Zum Konnektorengebrauch in der gesprochenen Wissenschaftssprache Deutsch durch fortgeschrittene Lerner/innen

ADRIANA SLAVCHEVA
Westsächsische Hochschule Zwickau / University of Helsinki

Abstract
Der internationale wissenschaftliche Diskurs ist heute stärker denn je durch akademische Mobilität geprägt. Das Gelingen der internationalen Wissenschaftskommunikation hängt dabei essentiell vom angemessenen Wissenschaftshandeln im akademischen Kontext des Zielsprachenlandes ab und setzt die Herausbildung einer wissenschaftssprachlichen Handlungskompetenz in der jeweiligen Zielsprache – insbesondere im Mündlichen – voraus. Als eine wesentliche Teilkomponente der mündlichen Sprechhandlungskompetenz wird die Textkompetenz erachtet. Diese wurde bisher allerdings keiner systematischen empirisch fundierten Beschreibung unterzogen, die eine gezielte fremdsprachliche Förderung ermöglichen könnte.
Der vorliegende Beitrag widmet sich aus fremdsprachendidaktischer Perspektive einem zentralen Teilaspekt der Textkompetenz – dem Gebrauch von Konnektoren zur Herstellung von Textkohäsion – in der gesprochenen Wissenschaftssprache Deutsch und stellt ein empirisches Untersuchungsdesign zur Aufdeckung von Lernschwierigkeiten beim Konnektorengebrauch fortgeschrittener Lerner/innen vor. Dieses ermöglicht erstmalig umfassende empirische Einblicke in die Verwendungspräferenzen der Konnektoren in der mündlichen Wissenschaftskommunikation bei L2-Sprecher/inne/n im Kontrast zu muttersprachlichen Sprachproduktionen.
Key words: Konnektoren, gesprochene Wissenschaftssprache, Korpuslinguistik, GeWiss-Korpus, Deutsch als Fremdsprache

1 Einleitung

Internationale Mobilität von Wissenschaftler/inne/n, Studierenden und Dozenten ist aus dem aktuellen wissenschaftlichen Diskurs nicht mehr wegzudenken. Das Gelingen der internationalen Wissenschaftskommunikation hängt dabei essentiell vom angemessenen Wissenschaftshandeln im akademischen Kontext des Zielsprachenlandes ab und setzt die Herausbildung einer wissenschaftssprachlichen Handlungskompetenz in der jeweiligen Zielsprache – insbesondere im Mündlichen – voraus (vgl. Fandrych et al. 2009: 18f.).

Aus der fremdsprachendidaktischen Perspektive des Unterrichts Deutsch-als-Fremdsprache stellt sich hier die Frage, wie die mündliche Handlungskompetenz in der fremden Wissenschaftssprache Deutsch beschaffen ist und wie sie aufgebaut werden kann.

Als eine wesentliche Teilkomponente der mündlichen Sprechhandlungskompetenz wird die Textkompetenz – die Fähigkeit, zusammenhängende Texte zu produzieren – erachtet (vgl. Bachman/Palmer 1996). Aus fremdsprachendidaktischer Perspektive ist dabei der Gebrauch von

Konnektoren als kohäsionsstiftenden sprachlichen Mitteln von besonderer Relevanz (vgl. bspw. Brinker 2006: 81). Der Konnektorengebrauch wurde für die gesprochene Wissenschaftssprache Deutsch bisher allerdings keiner systematischen empirisch fundierten Beschreibung unterzogen, die eine gezielte fremdsprachliche Förderung ermöglichen könnte. Begründet war dies unter anderem durch die fehlende Datengrundlage: Diese und vergleichbare Fragestellungen können kontrastiv nur auf einer breiten empirischen Basis in Form von methodisch stringent aufgebauten Korpora beschrieben werden. Für die empirische Untersuchung der gesprochenen deutschen Wissenschaftssprache wurde mit dem GeWiss-Korpus (vgl. Fandrych et al. 2014b) erstmalig eine verlässliche Datengrundlage geschaffen.[1]

Die vorliegende Untersuchung knüpft an das skizzierte Forschungsdesiderat an und geht aus fremdsprachendidaktischer Perspektive auf der Grundlage von Vergleichskorpora der Frage nach, wie sich die Verwendungspräferenzen der Konnektoren in monologischen Genres der mündlichen Wissenschaftskommunikation bei L1- und L2-Sprecher/inne/n unterscheiden und welche Konnektoren für Nichtmuttersprachler/innen besondere Schwierigkeiten bereiten und daher einer intensiveren Förderung im Fremdsprachenunterricht bedürfen.

Als Beispiel für ein wichtiges monologisches Genre wird dabei der studentische Vortrag ausgewählt, der im deutschen akademischen Kontext nach wie vor eine zentrale Rolle spielt (vgl. Guckelsberger 2006) und erfahrungsgemäß nicht nur Nichtmuttersprachler/inne/n Schwierigkeiten bereitet. Als vorbereitetes monologisches Genre liegt er zudem an der Schnittstelle zwischen Schriftlichkeit und Mündlichkeit, was erwarten lässt, dass Konnektoren als explizite Markierungen von Textkohärenz hier eine wichtige Rolle zukommt.

Die Untersuchung ist zielgruppenspezifisch angelegt für Studierende aus Bulgarien, die zahlenmäßig seit über 10 Jahren zu den 6 wichtigsten Herkunftsländern ausländischer Studierender an deutschen Hochschulen gehören (vgl. DAAD/DZHW 2017) und hat somit für den deutschen akademischen Kontext große Relevanz.

2 Untersuchungsgegenstand und –design

2.1 Untersuchungsgegenstand

Die Konnektoren gelten neben Pro-Formen als Kohäsionsmittel „par excellence" – sie verknüpfen explizit zwei Textsegmente und spezifizieren dabei die Natur des inhaltlichen Zusammenhangs zwischen den Segmenten. Für die Abgrenzung des Gegenstandsbereichs dieser syntaktisch heterogenen Klasse stellen Pasch et al. (2003: 331) fünf Merkmale auf: Konnektoren werden als nicht flektierbar, nicht kasusfordernd und als semantisch zweistellig definiert, die Argumente ihrer

[1] Das GeWiss-Korpus ist unter https://gewiss.uni-leipzig.de nach kostenloser Anmeldung öffentlich zugänglich. [Stand 24.01.2018]

Bedeutung sind propositionale Strukturen und müssen als Satzstrukturen ausgedrückt werden können.

Der Umfang des anvisierten Untersuchungsgegenstandes Konnektoren ist beachtlich: Im Handbuch der deutschen Konnektoren (vgl. Pasch et al. 2003) werden ca. 350 Formative gelistet. Allerdings basiert dieses Referenzwerk ausschließlich auf der Schriftsprache, weshalb seine Konnektorenliste eine Reihe stilistisch markierter Konnektoren mit ganz speziellen Verwendungsbedingungen beinhaltet. Viele davon kommen in der mündlichen Modalität äußerst selten vor und dürften von eher geringer Relevanz für den Fremdsprachenunterricht sein. Man denke dabei etwa an Konnektoren wie *alldieweil, einesteils, anderenteils, desungeachtet, in Sonderheit, nichtsdestominder, unbeschadet dessen, wofern* etc. Für die vorliegende Untersuchung gilt es also, eine begründete Auswahl der zu untersuchenden Konnektoren zu treffen, die für den DaF-Bereich von Relevanz ist.

Als Kriterium für die Bestimmung des Untersuchungsgegenstandes wird in der vorliegenden Untersuchung die Häufigkeit gewählt. Häufigkeit gilt nachweislich als zentrales Kriterium beim L1- und L2-Spracherwerb (in gesteuerten wie in ungesteuerten Situationen) in allen sprachlich relevanten Domänen: beim Lexikerwerb, beim Aufbau der fremdsprachlichen Phonologie und Syntax, beim Textverständnis und in der Textproduktion (vgl. Ellis 2002 für Hinweise zu zahlreichen Frequenzeffekten bei der sprachlichen Prozessierung in allen sprachrelevanten Domänen). Es gilt zudem als allgemein anerkannt, dass die häufigsten Wörter den Großteil eines geschriebenen und gesprochenen Textes abdecken. So machen lt. Tschirner (2008) im Allgemeinen die häufigsten 1000 Wörter des Deutschen ca. 73%, die häufigsten 2000 Wörter ca. 79%, die häufigsten 3000 Wörter ca. 82% und die häufigsten 4000 Wörter ca. 84% eines Textes aus.

Diese Vorüberlegungen führen zum folgenden Untersuchungsdesign für die vorliegende Untersuchung.

2.2 Untersuchungsdesign

Die Untersuchung wird als eine kontrastive Korpus-basierte Analyse bestehend aus drei aufeinander aufbauenden Studien realisiert:

In einer ersten Studie wird eine frequenzbasierte Liste der Konnektoren unter den 1000 häufigsten Wörtern in einem Referenzkorpus der gesprochenen deutschen Allgemeinsprache erstellt. Diese Liste dient als Grundlage für Studie 2, in der ein frequenzbasierter Korpusvergleich zweier mündlicher Korpora mit L1- und L2-Daten in der deutschen Wissenschaftssprache vorgenommen wird, um die Signifikanzunterschiede in den Häufigkeitsverteilungen der Konnektoren beider Sprechergruppen zu ermitteln. Die Konnektoren, deren Häufigkeitsverteilungen bei L1- und L2-Sprechern des Deutschen signifikant unterschiedlich sind, werden schließlich in Studie 3 hinsichtlich ihres Lernschwierigkeitsgrads untersucht.

3 Studie 1: Konnektoren in der gesprochenen deutschen Allgemeinsprache

3.1 Empirische Datengrundlage – das Herder/BYU-Korpus als Referenzkorpus

Die Ermittlung von Häufigkeitsverteilungen sprachlicher Elemente erfordert den Rückgriff auf sprachliche Korpora, die den Kriterien Aktualität, Repräsentativität und Ausgewogenheit genügen (vgl. Tschirner 2008: 195). Als Referenzkorpus für die Ermittlung der Konnektoren in der gesprochenen deutschen Allgemeinsprache wurde für die vorliegende Untersuchung das Herder/BYU-Korpus der deutschen Gegenwartssprache ausgewählt, welches aufgrund seiner Ausgewogenheit und Repräsentativität trotz seines im Vergleich zu schriftsprachlichen Korpora relativ kleinen Umfangs für die Erstellung von Frequenzlisten bestens geeignet ist (vgl. Jones 2004: 170).

Das Herder/BYU-Korpus wurde in Kooperation zwischen der Universität Leipzig und der Brigham Young University in den USA in Zusammenhang mit der Entwicklung eines Frequenzwörterbuchs des Deutschen (Jones/Tschirner 2006) aufgebaut und ist am Herder-Institut der Universität Leipzig für Forschungs- und Lehrzwecke intern zugänglich. Das Korpus hat einen Umfang von ca. 4,2 Millionen laufenden Wörtern und umfasst 5 Subkorpora: gesprochene Sprache, Zeitungssprache, literarische Sprache, Sach- und Fachtexte sowie Gebrauchstexte. Zudem wurden in allen Subkorpora deutsche, österreichische wie Schweizer Texte im Verhältnis 70:20:10 integriert, sodass das Korpus als repräsentativ für die deutschsprachigen Länder angesehen werden kann (vgl. Tschirner 2008: 196).

Das Subkorpus der gesprochenen Sprache umfasst mehr als 1 Million laufende Wörter, die auf vier verschiedene Genres verteilt sind: 700.000 Wörter wurden aus ca. 400 Interviews mit Muttersprachler/inne/n des Deutschen gewonnen, die als repräsentativ hinsichtlich soziodemografischer Parameter (Alter, Geschlecht, Bildungshintergrund), Themenvielfalt sowie der nationalen Varietät des Deutschen (deutsche Sprache in Deutschland, Österreich und der Schweiz) anzusehen sind. Ergänzend dazu wurde auch Fernsehmaterial in das Korpus aufgenommen – 100.000 laufende Wörter aus Fernsehserien, 140.000 Wörter aus Talkshows und Fernsehdiskussionen sowie 60.000 Wörter aus Berichten und längeren Monologen (vgl. Jones 2004: 170; Tschirner 2005: 138).

Die aufgenommenen Daten wurden orthographisch transkribiert und anschließend mit dem Stuttgart Tree Tagger (vgl. Schmid 1995) und dem Stuttgart-Tübingen-Tagset (STTS) (vgl. Schiller et al. 1999) nach Wortarten annotiert (Part-of-speech-Tagging), was die spätere grammatische Disambiguierung der Wortformen unterstützt.

Damit das ausgewählte Referenz-Korpus Herder/BYU als Grundlage für die Folgeuntersuchung in einem Spezialkorpus der gesprochenen Wissenschaftssprache dienen kann, muss die Vergleichbarkeit der in den Korpora enthaltenen Sprachdaten gewährleistet sein. Obwohl das Herder/BYU-Korpus keine Sprachdaten aus dem akademischen Bereich enthält, sind einzelne darin enthaltene Genres aufgrund der sie charakterisierenden allgemeinen

Kommunikationsbedingungen im Hinblick auf das Nähe-Distanz-Kontinuum (vgl. Koch/Oesterreicher 1985) ähnlich wie das im Spezialkorpus interessierende akademische Genre *studentischer Vortrag* einzuordnen (vgl. Tabelle 1). So sind die in dem Herder/BYU-Korpus enthaltenen Genres *Bericht* bzw. *längerer Monolog* aufgrund ihrer Monologizität, der Fremdheit der Kommunikationspartner und Öffentlichkeit der Kommunikation sowie der Themenfixierung und der Reflektiertheit der Äußerungen mit den Kommunikationsbedingungen *wissenschaftlicher Vorträge* im Spezialkorpus der gesprochenen deutschen Wissenschaftssprache vergleichbar. Die kommunikativen Parameter der *wissenschaftlichen Diskussionen* im Anschluss an einen Vortrag (Spezialkorpus) stimmen wiederum mit denen von *Fernsehdiskussionen* bzw. *Talkshows* (Herder/BYU-Korpus) weitestgehend überein – beide lassen sich durch die Kombination von Fremdheit der Kommunikationspartner und Öffentlichkeit der Kommunikation, Dialogizität, relative Spontaneität und freie Themenentwicklung sowie durch eine Face-to-Face-Interaktion und Situationsverschränkung charakterisieren.

Tabelle 1: Vergleichbarkeit der Daten des Referenzkorpus Herder/BYU und des Spezialkorpus der gesprochenen Wissenschaftssprache

Herder/BYU-Korpus		**Spezialkorpus der gesprochenen Wissenschaftssprache**	
Genre	**Konzeption**	**Genre**	**Konzeption**
Interviews	vorwiegend kommunikative Nähe	[keine Entsprechung]	
Talkshows**;** Fernsehdiskussionen	vorwiegend kommunikative Nähe	wissenschaftliche Diskussionen	vorwiegend kommunikative Nähe
Fernsehserien	fingierte Mündlichkeit	[keine Entsprechung]	
Berichte; längere Monologe	vorwiegend kommunikative Distanz	wissenschaftliche Vorträge	vorwiegend kommunikative Distanz

Die weitgehende Übereinstimmung der allgemeinen Kommunikationsbedingungen der Sprachdaten aus den hier zu untersuchenden Teilen des Spezialkorpus der gesprochenen Wissenschaftssprache mit den entsprechenden Referenzdaten aus dem Herder/BYU-Korpus stellt die Vergleichbarkeit beider Korpora sicher und erlaubt einen Korpusvergleich für die Zwecke der vorliegenden Untersuchung nach dem im Abschnitt 2.2 beschriebenen Untersuchungsdesign.

3.2 Häufigkeitsverteilung der Konnektoren im Referenzkorpus Herder/BYU

Als Ausgangspunkt der vergleichenden Korpusuntersuchung zum Gebrauch von Konnektoren durch L1- und L2-Sprecher/innen in der gesprochenen Wissenschaftssprache Deutsch (vgl. Abschnitt 4) dient eine Liste der häufigsten Konnektoren im mündlichen Subkorpus des Referenzkorpus Herder/BYU.

Für die Erstellung der Konnektorliste wurde in einem ersten Schritt eine Frequenzliste des gesamten Subkorpus mithilfe des WordSmith-Tools 4.0 (Scott 2004) generiert. Grundlage dafür waren die nach Wortarten getaggten Transkripte (insgesamt 24 Texte) mit einem Umfang von 1.011.528 laufenden Wörtern (Tokens).

Durch die morphosyntaktische Annotation des Herder/BYU-Korpus ist die Frequenzliste bereits grammatisch disambiguiert. So wurde beispielsweise bei der Form *da* zwischen drei verschiedenen Wortformen unterschieden – *da* als Adverb, als Konjunktion bzw. als Verbpartikel. Diese wurden beim automatischen POS-Tagging mit einem jeweils anderen Wortartenlabel (Tag) ausgezeichnet – *ADVB*, *KONJ* bzw. *VPRF* und bzgl. ihrer Frequenz getrennt gelistet. Somit unterstützt die grammatische Disambiguierung wesentlich die Identifikation der Konnektoren und die korrekte Ermittlung von deren Häufigkeitsverteilungen im Korpus. Im Falle von *da* beispielsweise werden nur *da [ADVB]* und *da [KONJ]* als Konnektoren gezählt, nicht aber *da [VPRF]*.

Die automatisch generierte Frequenzliste besteht aus individuellen Wortformen und wurde nicht lemmatisiert, da diese manuelle Aufbereitung der Liste keine Auswirkungen auf die Häufigkeitsverteilungen der Konnektoren nach sich ziehen würde.[2] Die Frequenzliste des mündlichen Subkorpus des Herder/BYU-Korpus umfasst somit 53.420 individuelle Wortformen (Types), 56,57% davon sog. Hapax Legomena, die im Korpus nur einmal vorkommen.

In einem zweiten Schritt wurden die Konnektoren unter den 1000 häufigsten Wortformen der allgemeinen Frequenzliste ermittelt. Die häufigsten 1000 Wörter ergeben 79,74% der laufenden Wörter in dem Subkorpus. Das erste Wort in der Liste kommt 31.969 Mal im mündlichen Subkorpus vor, das Wort mit der Rangnummer 1000 immerhin noch 82 Mal. Für die Identifikation der Konnektorkandidaten wurde daher auf die morphosyntaktische Annotation der

[2] Die Lemmatisierung ermöglicht, dass Flexionsformen ihren Grundformen zugeordnet werden, wobei die Häufigkeiten der einzelnen Formen zusammengerechnet werden. Auf diese Weise führt die Lemmatisierung zur Erhöhung der Häufigkeit bei Lemmata flektierbarer Wortarten und somit zu deren Aufstieg in der allgemeinen Frequenzliste. Bei nicht flektierbaren Wortarten hingegen führt die Lemmatisierung zu keinerlei Frequenzänderungen. Für die hier interessierende Wortklasse der Konnektoren, die ausschließlich nicht flektierbare Ausdrücke umfasst (vgl. Pasch et al. 2003: 1), bedeutet dies, dass eine Lemmatisierung zu keiner Änderung der frequenzbasierten absoluten Reihenfolge der Konnektoren untereinander führen wird. Zu erwarten wäre lediglich deren Verschiebung nach hinten in der Frequenzliste durch das Aufsteigen einzelner flektierbarer Lemmata, sodass einzelne niedrig frequentierte Konnektoren höchstens aus der Untersuchung ausgeschlossen werden müssten. Vor dem Hintergrund des hohen manuellen Aufwands dieses Aufbereitungsschrittes wurde daher auf eine Lemmatisierung der allgemeinen Häufigkeitsliste des mündlichen Subkorpus verzichtet.

Wortformen zurückgegriffen, indem alle Wortformen, die in der Frequenzliste als Adverbien [ADVB], Pronominaladverbien [PADVB] oder Konjunktionen bzw. Subjunktionen [KONJ] getaggt waren, auf der Grundlage des Handbuchs der deutschen Konnektoren (Pasch et al. 2003) auf eine konnektorale Verwendung hin untersucht wurden. Zudem wurde bei niedrig frequentierten polysemen Formativen, etwa *dagegen* (Frequenz 92, Rang 911 in der Frequenzliste) oder *allein* (Frequenz 156, Rang 607), sowie bei Formativen, die nur als Bestandteile mehrgliedriger Konnektoren eine konnektorale Lesart aufweisen, beispielsweise *davon* (Frequenz 395, Rang 290) als *davon abgesehen*, eine Disambiguierung anhand von Konkordanzanalysen durchgeführt, um zu überprüfen, ob die konnektoralen Verwendungen dieser Formative eine ausreichende Frequenz (mindestens 82 Vorkommen) aufweisen, um unter die 1000 häufigsten Wörter des gesamten Subkorpus gerechnet zu werden.[3] Bei den hochfrequentierten Formativen wurde hingegen auf eine Disambiguierung verzichtet, da ihre hohe Frequenz ein Herausfallen aus der Liste der 1000 häufigsten Wörter unwahrscheinlich macht.

Die so ermittelte Liste der Konnektoren unter den 1000 häufigsten Wörtern des mündlichen Teils des Herder/BYU-Korpus enthält insgesamt 76 Wörter: 47 Adverbien [ADVB], 16 Konjunktionen bzw. Subjunktionen [KONJ] sowie 12 Pronominaladverbien [PADVB] (vgl. Tabelle 2). Sie sind gleichmäßig in dem mündlichen Subkorpus vertreten und kommen in mindestens 68% der Korpustexte vor, 63 der Konnektoren sogar in mehr als 83% der Texte. Insgesamt machen die Vorkommen der 76 Konnektoren unter den 1000 häufigsten Wörtern 15,53% der laufenden Wörter des gesamten mündlichen Subkorpus des Herder/BYU-Korpus aus. Der häufigste Konnektor – *und* – ist mit einer Frequenz von 31.969 gleichzeitig auch das häufigste Wort im gesamten Subkorpus. Der letzte Konnektor in der Liste – *später* – kommt 83 Mal im mündlichen Subkorpus vor.

[3] Folgende 7 Konnektoren sind nach der Disambiguierung anhand von Konkordanzanalysen aus der Liste der 1000 häufigsten Wörter des mündlichen Teils des Herder/BYU-Korpus herausgefallen und wurden daher von der weiteren Untersuchung ausgenommen: *allein*, *dagegen*, *darüber hinaus*, *darum*, *davon abgesehen*, *genau gesagt*, *umso weniger als*.

Tabelle 2: Konnektoren unter den 1000 häufigsten Wörtern im mündlichen Teil des Herder/BYU-Korpus

Rang	Wort	Frequenz
1	UND	31969
10	AUCH	12670
11	SO	12659
13	DANN	10921
14	ALSO	10825
15	DA (ADVB)	9968
16	JA (ADVB)	8901
29	ODER	5961
32	NOCH	5425
34	WENN	5272
43	SCHON	4448
45	ABER (KONJ)	4287
54	WEIL	3111
56	EIGENTLICH	3005
59	NUR	2730
72	ABER (ADVB)	2167
74	DOCH (ADVB)	2141
80	WIEDER	1955
95	DENN (ADVB)	1559
124	GAR	1073
151	NUN	878
158	SONDERN	831
189	ALS (KONJ)	627
194	DENN (KONJ)	612
196	ZWAR	611
198	DAMALS	605
199	DAZU	604
204	SONST	580
207	DAMIT (PADVB)	572
233	DA (KONJ)	494
236	SELBST	491
252	DAFÜR	458
279	BESONDERS	404
289	ERSTMAL	396
293	DABEI	393
316	DARAUF	360
322	VORHER	353
329	NÄMLICH	334

Rang	Wort	Frequenz
364	OBWOHL	296
376	DADURCH	287
379	DESWEGEN	281
440	DAMIT (KONJ)	238
442	SOGAR	238
443	ALLERDINGS	237
456	BLOß	232
461	NACHHER	228
465	DESHALB	225
482	SOWIESO	216
486	ETWA	211
522	DOCH (KONJ)	197
525	TROTZDEM	197
543	JEDENFALLS	188
559	ZUERST	175
582	KAUM	164
596	DANACH	159
598	SOFORT	159
611	INSOFERN	155
622	GENAUSO	151
625	DAHER	150
629	DRAUF	148
645	BEVOR	144
659	INZWISCHEN	141
668	AUßERDEM	138
673	ANSONSTEN	137
705	NACHDEM	127
709	ENTWEDER	126
722	WOBEI	124
730	ÜBRIGENS	122
770	BEZIEHUNGSWEISE	113
787	MINDESTENS	111
800	WÄHREND (KONJ)	110
810	MITTLERWEILE	108
814	ZUNÄCHST	108
896	ZUMINDEST	95
945	EINMAL	87
988	SPÄTER	83

Die so ermittelte Liste der Konnektoren unter den 1000 häufigsten Wörtern des mündlichen Teils des Herder/BYU-Korpus dient als Grundlage für Studie 2, in der der

Konnektorengebrauch in der gesprochenen deutschen Wissenschaftssprache bei L1- und L2-Sprecher/inne/n des Deutschen untersucht wird.

4 Studie 2: Konnektoren in der gesprochenen deutschen Wissenschaftssprache

Zur Aufdeckung von sprachlichen Merkmalen, die Korpora voneinander unterscheiden, kann die Methode des Korpusvergleichs herangezogen werden. Voraussetzungen für zuverlässige Korpusvergleiche sind dabei in erster Linie die Homogenität und Vergleichbarkeit der untersuchten Korpora (vgl. Rayson/Garside 2000). Im Folgenden wird die empirische Datengrundlage der Studie 2 beschrieben und nachgewiesen, wie die Kriterien der Homogenität und Vergleichbarkeit der zugrunde gelegten Vergleichskorpora gewährleistet wurden.

4.1 Empirische Datengrundlage – das GeWiss-Korpus

Die Daten beider Vergleichskorpora in Studie 2 entstammen dem GeWiss-Korpus (vgl. Meißner/Slavcheva 2014)[4], welches im Rahmen des internationalen Forschungsprojekts „Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen" aufgebaut wurde. Es entstand in Kooperation zwischen dem Herder-Institut der Universität Leipzig, der Universität Wrocław/Polen und der Aston University/Großbritannien sowie mit weiteren assoziierten Partnern und stellt das einzige frei verfügbare Korpus der gesprochenen Wissenschaftssprache Deutsch im deutschsprachigen Raum dar. Das GeWiss-Korpus wurde gezielt nach kontrollierten Variablen und stringenten Kriterien aufgebaut, um kontrastive empirische Untersuchungen zur mündlichen Wissenschaftskommunikation in ausgewählten akademischen Kontexten zu ermöglichen. Es umfasst Daten in den Sprachen Deutsch (als L1 sowie als L2), Polnisch und Englisch zu zwei zentralen Genres der akademischen Kommunikation im universitären Bereich – dem wissenschaftlichen Vortrag und dem Prüfungsgespräch, die in den jeweiligen akademischen Kontexten Deutschlands, Polens, Großbritanniens, Bulgariens und Italiens erhoben wurden.

4.1.1 Probanden

Für die vorliegende Untersuchung wurden mündliche Produktionen in der fremden Wissenschaftssprache Deutsch von fortgeschrittenen L2-Sprecher/inne/n des Deutschen mit L1 Bulgarisch aus dem bulgarischen akademischen Kontext sowie vergleichbare Produktionen von L1-Sprecher/inne/n des Deutschen aus dem deutschen akademischen Kontext erhoben und kontrastiert.

Für das Lernerkorpus wurden insgesamt 19 Studierende im B.A.-Studiengang Germanistik aus dem bulgarischen akademischen Kontext aufgenommen, die ihre Schulbildung im bulgarischen Schulsystem absolviert haben. Es handelt sich dabei um mehrheitlich weibliche Sprecher/innen[5] im Alter zwischen 21 und 41 Jahren (Median des

[4] Vgl. auch Fußnote 1.

[5] Unter den 19 aufgenommenen Studierenden im bulgarischen akademischen Kontext gab es lediglich einen männlichen Sprecher; unter den 25 L1-Sprecher/inne/n aus dem deutschen akademischen Kontext 4. Diese geschlechtliche Verteilung der aufgenommen Personen entspricht der empirischen Beobachtung während der Datenerhebung, dass in germanistischen/DaF-Studiengängen an beiden

Alters 23 Jahre), deren L2-Sprachkompetenz im Deutschen mindestens auf das Niveau B2/C1 des Gemeinsamen europäischen Referenzrahmens für Sprachen[6] eingeschätzt werden kann.[7]

Für das Vergleichskorpus mit L1-Sprecherdaten wurden insgesamt 25 Studierende in Germanistik-/DaF-Studiengängen im deutschen akademischen Kontext aufgenommen, die eine Schulbildung im deutschen Schulsystem erfahren haben. Es waren mehrheitlich weibliche Sprecher/innen[8] im Alter zwischen 20 und 35 Jahren (Median des Alters 23 Jahre). 14 Sprecher/innen studierten in einem B.A.-Studiengang, 11 im Master.

4.1.2 Korpus[9]

Datengrundlage der vorliegenden Untersuchung sind Audio- bzw. Videoaufnahmen studentischer Seminarvorträge auf Deutsch, die für die folgende Analyse transkribiert, mit Metadaten angereichert und anschließend zu zwei multimodalen Korpora verknüpft wurden. Um die Ergebnisse der Untersuchung auf eine solide Datenbasis zu stützen, wurde beim Korpusaufbau ein Aufnahmevolumen von 5 Stunden je Teilkorpus festgesetzt. Zudem wurden beide Vergleichskorpora gezielt nach denselben Designkriterien erstellt und unterscheiden sich voneinander genau in den zwei zu untersuchenden Vergleichsparametern – *L1 der Probanden* (Deutsch vs. Bulgarisch) sowie *akademischer Kontext* (Deutschland vs. Bulgarien), in dem die Seminarvorträge erhoben wurden:

(1) **Korpus GEWISS-DEU-L1-SV** – ein Korpus mit deutschsprachigen studentischen Seminarvorträgen von **L1-Sprecher/inne/n** des Deutschen aus dem **deutschen akademischen Kontext**

Standorten mehrheitlich weibliche Studierende immatrikuliert sind. Diese Tatsache ist für den deutschen akademischen Kontext auch statistisch belegt – 2011 lag der Anteil der weiblichen Germanistik-Studierenden an deutschen Hochschulen bei 76,49 % (vgl. Statistisches Bundesamt 2018).

6 Vgl. http://www.europaeischer-referenzrahmen.de/. [Stand 24.01.2018]

7 Als Aufnahmevoraussetzung zum Germanistikstudium an bulgarischen Universitäten haben die in dieser Studie berücksichtigten Sprecher/innen mehrheitlich eine Aufnahmeprüfung in Deutsch auf dem Niveau B2 absolviert. Vier der aufgenommenen Sprecher/innen verfügen zusätzlich über ein Sprachzertifikat, das in Deutschland als sprachliche Hochschulzugangsberechtigung mind. auf dem Niveau C1 anerkannt wird (DSH, TestDaF bzw. Deutsches Sprachdiplom). Stichprobenartige Evaluationen der mündlichen Sprachkompetenz einzelner Sprecher/innen zum Zeitpunkt der Aufnahmen mit dem zertifizierten Testinstrument OPIc ergaben zudem Einstufungen auf den Niveaus Advanced Mid (~ B2 des GeR) bis Advanced High (~ C1 des GeR) des American Council on the Teaching of Foreign Languages (ACTFL) (vgl. Tschirner 2005 zu einem Vergleich der Skalen des Europarats und ACTFL). Bis zum Zeitpunkt der Aufnahmen haben die Sprecher/innen als Teil des Curriculums des B.A.-Germanistikstudiums an studienbegleitendem Deutschunterricht im Umfang von 510 Unterrichtseinheiten teilgenommen sowie Vorlesungen und Seminare zu verschiedenen germanistischen Teildisziplinen in deutscher Sprache mit einem Gesamtumfang von 645 Unterrichtseinheiten besucht. Das Curriculum sieht zudem für den Sprachunterricht wie für die theoretischen Disziplinen eine Modulabschlussprüfung auf Deutsch vor (vgl. Curriculum 2007).

8 Vgl. Fußnote 5.

9 Die methodischen Aspekte des Korpusaufbaus im Rahmen des GeWiss-Projekts (Datenerhebung, Datenaufbereitung, Metadatenanlage und Transkription) wurden von mir in Fandrych et al. (2012: 327f.) bzw. bei Meißner/Slavcheva (2014) ausführlich beschrieben.

(2) **Korpus GEWISS-DEU-L2-BG-SV** – ein Korpus mit deutschsprachigen studentischen Seminarvorträgen von **L2-Sprecher/inne/n** des Deutschen mit **L1 Bulgarisch** aus dem **bulgarischen akademischen Kontext**.

Das für die vorliegende Untersuchung zusammengestellte Korpus GEWISS-DEU-L1-SV umfasst 16 authentische studentische Einzel- bzw. Gruppenvorträge auf Deutsch, die im Rahmen von Bachelor- oder Masterseminaren im Fachgebiet Germanistik/DaF im deutschen akademischen Kontext gehalten wurden. Bei den meisten Vorträgen kommen verschiedene Medien zum Einsatz – in der Regel wird der Vortrag von einer eigens dafür vorbereiteten PowerPoint-Präsentation begleitet, häufig wird den Zuhörern zusätzlich ein Handout zur Verfügung gestellt, vereinzelt werden auch Audio-, Videobeispiele, ein Tafelbild oder das Internet genutzt. Die meisten Vorträge wurden frei vorgetragen, nur in zwei Fällen haben die Vortragenden zum Teil von einer Textvorlage abgelesen.

Die Gesamtdauer der aufgenommenen Kommunikationsereignisse beträgt ca. 7:24 h. Für die Zwecke der vorliegenden Untersuchung wurden davon nur die Sprecherbeiträge der Vortragenden mit L1 Deutsch ausgewertet, die von insgesamt 25 mehrheitlich weiblichen Sprecher/inne/n stammen. Die Sprecherbeiträge aller weiteren Beteiligten an den Kommunikationsereignissen – Seminarleiter/innen, Diskutant/inn/en sowie ggf. Mit-Vortragende mit Deutsch als L2[10] – wurden von den quantitativen Auswertungen ausgenommen. Die analysierten Vortragsteile umfassen somit größtenteils monologische Daten und haben einen Umfang von ca. 4:18 h (Durchschnittsdauer - ca. 16 Min. pro Sprecher/in).

Das für die vorliegende Untersuchung aufgebaute Korpus GEWISS-DEU-L2-BG-SV enthält 10 authentische deutschsprachige studentische Seminarvorträge (mehrheitlich Gruppenvorträge), aufgezeichnet in Bachelorseminaren im Fachgebiet Germanistik/DaF im bulgarischen akademischen Kontext. In den Vorträgen kommen verschiedene Medien zum Einsatz – alle Vortragenden verwenden ein Skript und setzen Handouts ein, in der Regel wird der Vortrag auch von einer eigens dafür vorbereiteten Präsentation begleitet; eine Vortragende verwendet zusätzlich die Tafel. Mit einer einzigen Ausnahme werden alle Vorträge teilweise von einer Textvorlage abgelesen.

Die aufgenommenen Kommunikationsereignisse haben eine Gesamtdauer von ca. 5:22 h. Wie auch beim Korpus GEWISS-DEU-L1-SV wurden für die vorliegende Untersuchung nur die überwiegend monologischen Sprecherbeiträge der Vortragenden ausgewertet – insgesamt 19 mehrheitlich weibliche Sprecher/innen mit L1 Bulgarisch. Der Umfang der analysierten Vortragsteile beträgt somit insgesamt ca. 4:27 h, durschnittlich 27 Min. pro Kommunikationsereignis.

Tabelle 3 gibt einen zusammenfassenden Überblick des Korpusdesigns und der Zusammenstellung der beiden Vergleichskorpora GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV und veranschaulicht deren weitestgehende Vergleichbarkeit und Homogenität.

[10] Sieben der aufgenommenen Vorträge wurden gemeinschaftlich von L1- und L2-Sprecher/inne/n des Deutschen gehalten. In diesen Fällen blieben die Sprecherbeiträge der L2-Vortragenden bei allen Auswertungen unberücksichtigt.

Tabelle 3: Übersicht der analysierten Korpora

Korpus **Designparameter**	**GEWISS-DEU-L1-SV**	**GEWISS-DEU-L2-BG-SV**
Sprache der Kommunikation	Deutsch	Deutsch
L1 der Probanden	Deutsch	Bulgarisch
Genre	Studentisches Referat	Studentisches Referat
Fachgebiet	Germanistik/DaF	Germanistik/DaF
Akademischer Kontext	Deutschland	Bulgarien
Anzahl analysierter Sprecher	25	19
Anzahl Kommunikationen	16	10
Gesamtdauer der Aufnahmen[11]	07:24:30 h	05:21:46 h
Dauer der analysierten Vortragsteile	04:18:04 h	04:26:32 h
Grad der Mündlichkeit	i.d.R. frei gesprochen	i.d.R. zum Teil abgelesen
Medieneinsatz	ja	ja
Transkriptionskonvention	GAT 2	GAT 2
Korpusgröße[12]	39.292 Tokens	31.060 Tokens

Beide Korpora wurden gezielt nach kontrollierten Variablen und stringenten Kriterien aufgebaut, sodass sie sich genau in den zwei zu untersuchenden Vergleichsparametern – *L1 der Probanden* (Deutsch vs. Bulgarisch) sowie *akademischer Kontext* (Deutschland vs. Bulgarien), in dem die Aufnahmen angefertigt wurden, voneinander unterscheiden. Sie umfassen eine vergleichbare *Anzahl* (16 bzw. 10) authentischer Aufnahmen von studentischen Referaten (*Genre*) auf Deutsch (*Sprache der Kommunikation*) im Fach Germanistik/DaF (*Fachbereich*) mit einer ausreichend großen *Sprecheranzahl* (25 bzw. 19). Die Aufnahmen beider Korpora wurden einheitlich nach der *Transkriptionskonvention* GAT 2 für die spätere Korpusanalyse verschriftlicht.

Neben den einheitlichen Designparametern wurde bei der Zusammenstellung der Korpora auch auf deren Größe geachtet. Als Kriterium für die Vergleichbarkeit wurde das Aufnahmevolumen festgesetzt, um somit die Datenerhebung an den beiden Standorten entsprechend planen zu können. Wie man Tabelle 3 entnehmen kann, konnte die *Dauer der analysierten Vortragsteile* beider Korpora auf diese Weise so gut wie identisch (ca. 4:20 h) gehalten werden. Vergleichbar sind darüber hinaus auch die statistisch ermittelten Korpusgrößen in laufenden Wörtern. Durch die Auswertung ausschließlich der Vortragsteile der Referent/inn/en der aufgenommenen studentischen Vorträge ist zudem gewährleistet, dass beide Vergleichskorpora größtenteils monologische Daten umfassen und daher als homogen angesehen werden können.

[11] Die Angaben zur Gesamtdauer des Korpus beziehen sich auf die kompletten Kommunikationsereignisse. Diese umfassen sowohl den eigentlichen Vortrag mit den Sprecherbeiträgen der Vortragenden als auch dessen Einleitung durch die Seminarleiter/innen sowie die Diskussion zum Vortrag.

[12] Die Angaben zur Korpusgröße beziehen sich ausschließlich auf die Sprecherbeiträge der Vortragenden. Herausgerechnet wurden dabei die Notation der Pausen, bspw. *(0.5)* oder *(.)* sowie sprachbegleitende para- und außersprachliche Handlungen und Ereignisse wie *((lacht))* oder *((räuspert sich))*.

Unterschiede zwischen den zu analysierenden Korpora ergaben sich vor allem hinsichtlich des Parameters *Grad der Mündlichkeit* – die Vorträge im Korpus GEWISS-DEU-L1-SV wurden i.d.R. frei gehalten, während diese im Korpus GEWISS-DEU-L2-BG-SV i.d.R. zum Teil von einer schriftlichen Vorlage abgelesen wurden. Diese Unterschiede sind in der Authentizität der erhobenen Daten begründet: Beide Korpora umfassen Aufnahmen studentischer Referate als Teil authentischer Lehrveranstaltungen aus dem deutschen bzw. bulgarischen akademischen Kontext, weshalb die allgemeine kommunikative Struktur der aufgenommenen Vorträge (etwa die Verteilung der Sprecherbeiträge der Vortragenden, Seminarleiter/inne/n und Diskutant/inn/en, das Vorhandensein und Frequenz der diskursiven Phasen etc.)[13], deren Grad der Mündlichkeit oder der Medieneinsatz darin ebenfalls nicht vorhergesehen bzw. kontrolliert werden konnte. Den Unterschieden hinsichtlich des Faktors *Grad der Mündlichkeit* ist bei der Interpretation der Korpusergebnisse durchaus Rechnung zu tragen. Dennoch beeinträchtigen sie nicht die allgemeine Vergleichbarkeit der zu analysierenden Korpora, die durch das einheitliche Korpusdesign und die vergleichbare Korpusgröße hinreichend gewährleistet ist.

4.2 Häufigkeitsverteilung der Konnektoren in studentischen Vorträgen von L1- und L2-Sprecher/inne/n des Deutschen

Als Grundlage für die Auswahl der zu untersuchenden Konnektoren in den hier zugrunde gelegten Teilen des Spezialkorpus GeWiss diente die frequenzbasierte Liste der Konnektoren unter den 1000 häufigsten Wörtern im mündlichen Teil des Referenzkorpus Herder/BYU, die als Ergebnis der Studie 1 ermittelt wurde (vgl. Abschnitt 3.2). Um die Verwendungspräferenzen beim Konnektorengebrauch von L1- und L2-Sprecher/inne/n des Deutschen zu ermitteln, wurden in einem ersten Schritt die Häufigkeitsverteilungen der 76 Konnektorkandidaten der Herder-BYU-Referenzliste in den Wortlisten der Vergleichskorpora[14] GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV ermittelt, welche mithilfe des EXMARaLDA Analyse- und Konkordanztools (EXAKT) (vgl. Schmidt/Wörner 2009) generiert wurden.[15]

---

[13] Zur kommunikativen Struktur studentischer Referate vgl. Guckelsberger (2006).

[14] Genau wie die Angaben zur Korpusgröße wurden auch die Wortlisten beider Korpora ausschließlich auf der Grundlage der Sprecherbeiträge der Vortragenden erstellt sowie um die Notation der Pausen, bspw. *(0.5)* oder *(.)*, sowie sprachbegleitender para- und außersprachlicher Handlungen und Ereignisse wie *((lacht))* oder *((räuspert sich))* bereinigt.

[15] Im Unterschied zum Herder/BYU-Korpus sind die Korpora GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV nicht nach Wortarten annotiert, was eine automatische grammatische Disambiguierung der Frequenzlisten anhand von Wortartentags sowie eine darauf basierende Ermittlung konnektoraler Verwendungen nicht erlaubt. Die in Tabelle 4 ausgewiesenen Häufigkeiten stellen also reine Tokenfrequenzen dar und können bei polysemen Formativen wie *ja*, *als* oder *während* sowohl konnektorale als auch nicht-konnektorale Verwendungen umfassen bzw. die Frequenz aller konnektoraler Lesarten bei Formativen, die verschiedene Lesarten zulassen (etwa bei *da* als *ADVB* und *KONJ* oder bei *damit* als *PADV* und *KONJ*). Da beide Vergleichskorpora in dieser Hinsicht gleich aufbereitet sind, sind die daraus generierten Frequenzlisten der Konnektorenkandidaten vergleichbar und können quantitativ ausgewertet werden. Für Folgeuntersuchungen basierend auf den GeWiss-Korpora, die speziell solche polyseme Formative betreffen, müsste in einem qualitativen Analyseschritt eine entsprechende Disambiguierung anhand von

In den beiden ausgewerteten GeWiss-Korpora kommen nicht alle Konnektoren aus der Referenzliste vor (vgl. die Angaben zur absoluten Frequenz in Tabelle 4). So finden sich im Korpus GEWISS-DEU-L1-SV 69 der Konnektoren der Referenzliste. Nicht enthalten sind darin *damals* und *inzwischen*. Deutlich weniger Konnektoren aus der Referenzliste enthält das Korpus mit nichtmuttersprachlichen Daten GEWISS-DEU-L2-BG-SV – lediglich 46 oder knapp 65% der Referenzliste, was auf eine viel geringere Varianz der Sprachmittel bei den L2-Sprecher/inne/n hindeutet. Zudem kommen hier mehr Formative nur einmal im Gesamtkorpus vor – insgesamt 8 Formative (*erstmals*, *doch*, *dadurch*, *nachher*, *besonders*, *während*, *obwohl*, *nun*), während das Korpus GEWISS-DEU-L1-SV lediglich zwei solche Hapax Legomena aufweist (*jedenfalls*, *sofort*).

Im Folgenden werden die Frequenzunterschiede in den Verteilungen der Konnektoren in beiden Vergleichskorpora näher beschrieben.

Die Häufigkeitsverteilungen von sprachlichen Elementen unterscheiden sich naturgemäß in Abhängigkeit von der Größe des zugrunde liegenden Korpus – in einem größeren Korpus ist prinzipiell mit einem häufigeren Vorkommen eines sprachlichen Elements zu rechnen als in einem kleineren. Daher ist bei Korpusvergleichen neben der absoluten auch die relative Frequenz der untersuchten Elemente von Bedeutung. Daneben stellt sich bei einem Korpusvergleich auch die Frage, ob unter Berücksichtigung der Korpusgröße die beobachteten Frequenzunterschiede in den Vergleichskorpora nicht zufällig zustande gekommen sind, d. h. ob die Korrelation zwischen der Frequenz der Elemente und der jeweiligen Korpusgröße signifikant ist. Dies kann mit einem Signifikanztest durch Messung des Abstands zwischen den in Abhängigkeit von der Korpusgröße erwarteten und den empirisch tatsächlich beobachteten Frequenzen geprüft werden. (vgl. Bubenhofer 2006-2018).

Aus diesen Gründen wurden die ermittelten Häufigkeitsverteilungen der Konnektoren in den beiden Vergleichskorpora GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV in einem nächsten Analyseschritt auf Signifikanz geprüft, um diejenigen Elemente zu identifizieren, in denen sich die Korpora statistisch nachweislich voneinander unterscheiden. Aufgrund der beobachteten niedrigen Frequenzen in den zu untersuchenden Korpora wurde dafür in Anlehnung an Rayson/Garside (2000)[16] der Log-likelihood-Test als geeignetes statistisches Verfahren ausgewählt, welcher auch unter solchen Bedingungen zuverlässige Werte ermittelt. Der Log-likelihood-Wert (LL) wurde für jeden Konnektor mit dem Log-likelihood-Kalkulator von Paul Rayson[17] berechnet und anschließend auf Signifikanz geprüft. Die Ergebnisse dieser Auswertung finden sich in Tabelle 4.

Der Signifikanztest hat ergeben, dass bei insgesamt 35 Konnektoren die Frequenzunterschiede in den Vergleichskorpora statistisch relevant sind, d. h., man kann

---

Konkordanzanalysen erfolgen, um die entsprechenden konnektoralen Verwendungen zu ermitteln.

[16] Der Überblick der Diskussion zur Reliabilität statistischer Verfahren bei Korpus-basierten Textanalysen in Rayson/Garside (2000) zeigt, dass Pearsons Chi-Quadrat-Signifikanztest bei erwarteten Häufigkeiten kleiner als 5 unzuverlässige Ergebnisse liefert und dass Dunnings Log-likelihood-Test eine bessere Alternative für kleinere Textkorpora darstellt.

[17] Vgl. http://ucrel.lancs.ac.uk/llwizard.html. [Stand: 24.01.2018]

mit einer Wahrscheinlichkeit von mindestens 95% davon ausgehen, dass die beobachteten Unterschiede nicht zufällig sind.[18] Darunter befinden sich sowohl hochfrequentierte Formative wie *dann, auch, und, also* als auch solche im Niedrigfrequenzbereich, etwa *deshalb, insofern, daher, übrigens*.

Tabelle 4: Signifikanzunterschiede in den Verteilungen der Konnektoren in den Vergleichskorpora[19]

LEGENDE: LL – LOG-LIKELIHOOD-WERT, P – IRRTUMSWAHRSCHEINLICHKEIT (PROBABILITY)

Korpus	**GEWISS-DEU-L1-SV**		**GEWISS-DEU-L2-BG-SV**		**Signifikanztest**	
Wort	**absolute Freqenz**	**relative Freqenz[20]**	**absolute Freqenz**	**relative Freqenz**	**LL**	**p**
DANN	439	112	82	26	191,91	< 0,0001
AUCH	604	154	219	71	108,16	< 0,0001
NOCH	231	59	47	15	93,32	< 0,0001
SO	396	101	130	42	85,65	< 0,0001
DA	231	59	63	20	66,61	< 0,0001
JA	377	96	173	56	37,13	< 0,0001
ZWAR	49	12	4	1	35,26	< 0,0001
UND	1100	280	1116	359	34,43	< 0,0001
DENN	56	14	7	2	32,73	< 0,0001
ERSTMAL	38	10	2	1	31,66	< 0,0001
ALSO	679	173	393	127	24,27	< 0,0001
DOCH	28	7	2	1	21,19	< 0,0001
ALS	131	33	169	54	17,90	< 0,0001
SCHON	139	35	61	20	15,66	< 0,0001
DESHALB	5	1	21	7	14,71	< 0,001
ZUMINDEST	12	3	0	0	13,98	< 0,001
WIEDER	41	10	10	3	13,63	< 0,001
GAR	19	5	2	1	12,20	< 0,001
NACHHER	13	3	1	0	9,58	< 0,01

[18] Bei den Signifikanztests werden vier Signifikanzniveaus unterschieden. Bei dem niedrigsten Signifikanzniveau beträgt die Irrtumswahrscheinlichkeit des statistischen Tests $p < 0,05$ d. h., es kann mit einer 95 %-igen Wahrscheinlichkeit davon ausgegangen werden, dass die beobachteten Frequenzunterschiede in den Korpora nicht zufällig sind; bei dem höchsten Signifikanzniveau ($p < 0.0001$) entsprechend mit einer 99,99%-igen Wahrscheinlichkeit. Die vier Signifikanzniveaus werden aufgrund kritischer Werte ermittelt. Diese betragen für die vorliegende Untersuchung: $p < 0,05$ ($H_0 = 5\%$) – kritischer Wert LL = 3,84; $p < 0.01$ ($H_0 = 1\%$) – kritischer Wert LL = 6,63; $p < 0.001$ ($H_0 = 0.1\%$) – kritischer Wert LL = 10,83; $p < 0.0001$ ($H_0 = 0.01\%$) kritischer Wert LL = 15,13.

[19] Im Vergleich zur Referenzliste (vgl. Tabelle 2) enthält diese Liste 71 Konnektoren. Dies ist damit zu erklären, dass die polysemen Formative *da*, *aber*, *doch*, *denn* und *damit*, die jeweils zwei konnektorale Lesarten zulassen (etwa als ADVB und KONJ bei *da* oder als PADV und KONJ bei *damit*) jeweils nur einmal gelistet sind, da eine automatische Disambiguierung der Korpora GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV aufgrund fehlender Wortartenannotation nicht möglich ist.

[20] Die relative Frequenz bezieht sich auf 10.000 Tokens.

ZUERST	8	2	21	7	9,50	< 0,01
ZUNÄCHST	16	4	2	1	9,35	< 0,01
ANSONSTEN	8	2	0	0	9,32	< 0,01
DAZU	60	15	25	8	7,79	< 0,01
ALLERDINGS	11	3	1	0	7,57	< 0,01
DADURCH	22	6	6	2	6,34	< 0,05
WEIL	111	28	60	19	5,81	< 0,05
ODER	267	68	167	54	5,72	< 0,05
DRAUF	14	4	3	1	5,37	< 0,05
NUR	109	28	60	19	5,22	< 0,05
DARAUF	18	5	5	2	5,06	< 0,05
ABER	214	54	133	43	4,82	< 0,05
INSOFERN	4	1	0	0	4,66	< 0,05
DAHER	4	1	0	0	4,66	< 0,05
ÜBRIGENS	4	1	0	0	4,66	< 0,05
DAFÜR	15	4	4	1	4,46	< 0,05
BLOß	7	2	1	0	3,76	
NÄMLICH	19	5	7	2	3,29	
WENN	156	40	98	32	3,23	
KAUM	6	2	1	0	2,88	
EINMAL	31	8	15	5	2,56	
BESONDERS	12	3	4	1	2,53	
SOWIESO	2	1	0	0	2,33	
MITTLERWEILE	5	1	1	0	2,05	
SONST	7	2	2	1	1,89	
DESWEGEN	29	7	15	5	1,85	
NUN	3	1	6	2	1,85	
TROTZDEM	16	4	7	2	1,82	
ETWA	4	1	7	2	1,69	
DABEI	27	7	15	5	1,23	
SOWEIT	1	0	0	0	1,16	
EIGENTLICH	92	23	85	27	1,07	
VORHER	9	2	4	1	0,98	
WÄHREND	8	2	10	3	0,94	
SPÄTER	7	2	3	1	0,84	
NACHDEM	5	1	2	1	0,72	
WOBEI	4	1	5	2	0,47	
ENTWEDER	3	1	4	1	0,47	
SONDERN	43	11	39	13	0,39	
GENAUSO	4	1	2	1	0,29	

AUßERDEM	7	2	7	2	0,19
SOGAR	8	2	5	2	0,17
OBWOHL	5	1	3	1	0,15
BEVOR	5	1	5	2	0,14
DAMIT	45	11	33	11	0,11
DANACH	10	3	9	3	0,08
MINDESTENS	2	1	2	1	0,06
BEZIEHUNGSWEISE	13	3	11	4	0,03
SOFORT	1	0	1	0	0,03
SELBST	27	7	22	7	0,01
DAMALS	0	0	0	0	0,00
INZWISCHEN	0	0	0	0	0,00

Bezüglich der syntaktischen Kategorienzugehörigkeit dominieren unter den 35 Konnektoren mit signifikanten Häufigkeitsunterschieden in den Vergleichskorpora ähnlich wie in der Referenzliste der Konnektoren aus dem Herder/BYU-Korpus (vgl. Abschnitt 3.2) Formative, die im Handbuch der deutschen Konnektoren (Pasch et al. 2003) als Adverbkonnektoren klassifiziert wurden – insgesamt 26 Kandidaten, die 55% der Adverbkonnektoren in der Referenzliste entsprechen. Weiterhin sind unter den 35 auch 2 der insgesamt 3 Konjunktoren aus der Referenzliste vertreten (*und*, *oder*), 2 der 6 Subjunktoren aus der Referenzliste (*weil*, *als*), sowie 3 Formative, die nach Pasch et al. (2003) potentiell eine Lesart als Adverbkonnektor und Konjunktor zulassen (*ja*, *denn*, *aber*), und zwei, die als Adverbkonnektor oder Subjunktor fungieren können (*so*, *da*).

Die aufgrund des Log-likelihood-Tests ermittelten 35 Konnektoren mit statistisch relevanten Frequenzunterschieden in den Vergleichskorpora GEWISS-DEU-L1-SV und GEWISS-DEU-L2-BG-SV wurden schließlich im Hinblick auf ihre „Lernschwierigkeit" für Nichtmuttersprachler/innen untersucht.

5 Studie 3: Hinweise auf Lernschwierigkeiten bei der Konnektorenverwendung in den L2-Korpusdaten

Aufschlüsse über Lernschwierigkeiten in Korpusdaten können sogenannte Overuse/Underuse-Daten geben (vgl. Zeldes et al. 2008). Diese stellen einen Vergleich der relativen Häufigkeiten bestimmter sprachlicher Elemente eines L2-Korpus mit deren relativen Häufigkeiten in einem L1-Korpus als Referenz dar. Der ermittelte relative Übergebrauch bzw. Mindergebrauch kann dabei mithilfe des Underuse/Overuse Add-In (V1-2) von Amir Zeldes[21] visualisiert werden, indem der Grad der Abweichung von den muttersprachlichen Häufigkeiten in zunehmend warmen oder kalten Tönen dargestellt wird Tabelle 5 zeigt den relativen Über- bzw. Mindergebrauch der Konnektoren im nichtmuttersprachlichen Korpus GEWISS-DEU-L2-BG-SV im Vergleich zu den muttersprachlichen Daten des Korpus GEWISS-DEU-L1-SV. In Dunkelblau wird dabei

[21] Vgl. https://github.com/amir-zeldes/XLAddIns/blob/master/UnderOverUse_V1.2.xla. [Stand 24.01.2018]

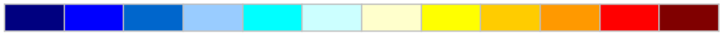
ein starker Mindergebrauch in den L2-Daten dargestellt – die relative Häufigkeit der entsprechenden Elemente beträgt in diesen Fällen lediglich ein Drittel oder weniger der relativen Häufigkeit im L1-Korpus. Analog dazu wird ein mindestens dreifacher Übergebrauch der relativen Häufigkeit im L1-Korpus in Dunkelrot dargestellt.

Tabelle 5: Relativer Über- bzw. Mindergebrauch der Konnektoren mit signifikanten Frequenzunterschieden im L2-Korpus GEWISS-DEU-L2-BG-SV im Vergleich zum L1-Korpus GEWISS-DEU-L1-SV

Korpus	**GEWISS-DEU-L1-SV**		**GEWISS-DEU-L2-BG-SV**	
Wort	**absolute Frequenz**	**relative Frequenz**[22]	**absolute Frequenz**	**relative Frequenz**
DANN	439	112	82	26
AUCH	604	154	219	71
NOCH	231	59	47	15
SO	396	101	130	42
DA	231	59	63	20
JA	377	96	173	56
ZWAR	49	12	4	1
UND	1100	280	1116	359
DENN	56	14	7	2
ERSTMAL	38	10	2	1
ALSO	679	173	393	127
DOCH	28	7	2	1
ALS	131	33	169	54
SCHON	139	35	61	20
DESHALB	5	1	21	7
ZUMINDEST	12	3	0	0
WIEDER	41	10	10	3
GAR	19	5	2	1
NACHHER	13	3	1	0
ZUERST	8	2	21	7
ZUNÄCHST	16	4	2	1
ANSONSTEN	8	2	0	0
DAZU	60	15	25	8
ALLERDINGS	11	3	1	0
DADURCH	22	6	6	2
WEIL	111	28	60	19
ODER	267	68	167	54
DRAUF	14	4	3	1
NUR	109	28	60	19

[22] Die relative Frequenz bezieht sich auf 10.000 Tokens.

DARAUF	18	5	5	2
ABER	214	54	133	43
INSOFERN	4	1	0	0
DAHER	4	1	0	0
ÜBRIGENS	4	1	0	0
DAFÜR	15	4	4	1



Underuse **Overuse**

Die Visualisierung der vorliegenden Overuse/Underuse-Daten der beiden Vergleichskorpora zeigt eindeutig, dass die überwiegende Mehrzahl (89%) der 35 Konnektorkandidaten mit statistisch relevanten Frequenzunterschieden in den Vergleichskorpora in den L2-Daten deutlich unterrepräsentiert sind. Dies deutet darauf hin, dass Konnektoren insgesamt einen schwierigen Lerngegenstand darstellen und selbst von fortgeschrittenen DaF-Lerner/inne/n mit L1 Bulgarisch generell „vermieden" werden.

Den stärksten Mindergebrauch weisen die Adverbkonnektoren *zumindest, ansonsten, insofern, daher* und *übrigens* auf. Sie alle werden von den nichtmuttersprachlichen Sprecher/inne/n des Korpus GEWISS-DEU-L2-BG-SV kein einziges Mal verwendet, obwohl von den L1-Sprecher/inne/n des Korpus GEWISS-DEU-L1-SV absolute Frequenzen von 4 bis 12 vorliegen. Hingegen zeigt sich bei den Formativen *als, deshalb* und *zuerst* ein starker Übergebrauch in den L2-Daten.

Beim näheren Betrachten der Semantik einzelner Formative fällt auf, dass L2-Sprecher/innen präferierte sprachliche Mittel zum Ausdruck bestimmter semantischer Relationen einsetzen, die sie im Vergleich zu den muttersprachlichen Sprecher/inne/n statistisch überproportional verwenden. Andere semantisch ähnliche Sprachmittel werden gleichzeitig deutlich mindergebraucht. So wird beispielsweise der konsekutive Konnektor *deshalb* in den L2-Daten stark übergebraucht, während andere semantisch verwandte Formative wie *also, daher, da, denn* oder *weil* deutlich seltener als im L1-Korpus oder gar nicht vorkommen. Diese empirisch ermittelten Diskrepanzen lassen sich nicht mit dem Sprachniveau der L2-Sprecher/innen im Korpus GEWISS-DEU-L2-BG-SV erklären, da die Konnektoren als Lerngegenstand bis zum B1-Niveaus nach dem GeR bereits eingeführt sind und die Sprachkompetenz der Probanden im Deutschen mindestens auf dem Niveau B2/C1 lag (vgl. Abschnitt 4.1.1).

Frühere empirische Untersuchungen zur Konnektorenverwendung auf der Grundlage des GeWiss-Korpus (vgl. Fandrych et al. 2014a zu *also* in studentischen Vorträgen bzw. Slavcheva/Meißner 2014 zu *also* und *so* in wissenschaftlichen Vorträgen) haben gezeigt, dass in der gesprochenen Wissenschaftssprache konzeptionell mündliche Verwendungen klar dominieren. Ausgehend von diesen Befunden lässt es sich auch für die Ergebnisse der vorliegenden Untersuchung vermuten, dass nicht die traditionellerweise als prototypisch angesehenen rein konnektiven Verwendungen dieser Formative, sondern vielmehr deren spezifischen gesprochensprachlichen Verwendungen, die von den L2-Sprecher/inne/n womöglich noch nicht sicher beherrscht werden, ursächlich für die

beobachteten Frequenzunterschiede in beiden Vergleichskorpora sind. Diese und ähnliche Fragen gilt es in qualitativen Folgeuntersuchungen noch detailliert zu prüfen.

6 Zusammenfassung und Ausblick

Die vorliegende Untersuchung stellt einerseits eine Methode zur Aufdeckung von Lernschwierigkeiten bei Nichtmuttersprachler/inne/n des Deutschen auf der Grundlage von Korpusvergleichen vor, die für ähnlich geartete Untersuchungen als Modell dienen kann. Gleichzeitig gibt sie umfassende empirisch fundierte Einblicke in die Verwendungspräferenzen der Konnektoren in der mündlichen Wissenschaftskommunikation bei L1- und L2-Sprecher/inne/n des Deutschen. Die quantitativen Untersuchungsergebnisse zeigen eindeutig, dass Konnektoren in den L2-Daten im Vergleich zu den muttersprachlichen Sprecher/inne/n deutlich unterrepräsentiert sind. Dies deutet darauf hin, dass selbst fortgeschrittene L2-Sprecher/innen des Deutschen eine viel geringere Varianz dieser Sprachmittel aufweisen und diese offenbar generell „vermeiden". Für den Ausdruck bestimmter semantischer Relationen scheinen sich bei ihnen zudem präferierte Formelemente mit starkem Übergebrauch im Vergleich zu Muttersprachler/inne/n herauszubilden.

Erklärungen für den beobachteten Minder- bzw. in vereinzelten Fällen auch Übergebrauch können Gegenstand qualitativer Folgeuntersuchungen zur kontrastiven Beschreibung des Konnektorengebrauchs bei L2-Lerner/inne/n des Deutschen bilden. Derartige empirisch fundierte Beschreibungen der gesprochenen Wissenschaftssprache, insbesondere unter der Berücksichtigung der spezifischen Aspekte der Mündlichkeit, stellen eine Grundvoraussetzung für ihre Vermittlung an muttersprachliche wie nichtmuttersprachliche Studierende dar und werden daher noch dringend gebraucht.

Literatur

Bachman, Lyle F, Adrian S Palmer (1996), *Language testing in practice*: *Designing and developing useful languages tests*. Oxford: Oxford Univ. Press. (=*Oxford Applied Linguistics*).

Brinker, Klaus (2006), "Vorstellung eines textlinguistischen Beschreibungsmodells als Basis des DaF-Unterrichts", in Foschi Albert, Marinne Hepp & Eva Neuland (Hrsg.), *Texte in Sprachforschung und Sprachunterricht*. *Pisaner Fachtagung 2004 zu neuen Wegen der italienisch-deutschen Kooperation*. München: Iudicium, 75–82.

Bubenhofer, Noah (2006-2018), *Einführung in die Korpuslinguistik*: *Praktische Grundlagen und Werkzeuge*. http://www.bubenhofer.com/korpuslinguistik/. [Stand 10.03.2018].

[Curriculum 2007] = Sofia University "St. Kliment Ohridsky" (2007): *Curriculum - Jahrgang 2007-2011*. *Studiengang: Deutsche Philologie*.

[DAAD/DZHW 2017] = Deutscher Akademischer Austauschdienst/Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (Hrsg.) (2017), *Wissenschaft Weltoffen 2017*. *Daten und Fakten zur Internationalität von Studium und Forschung in Deutschland*. Fokus: Akademische Mobilität. Bielefeld: Bertelsmann, W.

Ellis, Nick C. (2002). "Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition", *Studies in Second Language Acquisition* 24:143–188.

Fandrych, Christian, Erwin Tschirner, Cordula Meißner, Stefan Rahn & Adriana Slavcheva (2009), "Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen. Vorstellung eines gemeinsamen Forschungsvorhabens", in Błachut, Edyta, Lesław Cirko & Artur Tworek (Hrsg.), *Studia linguistica XXVIII*. Wrocław: Wydawnictwo Uniw. Wroclawskiego, 7–30.

Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (2012), "The GeWiss Corpus: Comparing Spoken Academic German, English and Polish", in Schmidt, Thomas & Kai Wörner (Hrsg.), *Multilingual corpora and multilingual corpus analysis*. Amsterdam: John Benjamins, 319–337.

Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (2014a), "Das Korpusprojekt 'Gesprochene Wissenschaftssprache kontrastiv' und seine Relevanz für die Vermittlung des Deutschen als Wissenschaftssprache", in Mackus, Nicole & Jupp Möhring (Hrsg.), *Wege für Bildung, Beruf und Gesellschaft - mit Deutsch als Fremd- und Zweitsprache. 38. Jahrestagung des Fachverberbandes Deutsch als Fremdsprache an der Universität Leipzig 2011*. Göttingen: Univ.-Verl. Göttingen, 141–160.

Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (Hrsg.) (2014b), *Gesprochene Wissenschaftssprache. Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron. (=*Wissenschaftskommunikation* 9).

Guckelsberger, Susanne (2006), "Zur kommunikativen Struktur von mündlichen Referaten in universitären Lehrveranstaltungen", in Ehlich, Konrad & Dorothee Heller (Hrsg.), *Die Wissenschaft und ihre Sprachen*. Bern: Lang, 147–173.

Jones, Randall L. (2004), "Corpus-based Word Frequency Analysis and the Teaching of German Vocabulary", *Fremdsprache Lehren und Lernen* 33:165–175.

Jones, Randall L. & Erwin Tschirner (2006), *A frequency dictionary of German*: *Core vocabulary for learners*. London: Routledge. (=*Routledge frequency dictionaries*).

Koch, Peter & Wulf Oesterreicher (1985) "Sprache der Nähe - Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte", *Romanistisches Jahrbuch* 36:15–43.

Meißner, Cordula & Adriana Slavcheva (2014), "Das GeWiss-Korpus - ein Vergleichskorpus der gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen. Design und Aufbau", in Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (Hrsg.), *Gesprochene Wissenschaftssprache*: *Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 15–38.

Pasch, Renate, Ursula Brauße & Eva Breindl (2003), *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*. Berlin [u.a.]: de Gruyter. (=*Schriften des Instituts für Deutsche Sprache* 9).

Rayson, Paul & Roger Garside (2000), *Comparing Corpora using Frequency Profiling*. http://www.comp.lancs.ac.uk/~paul/publications/rg_acl2000.pdf (19.09.2011).

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen, (1999), *Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset)*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf. [Stand 19.09.2011].

Schmid, Helmut (1995), "Improvements in Part-of-Speech Tagging with an Application to German" in *Proceedings of the ACL SIGDAT-Workshop*. Dublin.

Schmidt, Thomas & Kai Wörner (2009) "EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research", *Pragmatics* 19:565–582.

Scott, Mike (2004), *WordSmith Tools version 4*. Oxford: Oxford University Press.

Slavcheva, Adriana & Cordula Meißner (2014), "Also und so in wissenschaftlichen Vorträgen", in Fandrych, Christian, Cordula Meißner & Adriana Slavcheva (Hrsg.), *Gesprochene Wissenschaftssprache*: *Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 113–131.

Statistisches Bundesamt (2018), *Studierende*. *Studienfach Germanistik/Deutsch*. https://www.destatis.de/DE/ZahlenFakten/Indikatoren/LangeReihen/Bildung/lrbil04.html. [Stand 26.02.2018].

Tschirner, Erwin (2005), "Korpora, Häufigkeitslisten, Wortschatzerwerb", in Heine, Antje, Mathilde Hennig & Erwin Tschirner (Hrsg.), *Deutsch als Fremdsprache - Konturen und Perspektiven eines Faches*. Festschrift für Barbara Wotjak zum 65. Geburtstag. München: Iudicium, 133–149.

Tschirner, Erwin (2008), "Das professionelle Wortschatzminimum im Deutschen als Fremdsprache", *Deutsch als Fremdsprache* 4(45):195–208.

Zeldes, Amir, Anke Lüdeling & Hagen Hirschmann (2008), *What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data*. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/zeldes-et-al-1.pdf. [Stand 26.02.2018].