Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 1 (Juli 2023)
ISSN 2749-4411

GEWISS - EIN KORPUS DER GESPROCHENEN WISSENSCHAFTSSPRACHE

Franziska Wallner
Herder-Institut der Universität Leipzig

Abstract

GeWiss ist ein mehrsprachiges Vergleichskorpus der gesprochenen Wissenschaftssprache, das sich aus studentischen Referaten, Expertenvorträgen und Prüfungsgesprächen zusammensetzt. Im Beitrag werden mit besonderem Fokus auf die deutschsprachigen Daten die Struktur und Aufbereitung des Korpus erläutert. Zudem wird auf die verschiedenen Zugangswege zu *GeWiss* und auf Nutzungsmöglichkeiten eingegangen.

Keywords: gesprochene Wissenschaftssprache; *GeWiss*; Korpora in DaF/DaZ

Abstract

GeWiss is a multilingual comparative corpus of spoken academic language consisting of student presentations, research presentations and oral examinations. The paper explains the structure and further elaboration of the corpus with a special focus on the German language data. It also discusses the different ways of accessing *GeWiss* and its possible uses.

Keywords: spoken academic language; *GeWiss*; corpora in GFL/GSL

Das *GeWiss*-Korpus: Primär- und Metadaten

Das *GeWiss*-Korpus ist ein Vergleichskorpus der gesprochenen Wissenschaftssprache und beinhaltet Prüfungsgespräche, studentische Referate und Expertenvorträge in deutscher, englischer und polnischer Sprache. Zusätzlich sind italienischsprachige Expertenvorträge enthalten. Die Daten stammen von Sprecher:innen, die die jeweiligen Sprachen als Erstsprache (L1) sprechen; für das Deutsche und das Englische liegen daneben auch L2-Daten vor (vgl. hierzu ausführlich Fandrych / Wallner 2022). Erhoben wurden die Daten in authentischen akademischen Kommunikationssituationen in philologischen Fächern an Standorten in Großbritannien, Polen, Bulgarien, Italien, Finnland und Deutschland.

Mit dem Korpus werden die Audioaufnahmen zu den einzelnen Sprechereignissen sowie die dazugehörigen Transkriptionen bereitgestellt. Insgesamt umfasst das Korpus 1.205.306 Token und 146 Aufnahmestunden. Die deutschsprachigen Daten bilden mit 742.332 Token und ca. 92 Aufnahmestunden den Großteil des *GeWiss*-Korpus. Tab. 1 gibt einen Überblick über die Anzahl der im gesamten sowie im deutschsprachigen *GeWiss*-Korpus vertretenen Genres und die jeweils zugehörigen Tokenzahlen[1]:

[1] Die hier angegebenen Tokenzahlen wurden mit dem Tool *ZuRecht* (vgl. Frick / Helmer /Wallner 2023 in dieser Themenausgabe) ermittelt und umfassen ausschließlich die rein sprachlichen Token (also keine Pausen und andere nichtsprachliche transkribierte Einheiten).

Genre	*GeWiss* gesamt		*GeWiss* deutsch	
	Sprechereignisse	**Token**	**Sprechereignisse**	**Token**
Expertenvorträge	76	374.380	33	166.639
Studentische Vorträge	137	328.533	106	240.935
Prüfungsgespräche	223	502.393	137	334.800
gesamt	**436**	**1.205.306**	**276**	**742.374**

Tabelle 1
Überblick über die Genres im *GeWiss*-Korpus

Daneben liegen für das Korpus vielfältige Metadaten vor. Diese umfassen neben personenbezogenen Metadaten zu den Sprecher:innen (wie bspw. Alter, Geschlecht, Ausbildung, Sprecher:innenrolle) auch sprachbiografische Informationen wie etwa Erst- und Fremdsprachen, Auslandsaufenthalte und die jeweilige Aufenthaltsdauer im Ausland. Zudem werden umfassende Metadaten zum Sprechereignis bereitgestellt. Hierzu zählen neben Genre, Aufnahmeort und Sprache auch Informationen zum Setting (bspw. Beziehung der Sprecher:innen untereinander) und zum Grad der Mündlichkeit (frei gesprochen oder (teilweise) abgelesen).

Transkription und Aufbereitung

Die Audioaufnahmen des *GeWiss*-Korpus wurden aussprachenah transkribiert. Bei den deutschsprachigen Aufnahmen erfolgte dies zunächst nach dem GAT2 Minimaltranskript (vgl. Selting et al. 2009). Für die spätere korpuslinguistische Aufbereitung wurden die Transkripte an die Konventionen des cGAT (vgl. Schmidt / Schütte / Winterscheidt 2015) angepasst. Für jedes Transkript wurde zudem eine orthografisch normalisierte Fassung erstellt. Diese bildete die Grundlage für die Annotation von Wortarten und die Lemmatisierung. Hierfür wurde das um gesprochensprachliche Kategorien erweiterte Stuttgart-Tübingen-Tagset (STTS 2.0) (vgl. Westpfahl et al. 2017) genutzt. Neben diesen tokenbasierten Annotationen wurden in ausgewählten Teilkorpora auch tokenübergreifende Annotationen vorgenommen. Dabei handelt es sich um Sprachwechselphänomene[2], Diskurskommentierungen[3] sowie um Verweise und Zitate[4].

Zugriffsmöglichkeiten

Für die *GeWiss*-Daten existieren verschiedene Zugriffsmöglichkeiten, die in ihren Funktionalitäten sehr heterogen sind und unterschiedliche Nutzungsmöglichkeiten bieten. Zunächst wurden die *GeWiss*-Daten (mit Ausnahme der erst später aufgenommenen Daten aus Finnland) über das *GeWiss*-

[2] Annotiert wurden Wechsel von einer Sprache in eine andere Sprache innerhalb eines Sprechereignisses. Dies kann sowohl Einzeltoken als auch mehrere aufeinander folgende Token betreffen. Annotiert wurden die Sprachwechsel in den in Deutschland, Großbritannien, Polen und Bulgarien erhobenen Teilkorpora (vgl. Reershemius / Lange 2014).
[3] Diskurskommentierungen sind zentrale wissenschaftssprachliche Handlungen, die in wissenschaftlichen Vorträgen der Gliederung und Rezipientenorientierung dienen (vgl. Fandrych 2014) und in den deutschsprachigen L1-Expertenvorträgen des *GeWiss*-Korpus annotiert wurden (vgl. Baur et al. 2014).
[4] Zitate und Verweise beinhalten mündliche Bezugnahmen auf andere Forschungsarbeiten und wurden in den deutschsprachigen L1-Expertenvorträgen und in ausgewählten studentischen Vorträgen des *GeWiss*-Korpus mit Deutsch als L1 und L2 annotiert (vgl. Sadowski 2017).

Portal[5] zugänglich gemacht. 2017 erfolgte eine Integration der deutschsprachigen *GeWiss*-Daten in die Datenbank für gesprochenes Deutsch (DGD)[6]. Das *GeWiss*-Portal gestattet einen Zugriff auf die Audioaufnahmen und die aussprachenahen Transkriptionen. Zudem gibt es die Möglichkeit, Konkordanzsuchen durchzuführen. Dabei lassen sich die verschiedenen Teilkorpora des *GeWiss*-Korpus direkt ansteuern (bspw. alle Expertenvorträge mit Deutsch als L2). In Ergänzung dazu können weitere metadatenbezogene Filter aktiviert werden. Diese umfassen neben allgemeinen Informationen zum Sprechereignis (Genre, akademischer Kontext) auch detaillierte Informationen zu den Sprecher:innen inklusive sprachbiografischer Informationen (bspw. Erstsprache(n), Auslandsaufenthalte) sowie Informationen zum Setting (wie Grad der Mündlichkeit). Darüber hinaus sind über das *GeWiss*-Portal auch die Annotationen von Sprachwechseln, Diskurskommentierungen, Verweisen und Zitaten zugänglich. Tokenbasierte Annotationen wie die orthografische Normalisierung und das POS-Tagging können hingegen nicht abgerufen werden.

Die Nutzungsmöglichkeiten über die DGD sind aufgrund der dort bereits länger etablierten Datenstruktur etwas anders, aber zugleich auch deutlich umfassender als über das *GeWiss*-Portal. Neben den aussprachenahen Transkriptionen werden in der DGD auch die orthografisch normalisierten Fassungen der Transkripte sowie das POS-Tagging bereitgestellt und können bei der Konkordanzsuche mit einbezogenen werden. Aufgrund der Beschaffenheit des in der DGD implementierten Metadatenschemas ist es jedoch nicht möglich, auf alle mit dem *GeWiss*-Korpus verbundenen Metadaten zuzugreifen. So sind bspw. der Grad der Mündlichkeit oder auch sprachbiografische Informationen zu den Sprecher:innen über die DGD nicht vollständig zugänglich. Auch die *GeWiss*-Teilkorpora (bspw. alle Expertenvorträge mit Deutsch als L2) können nicht wie im *GeWiss*-Portal direkt angesteuert werden. Überdies sind tokenübergreifende Annotationen (bspw. Diskurskommentierungen, Verweise/Zitate) über die DGD nicht abrufbar.

Im Projekt *ZuMult*[7] wurden neue Zugangswege zu Korpora der gesprochenen Sprache geschaffen, die noch stärker an spezifischen Nutzungsinteressen ausgerichtet sind. Die dabei entwickelten Tools *ZuMal*[8], *ZuRecht*[9] und *ZuViel*[10] gestatten jeweils zielgruppenspezifische Zugriffsmöglichkeiten auf die *GeWiss*-Daten und bieten eine Vielfalt an neuen Forschungs- und Anwendungsmöglichkeiten (vgl. Fandrych / Wallner 2022, 2023; Fandrych et al. 2023; Schwendemann / Wallner im Dr.). Während die Tools *ZuMal* und *ZuViel* stärker an den Nutzungs- und Informationsbedürfnissen von Sprachdidaktiker:innen und -lernenden ausgerichtet sind, handelt es sich bei *ZuRecht* um ein mächtiges Suchinstrument, welches einen Großteil der bezüglich des *GeWiss*-Portals und der DGD angesprochenen Einschränkungen überwindet und beinahe sämtliche *GeWiss*-Daten (Primärdaten, Metadaten und Annotationen) inklusive der nichtdeutschsprachigen Daten gemeinsam zugänglich

[5] https://gewiss.uni-leipzig.de/ (25.05.2023).
[6] https://dgd.ids-mannheim.de/ (25.05.2023).
[7] Ausführliche Informationen zum Projekt finden sich unter https://zumult.org/ (25.05.2023) sowie im Fandrych et al. (2023 in dieser Themenausgabe). Die im Projekt entwickelten Tools können nach kostenloser Registrierung bei der DGD für Forschung und Lehre genutzt werden.
[8] *ZuMal* steht für *Zugang zu Merkmalsauswahl von Gesprächen*. Für ausführliche Informationen vgl. Fandrych et al. (2023 in dieser Themenausgabe). Zugang zu *ZuMal*: https://cht.ids-mannheim.de/ProtoZumult/prototype/dist/zuMal.jsp (25.05.2023).
[9] *ZuRecht* steht für *Zugang zur Recherche in Transkripten*. Für ausführliche Informationen vgl. Frick / Helmer / Wallner (2023 in dieser Themenausgabe). Zugang zu *ZuRecht*: https://zumult.ids-mannheim.de/ProtoZumult/jsp/zuRecht.jsp?lang=de (25.05.2023).
[10] *ZuViel* steht für *Zugang zu Visualisierungselementen für Transkripte*. Für ausführliche Informationen vgl. Schmidt / Schwendemann / Wallner (2023 in dieser Themenausgabe). Zugang zu *Zuviel*: https://zumult.ids-mannheim.de/ProtoZumult/jsp/zuViel.jsp?transcriptID=FOLK_E_00349_SE_01_T_01 (25.05.2023).

macht[11]. So ist es mit *ZuRecht* bspw. erstmals möglich, tokenbasierte und tokenübergreifende Annotationen innerhalb einer komplexen Suchanfrage abzurufen (vgl. Frick / Helmer / Wallner 2023 in dieser Themenausgabe).

Anwendung

Das Korpus *GeWiss* bietet vielfältige Möglichkeiten zur Erforschung und Vermittlung der mündlichen Hochschulkommunikation in philologischen Fächern. Insbesondere aus der Unterrichtspraxis stammende Fragestellungen – etwa bezüglich des Vorkommens und der Verbreitung von Mündlichkeitsphänomenen in der gesprochenen Wissenschaftssprache – lassen sich mit Hilfe dieses Korpus beantworten. So können bspw. mit dem Tool *ZuMal* für alle deutschsprachigen Sprechereignisse des *GeWiss*-Korpus Werte zum relativen Anteil einzelner Mündlichkeitsphänomene abgefragt werden. Zusätzlich ist es mit Hilfe der Filteroptionen von *ZuMal* möglich, die Auswahl der Sprechereignisse einzugrenzen. Abb. 1 zeigt dies am Beispiel studentischer Vorträge mit Deutsch als L1. In dem Ausschnitt aus der Ergebnisübersicht unterhalb des Streudiagramms wird für jeden Vortrag die Normalisierungsrate[12] sowie der Anteil an Modalpartikeln angezeigt. Anhand der Gesamtübersicht ist erkennbar, dass die Normalisierungsrate in den studentischen Vorträgen zwischen 7 % und 13 % liegt. Der Anteil an Modalpartikeln liegt zwischen 1,43 % und 2,95 %. Die Gesamtübersicht lässt sich kopieren und als Grundlage für weitere Auswertungen (bspw. Vergleiche mit anderen Sprechereignissen) nutzen (vgl. hierzu auch Schwendemann / Wallner i. Dr.).

[11] Einschränkungen existieren lediglich bezüglich ausgewählter Metadaten, die im Metadatenschema der DGD bislang nicht vorkommen (so etwa Grad der Mündlichkeit, der jedoch nur bezüglich der Vorträge erhoben wurde).

[12] Die Normalisierungsrate gibt an, wie viele sprachliche Einheiten in einem Sprechereignis abweichend von einem angenommenen schriftsprachlichen Standard realisiert werden (bspw. Reduktionen wie *hab* [habe] oder umgangssprachliche Formen wie *nee* [nein]) (vgl. Fandrych et al. 2023 in dieser Themenausgabe).

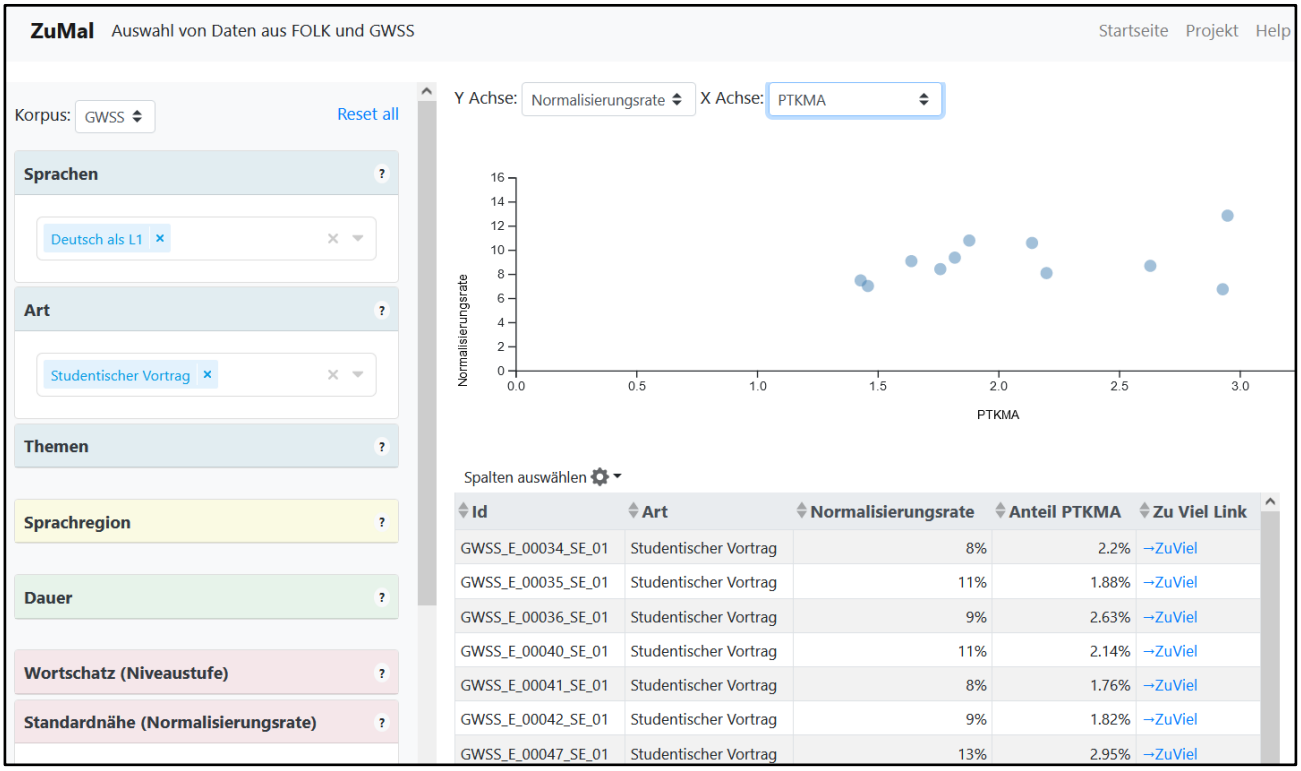

Abbildung 1
Das Tool *ZuMal* mit Filtereinstellungen für die Auswahl von studentischen Vorträgen mit Deutsch als L1

Um zu prüfen, welche sprachlichen Einheiten üblicherweise von Sprecher:innen mit Deutsch als L1 in studentischen Vorträgen des *GeWiss*-Korpus abweichend von einem angenommenen schriftsprachlichen Standard realisiert werden, kann das Tool *ZuRecht* mit der Anfrage (1) genutzt werden[13].

(1) ([word.type=".*diffNorm.*" & !pos="NGHES" & !pos="XY" & !pos="AB" & !pos="NE" & !pos="NN"] within <ses_sprachen_s="Deutsch \(L1\)"/> | <ses_sprachen_s="Deutsch \(L1\) ; .+"/>) within <e_se_art="Studentischer Vortrag"/>

Mit Hilfe einer quantitativen Auswertung (aufrufbar in *ZuRecht* über „Treffer gruppieren“) lassen sich die häufigsten „Normalisierungsfälle“ ermitteln. Hierzu zählen *is* [ist], *nich* [nicht] und *n* [ein] (vgl. Schwendemann/Wallner i. Dr.).

Mit Hilfe von Anfrage (2) lassen sich alle Modalpartikeln abrufen, die von Sprecher:innen mit Deutsch als L1 in studentischen Vorträgen produziert wurden.

(2) ([pos="PTKMA"] within <ses_sprachen_s="Deutsch \(L1\)"/> | <ses_sprachen_s="Deutsch \(L1\) ; .+"/>) within <e_se_art="Studentischer Vortrag"/>

Zu den häufigsten Modalpartikeln in diesem Genre gehören der quantitativen Auswertung zufolge *ja*, *eben* und *halt* (vgl. ebd.).

Die Ergebnisse derartiger empirischer Untersuchungen besitzen für sprachdidaktische Kontexte hohe Relevanz, da sich auf dieser Grundlage relevante Vermittlungsgegenstände identifizieren

[13] Mit Hilfe dieser Abfrage werden alle Einheiten (außer Häsitationen, Nichtwörter, Abbrüche, Eigennamen und Nomen) ermittelt, bei denen sich die aussprachenahe Transkription von der orthografischen Normalisierung unterscheidet und die von Sprecher:innen mit Deutsch als L1 innerhalb von studentischen Vorträgen produziert wurden.

lassen (vgl. auch Fandrych / Meißner / Wallner 2021). Überdies eröffnen die im Projekt *ZuMult* geschaffenen Tools *ZuMal* und *ZuViel* einen niederschwelligen Zugang für eine Veranschaulichung von Mündlichkeitsphänomenen in der gesprochenen Wissenschaftssprache anhand der *GeWiss*-Daten.

Literatur und Ressourcen

Baur, Benedikt / Gräfe, Karen / Lange, Daisy / Schmidt, Julia (2014): *Dokumentation zur Annotation der Diskurskommentierungen*. https://gewiss.uni-leipzig.de/fileadmin/documents/Annotationsdokumentation_GeWiss.pdf (25.05.2023).

Fandrych, Christian (2014): Metakommentierungen in wissenschaftlichen Vorträgen. In: Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (Hrsg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 95-111.

Fandrych, Christian / Meißner, Cordula / Wallner, Franziska (2021): Korpora gesprochener Sprache und Deutsch als Fremd- und Zweitsprache: Eine chancenreiche Beziehung. In: *Korpora Deutsch als Fremdsprache* 1: 2, 5-30. 10.48694/tujournals-76.

Fandrych, Christian / Wallner, Franziska (2022): Funktionale und stilistische Merkmale gesprochener fortgeschrittener Lerner:innensprache: Methodische und konzeptionelle Überlegungen am Beispiel von GeWiss. In: *Zeitschrift für Germanistische Linguistik* 50: 1, 202-239.

Fandrych, Christian / Wallner, Franziska (2023): Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation. In: Deppermann, Arnulf / Fandrych, Christian / Kupietz, Marc / Schmidt, Thomas (Hrsg): *Korpora in der germanistischen Sprachwissenschaft: Mündlich, schriftlich, multimedial*. Berlin / Boston: de Gruyter, 129-160.

Fandrych, Christian / Meißner, Cordula / Schwendemann, Matthias / Wallner, Franziska (2023): ZuMal: Zielgruppenspezifische Gesprächsauswahl aus Korpora gesprochener Sprache. In: *Korpora Deutsch als Fremdsprache* 3:1, 13-43.

Fandrych, Christian / Schmidt, Thomas / Wallner, Franziska / Wörner, Kai (Hrsg.) (2023): Zugänge zu multimodalen Korpora gesprochener Sprache für DaF und DaZ. Themenausgabe in *Korpora Deutsch als Fremdsprache* 3:1.

Frick, Elena / Helmer, Henrike / Wallner, Franziska (2023): ZuRecht: Neue Recherchemöglichkeiten in Korpora Gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache. In: *Korpora Deutsch als Fremdsprache* 3:1, 44-71.

Reershemius, Gertrud / Lange, Daisy (2014): Sprachkontakt in der mündlichen Wissenschaftskommunikation. In: Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (Hrsg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron, 57-74.

Sadowski, Sabrina (2017): Die Annotation von Zitaten und Verweisen im GeWiss-Korpus. In: Fandrych, Christian / Meißner, Cordula / Wallner, Franziska (Hrsg.): *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora*, Tübingen: Stauffenburg, 147-166.

Schmidt, Thomas / Schütte, Wilfried / Winterscheid, Jenny (2015): *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2).* Institut für Deutsche Sprache. Mannheim. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4616 (25.05.2023).

Schmidt, Thomas / Schwendemann, Matthias / Wallner, Franziska (2023): ZuViel: Transkriptvisualisierung und Arbeiten mit Transkripten. In: *Korpora Deutsch als Fremdsprache* 3:1, 72-91.

Schwendemann, Matthias / Wallner, Franziska (i. Dr.): Mündlichkeitsphänomene in der gesprochenen Wissenschaftssprache: Korpuslinguistische Befunde und didaktische Perspektiven. In: *InfoDaF* 2023.

Selting, Margret / Auer, Peter / Barth-Weingarten, Dagmar / Bergmann, Jörg / Bergmann, Pia / Birkner, Karin / Couper-Kuhlen, Elizabeth / Deppermann, Arnulf / Gilles, Peter / Günthner, Susanne / Hartung, Martin / Kern, Friederike / Mertzlufft, Christine / Meyer, Christian / Morek, Miriam / Oberzaucher, Frank / Peters, Jörg / Quasthoff, Uta / Schütte, Wilfried / Stukenbrock, Anja / Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, 353-402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf (25.05.2023).

Westpfahl, Swantje / Schmidt, Thomas / Jonietz, Jasmin / Borlinghaus, Anton (2017): *STTS 2.0. Guidelines für die Annotation von POS -Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS)*. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/6063 (25.05.2023).

Biographische Notiz: Dr. Franziska Wallner ist wissenschaftliche Mitarbeiterin am Herder-Institut der Universität Leipzig. Ihre Forschungsschwerpunkte sind unter anderen das Deutsche als fremde Bildungs- und Wissenschaftssprache, die korpusbasierte Erforschung der gesprochenen Sprache, Mündlichkeitsdidaktik sowie die Nutzung von Korpora im Kontext von Deutsch als Fremd- und Zweitsprache.

Kontaktanschrift:
Franziska Wallner
Herder-Institut
Universität Leipzig
Beethovenstr. 15
04107 Leipzig
Deutschland
f.wallner@uni-leipzig.de