Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 4, Nummer 1 (Juli 2024)
ISSN 2749-4411

KORPUS EINFACHES DEUTSCH (KED)

Daniel Jach, Southwest Jiaotong University
Gunther Dietz, Otto-Friedrich-Universität Bamberg

Abstract

Das *Korpus einfaches Deutsch* (KED) enthält bildungssprachliche Texte in einfacher Sprache, die sich an Lesende mit eingeschränkter Lesekompetenz richten. Das KED ist als Ressource für die datengeleitete Sprachvermittlung im Bereich Deutsch als Fremd- und Zweitsprache (DaF/DaZ) sowie für die korpuslinguistische Forschung zur sprachlichen Komplexität und Einfachheit konzipiert. Im vorliegenden Beitrag wird zunächst die Zusammensetzung des Korpus beschrieben. Im Anschluss wird in einer exemplarischen Anwendung des Korpus im DaF-Unterricht aufgezeigt, wie Belege für die *je-desto*-Konstruktion gewonnen und didaktisch aufbereitet werden können. Die Daten sind über die KorAP-Anwendung des Leibniz-Instituts für Deutsche Sprache zugänglich. Ein eigenes Suchportal mit sprachdidaktisch orientierten Such- und Filtermöglichkeiten ist in Planung.

Keywords: Korpus; Deutsch; DaF; DaZ; Sprachdidaktik; DDL; Komplexität; Korpus einfaches Deutsch

Abstract

The *Korpus einfaches Deutsch* (KED, ‘Corpus of simple German’) contains educational language texts in simple language aimed at readers with limited reading skills. The KED is designed as a resource for data-driven language learning in the field of German as a foreign and second language (GFL/GFS) and for corpus linguistic research on linguistic complexity and simplicity. This article first describes the composition of the corpus. The following section presents an exemplary application that illustrates how examples of the *je-desto* construction can be obtained from the corpus and utilized for GFL teaching. The data can be accessed via the KorAP application of the Leibniz Institute for the German Language. A separate search portal with language-didactically oriented search and filter options is currently under development.

Keywords: corpus; German; GFL; GFS; didactics; DDL; complexity; Corpus of simple German

1. Aufbau und Inhalt

Das *Korpus einfaches Deutsch* (KED) besteht aus 6.841 bildungssprachlichen Texten in einfachem Deutsch. Darunter werden hier solche Texte verstanden, die sich an Lesende mit (vermutlich) eingeschränkter Lesekompetenz richten und von denen daher zu erwarten ist, dass die Textproduzenten sie verständlicher und an die Bedürfnisse ihrer Adressaten angepasst formuliert haben (vgl. Bredel / Maaß 2016: 537). Lesende mit eingeschränkter Lesekompetenz sind zum Beispiel Kinder, Jugendliche und Erwachsene, deren literale Kompetenz schwach entwickelt ist. Tabelle 1 fasst das Korpus zusammen.

KED	
Quellen: n	14
Texte: n	6.841
Tokens: n	2.803.454
Tokens/Text: m (se)	409,8 (3,57)
Sätze: n	224.042
Sätze/Text: m (se)	32,75 (0,31)

n = Anzahl, m = Mittelwert (mean), se = Standardfehler

Tabelle 1
Korpus einfaches Deutsch im Überblick

Das Korpus umfasst authentische Texte mit fachlich-bildender oder informierender Absicht, die eine einfache, aber allgemein schriftsprachliche Varietät abbilden sollen. Texte, die eigens für die Sprachvermittlung oder für sprachdidaktische Zwecke im weitesten Sinn erstellt wurden, sind dagegen nicht enthalten. Ebenfalls nicht enthalten sind Texte in Leichter Sprache (vgl. Bredel / Maaß 2016). Leichte Sprache folgt bestimmten sprachlichen sowie ortho- und typografischen Prinzipien, die das Verstehen erleichtern sollen. Sie wird häufig verwendet, um juristische Texte verständlicher zu gestalten. Texte in Leichter Sprache können erheblich von schriftsprachlichen Normen abweichen und werden in der Regel von besonders geschulten Autoren verfasst. Das erschwert ihre korpuslinguistische Verarbeitung und entspricht nicht dem Ziel des KED. Zudem beinhaltet das KED ausschließlich Sachtexte. Fiktionale Texte wie Märchen sind bislang nicht Teil des Korpus. Tabelle 2 gibt einen Überblick über die Verteilung von Adressatengruppen, Textsorten, Vertextungsstrategien und Themen im KED.

(a) Adressaten	n Texte	%
Kinder	5.793	84,68
Jugendliche	525	7,67
Erwachsene	523	7,65

(b) Textsorte	n Texte	%
Lexikonartikel	4.424	64,67
Nachricht	943	13,78
Erklärtext	915	13,38
Empfehlung	287	4,2
Experiment	217	3,17
Argumentation	55	0,8

(c) Vertextungsstrategie	n Texte	%
Erklären	5.592	81,74
Berichten	943	13,78
Anweisen	251	3,67
Argumentieren	55	0,80

(d) Thema	n Texte	%
Geschichte und Kultur	2.193	32,06
Politik und Gesellschaft	2.180	31,87
Natur und Leben	1.880	27,48
Gesundheit und Krankheit	588	8,60

Tabelle 2
KED: Aufbau und Inhalt

2. Erhebung, Annotation, Metadaten

Das Korpus besteht aus Texten, die zum Zeitpunkt der Erhebung öffentlich online einsehbar waren. Die Quellwebseiten wurden automatisiert erfasst und mit Zustimmung der Rechteinhaber als statische Kopien im *Internet Archive*[1] archiviert und mit einem permanenten Link versehen. Anschließend wurden die Texte der archivierten Kopien (bzw. die Texte der nicht archivierten Quellwebseiten) automatisiert heruntergeladen, in Absätze, Sätze und Wortformen segmentiert und mit Lemmata und Wortarten nach dem Stuttgart-Tübingen-Tagset annotiert. Die Datenverarbeitung wurde mit den Programmiersprachen *R* und *Python* ausgeführt, für die Tokenisierung und Annotation der Texte mit Lemmata und POS wurde der Parser *spaCy* verwendet.

Die Texte wurden nach Einschätzung der Autoren nach Adressatengruppe, Textsorte und Vertextungsstrategie kategorisiert. Die von den Textproduzenten intendierte Adressatengruppe, Textsorte und Vertextungsstrategie waren in der Regel anhand der Einordnung der Texte auf der Quellwebseite erkennbar (z.B. *Nachrichten für Kinder*, *Experimente für Jugendliche*).

Die Themen der Texte wurden mit Hilfe von einem maschinellen Lernverfahren ermittelt ('topic modeling', vgl. Silge / Robinson 2017). Die Wortwolken in Abbildung 1 zeigen die wichtigsten Nomen jedes Themas. Schriftgröße und Farbe zeigen an, wie wichtig das Wort für das jeweilige Thema ist. Die Titel sind nachträglich hinzugefügt worden, um die Themen begrifflich zusammenzufassen.

Darüber hinaus enthält jeder Text Angaben zur Textdeckung. Die Textdeckung wurde auf der Grundlage eines Häufigkeitswörterbuchs des Deutschen ermittelt (vgl. Tschirner / Möhring 2020) und gibt an, welcher Anteil der Textwörter in den häufigsten 1.000, 2.000, 3.000, 4.000 und 5.000 Wörtern des Deutschen enthalten ist. Tabelle 3 zeigt einige relevante Metadaten im Überblick.

Metadatum	**Erläuterung**
corpusSigle	Identifikationsbezeichnung des Korpus (ked)
cover1kHerder	Textdeckung der 1.000 häufigsten Wörter des Deutschen
nToks	Anzahl der Tokens im Text
permalink	URL der Quellwebseite bzw. der archivierten Kopie der Quellwebseite
rcpnt	Adressatengruppe des Textes (kinder, jugendliche, erwachsene)
strtgy	Vertextungsstrategie (erklaeren, berichten, argumentieren, anweisen)
topic	Thema des Textes (politik_gesellschaft, geschichte_kultur, natur_leben, gesundheit_krankheit)
txttyp	Textsorte des Textes (lexikonartikel, nachricht, erklaertext, empfehlung, argumentation, experiment)

Tabelle 3
KED-Metadaten in Auswahl (Stand Mai 2024)

[1] https://archive.org/ (14.07.2024).

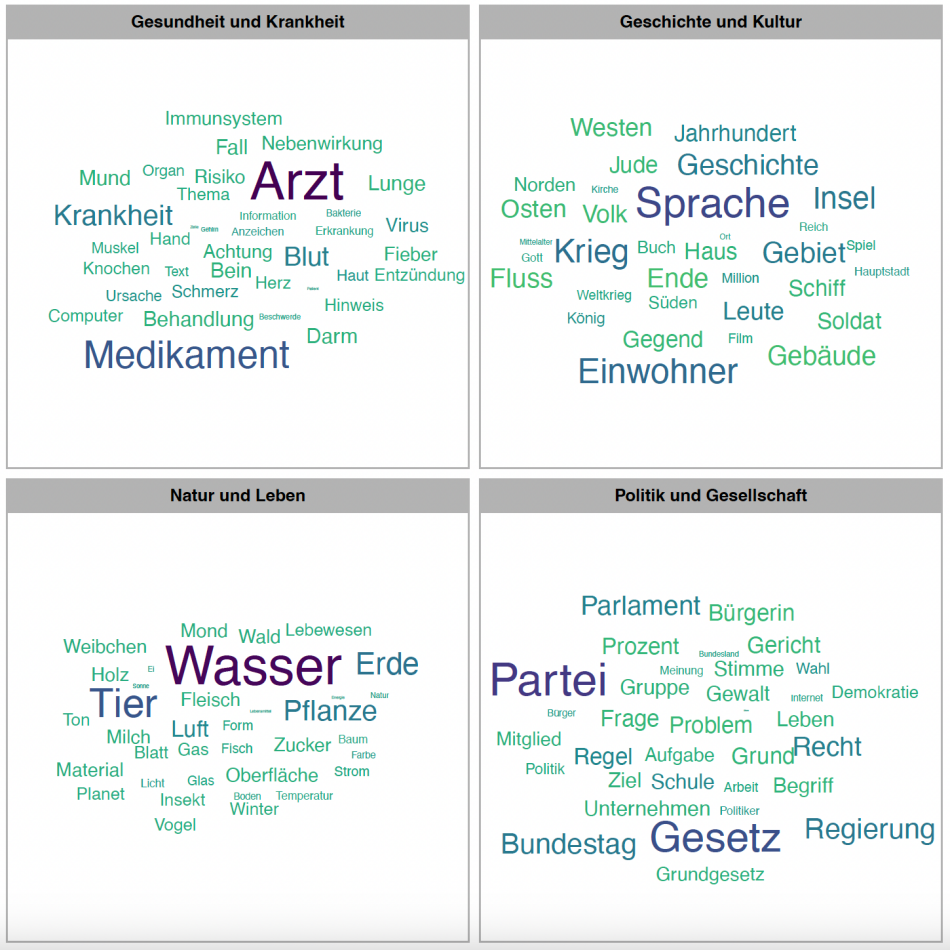


Abbildung 1
Wortwolken der Themen im KED

3. Zugang

Das KED wird Teil des Deutschen Referenzkorpus (DeReKo). Als Einzelinstanz ist es derzeit – nach erfolgter Registrierung und Anmeldung – über die KorAP-Plattform des Leibniz-Instituts für Deutsche Sprache (IDS)[2] durchsuchbar. Um den Zugang zu den Korpusdaten für Lehrpersonen und Lernende zu verbessern, ist die Einrichtung eines eigenen Suchportals für das KED mit vereinfachter Steuerung und didaktisch orientierten Suchmöglichkeiten vorgesehen.

4. Nutzungsbeispiel

Das KED fungiert unter anderem als Fundort für authentische Belege, die in der DaF- / DaZ-Sprachvermittlung im Rahmen von DDL-Aktivitäten (*data-driven learning*) genutzt werden können. Über die Nutzung für den Fremdsprachenunterricht hinaus gibt das KED auch Impulse für die (korpus-

[2] https://korap.ids-mannheim.de/instance/ked (14.07.2024).

linguistische) Erforschung sprachlicher Komplexität und Einfachheit, etwa durch vergleichende Analyse von Kindernachrichten und Nachrichten für Erwachsene[3].

Im Folgenden wird eine mögliche Anwendung des KED im Rahmen eines DaF-Kurses auf dem GER-Niveau B1 skizziert. Das Lernziel der Einheit wird aus konstruktionsdidaktischer Sicht (vgl. Herbst 2016; Amorocho / Pfeiffer 2023) bestimmt und besteht darin, die Lernenden zu befähigen, proportionale Entwicklungen und Veränderungen im Kontext von Sachtexten (z.B. Experimentbeschreibungen, Kochrezepten, Grafikbeschreibungen u.a.) mit Hilfe der *je-desto*-Konstruktion zu verstehen und auszudrücken. Hierfür wird zunächst ein Impuls in Form von ausreichendem Input gesetzt. Anschließend wird die Aufmerksamkeit der Lernenden mit Hilfe von DDL-Aktivitäten auf die Form und die Bedeutung der Konstruktion gelenkt, um den Erwerb zu unterstützen.

Hierzu sind zunächst entsprechende Belege aus dem Korpus zu gewinnen. Eine Suchanfrage im KED nach *je* im Kontext von *desto* oder *umso* (im Abstand von maximal 15 Wörtern vor und nach *je*) ergibt 293 Treffer. In 283 Fällen geht der *je*-Teil dem *desto/umso*-Teil voran, sodass dies als die prototypische Variante gelten kann. Lediglich 10 Belege mit vorangehendem *umso*-Teil und keinen einzigen mit vorangehendem *desto*-Teil finden sich im Korpus. Aus dieser Gesamt-Trefferliste kann nun für den Unterricht ein Auszug von ca. 12 Belegen vorbereitet werden (vgl. Abb. 2).

	Linker Kontext		rechter Kontext	Quelle
1.	… Für die 5%-Hürde gibt es einen Grund:	Je	***mehr*** Parteien im Bundestag sind, desto ***schwerer*** können Entscheidungen getroffen werden. …	KED/BPB/0004
2.	… 2. Gemüse kurz und bissfest garen	Je	kürzer Sie Gemüse garen, desto mehr Vitamine und Mineralstoffe bleiben erhalten. …	KED/BZE/00025
3.	… erscheinenden gedruckten Medien (sogenannte Printmedien) wie Zeitungen und Zeitschriften gebraucht.	Je	besser die Drucktechnik wurde, desto mehr Zeitungen konnten jeden Tag gedruckt und verkauft werden. …	KED/HAN/00583
4.	… Menschen, die angeblich furchtbare und oft unvorstellbare Ziele anstreben.	Je	unglaublicher diese Verschwörungstheorien klingen, umso faszinierender finden sie manche Menschen. …	KED/HAN/00185
5.	… In der Wirtschaft regelt einerseits die Nachfrage das Angebot:	Je	mehr Menschen eine Ware haben wollen, also eine Ware nachfragen, desto mehr wird diese Ware auch produziert …	KED/HAN/00265
11.	… Oft sind sie bedroht, weil ein Lebensraum verloren geht.	Je	kleiner dieser ist, desto größer ist auch die Gefahr.	KED/KLX/01703
12.	… So wurde zum Beispiel die alte römische Stadt Pompeji bei einem Ausbruch des Vesuv verschüttet.	Je	höher die Asche in die Atmosphäre aufsteigt, desto weiter kann sie sich verbreiten. …	KED/KLX/01333

Abbildung 2
Konkordanzliste zu *je-desto*-Belegen aus KED im Rahmen von DDL-Aktivitäten (Auszug)

Um die Aufmerksamkeit der Lernenden auf die *Formseite* der Konstruktion zu lenken, kommen etwa folgende Instruktionen in Frage:

1. Unterstreichen Sie in den Belegen die Signalwörter *je*, *desto* und *umso*.

[3] Siehe hierzu eine Vorstudie zu einer Pilotversion von KED (vgl. Jach 2022).

2. Markieren Sie alle Wörter hinter *je* und hinter *desto / umso*. Was fällt Ihnen auf? Um welche Wörter (Wortarten) handelt es sich? In welcher Form erscheinen die Wörter?
3. Wo steht das Verb im *je*-Teil, wo im *desto-/umso*-Teil?

Diese Aktivitäten sollen den Lernenden bewusst machen, dass sowohl der *je*-Teil als auch der *desto/umso*-Teil der Konstruktion jeweils eine Komparativform enthält, die Teile sich aber in der Verbstellung unterscheiden (Verbletztstellung im *je*-Teil, Verbzweitstellung im *desto/umso*-Teil).

Folgende Instruktionen fokussieren dagegen die *Bedeutungsseite* der Konstruktion:

4. Welche Information findet man im *je*-Teil, welche im *desto*-Teil?
5. Wie hängen die Informationen im *je*-Teil und im *desto*-Teil zusammen?

Diese Aktivitäten sollen die Funktion der *je-desto/umso*-Konstruktion als sprachliches Mittel zum Ausdruck von proportionalen Entwicklungen hervorheben und insbesondere den Wirkzusammenhang bewusst machen, dass graduelle Veränderungen im *je*-Teil zu graduellen Veränderungen im *desto/umso*-Teil führen.

Für eine stärkere unterrichtliche Steuerung könnte folgende Aufgabe (vgl. Abb. 3) unterstützend herangezogen werden.

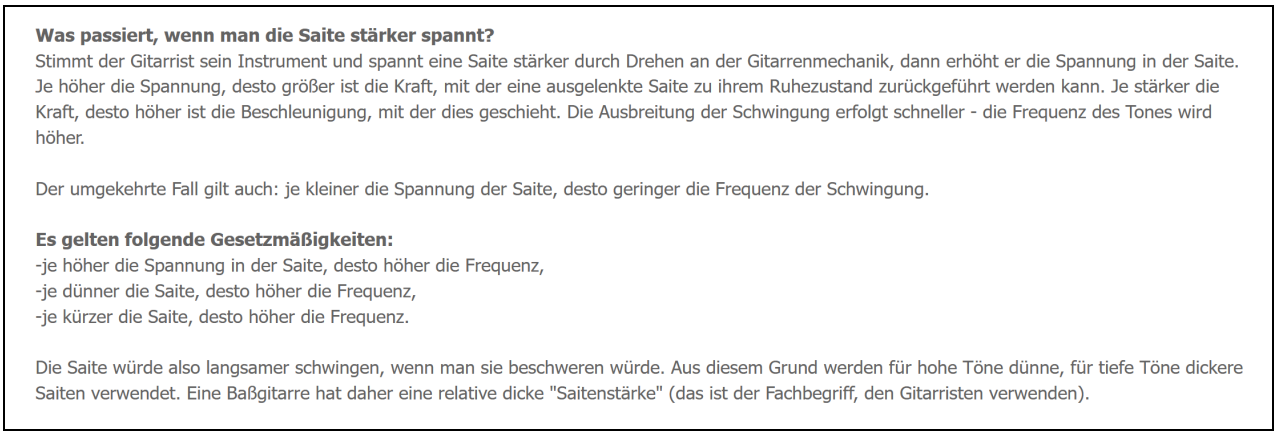
Welche Handlungen / Ereignisse werden im *je*- und im *desto/umso*-Teil beschrieben? Ergänzen Sie die Tabelle (wie im Beispiel für Beleg 1). Nutzen Sie die Informationen aus der Trefferliste oben.

	Graduelle Veränderung im *je*-Teil	→	**Graduelle Veränderung im *desto*-Teil**
1	Zahl der Parteien ***steigt an***	→	Entscheidungen zu treffen wird ***schwerer***
2	Gemüse: Garzeit wird	→	..... bleiben erhalten
3	Drucktechnik	→	..... Zeitungen .....
4	Verschwörungstheorien klingen	→	manche Menschen werden von ihnen ..... fasziniert
5		→	

Abbildung 3
Übung zur Bewusstmachung der Bedeutung von *je-desto*-Konstruktionen im Rahmen von DDL-Aktivitäten (Auszug)

Hieran anschließend sollten weitere, auch produktive Aktivitäten folgen, um Lernenden die Form- und Bedeutung dieser Konstruktion zu vermitteln.

Neben der Suche nach Belegen für didaktisch relevante sprachliche Phänomene können auch Volltexte des KED via Verlinkung im Kontext der Webseite genutzt werden. Der Nutzen dieser Möglichkeit zeigt sich etwa an folgendem Textausschnitt aus einem Erklärtext für Kinder, in dem auf dichtem Textraum sechs Vorkommen von *je-desto*-Konstruktionen (vgl. Abb. 4) zu finden sind.



Was passiert, wenn man die Saite stärker spannt?
Stimmt der Gitarrist sein Instrument und spannt eine Saite stärker durch Drehen an der Gitarrenmechanik, dann erhöht er die Spannung in der Saite. Je höher die Spannung, desto größer ist die Kraft, mit der eine ausgelenkte Saite zu ihrem Ruhezustand zurückgeführt werden kann. Je stärker die Kraft, desto höher ist die Beschleunigung, mit der dies geschieht. Die Ausbreitung der Schwingung erfolgt schneller - die Frequenz des Tones wird höher.

Der umgekehrte Fall gilt auch: je kleiner die Spannung der Saite, desto geringer die Frequenz der Schwingung.

Es gelten folgende Gesetzmäßigkeiten:
-je höher die Spannung in der Saite, desto höher die Frequenz,
-je dünner die Saite, desto höher die Frequenz,
-je kürzer die Saite, desto höher die Frequenz.

Die Saite würde also langsamer schwingen, wenn man sie beschweren würde. Aus diesem Grund werden für hohe Töne dünne, für tiefe Töne dickere Saiten verwendet. Eine Baßgitarre hat daher eine relative dicke "Saitenstärke" (das ist der Fachbegriff, den Gitarristen verwenden).

Abbildung 4
Auszug aus Erklärtext „Warum erhöht sich der Ton einer Gitarrensaite, wenn man sie spannt?“ (KED/KSE/00146[4])

Das Beispiel der *je-desto*-Konstruktion veranschaulicht, dass sich im KED ausreichend Belege auf (lexikalisch und grammatisch) angemessenem Niveau für die datengestützte DaFZ-Vermittlung unterschiedlicher Zielstrukturen finden lassen.

Literatur und Ressourcen

Amorocho, Simone / Pfeiffer, Christian (2023): Konstruktionsdidaktik – Grundzüge einer sprachdidaktischen Konzeption. In: *Deutsch als Fremdsprache* 60: 3, 131-147.

Bredel, Ursula / Maaß, Christiane (2016). *Leichte Sprache: theoretische Grundlagen, Orientierung für die Praxis*. Sprache im Blick. Berlin: Dudenverlag.

Herbst, Thomas (2016): Foreign language learning is construction learning – what else? Moving towards Pedagogical Construction Grammar. In: de Knop, Sabine / Gilquin, Gaëtanelle (Hrsg.): *Applied Construction Grammar*. Berlin: de Gruyter, 21-52.

Jach, Daniel (2022): Korpus Einfaches Deutsch. Materialgrundlage für die daten-getriebene Lehre von Deutsch als fremder Bildungssprache auf niedrigem Sprachniveau. In: Li, Yuan / Liu, Fang / Wang, Zhongxin (Hrsg.): *Didactica, Cultura, Lingua. Perspektiven des Deutschen*. München: iudicium, 231-244.

Silge, Julia / David Robinson (2017): *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O’Reilly. https://www.tidytextmining.com/ (17.07.2024).

Tschirner, Erwin / Möhring, Jupp (2020): *A frequency dictionary of German. Core vocabulary for learners*. 2nd ed. London / New York: Routledge.

[4] Quelle: https://web.archive.org/web/20231231064243/https://www.kids-and-science.de/kinderfragen/detailansicht/datum/2016/11/09/warum-erhoeht-sich-der-ton-einer-gitarrensaite-wenn-man-sie-spannt.html (14.07.2024).

Biographische Notiz: Daniel Jach studierte Linguistik an Universitäten in Deutschland, den USA und den Niederlanden und promovierte 2019 an der Universität Jena mit einer empirischen Arbeit zum gebrauchsbasierten Fremdsprachenerwerb. Seit 2019 arbeitet er als Dozent für deutsche Sprache und Linguistik an Universitäten in China, seit 2021 als DAAD-Lektor an der Southwest Jiaotong University in Chengdu. Seine Forschungsschwerpunkte sind Korpuslinguistik, gebrauchsbasierte Linguistik und Fremdsprachenerwerb.

Kontaktanschrift:
Daniel Jach
Southwest Jiaotong University
School of Foreign Languages
West Park of Hi-Tech Zone
611756 Chengdu, Sichuan
P. R. China
daniel.jach@outlook.com

Biographische Notiz: Gunther Dietz studierte und promovierte in Deutsch als Fremdsprache an der LMU München. Nach einem DAAD-Lektorat und einer Tätigkeit als Sprachdozent war er von 2009 bis 2024 wissenschaftlicher Mitarbeiter am Lehrstuhl für Deutsch als Zweit- und Fremdsprache und seine Didaktik der Universität Augsburg. Zurzeit vertritt er die Professur für Deutsche Sprachwissenschaft / Deutsch als Fremdsprache der Universität Bamberg. Seine Schwerpunkte sind die fremdsprachliche Hörverstehensvermittlung und die Nutzung von Korpora in der DaFZ-Vermittlungspraxis.

Kontaktanschrift:
Prof. Dr. Gunther Dietz
Otto-Friedrich-Universität Bamberg /
Professur für Deutsche Sprachwissenschaft / Deutsch als Fremdsprache
96047 Bamberg
gunther.dietz@uni-bamberg.de
gunther.dietz@dietz-und-daf.de