muss demnach als Adverb gewertet werden. Eine vergleichbare Zuverlässigkeit offenbart das POS-Tagging als Adverb: 82% der ADV-Belege sind tatsächlich Adverbien, 18% wurden hingegen als MPn klassifiziert. Dies gilt z. B. für die Äußerung *heiz ich den ofen schon mal vor* (E00329, T01, B0152), denn die Sprecherin, die mit der Essenszubereitung beschäftigt ist und gerade den Backofen einschaltet, bezieht sich mit ihrer Äußerung auf die „aktuelle Situation" (Westpfahl et al. 2017: 36).[23] Rechnet man die in den Stichproben ermittelten prozentualen Anteile korrekter Klassifizierungen auf die Gesamtheit der Vorkommen hoch, heben sich die Fehlerquoten weitgehend gegenseitig auf, i. e. es bleibt dabei, dass sich die Gesamtheit der Belege zu fast gleichen Teilen auf Adverb (51,6%) und MP (48,4%) verteilt (siehe Abbildung 7). Die Tokenrate der MP *mal* bleibt mit 0,34% nahezu unverändert (vorher 0,35%) und übersteigt diejenige im Brünjes-Korpus (0,03%) nach wie vor um mehr als den Faktor 10.

[23] Zu klären wäre, ob im vorliegenden Kontext auch das MP-Kriterium von Brünjes (2014: 148) zutrifft, wonach *mal* eine negierte Variante des vollzogenen Sprechakts präsupponiert, die als Standardverhalten gekennzeichnet wird. Möglicherweise signalisiert die Sprecherin durch *mal*, dass ihre Äußerung unter kommunikativen Gesichtspunkten nicht wirklich notwendig ist, z. B. weil für die anderen Anwesenden ohnehin ersichtlich ist, was sie tut.

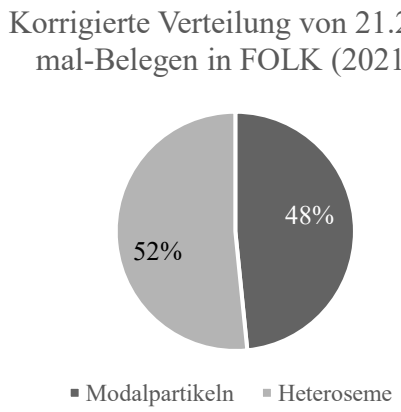

Abbildung 7: Korrigierter Anteil der MP *mal* am Gesamtvorkommen des Lexems in FOLK

Wie zu erwarten, scheidet das POS-Tagging als Quelle der beobachteten Unterschiede in der Frequenz der MP *mal* aus. Auch die jeweiligen MP-Definitionen für FOLK (cf. Westpfahl et al. 2017) bzw. bei Brünjes (2014) rechtfertigen, wie dargestellt, kaum Differenzen in der beschriebenen Größenordnung. Bleiben als Erklärung Unterschiede zwischen den Korpora. Interessant ist in diesem Zusammenhang, dass die Tokenraten des Adverbs *mal* deutlich weniger voneinander abweichen: 0,34% (korrigiert) in FOLK und 0,25% im Brünjes-Korpus. Dass nur die MP *mal* im Brünjes-Korpus so überaus selten auftritt, könnte zunächst einmal damit erklärt werden, dass Fernsehgespräche naturgemäß einen geringen Privatheitsgrad haben und somit insgesamt weniger MP-Belege aufweisen (siehe Abschnitt 2.2). Da MPn im Fernsehkorpus von Brünjes (2014) ungefähr halb so häufig sind wie in FOLK, wäre dies immerhin eine partielle Erklärung. Möglicherweise sind zudem die für die MP *mal* typischen direktiven Sprechakte in einem Korpus von Fernsehgesprächen eher unterrepräsentiert. Vorstellbar wäre demnach, dass sich die Teilnehmer:innen an einem Fernsehgespräch a) vergleichsweise selten zu etwas auffordern bzw. b) dass sie es ggf. mit anderen sprachlichen Mitteln tun (Konjunktiv II, *bitte* etc.). In Kontexten mit hoher Vertrautheit der Gesprächsteilnehmer:innen (Familie, Freunde, Schulklassen u. ä.), die in FOLK durchaus vertreten sind (z. T. mit kooperativen Aktivitäten wie Gesellschaftsspielen oder gemeinsamem Kochen) sind dagegen direktive Sprechakte öfter und informeller, i. e. unter Verwendung der MP *mal*, zu erwarten.

3.3 Die MP *halt*

Ähnlich wie bei *auch* unterscheidet sich FOLK hinsichtlich des Vorkommens der MP *halt* grundlegend von den beiden manuell ausgewerteten Korpora, aber im Gegensatz zu *auch* gehört *halt* mit 13% aller MP-Vorkommen zu den häufigsten MPn in FOLK (Platz 3), während dieselbe MP in den Korpora von Hentschel (1986) und Brünjes (2014) mit je 2% auf den hinteren Plätzen rangiert.

In FOLK kommt das Lemma *halt* 9.780 Mal vor (Tokenrate 0,33%); davon entfallen 9.694 Treffer (99,1%) auf die MP (Tokenrate 0,32%) und nur 86 Vorkommen (0,9%) auf die Interjektion. Das Korpus von Brünjes (2014) enthält mit 258 Treffern relativ zur Korpusgröße 11-mal weniger Vorkommen des Lemmas *halt* (Tokenrate 0,035%). Davon entfallen 223 Treffer (86,4%) auf die MP (Tokenrate 0,03%) und 35 (13,6%) auf die Heteroseme (siehe Abbildung

8).[24] Das Korpus von Hentschel enthält 9 Vorkommen der MP, was einer (mit Brünjes vergleichbaren) Tokenrate von 0,04% entspricht.

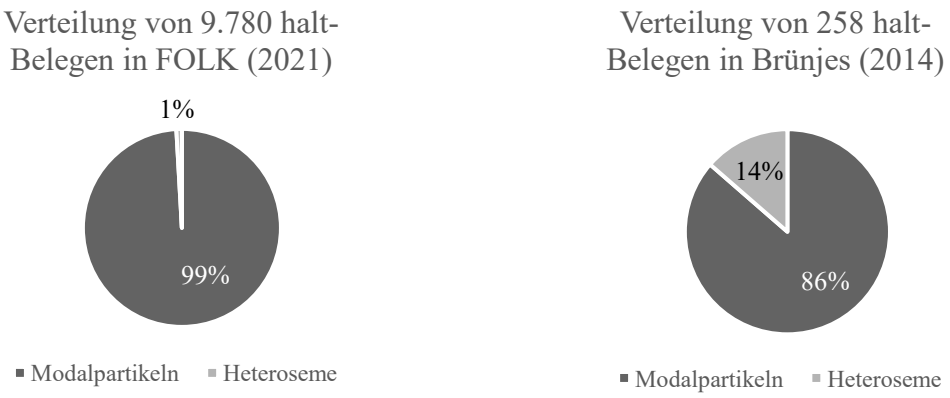

Abbildung 8: Vergleich des Anteils der MP *halt* am Gesamtvorkommen des Lexems

Da sich die MP *halt* und die Interjektion *halt* distributionell grundlegend unterscheiden, ist kaum zu erwarten, dass sich die unterschiedlichen Tokenraten aus Abgrenzungsproblemen bzw. unterschiedlichen Definitionen erklären lassen. Vermutlich aus eben diesem Grund enthält das STTS 2.0-Handbuch (cf. Westpfahl et al. 2017) auch keine speziellen Ausführungen zu *halt*, während Brünjes (2014: 124) anknüpfend an Diewald (1997: 92) den gemeinsamen Bedeutungsgehalt der MPn *halt* und *eben* wie folgt beschreibt: „*Eben* und *halt* verweisen auf eine pragmatisch präsupponierte Einheit, die aus einer mit der geäußerten Proposition identischen Proposition besteht und die dem Sprecher zugeordnet wird. Dabei wird offen gelassen, ob die präsupponierte Proposition auch für den Hörer gilt. Auf diese Weise entsteht die iterative Bedeutung." Mit der Hervorhebung des gemeinsamen Bedeutungsgehalts setzt sich Brünjes von Thurmair (1989: 119–128) ab, die unterschiedliche semantische Merkmale für *eben* (Evidenz) und *halt* (Plausibilität) gegeben sieht.[25] Die MP *halt* kann u. a. dazu dienen, den „Inhalt der Aussage […] als Wiederholung einer schon früher vertretenen Ansicht des Sprechers" (Brünjes 2014: 126) zu markieren. Eine solche vorhergehende Äußerung ist aber weder obligatorisch noch muss sie im näheren Kontext erfolgt sein, so dass sie in Korpusbelegen nur schwer nachzuweisen ist. So begründet im folgenden Beleg ein Schüler seinen Sprachaufenthalt in Spanien damit, dass er im Spanischunterricht an der deutschen Schule in Konkurrenz mit vielen *native speakers* steht. Er verwendet dabei die MP *halt*, ohne dass im vorhergehenden Kontext von den Muttersprachlern die Rede gewesen wäre: *[°h ja ich wollt das unbe]dingt machen weil wir halt viele °hh native speakers bei uns haben (.) also muttersprachler* (E00181, T01, B0481). Tatsächlich wird die Präsenz von Muttersprachlern nämlich erst nachfolgend zum Thema des Gesprächs. Auf jeden Fall impliziert die MP *halt* (gleiches gilt für *eben*) eine mit der aktuellen Äußerung inhaltlich identische Präsupposition, wodurch die iterative Bedeutung und – in zahlreichen Kontexten – der Charakter von Unabänderlichkeit des Mitgeteilten entsteht. In der

[24] Laut Brünjes (2014: 179) bezieht sich die Zahl 258 auf die „Gesamtbelegzahl des Lexems", aber an anderer Stelle (ibd.: 123) werden als Heteroseme der Imperativ *halt!* und die verkürzte Form *(ich) halt* genannt (die dem Lexem *halten* zuzuordnen wären), wohingegen die Interjektion *halt* überhaupt nicht erwähnt wird.

[25] Zur semantischen Differenzierung von *halt* und *eben* cf. auch den Beitrag von Thurmair (2026).

folgenden Äußerung verweist der Spielleiter eines Map-Task-Experiments z. B. auf die Unabänderlichkeit bestimmter Spielregeln: *also du darfst auch nachfragen […] die einzige einschränkung is halt das mit […] dem blick und dem zeigen* (E00094, T01, B0044-0049). In ähnlicher Funktion hätte hier auch *eben* eintreten können, obschon zwischen *eben* und *halt* regionale und temporale (s. unten) sowie emotionale Unterschiede (cf. Hentschel/Keller 2006: 86f.) bestehen mögen.

Könnte möglicherweise ein fehlerhaftes POS-Tagging die unterschiedlichen Häufigkeiten erklären? Auch das ist kaum zu erwarten, denn die distributionellen Unterschiede von MP und Interjektion sind natürlich auch eine gute Basis für die automatische Annotation. Dennoch soll hier wiederum eine Stichprobe überprüft werden: Von 100 zufällig ausgewählten Vorkommen von *halt* mit dem POS-Tag PTKMA erwiesen sich 98% tatsächlich als MP. Bei einem Beleg handelte es sich um die Interjektion *halt* und einmal lag die Imperativform des Verbs *halten* vor. Von den 86 Vorkommen der Interjektion *halt* waren (nach Ausschluss von 10 unklaren Belegen)[26] 75% korrekt getaggt, die restlichen 25% waren MPn. Aufgrund des vergleichsweise geringen Vorkommens der Heteroseme kommt es aber insgesamt kaum zu Verschiebungen in den relativen Anteilen von MPn und Heterosemen, wenn man diese Ergebnisse auf das Gesamtkorpus projiziert: 97,4% der Belege wären demnach MPn (statt zuvor 99,1%), nur 2,6% Heteroseme, i. e. Interjektionen bzw. Imperativformen von *halten* (siehe Abbildung 9). Wie erwartet, ist das POS-Tagging des Lexems *halt* sehr zuverlässig, i. e. bei den Tokenraten der MP *halt* bestätigen sich die extremen Unterschiede zwischen den drei Korpora: einem Wert von 0,32% in FOLK, stehen die 8- bis 11-mal geringeren Vorkommen im Korpus von Hentschel (0,03%) und Brünjes (0,04%) gegenüber.

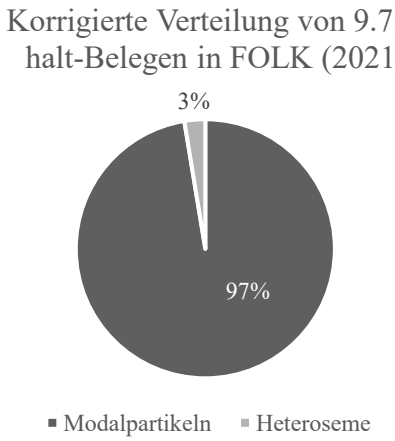

Abbildung 9: Korrigierter Anteil der MP *halt* am Gesamtvorkommen des Lexems in FOLK

Eine mögliche Erklärung für die beobachteten Unterschiede wäre eine unterschiedliche Gewichtung der konkurrierenden MPn *halt* und *eben*. Wird das geringe Vorkommen von *halt* in den Korpora von Hentschel (1986) und Brünjes (2014) vielleicht durch ein stärkeres

[26] Überwiegend ist in diesen Fällen das Lexem *halt* trotz mehrfachen Abhörens nicht wahrnehmbar, vor allem aufgrund der Überlappung von mehreren Sprecherbeiträgen. In der Regel handelt es sich außerdem um Beiträge, die nur aus dem Wort *halt* bestehen und somit keine Anhaltspunkte für eine nähere Bestimmung liefern.

Vorkommen von *eben* kompensiert? Ein Blick auf Abbildung 10 bestätigt diese Vermutung, wenn auch nur teilweise.

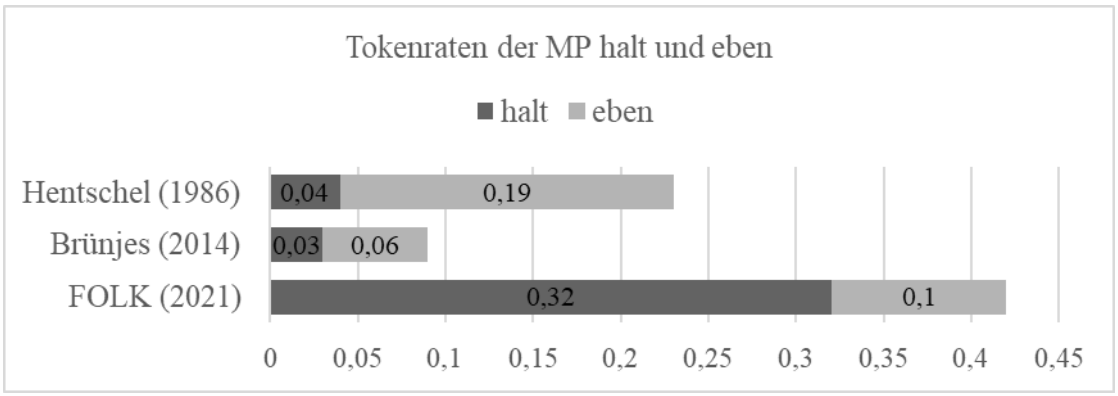

Abbildung 10: Vergleich der Tokenraten der MPn *halt* und *eben* (Angaben in %)

In den Korpora von Hentschel (1986) und Brünjes (2014) sind die Tokenraten der MP *eben* mit 0,19% bzw. 0,06% deutlich höher als diejenigen der MP *halt*, während sich in FOLK die Verhältnisse umkehren: Hier kommt *halt* ca. 3-mal so häufig vor wie *eben*. Zwar kann *eben* das geringe Vorkommen von *halt* in den manuell ausgewerteten Korpora nicht gänzlich kompensieren, aber die Häufigkeitsunterschiede reduzieren sich doch deutlich: Betrachtet man beide MPn zusammen, fällt die Tokenrate in Hentschel mit 0,23% etwa halb so groß aus wie in FOLK (0,42%). Im Vergleich von Brünjes und FOLK reduziert sich der Abstand von 1:11 (für *halt*) auf ca. 1:5 (für *halt/eben*), wobei zu bedenken ist, dass im Korpus von Brünjes (2014) insgesamt weniger MPn auftreten als in FOLK (siehe Tabelle 6). Da bei *halt* (und analog auch bei *eben*) weder definitorisch noch bei der (automatischen) Annotation größere Probleme bei der Abgrenzung von MPn und Heterosemen anzunehmen sind, lassen sich die verbliebenen Häufigkeitsunterschiede kaum anders als durch unterschiedliche Korpusmerkmale erklären.

Eine Frage, die hier nur am Rande behandelt werden kann, betrifft die Gründe für die Verschiebungen zwischen *halt* und *eben*. Bezüglich einer möglichen diatopischen Variation entlang der Achse Nord (*eben*) vs. Süd (*halt*) stellt schon Hentschel (1986: 178) fest: Die „regionalen Grenzen verschieben sich zusehends oder lösen sich gänzlich auf“ (zit. nach Brünjes 2014: 132; cf. Hentschel/Keller 2006: 84–87). Anders wäre auch nicht zu erklären, warum gerade bei Hentschel (1986), deren Korpus zum Großteil aus Gesprächen des Freiburger Korpus (FR) besteht, der Anteil von *halt* verglichen mit *eben* relativ gering ausfällt. Auch Thurmair (1989: 123f.) bestätigt, dass *eben* und *halt* sowohl im süddeutschen wie im norddeutschen Raum nebeneinander bestehen. Allerdings hat sich *halt* in Norddeutschland erst ab Ende der 1970er Jahre ausgebreitet (cf. Dittmar 2000: 212), wobei der Prozess – wie Dittmar (2000) am Beispiel des 1993–1996 erhobenen *Berliner Wendekorpus* zeigt – bei Ostberliner Sprecher:innen im Erhebungszeitraum noch im Gange ist und als Indikator eines gesellschaftlichen Umbruchs gedeutet werden kann (cf. auch Cognola/Moroni 2022: 112–115). Verschiebungen im Zeitablauf sind jedenfalls auch plausibel, wenn man die drei hier berücksichtigten Korpora vergleicht, denn der Anteil der MP *halt* am jeweiligen Gesamtvorkommen von *halt+eben* steigt parallel zur Entstehungszeit der jeweils ins Korpus aufgenommenen Gespräche (siehe Tabelle 11).

	Tokenrate MP *halt*	Tokenrate MP *eben*	Tokenrate MP *halt* + *eben*	Anteil von *halt* am Gesamtvorkommen
Hentschel (1986) Texte 1960–1974[27]	0,04%	0,19%	0,23%	17%
Brünjes (2014) Texte 1989–2003	0,03%	0,06%	0,09%	33%
FOLK (2021) Texte 2003–2020	0,33%	0,10%	0,43%	78%

Tabelle 11: Vergleich des Anteils von *halt* am Gesamtvorkommen von *halt*+*eben*

3.4 Korrelation der MP-Rangfolgen in den untersuchten Korpora

Abschließend seien die Unterschiede zwischen den MP-Häufigkeitsrangfolgen in den drei untersuchten Korpora anhand des Pearson-Korrelationskoeffizienten dargestellt. Mit den ursprünglichen (nicht korrigierten) Werten der absoluten Häufigkeit (Spalte a, b, c in Tabelle 7) ergeben sich die in Tabelle 12 dargestellten Korrelationen.

Korrelierte Korpora	Korrelationskoeffizient Pearson's r
Hentschel (1986) – Brünjes (2014)	0,812
Hentschel (1986) – FOLK (2021)	0,762
Brünjes (2014) – FOLK (2021)	0,638

Tabelle 12: Korrelation der absoluten Zahlen der MP-Belege (Spalte a, b, c in Tabelle 7)

Unschwer lässt sich erkennen, dass die beiden manuell annotierten Korpora, i. e. Hentschel (1986) und Brünjes (2014), trotz unterschiedlicher Korpusmerkmale mit r=0,812 die höchste Korrelation aufweisen. Erst an zweiter Stelle kommen, trotz einer ähnlich breiten Mischung von Sprechereignissen, Hentschel (1986) und FOLK (2021) mit einem Koeffizienten von r=0,762, wobei die Probleme mit dem automatischen POS-Tagging ausschlaggebend sein dürften. Die geringste Korrelation von r=0,638 besteht zwischen den MP-Häufigkeiten von Brünjes (2014) und FOLK (2021), die sich sowohl in puncto Korpuseigenschaften wie Annotationsmethode unterscheiden.

Ersetzt man die absoluten FOLK-Häufigkeitswerte für *auch*, *mal* und *halt* in Tabelle 7 durch die aufgrund der Stichprobenanalysen korrigierten (geschätzten) Werte, gelangt man zu den in Tabelle 13 angegebenen Pearson-Korrelationskoeffizienten. Die weitgehende Angleichung der Werte, die in erster Linie auf die korrigierte Beleganzahl von *auch* zurückgeht, zeigt, dass das automatische POS-Tagging in FOLK (nach Korrektur der Werte von *auch*, *mal*, *halt*) eine Häufigkeitsrangfolge der MPn liefert, die mit denen anderer Korpora ebenso gut vergleichbar ist wie die Rangfolgen manuell ausgezählter Korpora untereinander.

[27] Die angegebene Zeitspanne bezieht sich auf die Gespräche des Freiburger Korpus. Die zusätzlich verwendeten Gespräche, die Studierende der Freien Universität Berlin aufgenommen haben (cf. Hentschel 1986: 240) sind vermutlich in den darauffolgenden Jahren entstanden.

Korrelierte Korpora	Korrelationskoeffizient Pearson's r
Hentschel (1986) – Brünjes (2014)	0,812
Hentschel (1986) – FOLK (2021, mit Korrektur)	0,811
Brünjes (2014) – FOLK (2021, mit Korrektur)	0,823

Tabelle 13: Korrelation der absoluten Zahlen der MP-Belege (korrigierte FOLK-Werte für *auch*, *mal*, *halt*)

4 Resümee

Der erste Teil der hier vorgelegten Analyse (Abschnitt 2.1) ging der Frage nach, ob die Liste der in FOLK als MPn ausgewiesenen Lexeme mit dem Repertoire allgemein anerkannter MPn kompatibel ist. Zweifelhaft erschienen die Lemma-Types *ausgerechnet* (12 Belege), *echt* (110), *einmal* (2), *fei* (10), *glatt* (21), *nun* (5), *überhaupt* (308), *wirklich* (43) und *zu* (3). Allerdings erwiesen sich die relativ frequenten Partikeln (*überhaupt*, *echt*, *wirklich*, *glatt*) sowie das dialektale *fei* als durchaus ernst zu nehmende MP-Kandidaten, während umgekehrt die offensichtlichen Fehlklassifikationen (*ausgerechnet*, *einmal*, *nun*, *zu*) quantitativ kaum ins Gewicht fallen. In FOLK nicht erfasst werden dagegen die allgemein anerkannten MPn *eigentlich*, *etwa* und *vielleicht*. Problematisch erscheint besonders das Fehlen der MP *eigentlich*, die in anderen Korpora immerhin auf einen Anteil von 2–3% kommt (siehe Tabelle 7). Wie gezeigt, liegt das Problem in diesem Fall in einer inkorrekten manuellen Annotation des „Goldstandards", mit dem der POS-Tagger trainiert wurde. Die anderen beiden MPn kommen dagegen im Trainingsmaterial gar nicht vor (*etwa*) oder nur in ambigen Kontexten (*vielleicht*), so dass der POS-Tagger zwangsläufig keine adäquaten Parameter ausbilden konnte. Um die dargestellten Probleme bei der Auswahl der MP-Lexeme zu lösen, wäre es vermutlich erforderlich, einerseits mittels einer geschlossenen MP-Wortliste die Lexeme *ausgerechnet*, *einmal*, *nun* und *zu* auszuschließen, andererseits durch Revision der Annotationen von *eigentlich* im Trainingskorpus und durch dessen gezielte Erweiterung um Belege für die MP *etwa* und *vielleicht* diese MP einzubeziehen.

Die zweite in diesem Beitrag untersuchte Fragestellung (Abschnitt 2.2) betraf die Häufigkeit der MPn, sowohl als Klasse als auch einzeln. Wie dargestellt, ergibt die automatische POS-Annotation in FOLK eine Tokenrate aller MPn von zusammen 2,55%, i. e. jedes 39. laufende Wort in FOLK ist als MP getaggt. Dieser Wert kann als durchaus plausibel gelten, wenn man ihn mit dem Korpus von Hentschel (1986) vergleicht, das hinsichtlich des Privatheitsgrads ebenfalls gemischt ist und eine händisch ausgezählte Tokenrate von 2,20% (jedes 45. Wort) aufweist. Die deutlich geringere Tokenrate (1,36%, jedes 73. Wort) im Fernsehgesprächskorpus von Brünjes (2014) erscheint als durch den öffentlichen Charakter der von ihr untersuchten Gespräche erklärbar. Die MP-Frequenz in FOLK liegt also im Bereich des Erwartbaren und bietet als solche keinen Anlass, an der Zuverlässigkeit der automatischen Annotation zu zweifeln. Beim Blick auf die Frequenz der einzelnen MPn hingegen ergeben sich, einmal abgesehen von der unangefochtenen Dominanz der MP *ja* in allen drei Korpora, einige Abweichungen. Geringere Frequenzunterschiede, z. B. die Tatsache, dass *doch* in allen drei Korpora zu den 5 häufigsten MPn gehört, aber bei Hentschel (1986) nach *ja* auf Platz 2 steht, bei Brünjes (2014) nach *ja* und *auch* auf Platz 3 und in FOLK nach *ja*, *mal* und *halt* auf Platz 4, können wohl als Ausdruck der natürlichen Variation von (hinsichtlich Diskurstypen, regionalen Varietäten,

Erfassungszeitraum etc.) unterschiedlich zusammengesetzten Korpora interpretiert werden. Nur Fälle mit sehr auffälligen Diskrepanzen in Bezug auf die Tokenrate bzw. den prozentualen Anteil an allen MPn, nämlich *auch*, *mal* und *halt*, wurden daher einer genaueren Prüfung unterzogen.

Die MP *auch* steht bei Brünjes (2014) mit einem Anteil von 24% auf Platz 2 ihrer Rangliste, bei Hentschel (1986) mit 9% immerhin auf Platz 4, aber bildet mit nur 29 von 75.817 Belegen (entsprechend ca. 0%) das Schlusslicht in der MP-Häufigkeitshierarchie von FOLK (2021). Auch die Tokenrate weicht in FOLK (0,00%) fundamental von den beiden manuell ausgezählten Korpora ab: 0,20% bei Hentschel (1986), 0,33% bei Brünjes (2014). Anhand von Stichproben aus FOLK konnte gezeigt werden, dass die automatische Annotation von *auch* wenig zuverlässig ist. Rechnet man die Ergebnisse der Nachprüfung auf das Korpus hoch, kommt man auf eine Tokenrate der MP *auch* in FOLK von 0,33%, also praktisch auf dasselbe Ergebnis wie Brünjes (2014). Anzumerken ist, dass schon die manuelle Unterscheidung der MP *auch* von ihren Heterosemen (Adverb, Fokuspartikel) schwierig und zeitaufwändig ist, weil sie oft die Berücksichtigung der Prosodie und eines größeren Kontextes erfordert. Ein automatisches POS-Tagging stößt hier an Grenzen, da es keine prosodische Information verarbeitet und vermutlich eher kleinere Kontexte analysiert. Möglicherweise liegen die Probleme z. T. auch in einer wenig hilfreichen Handreichung für die Annotator:innen des Trainingskorpus, welche zu einer weitgehenden Ausblendung der MP *auch* in Assertiva geführt haben könnte.

Die MP *mal* ist bei Hentschel (1986) und in FOLK (2021) sehr häufig vertreten, dagegen bei Brünjes (2014) unterrepräsentiert. Mit 17% aller MP-Vorkommen steht sie bei Hentschel (1986) auf Platz 3, in FOLK (2021) mit 14% auf Platz 2, bei einer identischen Tokenrate von 0,35%. Bei Brünjes (2014) hingegen kommt *mal* nur auf 2% aller MP-Belege und eine ca. 12-mal geringere Tokenrate von 0,03%. Da auch im Falle von *mal* die MP nicht immer eindeutig vom (temporalen) Adverb abgegrenzt werden kann, ist denkbar, dass unterschiedliche Definitionen bzw. Grenzziehungen zwischen MP und Adverb den dargestellten Diskrepanzen zugrunde liegen – eine Hypothese, für die aber letztlich keine plausiblen Belege gefunden wurden. Auch die Fehlerquoten der automatischen Annotation von *mal* in FOLK scheiden als Erklärung aus. Zwar waren in den Stichproben 18% der als Adverb getaggten Belege bzw. 21% der MP-Belege falsch getaggt, aber die Fehlerraten heben sich bei Hochrechnung auf das Gesamtkorpus auf. Als einzig plausible Erklärung der fundamentalen Häufigkeitsunterschiede bleiben unterschiedliche Korpusmerkmale: Der geringere Privatheitsgrad der Fernsehgespräche von Brünjes (2014) und das dementsprechend allgemein geringere Vorkommen von MP erklärt einen Teil der Differenzen bei den Tokenraten, während das tendenziell geringere Auftreten von Handlungsaufforderungen in Fernsehgesprächen bzw. ihre Realisierung in einem formelleren Register (*bitte*) den unterschiedlichen Anteil von *mal* an der Gesamtheit der MPn begründen könnte.

Mit 13% aller MP-Vorkommen (und einer Tokenrate von 0,32%) steht *halt* in FOLK (2021) auf Platz 3 der Häufigkeitsrangfolge, während dieselbe MP in den manuell ausgewerteten Korpora eher selten auftritt: Mit je 2% Anteil an allen MPn und Tokenraten von 0,03% (Brünjes) bzw. 0,04% (Hentschel) rangiert sie dort auf den hinteren Rängen. Da die Abgrenzung der MP *halt* von der gleichlautenden, nicht in den Satz integrierten Interjektion keine Probleme aufwirft, lassen sich die dargestellten Diskrepanzen erwartungsgemäß weder aus unterschiedlichen

Definitionen noch aus der (mit 2% äußerst geringen) Fehlerquote des automatischen POS-Taggings der MP, sondern nur aus unterschiedlichen Korpuseigenschaften erklären. Eine partielle Erklärung liefert eine gemeinsame Betrachtung der gleichbedeutenden MPn *halt* und *eben:* Offenbar hat sich die Gewichtung der beiden MPn im Zeitablauf (zwischen 1960 und 2020) zugunsten von *halt* verschoben: Während *eben* bei Hentschel (1986) ca. 5-mal häufiger ist als *halt*, bei Brünjes (2014) immerhin noch ca. 2-mal häufiger, dominiert bei FOLK umgekehrt die MP *halt* im Verhältnis 3:1. Summiert man die Vorkommen von *halt* und *eben* und berücksichtigt zudem die insgesamt geringere MP-Frequenz im Fernsehkorpus von Brünjes (2014), so relativieren sich die eingangs genannten tiefgreifenden Differenzen auf ein Verhältnis von 2:1 zugunsten von FOLK – eine Größenordnung, die man möglicherweise als „natural frequency variation" (Hentschel/Keller 2006: 83) zwischen unterschiedlichen Korpora einstufen kann.

Die drei Einzelanalysen haben gezeigt, dass Probleme des POS-Taggings in FOLK nur zum Teil für die eklatanten Abweichungen bei relativer Häufigkeit und Tokenraten der betrachteten MPn verantwortlich sind. Lediglich im Falle von *auch* hat die Überprüfung der Zufallsstichproben Fehlerraten ans Licht gebracht, die einen relevanten Einfluss auf die MP-Häufigkeitswerte haben. Bei *mal* dagegen kompensieren sich die inkorrekten POS-Tags und bei *halt* bleiben sie ohne Einfluss auf das Gewicht der MP. Nach entsprechender Korrektur (vor allem für *auch*) weist die Häufigkeitsrangfolge der MPn in FOLK eine hohe Korrelation mit den manuell ausgezählten Rangfolgen von Hentschel (1986) und Brünjes (2014) auf, was für die prinzipielle Plausibilität der MP-Häufigkeitsangaben im automatisch annotierten FOLK-Korpus spricht. Die verbleibenden Abweichungen bewegen sich im Rahmen der erwartbaren Variabilität unterschiedlich zusammengesetzter Korpora.

Damit lassen sich die beiden Hauptfragen der Untersuchung wie folgt beantworten: 1. Die Auswahl der MP-Lexeme in FOLK ist im Großen und Ganzen überzeugend, allerdings sind Nachbesserungen bei Umfang und manueller Annotation des Trainingskorpus notwendig, um die MPn *eigentlich*, *etwa* und *vielleicht* zu erfassen. 2. Die in FOLK ermittelten MP-Häufigkeiten und -Tokenraten sind überwiegend plausibel, wenn man sie mit manuell annotierten Korpora vergleicht. Eine Ausnahme jedoch bildet das Lexem *auch*, das einen überraschend geringen MP-Anteil aufweist. Hier wäre es dringend geboten, das Trainingskorpus dahingehend zu prüfen, ob es genug Belege für die MP *auch* umfasst und ob diese korrekt (manuell) annotiert sind.

Die Untersuchung hat darüber hinaus gezeigt, und zwar beim Lexem *mal*, dass auch plausible Häufigkeitsangaben auf problematischen POS-Taggings beruhen können, nämlich dann, wenn sich Fehlerraten von MPn und Heterosemen gegenseitig kompensieren. Da aber in der Regel nicht nur plausible Frequenzangaben, sondern ebenso korrekte Annotationen gefragt sind, sind weitere Untersuchungen notwendig. Wünschenswert wären z. B. detaillierte Analysen von größeren Stichproben aller MPn in FOLK, wobei die zugrundegelegten MP-Definitionen bzw. Abgrenzungen von den jeweiligen Heterosemen offengelegt und mit den im Annotationshandbuch (cf. Westpfahl et al. 2017) festgelegten Regeln abzugleichen wären. Die manuelle Neu-Annotation müsste von mindestens zwei Beurteiler:innen unabhängig voneinander und ohne Kenntnis der FOLK-POS-Tags erfolgen (mit Angabe der Beurteiler-Übereinstimmung).

Dort, wo sich bei einer solchen Analyse Defizite der automatischen Annotierung offenbaren, wäre als nächstes eine Prüfung des Trainingskorpus bezüglich derselben MPn sinnvoll. Dabei

wäre es von Vorteil, wenn das Goldstandard-Trainingskorpus über das Abfragetool der DGD zugänglich gemacht würde, so dass problematische manuelle Annotationen (wie etwa bei der MP *eigentlich*, siehe Abschnitt 2.1.2) leichter aufgefunden werden könnten. Zugleich würde so die Möglichkeit geschaffen, unterrepräsentierte MPn bzw. Verwendungsweisen (z. B. in bestimmten Satzmodi) aufzudecken und ggf. Ergänzungen vorzuschlagen, um das Training des POS-Taggers zu verbessern. Letztlich ginge es darum, jene Defizite der automatischen Annotation zu überwinden, die auf einer inkonsistenten manuellen MP-Annotation oder auf einer zu geringen Anzahl von Belegen im Trainingskorpus beruhen.

Literaturverzeichnis

Autenrieth, Tanja (2002): *Heterosemie und Grammatikalisierung bei Modalpartikeln. Eine synchrone und diachrone Studie anhand von eben, halt, echt, einfach, schlicht und glatt*. Tübingen: Niemeyer.

Brünjes, Lena (2014): *Das Paradigma deutscher Modalpartikeln: Dialoggrammatische Funktion und paradigmeninterne Oppositionen*. Berlin/New York: De Gruyter.

Cognola, Federica/Moroni, Manuela Caterina (2022): *Le particelle modali del tedesco. Caratteristiche formali, proprietà pragmatiche ed equivalenti funzionali in italiano*. Roma: Carocci.

Cognola, Federica/Coniglio, Marco (2026): "On *etwa* as a modal particle at the syntax-semantics interface". *Linguistik online* 146, 5/26: 175–191. doi.org/10.13092/9f4s6h77.

Cognola, Federica/Moroni, Manuela Caterina/Bidese, Ermenegildo (2022): "A comparative study of German auch and Italian anche: functional convergences and structural differences". In: Gergel, Remus/Reich, Ingo/Speyer, Augustin (eds.): *Particles in German, English, and Beyond*. Amsterdam, Benjamins: 209–242.

Coniglio, Marco (2011): *Die Syntax der deutschen Modalpartikeln. Ihre Distribution und Lizenzierung in Haupt- und Nebensätzen*. Berlin: Akademie-Verlag.

Diewald, Gabriele (1997): *Grammatikalisierung. Eine Einführung in Sein und Werden grammatischer Formen*. Tübingen: Niemeyer.

Dittmar, Norbert (2000): „Sozialer Umbruch und Sprachwandel am Beispiel der Modalpartikeln *halt* und *eben* in der Berliner Kommunikationsgemeinschaft nach der ‚Wende'". In: Auer, Peter/Hausendorf, Heiko (eds.): *Kommunikation in gesellschaftlichen Umbruchssituationen: Mikroanalytische Aspekte des sprachlichen und gesellschaftlichen Wandels in den neuen Bundesländern*. Tübingen, Niemeyer: 199–234.

Engel, Ulrich (1991): *Deutsche Grammatik*. Heidelberg: Groos.

FOLK_E_00042: IDS, Datenbank für Gesprochenes Deutsch (DGD). hdl.handle.net/10932/00-0332-C1D6-9CC3-D201-9 [11.03.2026].

Fuchs, Harald/Schank, Gerd (1975): *Texte gesprochener deutscher Standardsprache. 3: Alltagsgespräche*. München: Hueber.

Heinrich, Wilma (2007): „‚Glatt' – ein abtönungsverdächtiges Lexem". In: Thüne, Eva-Maria/Ortu, Franca eds.): *Gesprochene Sprache – Partikeln. Beiträge der Arbeitsgruppen der 2. Tagung Deutsche Sprachwissenschaft in Italien Rom*. Frankfurt a. M., Lang: 167–176.

Helbig, Gerhard/Buscha, Joachim (2005): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. 5. Auflage. Berlin etc.: Langenscheidt.

Hentschel, Elke (1986): *Funktion und Geschichte deutscher Partikeln*. Tübingen: Niemeyer.

Hentschel, Elke (2013): „Verschiedene Wege, verschiedene Ziele“. *Germanistische Mitteilungen* 39: 63–78. doi.org/10.33675/GM/2013/1/5.

Hentschel, Elke/Keller, Heidi (2006): “Cultural Concepts of Parenting. A Linguistic Analysis”. *Linguistik online* 29, 4/06: 73–95. doi.org/10.13092/lo.29.558.

Hentschel, Elke/Weydt, Harald (2002): „Die Wortart ‚Partikel‘“. In: Wiegand, Herbert Ernst (ed.): *Handbücher zur Sprach- und Kommunikationswissenschaft.* Bd. 21*: Lexikologie*. Berlin/New York, De Gruyter: 646–653.

Kwon, Min-Jae (2005): *Modalpartikeln und Satzmodus: Untersuchungen zur Syntax, Semantik und Pragmatik der deutschen Modalpartikeln*. Dissertation, LMU München. edoc.ub.uni-muenchen.de/4877/1/Kwon_Min-Jae.pdf [12.02.2026].

Moroni, Manuela (2010): *Modalpartikeln zwischen Syntax, Prosodie und Informationsstruktur*. Frankfurt a. M. etc.: Lang.

Moroni, Manuela/Bidese, Ermenegildo (2021): „Deutsches *auch* und italienisches *anche* im Vergleich. Gemeinsame Funktion und sprachspezifischer Gebrauch“. *Linguistik online* 111, 6/21: 187–208. doi.org/10.13092/lo.111.8247.

Müller, Sonja (2018): *Distribution und Interpretation von Modalpartikel-Kombinationen*. Berlin: Language Science Press.

Schoonjans, Steven (2018): *Modalpartikeln als multimodale Konstruktionen: Eine korpusbasierte Kookkurrenzanalyse von Modalpartikeln und Gestik im Deutschen*. Berlin/Boston: De Gruyter. doi.org/10.1515/9783110566260.

Silberstein, Dagmar (2021): „Die Verwendung von Modalpartikeln in Gesprächen – korpusbedingte Unterschiede, häufige Muster und didaktische Zugänge“. *Informationen Deutsch als Fremdsprache* 48: 697–710. doi.org/10.1515/infodaf-2021-0080.

Silberstein, Dagmar (2024): *Modalpartikeln als Lerngegenstand. Partikelprofile für die Vermittlung von* aber, ja, doch, mal, denn, eigentlich *und* etwa *im DaF-Unterricht.* Berlin: Erich Schmidt.

Thurmair, Maria (1989): *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer.

Thurmair, Maria (2026): „Überlegungen und Anregungen zur Didaktik von Modalpartikeln“. *Linguistik online* 146, 5/26: 11–36. doi.org/10.13092/w63zyj27.

Weinrich, Harald (1993): *Textgrammatik der deutschen Sprache*. Mannheim etc.: Dudenverlag.

Westpfahl, Swantje (2020): *POS-Tagging für Transkripte gesprochener Sprache: Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)*. Tübingen: Narr Francke Attempto.

Westpfahl, Swantje/Schmidt, Thomas (2016): “FOLK-Gold – A GOLD standard for Part-of-Speech Tagging of Spoken German”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). European Language Resources Association: 1493–1499. lrec-conf.org/proceedings/lrec2016/pdf/397_Paper.pdf [12.07.2022].

Westpfahl, Swantje et al. (2017): *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS) [Online-Publikation]*. Mannheim: Institut für Deutsche Sprache. ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063 [11.11.2021].

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): *Grammatik der deutschen Sprache*. Berlin/New York: De Gruyter.

Korpora und Online-Ressourcen

DWDS = Digitales Wörterbuch der deutschen Sprache, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. dwds.de/d/wb-dwdswb [07.07.2022].

FOLK = Institut für Deutsche Sprache: Datenbank für Gesprochenes Deutsch (DGD). Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK), Version 2.16 vom 17.05.2021. Beschreibung des Korpus: agd.ids-mannheim.de/folk.shtml [01.07.2022]. Zugang zu Korpusdaten für registrierte NutzerInnen über: dgd.ids-mannheim.de/ [06.05.2022].

FR = Institut für Deutsche Sprache: Datenbank für Gesprochenes Deutsch (DGD), Grundstrukturen: Freiburger Korpus (FR). agd.ids-mannheim.de/FR--_extern.shtml [11.11. 2021].

GF = Institut für Deutsche Sprache: Gespräche im Fernsehen: Talkshows, Diskussionen, Interviews (GF). Informationen unter agd.ids-mannheim.de/korpus_index.shtml [11.11.2021].

GRAMMIS = Leibniz-Institut für Deutsche Sprache: Grammatisches Informationssystem grammis, „Systematische Grammatik“. grammis.ids-mannheim.de/systematische-grammatik/ [19.06.2022]

Anhang: Ausgewählte STTS 2.0-POS-Tags

In der folgenden Tabelle sind nur die im Beitrag verwendeten Kürzel aufgeführt. Eine vollständige Auflistung findet sich in Westpfahl et al. (2017: 8f.).

ADJD	adverbiales oder prädikatives Adjektiv	*[er fährt] schnell, [er ist] schnell*
ADV	Adverb	*hier, bald, gestern*
ART	bestimmter oder unbestimmter Artikel	*der, die, das, ein, eine*
KOUS	unterordnende Konjunktion mit Satz	*weil, dass, damit, wenn, ob*
NGIRR	Interjektionen, Responsive und Rezeptionssignale	*mhm, ach, tja*
NN	Appellativa	*Tisch, Herr, [das] Reisen*
PPER	irreflexives Personalpronomen	*ich, er, ihm, mich, der*
PRF	reflexives Personalpronomen	*sich, einander, dich, mir*
PTKIFG	Intensitäts-, Fokus- und Gradpartikeln	*sehr, nur, ziemlich*
PTKMA	Modal- und Abtönungspartikeln	*halt, ja, schon*
PTKNEG	Negationspartikel	*nicht*
VAFIN	finites Verb, Auxiliar	*[du] bist, [wir] werden*
VVFIN	finites Verb, voll	*[du] gehst, [wir] kommen [an]*