



Vilniaus
universiteto
leidykla

Kalbotyra

ISSN 1392-1517
eISSN 2029-8315

VILNIUS UNIVERSITY

Academic Journal

Linguistics

2025
Volume 78

Published since 1958

*Towards an accessible and sound corpus methodology
in research and teaching*

Vilnius University Press
Vilnius, 2025

VILNIAUS UNIVERSITETAS

Mokslo žurnalas
Kalbotyra
2025
78 tomas

Leidžiamas nuo 1958 metų

*Tekstynų teorija ir praktika:
prieinami ir patikimi metodai ir įrankiai*

The journal is published once a year.

Editors-in-Chief

Jolanta Šinkūnienė Vilnius University (04H)
E-mail: jolanta.sinkuniene@flf.vu.lt

Vaiva Žeimantienė Vilnius University (04H)
E-mail: vaiva.zeimantiene@flf.vu.lt

Executive secretary

Justina Daunorienė Vilnius University (04H)

Editorial board

Diego Ardoino	Universität Bern (04H), Vilniaus universitetas (04H)
Johan van der Auwera	Universiteit Antwerpen (04H)
Ineta Dabašinskienė	Vytauto Didžiojo universitetas (04H)
Klaus Geyer	Syddansk Universitet i Odense (04H)
Axel Holvoet	Vilniaus universitetas (04H)
Djuddah A. J. Leijen	Tartu University (04H)
Aurelija Leonavičienė	Vytauto Didžiojo universitetas (04H)
Rosa Lorés	Universidad de Zaragoza (04H)
Nicole Nau	Uniwersytet im. Adama Mickiewicza w Poznaniu (04H)
Sonia Oliver del Olmo	Universitat Autònoma de Barcelona (04H)
Wolfgang Pöckl	Universität Innsbruck (04H)
Jürgen Schiewe	Universität Greifswald (04H)
Daniel Van Olmen	Lancaster University (04H)
Heike Zinsmeister	Universität Hamburg (04H)

Editorial Address

Vilnius University
Faculty of Philology
Universiteto g. 5
LT-01131 Vilnius, Lithuania
Email: kalbotyra@flf.vu.lt

Website address

<http://www.journals.vu.lt/kalbotyra>

Kalbotyra 78 (2025) is published with financial support from the Research Council of Lithuania under the Lithuanian Studies Programme 2025–2030 (Contract No. P-LISs-25-62) and from the German Academic Exchange Service (DAAD), funded by the German Federal Foreign Office, within the framework of the project “KoDi-FS” of the Departments of German Studies at Vilnius University and the University of Hamburg.

Copyright © 2025 Authors. Published by Vilnius University Press.

This is an Open Access journal distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Žurnalas leidžiamas vieną kartą per metus.

Atsakingosios redaktorės

Jolanta Šinkūnienė Vilniaus universitetas (04H)
El. paštas: jolanta.sinkuniene@ffl.vu.lt

Vaiva Žeimantienė Vilniaus universitetas (04H)
El. paštas: vaiva.zeimantiene@ffl.vu.lt

Atsakingoji sekretorė

Justina Daunorienė Vilniaus universitetas (04H)

Redaktorių kolegija

Diego Ardoino	Berno universitetas (04H), Vilniaus universitetas (04H)
Johan van der Auwera	Antverpeno universitetas (04H)
Ineta Dabašinskienė	Vytauto Didžiojo universitetas (04H)
Klaus Geyer	Pietų Danijos universitetas (04H)
Axel Holvoet	Vilniaus universitetas (04H)
Djuddah A. J. Leijen	Tartu universitetas (04H)
Aurelija Leonavičienė	Vytauto Didžiojo universitetas (04H)
Rosa Lorés	Saragosos universitetas (04H)
Nicole Nau	Poznanės Adomo Mickevičiaus universitetas (04H)
Sonia Oliver del Olmo	Barselonos autonominis universitetas (04H)
Wolfgang Pöckl	Insbruko universitetas (04H)
Jürgen Schiewe	Greifsvaldo universitetas (04H)
Daniel Van Olmen	Lankasterio universitetas (04H)
Heike Zinsmeister	Hamburgo universitetas (04H)

Redakcijos adresas

Vilniaus universitetas
Filologijos fakultetas
Universiteto g. 5
LT-01131 Vilnius, Lietuva
El. paštas: kalbotyra@ffl.vu.lt

Tinklapio adresas

<http://www.zurnalai.vu.lt/kalbotyra>

Kalbotyra 78 (2025) leidybą finansuoja Lietuvos mokslo taryba pagal Lituanistikos 2025–2030 m. programą (sutarties Nr. P-LISs-25-62) ir Vokietijos akademinių mainų tarnyba (DAAD) Vokietijos Federalinės užsienio reikalų ministerijos lėšomis pagal Vilniaus ir Hamburgo universitetų germanistikos katedrų projektą „KoDi-FS“.

Autorių teisės © 2025 Autoriai. Publikavo Vilniaus universiteto leidykla.

Tai yra atvirosios prieigos žurnalas. Žurnalas platinamas vadovaujantis Kūrybinių bendrijų licencija Priskyrimas, kuri leidžia laisvai ir neribotai naudoti, kaip mano esant reikalinga, be kita ko, kopijuoti, daryti pakeitimus ir kurti išvestinius kūrinius, naudoti komerciniais ir nekomerciniais tikslais nurodant informacijos šaltinį ir autoriumi.

Contents

Preface. <i>Towards an accessible and sound corpus methodology in research and teaching</i>	7
Vorwort. <i>Perspektiven einer zugänglichen und fundierten Korpusmethodologie in Forschung und Lehre</i>	9
Pratarmē. <i>Tekstynų teorija ir praktika: prieinami ir patikimi metodai ir įrankiai</i>	11
Papers	
Eglē Kontutytē. <i>Was können Korpora besser als andere digitale Hilfsmittel? Korpusarbeit im fachbezogenen DaF-Unterricht am Beispiel der Rechtssprache</i>	
Abstract.....	13
Marie Flüh. <i>Digitale Literaturwissenschaft und pädagogische Praxis</i>	
Abstract.....	36
Carla Sökefeld, Anastasiia Stulen. <i>Korpusbasierte Übungsmaterialien für die Vermittlung von Korpuskompetenzen im DaF-Unterricht</i>	
Abstract.....	63
Baiba Egle, Dzintra Lele-Rozentāle, Agnese Dubova, Gints Jēkabsons. <i>Building and annotating a bachelor paper abstract corpus: First findings about Latvian student abstracts</i>	
Abstract.....	89
Bernhard Fisseni, Deniz Sarikaya, Bernhard Schröder. <i>How to annotate a corpus for research on frames in the language of mathematics</i>	
Abstract.....	111
Skaistē Volungevičienē, Maximilian Arndt. <i>KI als Hilfsmittel für die Formulierung von Suchanfragen in Korpora</i>	
Abstract.....	138
Patrick Grommes, Anastasiia Stulen. <i>Vom Seminar zur Praxis: Korpusgestützt lehren lernen</i>	
Abstract.....	156
Veslava Čižik-Prokaševa. <i>Daiktavardžio pykčis valentingumas Dabartinės lietuvių kalbos tekstyne</i>	
Abstract.....	179
Book review	
Melanie Andresen. <i>Computerlinguistische Methoden für die Digital Humanities: Eine Einführung für Geisteswissenschaftler:innen.</i> Tübingen: Narr Francke Attempto, 2024, 240 S. ISBN 978-3-8233-8579-0	
Recenzavo/ Reviewed by Christa Dürscheid	201

Towards an accessible and sound corpus methodology in research and teaching

This special issue broadens the scope of the *Kalbotyra* journal beyond linguistic analyses to encompass broader uses of corpora within German philology, or *Germanistik*, including literary analysis and second language teaching. A particular emphasis is placed on the use of corpora within an academic setting where the corpus language is studied as a second language, while recognising the challenges faced by learners at various levels of proficiency. To reflect the rapidly evolving digital landscape, the scope of this issue is also expanded by including studies examining the role of generative artificial intelligence, and specifically evaluating the application of large language models for research and teaching in philology, as well as studies addressing other languages such as English, Lithuanian, or Latvian.

For linguists, literary scholars, and experts of German as a foreign language, working with data in the form of empirical or literary corpora constitutes an indispensable resource for both research and teaching. Corpora offer structured access to language data across different topics and text types, including literary texts. Their language data can be enriched by linguistic and other interpretative annotation at multiple levels, which makes latent characteristics retrievable, thereby enabling generalisations and insights beyond the text surface. However, the practical use of corpora is far from straightforward. Researchers, teachers and students alike are often confronted with the challenge of mastering corpus access in terms of interfaces and query languages, while pedagogical resources for users with limited prior knowledge remain scarce. Another challenge they have to master is critical data literacy, for example, the capacity of understanding that every corpus compilation process has inherent limitations that are reflected in the findings derived from the resulting corpora.

This special issue addresses these challenges by presenting contributions that explore the use of corpora in research and teaching, reflect on the necessary corpus literacy, and consequences for corpus didactics. It brings together case studies demonstrating the application of specific corpora and tools to concrete research questions, teaching

scenarios that integrate corpus methods either in the preparation of materials or in direct student engagement, as well as discussions of the methodological and technical foundations required for corpus-based inquiry. Furthermore, it includes reflections on challenges in compiling and annotating corpora for specialised purposes, and considers the potential role of generative artificial intelligence in supporting linguistic and literary analyses and corpus development.

Although initiated by an open call, a majority of the contributions in this volume originate from talks or practical sessions at workshops, research and teaching stays organised within the framework of the project “Corpus competence for formulaic language / Korpusdidaktik für formelhafte Sprache” (KoDi-FS), an institutional partnership project between Vilnius University and the University of Hamburg. Each paper has undergone thorough anonymous peer reviewing by at least two experts. We would like to thank the reviewers for their valuable time and constructive feedback, which has greatly contributed to the scientific quality of this issue. We also gratefully acknowledge the support of the German Academic Exchange Service (DAAD) with funds of the German Federal Foreign Office, which has funded the project KoDi-FS since 2022, and has also supported this special issue. Funding to this publication has also been provided by the Research Council of Lithuania under the Lithuanian Studies Programme 2025–2030 (Agreement No. P-LISs-25-62).

Heike Zinsmeister, Vaiva Žeimantienė, Skaistė Volungevičienė, and Carla Sökefeld
Hamburg and Vilnius, 2025

Perspektiven einer zugänglichen und fundierten Korpusmethodologie in Forschung und Lehre

Dieses Themenheft erweitert den thematischen Horizont von *Kalbotyra* über die rein linguistische Analyse hinaus und richtet den Blick auf die vielfältigen Einsatzmöglichkeiten von Korpora innerhalb der Germanistik, einschließlich literaturwissenschaftlicher Fragestellungen und des Fremdsprachenunterrichts. Ein besonderes Augenmerk liegt dabei auf dem Einsatz von deutschen Korpora in der Auslandsgermanistik, wo die Lernenden neben den Fachinhalten auch Deutsch als Fremdsprache lernen. Dabei werden die Herausforderungen berücksichtigt, denen die Lernenden aufgrund unterschiedlicher sprachlicher Niveaus begegnen. Um zugleich den sich rasant wandelnden digitalen Entwicklungen Rechnung zu tragen, wird das Themenspektrum dieses Heftes durch Beiträge ergänzt, die die Rolle generativer Künstlicher Intelligenz untersuchen, insbesondere hinsichtlich der Anwendung großer Sprachmodelle in philologischer Forschung und Lehre, sowie Beiträge, die sich mit anderen Sprachen wie Englisch, Litauisch oder Lettisch beschäftigen.

Für Sprachwissenschaftler:innen, Literaturwissenschaftler:innen und Expert:innen für Deutsch als Fremdsprache stellen Daten in Form empirischer oder literarischer Korpora eine unverzichtbare Ressource für Forschung und Lehre dar. Korpora ermöglichen einen strukturierten Zugang zu Sprachdaten unterschiedlicher Themen und Textsorten, einschließlich literarischer Texte. Die Daten können dabei durch linguistische oder andere interpretative Annotationen auf verschiedenen Ebenen angereichert werden, sodass latente Merkmale erschließbar werden, die über die Textoberfläche hinausgehende Generalisierungen und Einsichten erlauben. Die praktische Nutzung von Korpora gestaltet sich jedoch keineswegs einfach. Forschende, Lehrende und Lernende sind gleichermaßen mit der Herausforderung konfrontiert, sich Löschen den Zugang zu den Korpusdaten über Interfaces und Abfragesprachen zu erschließen, wofür es nach wie vor zu wenig didaktische Materialien für Nutzer:innen mit begrenzten Vorkenntnissen gibt. Eine weitere zentrale Herausforderung besteht in der Entwicklung kritischer Datenkompetenz – etwa dem Bewusstsein, dass jede Korpuskompilierung bestimmte

Beschränkungen aufweist, die sich später in den Ergebnissen niederschlagen, die aus dem Korpus gewonnen werden.

Das Themenheft greift die genannten Herausforderungen auf, indem es Beiträge versammelt, die den Einsatz von Korpora in Forschung und Lehre beleuchten, die Bedeutung von „Corpus Literacy“ reflektieren und Folgen für eine fundierte Korpusdidaktik ableiten. Es vereint Fallstudien, die die Anwendung spezifischer Korpora und Werkzeuge auf konkrete Forschungsfragen demonstrieren, didaktische Szenarien entwickeln, bei denen die Korpusmethoden entweder in die Materialerstellung oder unmittelbar in den Unterricht integriert werden, sowie Diskussionen zu den methodischen und technischen Grundlagen korpusbasierter Forschung. Darüber hinaus enthält es Reflexionen über die Herausforderungen des Aufbaus und der Annotation von spezialisierten Korpora und diskutiert den Einsatz generativer Künstlicher Intelligenz zur Unterstützung linguistischer und literaturwissenschaftlicher Analysen sowie zur Korpusentwicklung.

Auch wenn das Themenheft durch einen offenen Call initiiert wurde, stammen die meisten Beiträge dieses Bandes aus Vorträgen oder praxisorientierten Sitzungen, die im Rahmen von Workshops sowie Forschungs- und Lehraufenthalten im Projekt *Korpusdidaktik für formelhafte Sprache* (KoDi-FS) entwickelt wurden – einem institutionellen Partnerschaftsprojekt zwischen der Universität Vilnius und der Universität Hamburg. Jeder Beitrag wurde in einem anonymen Peer-Review-Verfahren von mindestens zwei Expert:innen sorgfältig begutachtet. Unser besonderer Dank gilt den Gutachter:innen für ihre Zeit und ihr konstruktives Feedback, das entscheidend zur wissenschaftlichen Qualität dieses Heftes beigetragen hat. Ebenso möchten wir dem Deutschen Akademischen Austauschdienst (DAAD) für die Unterstützung mit Mitteln des Auswärtigen Amts danken. Der DAAD fördert das Projekt KoDi-FS seit 2022 und hat in diesem Rahmen auch die Veröffentlichung dieses Themenhefts ermöglicht. Die Veröffentlichung wird auch über das Programm für lituanistische Forschung 2025–2030 (Projekt-Nr. P-LISs-25-62) vom Forschungsrat Litauens finanziert.

Heike Zinsmeister, Vaiva Žeimantienė, Skaistė Volungevičienė und Carla Sökefeld
Hamburg und Vilnius, 2025

Tekstynų teorija ir praktika: prieinami ir patikimi metodai ir įrankiai

Šis teminis numeris praplečia mokslo darbų žurnalo „Kalbotyra“ tematiką už lingvistinės analizės ribų ir išsamiau pristato tekstynų naudojimo galimybes atliekant germanistinius, lietuvių ir kitų kalbų tyrimus, iškaitant literatūros analizę ir svetimosios kalbos mokymo didaktiką. Ypatingas dėmesys skiriamas tekstynų naudojimui akademiniėje aplinkoje, kai tekstyno kalba yra užsienio kalba, pripažistant iššūkius, su kuriais susiduria įvairių lygių besimokantieji. Šiame numeryje taip pat publikuojami straipsniai, kuriuose nagrinėjamas generatyvinio dirbtinio intelekto vaidmuo, konkrečiai vertinant didžiųjų kalbos modelių taikymą filologijos moksliniams tyrimams ir kalboms mokyti.

Lingvistams, literatams ir vokiečių ar kitos kalbos kaip užsienio kalbos ekspertams darbas su empiriniais kalbos duomenimis ar literatūrinių tekstu tekstynais yra nepakeičiamas išteklius tiek moksliniams tyrimams, tiek mokymui. Tekstynai suteikia struktūriuotą prieigą prie kalbos duomenų įvairiomis temomis, prie įvairių tekstu tipų, iškaitant literatūrinius tekstus. Tekstynų kalbiniai duomenys gali būti įvairiais lygiais papildyti lingvistinėmis ir kitomis interpretacinėmis anotacijomis, o tai leidžia ižvelgti teksto paviršiuje įprastai nematomus dėsningumus ir pateikti apibendrinimus bei naujas įžvalgas. Vis dėlto naudotis tekstynais nėra paprasta. Tyrėjai, dėstytojai, mokytojai ir studentai dažnai susiduria su iššūkiais, kai reikia susipažinti su specifine skirtingų tekstynų užklausų formulavimo sintakse ir interpretuoti užklausų rezultatus. Be to, didaktiniai ištekliai vartotojams, turintiems ribotas žinias, vis dar reti. Kitas iššūkis, kurį tenka įveikti, yra kritinis tekstynų duomenų vertinimas, pavyzdžiui, supratimas, kad kiekvienu atveju sudarant tekstynus susiduriama su būdingais apribojimais, kurie atsispindi iš tekstynų gautuose rezultatuose.

Šio numero straipsniuose nagrinėjamas tekstynų naudojimas atliekant mokslinius tyrimus ir mokant kalbų, aptariami tekstynų raštingumo ir didaktikos klausimai. Pateikiami pavyzdžiai, iliustruojantys konkrečių tekstynų naudojimą ir skaitmeninių įrankių taiky-

mą, mokymo scenarijai, kuriuose tekstynų metodai integruojami į mokomąjį medžiagą, taip pat ir tiesiogiai įtraukiant studentus, pristatomos diskusijos apie metodologiją ir techninius pagrindus, kurie reikalingi atliekant tekstynais grįstą tyrimą. Be to, straipsniuose aptariami klausimai, susiję su tekstynų sudarymu ir jų specifiniu anotavimu, svarstomas galimas generatyvinio dirbtinio intelekto vaidmuo atliekant lingvistines ir literatūrines analizes ir kuriant tekstynus.

Dėl straipsnių į šį teminį „Kalbotyros“ numerį buvo paskelbtas atviras kvietimas. Dauguma publikuojamų straipsnių parengti pagal pranešimus, skaitytus seminaruose ir konferencijose, organizuotose įgyvendinant Vilniaus ir Hamburgo universitetų germanistikos katedrų partnerystės projektą „Tekstynų didaktika ir tekstynais grįsti stabiliųjų junginių tyrimai / Korpusdidaktik für formelhafte Sprache“ (KoDi-FS). Kiekvienas straipsnis anonimiškai recenzuotas mažiausiai dviejų ekspertų.

Norėtume padėkoti recenzentams už jų brangų laiką ir konstruktyvius atsiliepimus, kuriie labai prisdėjo prie šio leidinio mokslinės kokybės. Taip pat dėkojame Vokietijos akademinių mainų tarnybai (DAAD), kuri Vokietijos federalinės užsienio reikalų ministerijos lėšomis nuo 2022 m. iki šiol finansuoja projektą „KoDi-FS“ ir parėmė šio numerio leidybą. Leidybą taip pat finansavo Lietuvos mokslo taryba pagal Lituanistikos 2025–2030 m. programą (sutarties Nr. P-LISs-25-62).

Heike Zinsmeister, Vaiva Žeimantienė, Skaistė Volungevičienė ir Carla Sökefeld
Hamburgas ir Vilnius, 2025

Was können Korpora besser als andere digitale Hilfsmittel? Korpusarbeit im fachbezogenen DaF-Unterricht am Beispiel der Rechtssprache

Eglė Kontutytė

Lehrstuhl für Deutsche Philologie
Institut für Sprachen und Kulturen des Ostseeraums
Philologische Fakultät
Universität Vilnius
Universiteto g. 5
LT-01131 Vilnius, Litauen
E-Mail: egle.kontutyte@ffl.vu.lt
ORCID iD: 0000-0002-9097-6124
<https://ror.org/03naddee84>

Abstract. Als ein digitales Medium bei der Spracharbeit im praktischen Umfeld und deshalb auch als digitales Lernmittel beim Fremdsprachenlernen können heutzutage auch Korpora dienen. Deshalb hat sich seit einiger Zeit ein relativ neuer Bereich der Mediendidaktik im DaF-Unterricht die Korpusdidaktik etabliert.

Das Ziel dieses Beitrags ist es, Möglichkeiten der Korpusarbeit beim Erlernen der Fachsprache als Fremdsprache auszuloten und eventuell die Vorteile der Korpusarbeit im Vergleich zu anderen digitalen Medien zu zeigen. Für die korpusdidaktische Arbeit wird das GeLeCo-Korpus der deutschen Rechtssprache herangezogen, so dass für die korpusbasierte Fachsprachendidaktik die Rechtssprache als Beispiel dient.

Anhand eines Beispiels der sprachlichen Markierung der Intertextualität in einer Gerichtsentscheidung wurde gezeigt, ob und eventuell welche Vorteile Korpora im Vergleich zu den Übersetzungsmaschinen oder KI-Chatbots beim Übersetzen der Rechtstexte haben. Die Analyse der sprachlichen Mittel im GeLeCo hat gezeigt, dass sich die in Gerichtsentscheidungen verwendeten sprachlichen Mittel zur Markierung der

Submitted: 27/06/2025. **Accepted:** 18/09/2025

Copyright © 2025 Eglė Kontutytė. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Bezüge auf andere Rechtstexte in ihrer Verwendungshäufigkeit von den sprachlichen Mitteln, die bei der maschinellen Übersetzung vorkommen, unterscheiden. Das stellt die Grundlage für die didaktische Überlegungen und die Entwicklung von Unterrichtsszenarien und Aufgaben dar, damit die Lernenden im rechtsbezogenen DaF-Unterricht dafür sensibilisiert werden, dass die Übersetzungstools und Chatbots grob den Inhalt des Textes wiedergeben, aber die präzisere, die für juristische Handlungen relevanten Nuancen der Rechtstexte nicht ausdrücken.

Schlüsselwörter: Chatbot, DaF-Unterricht, Fachsprache, Fachsprachendidaktik, Korpusdidaktik, maschinelle Übersetzung, Rechtssprache

What can corpora do better than other digital tools?

Corpus-based language learning for specific purposes (LSP) in German as a Foreign Language (GFL), using the example of legal language

Abstract. Corpora are being increasingly used as a digital medium for language work in practical environments and, consequently, as a digital learning tool in foreign language learning. This has given rise to the relatively new area of corpus didactics within media pedagogy for German as a Foreign Language.

This article explores the possibilities of the corpus-based approach to learning languages for specific purposes and its potential advantages in comparison to other digital tools, with a particular focus on legal language. The corpus GeLeCo (German Legal Corpus) is used to illustrate corpus-based methods in LSP didactics.

By using the example of intertextual markers in court decision, the study investigates whether and in what ways corpora might offer advantages over machine translation tools and AI chatbots when translating legal texts. The analysis of linguistic means in the corpus GeLeCo shows that the linguistic means used in court decisions to reference other legal texts differ in frequency from those produced by translation engines.

These findings provide the foundation for didactic considerations and the development of teaching scenarios and tasks designed to raise awareness of learners in law-related GFL learning of the limitations of automated translation tools and chatbots. While such tools might roughly reproduce the general content of a legal text, they often fail to capture the nuanced legal formulations that are critical for professional legal communication.

Keywords: chatbot, corpus didactics, language for specific purposes (LSP), LSP didactics, GFL learning, machine translation, legal language

1 Einleitung

Viele Akteure der Bildung sind sich einig, dass das Lernen heute ohne digitale Technologien, ohne künstliche Intelligenz nicht mehr vorstellbar ist. Laut Brockhorst (2022) ist „KI-basierte Technologie [...] allgegenwärtig und auch aus dem Kontext des Fremdsprachenunterrichts kaum noch wegzudenken“. Köbis (2023, 32–33) hebt aber hervor, dass viele Deutschlehrkräfte in Bezug auf die Rolle der KI im DaF-Unterricht unsicher und skeptisch sind, weil Forschungen und empirische Studien über den didaktischen Einsatz der KI-Tools im DaF-/DaZ-Bereich sowie Best-Practice-Beispiele immer noch fehlen.

Seit einiger Zeit werden in der Didaktik, darunter auch in der Fremdsprachendidaktik, neue Begriffe in Bezug auf den Einsatz der modernen Technologien im Unterricht verwendet. Im Buch von Kerres „Mediendidaktik“ kommen neben dem Begriff „Mediendidaktik“ auch die Begriffe „E-Learning“, „digitales Lernen“ und „computergestütztes Lernen“ vor (Kerres 2024). Der Begriff Mediendidaktik wird schon im Titel des Buches von Kerres (2024) definiert, der „Mediendidaktik – Lernen in der digitalen Welt“ lautet. In der Einleitung beschreibt Kerres (2024, 1) die Aufgabe der Mediendidaktik, die er als „gestaltungsorientierte Mediendidaktik“ präzisiert, folgenderweise: „Der Mediendidaktik geht es nicht um das Digitale an sich, sondern um das didaktische Design von Lernangeboten, die pädagogischen Anliegen gerecht werden“ (Kerres 2024, 1). Aus dieser Definition geht hervor, dass es die Aufgabe der Didaktiker:innen ist, Lernszenarien anhand der Inhalte, die die vielfältige Medienlandschaft bietet, für das Lernen zu entwickeln. Die Fremdsprachendidaktiker:innen konzipieren und erarbeiten die mediengestützten Lernangebote für das effizientere Erlernen von Fremdsprachen. Somit wird der Fremdsprachenerwerb mit dem Erwerb der Medienkompetenz kombiniert. Außerdem steigt laut Flinz (2019, 11) einerseits der Bedarf der Fachsprachen im DaF- und DaZ-Bereich, andererseits haben die Deutschlehrenden nicht immer die Kompetenz in entsprechenden Fächern und fachsprachliche DaF-Lehrwerke sind nicht immer ideal. Die Korpusarbeit als ein Teil der Medienkompetenz bietet im fachbezogenen DaF-Unterricht also die Möglichkeit, die oben genannten Lücken zu schließen.

Das Ziel dieses Beitrags ist es, Möglichkeiten der Korpusarbeit beim Erlernen der Fachsprache als Fremdsprache auszuloten und eventuell die Vorteile der Korpusarbeit im Vergleich zu anderen digitalen Medien zu zeigen. Für die korpusdidaktische Arbeit wird das Korpus der Rechtssprache GeLeCo herangezogen, so dass für die korpusbasierte Fremdsprachendidaktik die Fachsprache Jura als Beispiel dient.

2 Mediendidaktik und Korpusdidaktik im (fachbezogenen) DaF-Unterricht

Obwohl die neuen Medien auch kritisch bewertet werden, wird ihnen in der Fremdsprachendidaktik für die Vermittlung von Fremdsprachen eine wichtige Rolle zugeschrie-

ben (vgl. Beißwenger & Knopp 2019, 10–11). Als Mehrwert der Mediendidaktik im fachbezogenen Fremdsprachenunterricht nennt Roche folgende Punkte: „Optimierung logisch-administrativer Aspekte, die Förderung des Wissenstransfers, die Vermittlung von Handlungskompetenzen, die Förderung von Lernkompetenzen, die Individualisierung und Intensivierung des Lernens durch die Modellierung natürlicher kognitiver Prozesse der Sprachverarbeitung“ (Roche 2018, 241). Laut Efing und Kiefer (2018, 189) ermöglichen neue Medien im fachbezogenen DaF-Unterricht einen „eigenverantwortlichen Lernprozess und individuelle Lernpfade [...]\“. Roche (2008, 18–19) nennt auch neben dem Erwerb von Kompetenzen und der Organisation der Lernprozesse die Bedeutung der neuen Medien bei der Darbietung der Inhalte in einem entsprechenden Fach, über die die Deutschlehrkräfte nicht verfügen, am Beispiel der Fach- und Berufssprachenmodule der Lernplattform Deutsch-Uni Online (DUO). Eine wichtige Rolle ordnet er auch digitalen Wörterbüchern zu, die als Lernwerkzeuge, aber auch später im beruflichen Alltag als Arbeitswerkzeuge eingesetzt werden können (Roche 2008, 26).

Als ein neues digitales Medium bei der Spracharbeit im praktischen Umfeld und deshalb auch als ein digitales Lernmittel können heutzutage Korpora dienen. Als ein relativ neuer Bereich der heutigen Mediendidaktik im DaF-Unterricht hat sich seit einiger Zeit die Korpusdidaktik etabliert. Von einem selbstständigen Zweig und einer wichtigen Rolle der Korpusarbeit in der DaF-Didaktik zeugt die 2021 gegründete Zeitschrift „Korpora Deutsch als Fremdsprache“ (KorDaF)¹. In dem einleitenden Beitrag der 1. Nummer dieser Online-Zeitschrift umreißen die Begründerinnen der Fachzeitschrift Flinz, Hufeisen, Luppi und Mell (2021) die Rolle der Korpora im DaF-Unterricht folgendermaßen:

So bieten Korpora sowohl ein großes wissenschaftliches, aber auch didaktisch-pädagogisches sowie darüber hinaus unterrichtsmethodisches Potenzial: Sie bieten interessante und neue Zugänge für die Sprachbeschreibung, für die Fremd- und Zweitsprachenerwerbsforschung, für die Entwicklung von Nachschlagewerken, von Lehr- und Unterrichtsmaterialien sowie für die Anwendung und Unterrichtspraxis.

Im Weiteren nennen sie die Vorteile der Korpora in der DaF-Didaktik: Die Korpora helfen bei der Produktion und der Rezeption, beim präzisen und normkonformen Sprachgebrauch, bei Formulierungsschwierigkeiten. Ebenso dienen die Korpora im Lernprozess zur Erweiterung des Wortschatzes, zur Förderung der Sprachenbewusstheit, der Lernerautonomie und der Motivation beim Sprachenlernen (Flinz et al. 2021). Außerdem stellt laut Krekeler (2021, 191) die Korpusarbeit im fachbezogenen DaF-Unterricht ein entdeckendes Lernen dar, welches die Lernenden motiviert, neue Lernstrategien vermittelt und ein anderes Deutschlernen ermöglicht. Lüdeling und

¹ Mehr Information über die Zeitschrift unter: <https://kordaf.tujournals.ulb.tu-darmstadt.de/site/about/>. Zugriff: 07.08.2025.

Walter greifen den Begriff *corpus literacy* von Mukherjee (2002, 179–180) auf und stufen sie als eine Form der Medienkompetenz ein, die im Fremdsprachenunterricht erworben wird (Lüdeling & Walter 2009, 2). Sie nennen auch Möglichkeiten der qualitativen und quantitativen Analyse der Korpora für Lehrende und Lernende im DaF-Unterricht. In qualitativer Hinsicht können Korpora für Lehrende als Lernstoffressource und vor allem für nichtmuttersprachliche Lehrende als Hilfsmittel zur Überprüfung sprachlicher Strukturen dienen (Lüdeling & Walter 2009, 6). Die quantitative Analyse hilft den Lehrenden, Häufigkeitslisten der lexikalischen Einheiten zu erstellen sowie typische Wörterkombinationen und grammatische Strukturen zu ermitteln (Lüdeling & Walter 2009, 10–11). Die Lernenden können mithilfe von Korpora ihre produktiven Fertigkeiten verbessern und für unterschiedliche Varietäten einer Sprache sensibilisiert werden, indem sie durch die Bekanntmachung mit verschiedenen Typen von Korpora und mit der Abfrage der Korpusdaten die *corpus literacy* erwerben (Lüdeling & Walter 2009, 8). Laut Lüdeling und Walter (2009, 12–13) kann die Erstellung und die quantitative Analyse von Listen sprachlicher Phänomene in vorhandenen oder selbst zusammengestellten Korpora den Lernenden insbesondere beim Erwerb von Fachsprachen ein wichtiges Werkzeug für die Ermittlung der für eine Fachsprache typischen sprachlichen Phänomene sein. Lüdeling und Walter (2009, 24) weisen darauf hin, dass die Korpusarbeit in der modernen Sprachvermittlung bei der Sensibilisierung der Lernenden für den authentischen Sprachgebrauch eine wichtige Rolle spielt.

Eine der Vermittlung der Fachsprache gewidmete Ausgabe der Zeitschrift „Korpora Deutsch als Fremdsprache“ zeugt auch davon, dass die Korpusdidaktik ein wichtiges Lernmittel im Fachsprachenunterricht sein kann², was Mell (2022, 1) als Herausgeberin dieser thematischen Ausgabe in ihrem einleitenden Beitrag betont. Auch Meier-Vieracker (2022, 11–12) weist am Beispiel der Fußballsprache auf den Mehrwert der Korpusdidaktik im Fachsprachenunterricht bei der „Ermittlung der gebräuchlichen Ausdrücke“ hin, indem den Lernenden Techniken dazu vermittelt werden, wie „mit dem tatsächlichen Fach- und Sondersprachgebrauch rezeptiv und produktiv sicher umzugehen“ ist. Krekeler (2021, 162) macht auf den Nutzen der Korpusarbeit im fachbezogenen DaF-Unterricht aufmerksam, indem durch ein angemessen erstelltes Korpus die Lernenden mit dem authentischen Sprachgebrauch und den für ein Fach typischen sprachlichen Mustern bekannt gemacht werden, was er am Beispiel des Unterrichts Wirtschaftsdeutsch zum Thema „Influencer Marketing“ illustriert. Luppi (2024) befasst sich in ihrem Beitrag mit der korpusbezogenen Vermittlung von Fachsprache im DaF-Unterricht mit dem Schwerpunkt Kollokationenkompetenz am Beispiel des von den Lernenden erstellten Korpus zum Covid-19-Diskurs. Flinz (2019, 11–14) zeigt, wie die von Lehrenden oder Lernenden erstellten Korpora in Forschung

² Mehr Informationen über die Ausgabe der KorDaF-Zeitschrift 2022, Nummer 1. <https://kordaf.tujournals.ulb.tu-darmstadt.de/issue/83/info/>. Zugriff: 07.08.2025.

und Lehre genutzt werden können. Somit werden die Möglichkeiten und der Nutzen der Korpora im fachbezogenen Fremdspracheunterricht aufgezeigt.

Wiesmann (2022, 46) stellt heraus, dass bei der Vermittlung der Rechtssprache und bei der Rechtsübersetzung, die sich durch die Besonderheit auszeichnet, dass die einzelnsprachlichen Rechtstexte an das jeweilige Rechtssystem gebunden sind (vgl. de Groot 1999, 204–206; Wiesmann 2022, 44–45), die Korpusanalyse sowohl der vorhandenen als auch der selbst zusammengestellten Korpora eine wichtige Rolle spielt. Sie stellt das GeLeCo, ein Korpus der Rechtssprache, und seine Einsatzmöglichkeiten bei der Übersetzungsausbildung vor (Wiesmann 2022).

3 Bezugnahme in der Rechtssprache und ihre Recherche im GeLeCo als Korpusarbeit

Das GeLeCo als Korpus der Rechtssprache kann auch als ein Mittel zur Ermittlung von sprachlichen Phänomenen im DaF-Unterricht der Rechtssprache dienen. Im folgenden Kapitel werden Überlegungen angeführt, wie das GeLeCo im fachbezogenen Fremdsprachenunterricht eingesetzt werden kann, damit Studierende mit der Korpusarbeit vertraut werden und lernen, eine kritische Datenrecherche durchzuführen, sowie die Vorteile der Korpusarbeit kennenlernen.

3.1 GeLeCo als Korpus der deutschen Rechtssprache

Das GeLeCo³ (German Legal Corpus) ist ein Korpus der deutschen Rechtssprache, das an der Universität Bologna im Jahr 2021 erstellt wurde. Es umfasst 196.107.719 Tokens, 144.462.122 Wörter, 11.260.450 Sätze und 62.695 Texte. Das Korpus enthält drei Textsorten der bundesdeutschen Rechtssprache, die dem Internet entnommen sind: Gesetze, Verwaltungsvorschriften und Gerichtsentscheidungen. Die Korpustexte sind über die Online-Plattform von NoSketchEngine öffentlich zugänglich (Wiesmann 2022).

3.2 Bezugnahme als Phänomen der Rechtssprache

Als Gegenstand der Recherche im GeLeCo wird das Phänomen der Intertextualität gewählt, durch die sich Rechtstexte auszeichnen. Wie Morlok hervorhebt, ist „die Arbeit des Juristen [...] durch und durch intertextuell. Das normale Geschäft des Juristen besteht darin, Texte auf andere Texte zu beziehen und daraus einen neuen Text zu fertigen [...]“ (Morlok 2015, 69). Er weist darauf hin, dass die „Bausteintechnik“ der Rechtstexte, indem auf Vor-Texte Bezug genommen wird, eine übliche Praxis der Rechtssprache ist, zur „dogmatische[n] Durchdringung des Rechts“ beiträgt und somit die Qualität der juristischen Entscheidungen verbessert (Morlok 2015, 78). Wie Ballansat

³ Das GeLeCo ist abrufbar unter: <https://bellatrix.sslmit.unibo.it/noske/public/#dashboard?corpname=geleco>. Zugriff: 07.08.2025.

und Perrin (2007, 273) betonen, stellen ausdrückliche Verweise auf andere Rechtstexte eine der Ebenen der Intertextualität in Rechtstexten dar. Da laut Morlok (2015, 78) für Rechtssprache Standardisierungen wichtig sind und sie die Rechtssicherheit gewährleisten, werden im „Handbuch der Rechtsförmlichkeit des Bundesministeriums der Justiz“ (2024, 160) vor allem die Formulierungen mit den Präpositionen *nach* und *gemäß* als sprachliche Mittel angeführt, die Bezüge auf andere Rechtstexte markieren, was die Beispiele aus Gerichtsentscheidungen im GeLeCo veranschaulichen:

- (1) *Auch die richterlichen Hinweis- und Aufklärungspflichten **gemäß** § 139 ZPO sollten regelmäßig analog herangezogen und beachtet werden.* (GeLeCo)
- (2) *Dabei ist zunächst ein Hausbesuch und später laut Abänderungsbeschluss vom 22.6.2012 eine ambulante Untersuchung vorgesehen gewesen.* (GeLeCo)

Im Hintergrund der standardisierten Formulierungen der Rechtssprache wird aus der Perspektive der Didaktik der Rechtssprache als Fremdsprache nicht nur die Fragestellung aufgeworfen, welche sprachlichen Markierungen der Intertextualität in deutschen Rechtstexten vorkommen, sondern auch wann und in welchen Rechtstextsorten. Dieses Wissen ist für Nichtmuttersprachler:innen wichtig, die z. B. im Studium⁴ auf die Beschäftigung mit deutschen Rechtstexten vorbereitet werden. Da heutzutage im digitalen Raum viele verschiedene Medien und Tools für die Übersetzung und andere Arbeit mit Texten angeboten werden, stellt sich bei der Vermittlung der Rechtssprache als Fremdsprache die Frage, wie ein Fachsprachenkorpus wie das GeLeCo genutzt werden kann und welche Vorteile es im Vergleich zu maschinellen Übersetzungsmöglichkeiten, beispielsweise auch durch KI-Chatbots, bietet.

4 Exemplarische Beispielanalyse der intertextuellen Markierung

In diesem Kapitel werden anhand des Beispiels der sprachlichen Markierung der Intertextualität didaktische Überlegungen angestellt, ob und eventuell welche Vorteile Korpora im Vergleich zur maschinellen Übersetzung beim Übersetzen von Rechtstexten haben. Als exemplarischer Beleg dient ein Auszug aus der Entscheidung des Verfassungsgerichts der Republik Litauen zur Verfassungskonformität eines Satzes im Waffenbesitzgesetz, indem der Bezug auf das Gesetz zur Waffenhaltung mit einer in der litauischen Sprache typischen Formulierung *remdamasi* ‘bezugnehmend’ explizit markiert wird:

⁴ An der Universität Vilnius besteht z. B. seit 2006 der interdisziplinäre Masterstudiengang „Fachsprache Jura“ (Unterrichtssprache Deutsch oder Englisch), wo Linguist:innen mit der Unterstützung der Jurist:innen der Universität Vilnius den Studierenden vielfältige rechtssprachliche Kompetenzen vermitteln (vgl. <https://www.vu.lt/studijos/stojantiesiems/magistro-studiju-sarasas/dalykine-teises-kalba>. Zugriff: 07.08.2025).

(3) Pareiškėjas pateikė prašymą atitinkamai policijos įstaigai pratęsti jam išduoto leidimo laikyti (nešiotis) ginklus galiojimą, tačiau policijos įstaiga, remdamasi (‘bezugnehmend’ → **Partizipialphrase in der adverbialen Funktion**) Ginklų ir šaudmenų kontrolės įstatymo (2019 m. birželio 27 d. redakcija) (toliau – *Istatymas*) 40 straipsnio 1 dalies 10 punktu, nusprendė neperregistruoti jam išduoto leidimo. (<https://lrkt.lt/lt/teismo-aktai/paieska/135/ta2803/content>, Zugriff: 07.08.2025)

‘Der Beschwerdeführer beantragte bei der zuständigen Polizeibehörde die Verlängerung seiner Erlaubnis zum Besitz (zum Führen) von Waffen, aber die Polizeibehörde beschloss gemäß Art. 40 Abs. 1 Nr. 10 des Gesetzes über die Kontrolle von Waffen und Munition (die Fassung vom 27. Juni 2019) (im Folgenden „das Gesetz“), die ihm ausgestellte Erlaubnis nicht erneut zu registrieren.⁵’

Die Abbildungen 1 und 2 zeigen die Übersetzung des Auszugs ins Deutsche durch die Übersetzungsmaschinen „Google Translator“ und „DeepL Translator“, wo in beiden Übersetzungen die maschinelle Übersetzung das sprachliche Mittel der Markierung der Intertextualität „auf (der) Grundlage“ bietet⁶.

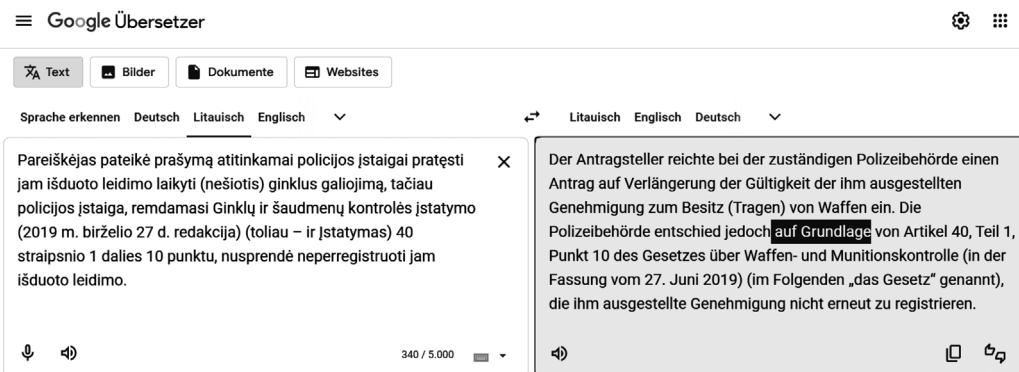


Abb. 1. Übersetzung von „Google translator“ (Zugriff: 26.02.2025)

⁵ Übersetzung der Autorin des Beitrags.

⁶ An dieser Stelle kann darauf hingewiesen werden, dass 2023 in der durch „Google Translator“ übersetzten Variante dieses Auszugs aus der Entscheidung des Verfassungsgerichts der Republik Litauen die Bezugnahme auf das entsprechende Gesetz mit dem Verb *sich stützen* übersetzt wurde. Die Formulierung *auf (der) Grundlage* im Jahr 2025 könnte ein Indiz für die gestiegene Übersetzungsqualität der Übersetzungsmaschinen sein.

The screenshot shows the DeepL Translator website. At the top, there are navigation links for 'Translator', 'Products', 'Solutions', 'Pricing', 'Apps', and user options like 'Log in' and 'Start free trial'. Below the header are three buttons: 'Translate text' (33 languages), 'Translate files' (.pdf, .docx, .pptx), and 'DeepL Write' (AI-powered edits). The main content area shows a translation from Lithuanian to German. The Lithuanian text discusses a police officer's request for a permit extension, mentioning Article 40, Section 1, Article 27, and Article 10. The German translation states that the applicant requested an extension at the police station, but the station decided not to renew it based on Article 40, Section 1, Article 10, and Article 27 of the Weapons Control Act (GKMG) from June 2019.

Abb. 2. Übersetzung von „DeepL Translator“ (Zugriff: 26.02.2025)

Die Phrasen-Recherche im GeLeCo bei der Textsorte „Gerichtsentscheidung“ ergab, dass die Wortgruppe *auf der Grundlage* unter anderem auch intertextuelle Bezüge auf die Gesetzestexte markiert, was die Paragraphenzeichen, Abkürzungen von Gesetzen oder von anderen Rechtequellen zeigen (z. B. *auf der Grundlage von § 21 TKG, auf der Grundlage von § 9 BV*) (s. Abb. 3).

The screenshot shows the GeLeCo CONCORDANCE search interface. The search term 'phrase auf der Grundlage' is entered, with a count of 27,165 occurrences and a rate of 138.52 per million tokens. The results list 12 entries, each with a checkbox and a preview of the text context. The results show various legal contexts, such as 'auf der Grundlage' of specific articles (e.g., § 21 TKG, § 9 BV) or 'auf der Grundlage' of entire laws (e.g., 'auf der Grundlage' of the BGB). The interface includes a sidebar with various search filters and a toolbar with icons for search, download, and other functions.

Abb. 3. Ergebnisse der Recherche der Phrase *auf der Grundlage* in Gerichtsentscheidungen im GeLeCo

Eine detaillierte Recherche im Korpus zeigt, dass beim Zitieren der Rechtsquellen auch die Präposition *nach* verwendet werden kann: In einem Korpusauszug mit dem Intertextualitätshinweis *auf der Grundlage* kommt die Präposition *nach* im Satz *Nach § 106 Satz 1 GewO hat der Arbeitgeber sein Weisungsrecht nach billigem Ermessen auszuüben* als Markierung des Verweises auf eine Rechtsquelle vor (s. Abb. 4). Die

Präposition *nach* wird auch als ein sprachliches Mittel des Zitierens im „Handbuch der Rechtsförmlichkeiten“ (2024) angegeben.

The screenshot shows the CONCORDANCE software interface. At the top, the title 'CONCORDANCE' is displayed, followed by the search term 'GeLeCo' and a search icon. Below the search bar, a message box shows 'Text types 1 (1) ... phrase auf der Grundlage • 27,165 138.52 per million tokens • 0.014%' with a help icon. The main interface features a table with columns: 'Left context', 'KWIC', and 'Right context'. The 'Left context' column contains numbered rows 1 through 9, each with a checkbox and a detailed description of a sentence from a document. The 'KWIC' and 'Right context' columns are currently empty. On the right side of the interface, there is a vertical sidebar with a list of terms and their definitions, such as 'freigest', 'tgleit für', 'üblichen', 'TKG (m)', 'eklagte zi', 'ungen ke', 'echnunge', and 'von der'. The bottom of the interface shows a navigation bar with icons for file operations.

Abb. 4. Ein exemplarisches Beispiel der Phrase *auf der Grundlage* im Kontext im GeLeCo

In einem weiteren Schritt wurde eine CQL-Recherche im GeLeCo nach einem breiteren Kontext durchgeführt, indem durch die Abfrage [word=„auf“][word=„der“] [word=„Grundlage“]{} {2,5}[tag=„N.*“] nach einer nominalen Phrase gesucht wurde, die mit einem Wortabstand von zwei bis fünf zusammen mit der Phrase *auf der Grundlage* auftritt. Wie aus den Konkordanzen in Abbildung 5 zu erschließen ist, verdeutlichen die Paragraphenzeichen und die Abkürzungen von Rechtsquellen wie TKG (Telekommunikationsgesetz) und BV (Betriebsvereinbarung), dass die Wortgruppe *auf der Grundlage* den Bezug auf Gesetze und andere Rechtsquellen markiert.

CONCORDANCE		GeLeCo	?	
Text types 1 (1) ...		CQL [word="auf"] [word="der"] [word="Grundlage"] [[(2,5)[tag... • 30,124 153.61 per million tokens • 0.015%	?	
		?	?	
□	Details	Left context	KWIC	
			Right context	
1	□	① #667673 affelungsregelungen	auf der Grundlage dieser zum 1.1.2009 in Kraft	getretenen Vorschrift
2	□	① #677936 rmel) und unterwarf	auf der Grundlage von § 9 Abs	.2 , § 13 Abs . 1 und :
3	□	① #678591 s>Sie kann vielmehr	auf der Grundlage des Wortlauts der §§	21 und 25 TKG mit Hi
4	□	① #678623 lesnetzagentur kann	auf der Grundlage von § 21 TKG	(i.V.m. § 9 Abs . 2 un
5	□	① #684878 /s><s>Während der	auf der Grundlage von § 9 BV	2011 angeordneten P
6	□	① #685074 ier Kläger hat für die	auf der Grundlage von § 9 BV	2011 angeordneten A
7	□	① #687018 e von der Beklagten	auf der Grundlage von § 9 BV	2011 angeordneten A
8	□	① #690263 sen Betrag habe sie	auf der Grundlage einer eigenen Berechnung	unter Bezugnahme au
9	□	① #690263 sen Betrag habe sie	auf der Grundlage einer eigenen Berechnung unter Bezugnahme	auf von dem Sachver
10	□	① #690491 schädigungsangebot	auf der Grundlage einer beschränkten Berechnung	der Klägerin erstellt h
11	□	① #690491 schädigungsangebot	auf der Grundlage einer beschränkten Berechnung der Klägerin	erstellt habe , nämlich
12	□	① #695202 Januar 2011 werden	auf der Grundlage des im Beschwerdeverfahren	eingeschränkten War

Abb. 5. Ergebnisse der COL-Recherche *auf der Grundlage* im GeLeCo

Die Frequenzeinstellung der Phrase *auf der Grundlage* macht deutlich, dass diese Phrase nicht nur Bezüge auf konkrete Rechtstexte markiert: In der Lemma-Liste ab dem 3. Platz rangieren Muster ohne Paragraphenzeichen wie *auf die Grundlage die bisherig Feststellung*, *auf die Grundlage die materiellrechtlichen Standpunkt* u. a., obwohl die Muster *auf Grundlage von § @card@ Abs* und *auf Grundlage die § @card@ Abs* relativ häufig beim Zitieren der Gesetze oder anderer Rechtsvorschriften verwendet werden, da sie jeweils mit 699 bzw. 666 Belegen den ersten und zweiten Platz in der Liste belegen und somit Bezüge mit Paragraphenzeichen deutlich häufiger als die Belege ohne Paragraphenzeichen ausdrücken. Solche relativ frequent vorkommenden Muster wie *auf die Grundlage die einschlägig Rechtsprechung* oder *auf die Grundlage eine vorherig Entscheidungspraxen* markieren zwar die Intertextualität, sie verweisen aber nicht auf konkrete Texte, sondern eher auf Textwelten⁷ (s. Abb. 6).



Abb. 6. Ergebnisse der Frequenz der Phrase *auf der Grundlage* im GeLeCo

Als ein weiteres sprachliches Mittel zur Markierung von Zitaten wird im „Handbuch der Rechtsförmlichkeiten“ (2024) die Präposition *gemäß* genannt. Wenn man aber den Gebrauch der Präposition *gemäß* im GeLeCo überprüft, so zeigen die CQL-Abfrageergebnisse, dass diese Präposition in den Gerichtsentscheidungen meistens Verweise auf Gesetze oder andere Rechtsvorschriften zum Ausdruck bringt: Das zeigen die Phrasen *gemäß §§ 64 Abs, gemäß § 8 Abs, gemäß § 8 Abs . 2 Nr, gemäß Art 3 Abs 1 GG* u. ä. in den Konkordanzen der Korpusabfrage.

Im Vergleich zu dem Intertextualitätsmarker *auf der Grundlage* stellen die Muster mit der Präposition *gemäß* eine größere Vielfalt innerhalb der Musterliste mit dem Para-

⁷ Fix unterscheidet als Typen der Intertextualität die Text-Text-Beziehung, die Text-Text-muster-Beziehung sowie die Text-Textwelt-Beziehung, die den Verweis in einem Text auf andere, nicht eindeutig identifizierbare Texte wiedergibt (Fix 2000, 449).

phenzeichen dar. Außerdem weisen die *gemäß*-Muster mit 254.571 Belegen eine viel größere Frequenz in den Gerichtsentscheidungen im GeLeCo gegenüber den Mustern mit *auf der Grundlage* auf, die mit 30.124 Belegen im GeLeCo vertreten sind (s. Abb. 7).

CONCORDANCE GeLeCo 🔍 ⓘ

Text types 1 (1) ... CQL [word="gemäß"][(2,5)][tag="N."] • 254,571 1,298.12 per million tokens • 0.13%

Details Left context KWIC Right context

1	□ ⓘ #663852 <> Die Beschwerde der IR-Markeninhaberin ist	gemäß §§ 64 Abs	. 6 , 66 MarkenG zulässig und hat auch in der S:
2	□ ⓘ #664656 ten Anforderungen an die Unterscheidungskraft	gemäß § 8 Abs	. 2 Nr . 1 MarkenG genügt die IR-Marke</s><s>
3	□ ⓘ #664656 ten Anforderungen an die Unterscheidungskraft	gemäß § 8 Abs . 2 Nr	. 1 MarkenG genügt die IR-Marke</s><s>.</s><s>
4	□ ⓘ #666779 itung zu entscheiden</s><s> (Zulassungsgrund	gemäß § 160 Abs	2 Nr 1 SGG)</s><s>.</s><s>5</s><s>II . Die B
5	□ ⓘ #666779 itung zu entscheiden</s><s> (Zulassungsgrund	gemäß § 160 Abs 2 Nr	1 SGG)</s><s>.</s><s>5</s><s>II . Die Besch
6	□ ⓘ #668419 ie , ob es mit dem Gleichbehandlungsgrundsatz	gemäß Art 3 Abs	1 GG und speziell mit dem Grundsatz der Honor
7	□ ⓘ #668419 ie , ob es mit dem Gleichbehandlungsgrundsatz	gemäß Art 3 Abs 1 GG	und speziell mit dem Grundsatz der Honorarvert
8	□ ⓘ #670150 die Eingliederungshilfe in seiner Eigenschaft als	gemäß § 97 Abs	. 1 i.V.m.</s><s>Abs . 2 SGB XII i.V.m. § 2 Abs .
9	□ ⓘ #670150 die Eingliederungshilfe in seiner Eigenschaft als	gemäß § 97 Abs . 1 i.V.m.	</s><s>Abs . 2 SGB XII i.V.m. § 2 Abs . 1 Nr . 1
10	□ ⓘ #670507 <s>Internat im Sinne des § 6 Abs . 1 HärteV ist	gemäß § 6 Abs	. 2 Satz 1 HärteV ein der besuchten Ausbildungs
11	□ ⓘ #670507 <s>Internat im Sinne des § 6 Abs . 1 HärteV ist	gemäß § 6 Abs . 2 Satz	1 HärteV ein der besuchten Ausbildungsstätte ar
12	□ ⓘ #677007 Ihrerin beantragt , die Erstattung Ihrer Auslagen	gemäß § 34a Abs	. 3 BVerfGG der Bundesrepublik Deutschland au

Abb. 7. Belege für den Gebrauch von *gemäß* in Gerichtsentscheidungen im GeLeCo

Auch die Frequenzanalyse bestätigt, dass die Präposition *gemäß* häufiger beim Zitieren der Gesetze auftritt, da in der Lemma-Liste die sprachlichen Muster mit Paragraphenzeichen und Abkürzungen der Strukturteile der Gesetze (z. B. *Abs*, *Satz*, *nr*) und Bezeichnungen der Gesetze oder anderer Vorschriften (z. B. ZPO) auf den ersten 15 Plätzen rangieren (s. Abb. 8).



Abb. 8. Ergebnisse der Frequenzanalyse der Muster mit der Präposition *gemäß* im GeLeCo

Die Präposition *nach* ist polysem und kann in unterschiedlichen Kontexten verwendet werden. Auch in den Gerichtsentscheidungen im GeLeCo dient *nach* nicht nur zur Markierung der Bezüge auf andere Texte wie *nach § 8 Abs*, sondern tritt auch in der modalen Bedeutung *entsprechend* (z. B. *ihrer Herkunft nach*), in der lokalen Bedeutung (z. B. *nach einem syntaktisch vollständigen Aussagesatz*) oder in der temporalen Bedeutung (z. B. *nach Inkrafttreten entsprechender Regelung*, s. Abb. 9) auf.

The screenshot shows a search interface for the GeLeCo corpus. The search term is 'nach'. The results table has four columns: 'Left context', 'KWIC', and 'Right context'. The 'Left context' column contains numbered entries with checkboxes. The 'KWIC' and 'Right context' columns show the surrounding text for each entry. The interface includes various search filters and a KWIC dropdown menu.

	Left context	KWIC	Right context
1	□ ① #664523 er Dienstleistungen ihrer Herkunft	nach zu unterscheiden , solchen Angaben	nicht zu , die aus gebräuchlichen W
2	□ ① #665137 ernis der beschreibenden Angabe	nach § 8 Abs	. 2 Nr . 2 MarkenG gilt dies entspre
3	□ ① #665137 ernis der beschreibenden Angabe	nach § 8 Abs . 2 Nr	. 2 MarkenG gilt dies entsprechend
4	□ ① #665271 's><s>im Regelfall steht ein Punkt	nach einem syntaktisch vollständigen Aussagesatz	, nach Satzverbindungen und Satzgefügen
5	□ ① #665277 -isch vollständigen Aussagesatz ,	nach Satzverbindungen und Satzgefügen	.</s><s>Der Punkt hat in der Werbu
6	□ ① #665277 -isch vollständigen Aussagesatz ,	nach Satzverbindungen und Satzgefügen .</s><s>Der Punkt	hat in der Werbung die Aufgabe , ei
7	□ ① #666056 R-Marke auch ein Schutzhindernis	nach § 8 Abs	. 2 Nr . 2 MarkenG nicht bejaht wer
8	□ ① #666056 R-Marke auch ein Schutzhindernis	nach § 8 Abs . 2 Nr	. 2 MarkenG nicht bejaht werden. <
9	□ ① #666449 streitbefangene Quartal setzte sie	nach den Feststellungen des LSG	auf insgesamt 76 362 , 61 Euro fes
10	□ ① #667615 Ärzten zu differenzieren , die erst	nach Inkrafttreten entsprechender Regelung	ihre Praxistätigkeit so ausweiten ,

Abb. 9. Ergebnisse des Gebrauchs der Präposition *nach* in Gerichtsentscheidungen im GeLeCo

Die Frequenzanalyse der Verwendung der Präposition *nach* ergibt, dass *nach* jedoch in Gerichtsentscheidungen häufig Zitate der Gesetze oder anderer Rechtsvorschriften markiert, was Paragraphenzeichen und Abkürzungen der Strukturteile der Gesetze wie *Abs* oder Gesetzesabkürzungen verdeutlichen, z. B. *nach § @card@ Abs*, *nach § @cadr@ Abs . @cadr@ Satz*, *nach @cadr@ BGB*. Solche sprachlichen Muster belegen in der Lemma-Liste die Plätze 1 bis 13 (s. Abb. 10).

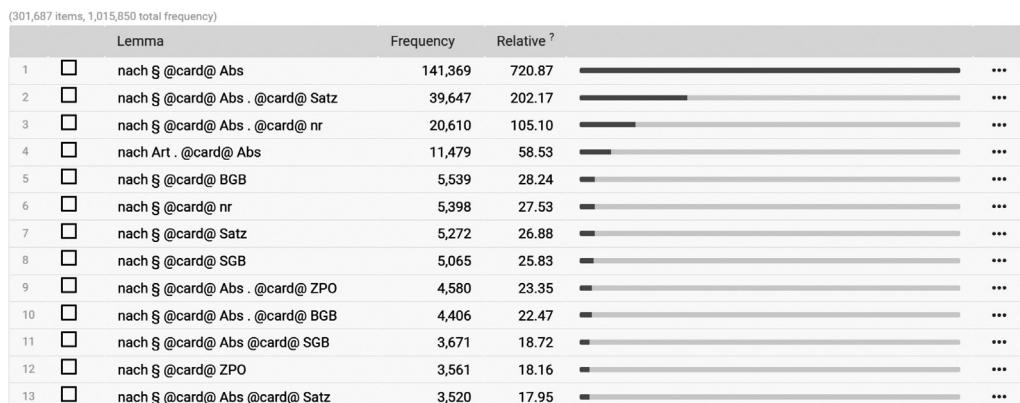


Abb. 10. Ergebnisse der Frequenzanalyse der Muster mit der Präposition *nach* im GeLeCo

Der Vergleich des Gebrauchs der Präpositionen *gemäß* und *nach* in Bezug auf ihre Funktion in Gerichtsentscheidungen zeigt, dass sie beide meistens beim Verweisen auf Gesetzestexte oder ähnliche Rechtsvorschriften vorkommen: Die sprachlichen Muster *nach § @card@ Abs*, *gemäß § @card@ Abs*, *nach § @card@ Abs . @card Satz*, *gemäß § @card@Abs . @card Satz* und eine ähnliche Struktur aufweisende Muster machen die ersten 11 Positionen der Lemma-Liste aus (s. Abb. 11). Den Daten der Frequenzanalyse ist zu entnehmen, dass die Präposition *nach* sogar häufiger in der Funktion des Zitierens als die Präposition *gemäß* auftritt: Das auf dem ersten Platz rangierende sprachliche Muster *nach § @card@ Abs* mit der Frequenz 141.369 kommt fast doppelt so häufig in Gerichtsentscheidungen des GeLeCo vor wie das sprachliche Muster *gemäß § @card@ Abs*, dessen absolute Häufigkeit 73.974 ist. Das sprachliche Muster *nach § @card@ Abs . @card Satz* mit der Frequenz 39.647 wird auch fast zweimal häufiger als die gleiche Konstruktion mit der Präposition *gemäß* verwendet, deren Frequenz 23.536 in Gerichtsentscheidungen ist. Ein ähnliches Verhältnis weisen auch die sprachlichen Muster *nach § @card@ Abs . @card@ nr* mit der Häufigkeit 20.610 und *gemäß § @card@ Abs . @card@ nr* mit der Häufigkeit 11.570 auf. Das sprachliche Muster *nach Art . @card@ Abs*, dessen Frequenz 11.479 ist, kommt auch halb so oft wie das gleiche Muster mit der Präposition *gemäß* mit der Häufigkeit 5.983 vor. Die Plätze 9, 10 und 11 belegen die sprachlichen Muster mit der Präposition *nach*: *nach § @card@ BGB* (5.539 Belege), *nach § @card@ nr* (5.398) und *nach § @card@ Satz* (5.272), die auch ihrer Struktur nach die Bezüge auf Gesetze und andere Vorschriften ausdrücken (s. Abb. 11).

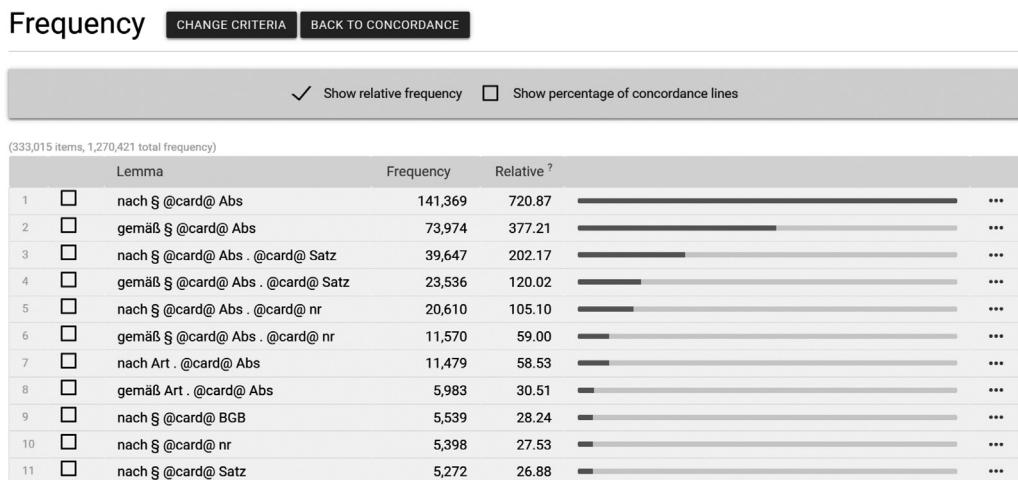


Abb. 11. Frequente Muster mit den Präpositionen *gemäß* und *nach* in Gerichtsentscheidungen im GeLeCo

Eine vergleichende Häufigkeitsanalyse der die Verweise auf andere Texte markierenden Redemittel zeigt, dass in Gerichtsentscheidungen die Phrase *auf der Grundlage* 2,59 %

der drei analysierten Redemittel ausmacht, während die Phrasen mit der Präposition *nach* mit großem Abstand mit 81,2 % auf Platz 1 rangieren. Die Präposition *gemäß* bildet 16,13 % der analysierten Fälle (s. Abb. 12). Dabei sollte man in Betracht ziehen, dass die Präposition *nach* nicht nur Verweise auf andere Texte markiert, sondern auch in anderen Funktionen auftritt, so dass ihre Häufigkeit in der Intertextualitätsfunktion anhand einer Stichprobe überprüft werden sollte. Die Wortgruppe *auf der Grundlage* bildet nur einen kleinen Teil der drei analysierten Redemittel, die Verweise auf andere Texte ausdrücken.

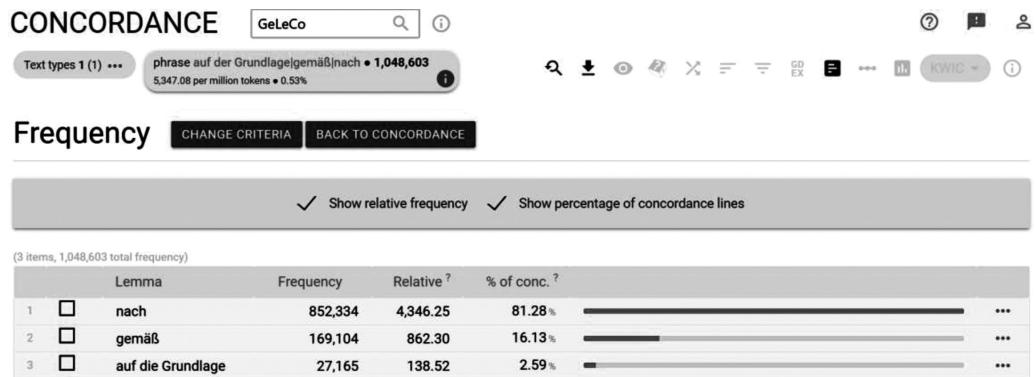


Abb. 12. Vergleich der Verwendung von *nach*, *gemäß*, *auf der Grundlage* in Gerichtsentscheidungen im GeLeCo

Die ersten 500 Belege mit der Präposition *nach* aus den Gerichtsentscheidungen im GeLeCo wurden einer Stichprobenanalyse in Bezug auf die Funktion der Präposition *nach* unterzogen. Die manuelle Untersuchung der Belege ergab, dass nur etwas mehr als Hälfte (277 Belege, rund 55 %) der untersuchten Phrasen die Verweisfunktion haben. Die Textsorten, auf die mit der Präposition *nach* Bezug genommen wird, sind Gesetz, Verordnung, Anspruch, Antrag, Plan, polizeiliche Angaben, Ausführungen des Sachverständigen, Arztattest, Bescheinigung, Gutachten, Satzung, Patentschrift, Verfügung, Beschluss, Vorbringen des Klägers, Urteil, Prognose, Entscheidung, Vertrag und seine Anlage, Bestimmung oder Rechtsprechung und Recht, Theorie oder Erkenntnisse als Hyperonyme. Diese Analyse zeigt, dass die Präposition *nach* häufig zum Ausdruck anderer Funktionen in den Gerichtsentscheidungen verwendet wird. Die anderen Funktionen, die mit der Präposition *nach* ausgedrückt werden, sind die Temporalität (z. B. **nach fachkundiger Beratung**, **nach Inkrafttreten entsprechender Regelung**, **nach dem Zeitpunkt**, **nach 1:45 Uhr**, **nach der Tatnacht**), die Lokalität (z. B. **nach Hause**, **nach Deutschland**, **nach Satzverbindungen**) und die Modalität (z. B. **ihrer Herkunft nach**, **je nach Lage ihres Wohnortes**, **nach allgemeinen Maßstäben**, **nach Quadratmeter und Verwendungszweck**, **nach Treu und Glauben**, **nach Maßgabe**, **nach Grund und Höhe**, **nach Ansicht**). Letztere stellt häufig feste Wortverbindungen dar. Außerdem wird *nach* als

Präposition eines präpositionalen Objekts verwendet (z. B. *nach Ärzten mit über- und mit unterdurchschnittlichen Fallzahlen zu differenzieren*) oder als trennbares Präverb (z. B. *Kommt er dem nicht fristgerecht nach*, [...]). Da etwa in der Hälfte der untersuchten Phrasen *nach* die Verweisfunktion und in der anderen Hälfte der Belege andere Funktionen aufweist, dient die Präposition *nach* doch relativ frequent als Intertextualitätsmarker in Gerichtsentscheidungen im Vergleich zu ihren anderen Funktionen und zu den anderen untersuchten Intertextualitätsmarkern, da *gemäß* 16,13 % und *auf der Grundlage* nur 2,59 % aller Belege mit den drei Redemitteln ausmacht.

Da in der letzten Zeit immer häufiger für die Übersetzung von verschiedenen Texten generative KI-Modelle als Hilfsmittel herangezogen werden, sollte im nächsten Schritt auch ihre Zuverlässigkeit aufgrund der aus dem GeLeCo gewonnenen Daten überprüft werden. Dazu wurde ein Prompt als Übersetzungsauftrag für zwei Chatbots, ChatGPT und Copilot, formuliert. In der laut ChatGPT juristisch präziseren Variante der Übersetzung wurde *unter Berufung auf* als ein den Verweis auf andere Texte markierendes Redemittel in folgendem Satz gebraucht: *Die Polizeibehörde lehnte jedoch die Verlängerung ab und verweigerte die Neuregistrierung der Erlaubnis **unter Berufung auf** Artikel 40 Absatz 1 Punkt 10 des Gesetzes über die Kontrolle von Waffen und Munition in der Fassung von 27. Juni 2019*⁸ (s. Abb. 13).

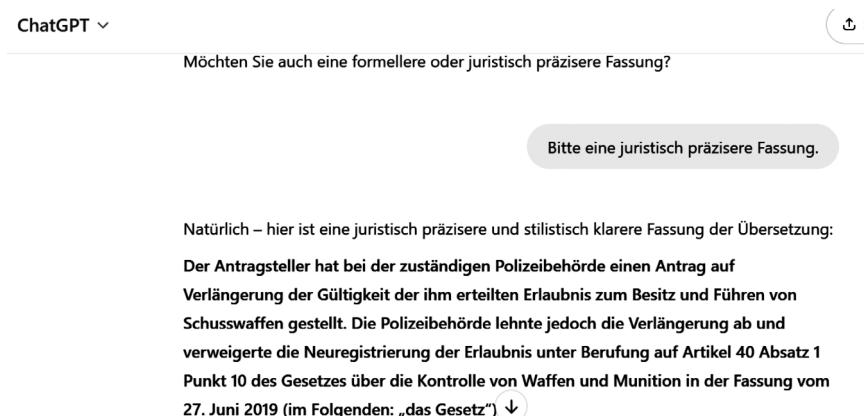


Abb. 13. Übersetzung von ChatGPT des Abschnittes aus der litauischen Gerichtsentscheidung ins Deutsche

Das Chatbot Copilot verwendet das Verb *sich berufen* für die Markierung der Bezugnahme auf das Gesetz im Satz: *Die Polizeidienststelle entschied jedoch, **sich auf** Artikel 40 Absatz 1 Punkt 10 des Waffenkontrollgesetzes [...] zu berufen und die erteilte Genehmigung nicht erneut zu registrieren* (s. Abb. 14).

⁸ Andere sprachliche Mittel und Termini im übersetzten Satz werden nicht kommentiert.

Hier ist die Übersetzung des Satzes ins Deutsche:

Der Antragsteller stellte bei der zuständigen Polizeidienststelle einen Antrag auf Verlängerung der ihm erteilten Genehmigung zur Aufbewahrung (Tragen) von Waffen. Die Polizeidienststelle entschied jedoch, sich auf Artikel 40 Absatz 1 Punkt 10 des Waffenkontrollgesetzes (Fassung vom 27. Juni 2019) (nachfolgend – das Gesetz) zu berufen und die erteilte Genehmigung nicht erneut zu registrieren.

Falls du eine Anpassung oder eine genauere Formulierung brauchst, lass es mich wissen!

Message Copilot

Abb. 14. Übersetzung von Copilot des Abschnittes aus der litauischen Gerichtsentscheidung ins Deutsche

Die Abfragen in den Gerichtsentscheidungen des GeLeCo zeigen, dass das Verb *sich berufen* selten Bezüge auf andere Texte markiert, denn dieses Verb wird meistens mit den Nominalphrasen *die Kläger, die Klägerin, die Arbeitgeber, die Beklagte, die Revision, ohne Erfolg* sowie mit einem Verb oder einer Verbkonstruktion *können, können, um eine identisch Entscheidung, müssen eine Rechtsfrage* im sprachlichen Muster verwendet. Die sprachlichen Muster, in denen die Funktion der Bezugnahme durch Paragraphenzeichen und die Abkürzungen der Strukturteile der Gesetze zu erkennen ist, (z. B. *berufen . @card@ aa* und *berufen (§ @card@ Abs)*), stehen in der Liste der Abfrageergebnisse im GeLeCo jeweils auf Platz 9 und 12 (s. Abb. 15).

(14,161 items, 22,567 total frequency)

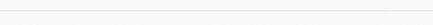
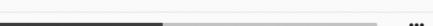
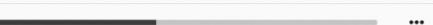
	Lemma	Frequency	Relative ?	% of conc. ?	
1	□ berufen sich die Kläger	179	0.91	0.79 %	
2	□ berufen sich die Klägerin	173	0.88	0.77 %	
3	□ berufen sich die Arbeitgeber	162	0.83	0.72 %	
4	□ berufen sich die Beklagte	161	0.82	0.71 %	
5	□ berufen können (vgl	128	0.65	0.57 %	
6	□ berufen sich die Revision	111	0.57	0.49 %	
7	□ berufen können , um eine identisch Entscheidung	106	0.54	0.47 %	
8	□ berufen , um eine identisch Entscheidung	86	0.44	0.38 %	
9	□ berufen . @card@ aa	79	0.40	0.35 %	
10	□ berufen, müssen eine Rechtsfrage	79	0.40	0.35 %	
11	□ berufen sich ohne Erfolg	76	0.39	0.34 %	
12	□ berufen (§ @card@ Abs	57	0.29	0.25 %	

Abb. 15. Frequente Muster mit dem Verb *sich berufen* in Gerichtsentscheidungen im GeLeCo

Die präpositionale Nominalisierung *unter Berufung* wird eher häufiger als sprachlicher Verweis auf Gesetze oder ähnliche Rechtsvorschriften im Vergleich zu ihrer Verbvariante gebraucht, was aus den Korpusdaten ersichtlich ist, wo das sprachliche Muster mit Paragraphenzeichen und Abkürzungen der Strukturteile der Gesetze *unter Berufung auf § @card@ Abs* 5,48 % aller abgerufenen Belege der Wortgruppe *unter Berufung* ausmacht und auf Platz 1 rangiert. Ein zweites sprachliches Muster des Zitierens ist *unter Berufung auf Art. @card@ Abs*, es steht auf Platz 4 und bildet 1,76 % aller Belege mit *unter Berufung* (s. Abb. 16).

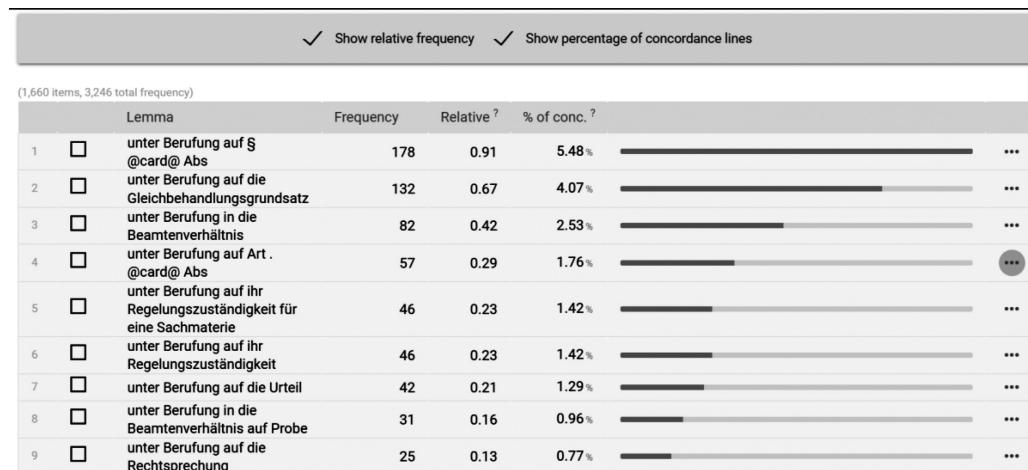


Abb. 16. Frequente Muster mit der Phrase *unter Berufung* in Gerichtsentscheidungen im GeLeCo

Für den Vergleich der durch die Chatbots und der im GeLeCo verwendeten sprachlichen Mittel wurden im GeLeCo nur die Wortgruppe *unter Berufung* und die Präposition *gemäß* herangezogen, da das Verb *sich berufen* und die Präposition *nach* auch andere Funktionen in den Gerichtsentscheidungen des Korpus haben und nicht nur Verweise auf andere Texte markieren.

Aus den verglichenen Daten der sprachlichen Mittel *gemäß* und *unter Berufung* geht hervor, dass die Verwendung der Präposition *gemäß* mit der Frequenz 169.104 im Vergleich zu *unter Berufung* mit der Frequenz 2.299 viel häufiger für die Markierung der Verweise vorkommt (Abb. 17).

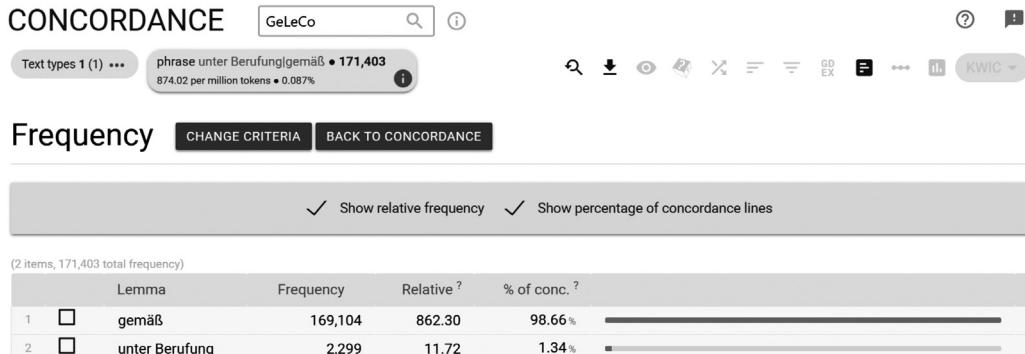


Abb. 17. Vergleich der Verwendung von *gemäß* und *unter Berufung* in Gerichtsentscheidungen im GeLeCo

Die Analyse der sprachlichen Mittel im GeLeCo zeigt, dass sich die in Gerichtsentscheidungen verwendeten sprachlichen Mittel zur Markierung der Bezüge auf andere Rechtstexte in ihrer Verwendungshäufigkeit von den sprachlichen Mitteln, die bei der maschinellen Übersetzung vorkommen, unterscheiden. Das stellt die Grundlage für die didaktischen Überlegungen und die Entwicklung von Unterrichtsszenarien und Aufgaben dar, damit die Lernenden im rechtsbezogenen DaF-Unterricht dafür sensibilisiert werden, dass die Übersetzungstools und Chatbots grob den Inhalt des Textes wiedergeben, aber die präziseren, für juristische Handlungen relevanten Nuancen der Rechtstexte nicht ausdrücken.

5 Schlussfolgerungen und Konsequenzen für den fachbezogenen DaF-Unterricht der Rechtssprache

Die maschinelle Übersetzung des Auszugs aus einer Entscheidung des Verfassungsgerichts der Republik Litauen und die Recherche der Korpusdaten im GeLeCo hat ergeben, dass sich die sprachlichen Mittel der Bezugnahme auf andere Texte in den von den Übersetzungs- und KI-Tools übersetzten Rechtstexten von den in den Gerichtsentscheidungen des Korpus verwendeten typischen sprachlichen Mitteln unterscheiden. Daraus könnte man den Schluss ziehen, dass maschinelle Übersetzungen für den Bereich der Verweise nicht die notwendigen rechtssicheren Formulierungen liefern und die maschinelle Übersetzung momentan die professionellen Übersetzer:innen nicht immer ersetzen kann, denn Rechtstexte haben häufig in der juristischen Arbeit die gleiche Funktion wie Originale und die Übersetzungen dienen nicht zur Information über den Inhalt des Textes, wie das häufiger der Fall in anderen Bereichen ist, sondern sie haben auch Rechtswirkung. Eine ungenaue Inhaltswiedergabe könnte somit in bestimmten Situationen entsprechende Rechtsfolgen haben, wenn ein Rechtstext nicht präzise übersetzt wird.

Da Wörterbücher als Hilfsmittel bei der Übersetzung von Rechtstexten nicht mehr die wichtigste Rolle spielen, sondern immer häufiger digitale Tools herangezogen werden, sollten die Übersetzer:innen bzw. Linguist:innen, die sich mit Rechtstexten beschäftigen werden, bei der Ausbildung für die Besonderheiten der Rechtssprache und ihre Übersetzungsspezifik sensibilisiert werden. Die Lernenden sollten auch im rechtsbezogenen Fremdsprachenunterricht auf die Ungenauigkeiten der maschinellen Übersetzung der Rechtstexte aufmerksam gemacht werden. Die Aufgaben und Lernszenarien der Korpusarbeit und somit die Korpusdidaktik könnten dabei eine wichtige Rolle im rechtsbezogenen Fachsprachenunterricht spielen. Die kritische Bewertung der mithilfe der Übersetzungsmaschinen produzierten oder mit KI-Tools generierten Texte und der Vergleich der sprachlichen Mittel in diesen Texten mit den entsprechenden sprachlichen Mitteln in Korputexten sollten auf einer korpusbezogenen Recherche basieren und somit die *corpus literacy* der Lernenden aufbauen und fördern. Die korpusbasierten Aufgaben könnten auch dabei helfen, die Qualität der beim Übersetzen eingesetzten digitalen Tools zu überprüfen.

Trotz einiger Nachteile der Korpusrecherche, einer komplizierten Abfragesprache und des Zeitaufwands, ermöglicht die Arbeit mit Korpora den Zugang zur authentischen Sprachverwendung. Die Korpusarbeit mit dem GeLeCo der deutschen Rechtssprache kann nicht nur die authentische Verwendung der deutschen Rechtssprache, sondern auch die sprachliche Spezifik oder die sprachlichen Unterschiede der einzelnen Textsorten zeigen, da das GeLeCo Gesetze, Gerichtsentscheidungen und Verwaltungsvorschriften enthält. Als Voraussetzungen für die Korpusarbeit im rechtsbezogenen Fremdsprachenunterricht sind Sprachkenntnisse auf einem relativ hohen Niveau nötig sowie Textsortenkompetenz und juristisches Basiswissen, so dass die Korpusarbeit mit Rechtstexten vorwiegend im Masterstudium der Auslandsgermanistik ihren Platz finden könnte.

Abschließend sollte die Vermutung ausgedrückt werden, dass bis jetzt die Korpusarbeit immer noch die zuverlässigste Methode für die Sensibilisierung der Studierenden für authentische Sprachverwendung ist und deshalb die Korpusrecherche ihren Platz im fachbezogenen Fremdspracheunterricht finden sollte.

Als weitere Aktivitäten und Forschungsfragen, die auf der in diesem Beitrag vorgestellten Untersuchung und auf den didaktischen Überlegungen für den fachbezogenen Fremdsprachenunterricht basieren, könnten die Entwicklung der Unterrichtsszenarien und Aufgaben sowie ihr Einsatz im Unterricht und das Feedback der Lehrenden und Lehrkräfte genannt werden.

Quellenverzeichnis

Lietuvos Respublikos Konstitucinio Teismo nutarimas dėl Lietuvos Respublikos Ginklų ir šaudmenų kontrolės įstatymo (2019 m. birželio 27 d. redakcija) 40 straipsnio 1 dalies 10 punkto atitinkies Lietuvos Respublikos Konstitucijai. 2023 m. vasario 9 d. Nr. KT14-A-N2/2023 (‘Entscheidung des Verfassungsgerichts der Republik Litauen zur Verfassungskonformität von Art. 40 Abs. 1 Nr. 10 des Waffenbesitzgesetzes vom 9. Februar 2023, Nr. KT14-A-N2/2023’). Abrufbar unter: <https://lrkt.lt/lt/teismo-aktai/paieska/135/ta2803/content>. Zugriff: 07.08.2025.

Korpus GeLeCo (German Legal Corpus). Abrufbar unter: <https://bellatrix.sslmit.unibo.it/noske/public/#dashboard?corpname=geleco>. Zugriff: 07.08.2025.

Literaturverzeichnis

- Ballansat, Suzanne & Gunhilt Perrin. 2007. Kombinierte Analyse von Inhalt und Sprache juristischer Texte als Vorbereitung auf den Übersetzungsunterricht. *Studien zur Rechtskommunikation*. Dorothee Heller & Konrad Ehlich, Hrsg. Berlin u. a.: Peter Lang. 259–286.
- Beißwenger, Michael & Matthias Knopp. 2019. Soziale Medien in Schule und Hochschule. Zur Einführung. *Soziale Medien in Schule und Hochschule: Linguistische, sprach- und mediendidaktische Perspektiven*. Michael Beißwenger & Matthias Knopp, Hrsg. Berlin u. a.: Peter Lang. 9–21. (Forum Angewandte Linguistik 63)
- Brockhorst, Tobias. 2022. Künstliche Intelligenz im Fremdsprachenunterricht. Wie KI basierte Software Einzug in den Fremdsprachenunterricht hält. *Magazin Sprache*. Goethe Institut. <https://www.goethe.de/ins/be/de/spr/mgs/24202694.html>. Zugriff: 07.08.2025.
- Bundesministerium der Justiz. 2024. *Handbuch der Rechtsförmlichkeit*. 4., vollständig überarbeitete Auflage. <https://www.legistik.de/HdR-2024.pdf>. Zugriff: 07.08.2025.
- Efing, Christian & Karl-Hubert Kiefer. 2018. Fach- und Berufssprachenvermittlung. *Berufs-, Fach- und Wissenschaftssprachen. Didaktische Grundlagen*. Jörg Roche & Sandra Drumm, Hrsg. Tübingen: Narr Francke Attempto. 167–196.
- Fix, Ulla. 2000. Aspekte der Intertextualität. *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung. 1. Halbband*. Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager, Hrsg. Berlin & New York: Walter de Gruyter. 449–457.
- Flinz, Carolina. 2019. Fachsprachen – aktuelle Fragen zu Forschung und Lehre. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 24 (1), 1–20. <https://zif.tu-journal.ulb.tu-darmstadt.de/article/id/3172/>. Zugriff: 07.08.2025.

- Flinz, Carolina, Britta Hufeisen, Rita Luppi & Ruth M. Mell. 2021. Zeit für eine neue Zeitschrift? – die elektronische Zeitschrift KorDaF – Korpora Deutsch als Fremdsprache stellt sich vor. *Korpora Deutsch als Fremdsprache* 1 (1), 1–5. <https://kordaf.tujournals.ulb.tu-darmstadt.de/article/id/36/>. Zugriff: 07.08.2025.
- Groot, Gerard-Rene, de. 1999. Zweisprachige juristische Wörterbücher. *Übersetzen von Rechtstexten*. Peter Sandrini, Hrsg. Tübingen: Narr. 203–227.
- Kerres, Michael. 2024. *Mediendidaktik. Konzeption und Entwicklung neuer Lernangebote*. 6. Auflage. Oldebourg: de Gruyter.
- Köbis, Laura. 2023. Künstliche Intelligenz in der DaF/DaZ-Lehre: Theoretischer Überblick und praktischer Einsatz. *German as a Foreign Language* 2, 31–53. <https://gfl-journal.de/article/kunstliche-intelligenz-in-der-daf-lehre-theoretischer/>. Zugriff: 07.08.2025.
- Krekeler, Christian. 2021. Korpustraining und datengestütztes Lernen im Wirtschaftsdeutsch-Unterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 161–195. <https://zif.tujournals.ulb.tu-darmstadt.de/article/id/3326/>. Zugriff: 07.08.2025.
- Lüdeling, Anke & Maik Walter. 2009. Korpuslinguistik für Deutsch als Fremdsprache. *Sprachvermittlung und Spracherwerbsforschung*, 1–37. <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/LuedelingWalterDaF.pdf>. Zugriff: 07.08.2025.
- Luppi, Rita. 2024. Vergleichbare Fachsprachenkorpora und die Vermittlung von Kollokationen in DaF am Beispiel des Projekts *TermiKnowlege*. *Korpora Deutsch als Fremdsprache* 4 (2), 236–255. <https://doi.org/10.48694/kordaf.4127>. Zugriff: 07.08.2025.
- Meier-Vieracker, Simon. 2022. Fußballwortschatz digital. Korpuslinguistische Ressourcen für den Sprachunterricht. *Korpora Deutsch als Fremdsprache* 2 (1), 7–24. <https://kordaf.tujournals.ulb.tu-darmstadt.de/article/id/3485/>. Zugriff: 07.08.2025.
- Mell, Ruth Maria. 2022. Fachsprachenbezogene Wissensvermittlung und Korpora. *Korpora Deutsch als Fremdsprache* 2 (1), 1–6. <https://kordaf.tujournals.ulb.tu-darmstadt.de/article/id/3487/>. Zugriff: 07.08.2025.
- Morlok, Martin. 2015. Intertextualität und Hypertextualität im Recht. *Zugänge zur Rechtssemantik. Interdisziplinäre Ansätze im Zeitalter der Mediatisierung*. Friedemann Vogel, Hrsg. Berlin/Boston: Walter de Gruyter. 69–90.
- Mukherjee, Joybrato. 2002. *Korpuslinguistik und Englischunterricht: Eine Einführung*. Berlin u. a.: Peter Lang.
- Roche, Jörg. 2008. *Handbuch Mediendidaktik. Fremdsprachen*. Ismaning: Hueber.
- Roche, Jörg. 2018. Mediale Lernmehrwerte durch Aufgaben- und Handlungsorientierung – Berufs-, Geschäfts- und Wissenschaftssprachen. *Berufs-, Fach- und Wis-*

senschaftssprachen. Didaktische Grundlagen. Jörg Roche & Sandra Drumm, Hrsg. Tübingen: Narr Francke Attempto. 240–250.

Wiesmann, Eva. 2022. Rechtsbezogene Wissensvermittlung. Zum Einsatz von GeLeCo in der Übersetzausbildung. *Korpora Deutsch als Fremdsprache* 2 (1), 42–60. <https://cris.unibo.it/retrieve/e038653d-c674-42ad-a5f6-50b030f45fc2/kordaf-56-wiesmann.pdf>. Zugriff: 07.08.2025.

Digitale Literaturwissenschaft und pädagogische Praxis

Marie Flüh

Institut für Germanistik

Fachbereich: Sprache, Literatur und Medien I

Universität Hamburg

Von-Melle-Park 6

20146 Hamburg, Deutschland

E-Mail: marie.flueh@uni-hamburg.de

ORCID iD: <https://doi.org/0000-0002-1707-284X>

<https://ror.org/00g30e956>

Abstract. Im Rahmen dieses Beitrags soll ein Lehrkonzept für die Professionalisierung angehender Lehrkräfte vorgestellt werden, dessen Schwerpunkt auf der Synchronisierung von digitalen Grundkompetenzen im Umgang mit Methoden der digitalen Textanalyse – wozu auch literaturwissenschaftlich ausgerichtete Korpusanalysen zählen – mit einer literaturdidaktischen Kompetenzentwicklung liegt. Nach einer kurzen terminologischen Verortung werden im ersten Teil des Beitrags Motivation und Relevanz für den Brückenschlag zwischen Digital Humanities und Lehrkräfteausbildung herausgestellt. Der zweite Teil fokussiert konzeptuelle Grundlagen, Wissensbereiche und den groben Verlauf des Seminars *Digitale Literaturwissenschaft und Literaturdidaktik* (durchgeführt im WiSe 23/24 und WiSe 2024/25 am Institut für Germanistik der Universität Hamburg, gefördert durch die Stiftung Innovation in der Hochschullehre) unter besonderer Berücksichtigung der Einheit zur Korpusanalyse. Abschließend werden Ergebnisse, Chancen und Grenzen des Lehrkonzepts beleuchtet.

Schlüsselwörter: Digital Humanities, Computational Literary Studies, Germanistik, Didaktik, Hochschullehre, Lehrkräfteausbildung

Computational literary studies and pedagogical practice

Abstract. This article presents a teaching concept for the professionalization of prospective teachers. The concept is focused on the synchronization of basic digital skills in dealing with different methods of computational text analysis – which also includes literary studies-oriented corpus analyses – with the development of didactic skills. After a brief introduction, the first part of the article highlights the motivation and relevance of bridging the gap between digital humanities and teacher training. The second part focuses on conceptual foundations, areas of knowledge, and the rough outline of the seminar “Digital Literary Studies and Literary Didactics” (held in the Winter semester of Study year 2023/24 and the Winter semester of 2024/25 at the Institute of German Studies at the University of Hamburg) with special consideration of the session on corpus analysis. Finally, the results, opportunities and limitations of the teaching concept are examined.

Keywords: digital humanities, computational literary studies, German language and literature, didactics, pedagogy, teacher training

1 Einleitung: Digitalität in der Schule und der Literaturwissenschaft

Mittlerweile hat die Digitalisierung Einzug in viele Bereiche des öffentlichen und privaten Lebens gehalten und beeinflusst die Praxis des Forschens, Lernens und Lehrens an Schulen sowie Hochschulen. Ausgehend von einer in den 1990er Jahren noch widersprüchlichen fachdidaktischen Debatte – entweder grundlegend ablehnend, distanziert-kritisch oder affirmativ verlaufend – ist das Verständnis des Computers als reines Schreibmedium längst ersetzt durch das Bewusstsein für dessen große Bedeutung als Informations- und Kommunikationsmedium und Forschungsinstrument (Albrecht 2014, 138 f.). Das Populärwerden von generativer Künstlicher Intelligenz (gKI) in Form von niedrigschwellig anwendbaren Tools wie *Gemini* oder *ChatGPT* hat die große Bedeutung und das Bewusstsein für den Einfluss, den die digitale Transformation auf Leben, Arbeiten sowie Bildungs- und Lernprozesse hat, weiter erhöht.

Warum ist ein Brückenschlag zwischen Digital Humanities (DH) und Didaktik bzw. Lehrkräfteausbildung sinnvoll? Dieser Frage soll vor allem im folgenden Abschnitt nachgegangen werden, bevor ein konkretes Lehrkonzept zur Brückengestaltung zwischen den beiden Fachdisziplinen vorgestellt wird. Im Rahmen einer digitalisierungsbezogenen Bestandsaufnahme wird zunächst erläutert, welche Rolle die Digitalisierung im schulischen und im literaturwissenschaftlichen Kontext spielt. Dabei werden Leerstellen sichtbar, die sich – so argumentiere ich in diesem Beitrag – durch eine engere Verzahnung von Inhalten der digitalen Literaturwissenschaft und Lehrkräfteausbildung für das Unterrichtsfach Deutsch füllen lassen.

1.1 Digitalität und Schule

Dass technische Medien im Deutschunterricht einen Platz einnehmen sollten, ist eine Erkenntnis, die sich ausgehend vom didaktischen Diskurs der 1970er Jahre ihren Weg in den Schulunterricht gebahnt hat (Frederking & Krommer 2014, 150 f.).¹ Im schulischen Bereich stehen Lehrkräfte in der Verantwortung, Schüler:innen zu einer reflektierten Teilhabe am Leben in einer digitalisierten Welt zu befähigen. Um eine verantwortungsvolle Teilhabe von Schüler:innen an einer digitalisierten Gesellschaft zu garantieren, ist es notwendig, die Lehrer:innen erst einmal selbst mit den entsprechenden Kompetenzen auszustatten. Im Rahmen der Strategie für „Bildung in der digitalen Welt“, einer Verordnung der Kultusministerkonferenz (KMK) der Länder, wird Medienerziehung als fächerübergreifendes Prinzip vorgestellt, das den verbindlichen Kompetenzerwerb in sechs Kompetenzbereichen („Suchen, Verarbeiten und Aufbewahren“, „Kommunizieren und Kooperieren“, „Produzieren und Präsentieren“, „Schützen und Agieren“, „Problemlösen und Handeln“, „Analysieren und Reflektieren“) in allen Unterrichtsfächern vorschreibt (KMK 2026, 16 f., 19). Trotzdem gelten vorzugsweise Deutsch, Kunst, Musik und Sozi-

¹ Unter kulturgeschichtlicher Perspektive auf Medientypen und Mediennutzung lassen sich vier Paradigmen unterscheiden, in denen sich unterschiedliche Medientypen als Leitmedien herausgebildet haben. Das orale Paradigma bestimmt die gesprochene Sprache. Die Stimme fungiert als Primärmedium des vorliteralen Stadiums (Frederking, Krommer & Maiwald 2018, 27). Die Erfindung der Schrift führte zum ersten medialen Paradigmenwechsel und zum Beginn des literalen Paradigmas. Leitmedien waren skripturale Medien und typographische Medien, also an verschiedene Trägermedien gebundenen Formen von Schriftlichkeit (Frederking, Krommer & Maiwald 2018, 29). Durch die Erfindung des Buchdrucks im Jahr 1450 gewinnen neben handschriftlichen Erzeugnissen (skripturale Medien) druckmaschinelle Massenfertigungen (typographische Medien) an Bedeutung; das gedruckte Buch wird zum Leitmedium. In der zweiten Hälfte des 19. Jahrhunderts endete mit der industriellen Revolution die Vorherrschaft des literalen Paradigmas. Die medialen Optionen erweitern sich um akustisch-auditive Medien (Grammophon, Telefon, Rundfunk / Radio, Schallplatte, Tonbandvorräufer), optisch-visuelle Medien (Fotografien als optische Speichermedien, Film) und audiovisuelle Medien (Verbindung von Bild und Ton im Tonfilm, Fernsehen) (Frederking, Krommer & Maiwald 2018, 41). Ein vierter Paradigmenwechsel hin zum digital-multimedialen Paradigma ereignet sich in den 1980er Jahren durch die Entwicklung des Digitalmediums. Die Entwicklung des Internets in den 1980er Jahren bildet die Grundlage für weitere Meilensteine des digital-multimedialen Paradigmas, beispielsweise das Aufkommen sozialer Netzwerke und Cloud-Technologien (Frederking, Krommer & Maiwald 2018, 58) sowie die damit einhergehende immens steigende Datenproduktion, die wiederum die Grundlage für große Sprachmodelle und künstliche Intelligenz bildet. Dieses Paradigma zeichnet eine Medienverschmelzung aus. Leitmedien des digital-multimedialen Paradigmas stellen multimediale Integrationsmedien / Digitalmedien dar. Der Computer kann als Trägermedium alle medialen Präsentationsformen bedienen. Er ermöglicht eine Verbindung von schriftsprachlichen Texturen mit piktoralen, auditiven und audiovisuellen Elementen. In diesem Sinne wird er als Symmedium beschrieben (Frederking, Krommer & Maiwald 2018, 53). Die schulische Medienbildung adressiert unterschiedliche Medientypen und Medientechnologien. Heute liegt der Fokus besonders auf der Auseinandersetzung mit Digitalmedien und damit verbundenen Medientechnologien. In diesem Beitrag geht es vorrangig um den Computer (= Trägermedium), der als Forschungsinstrument eingesetzt wird, um literarische Texte zu analysieren.

alkunde als Leitfächer der Medienerziehung (Lecke 2008, 49). In einer digital vernetzten Welt, in der die sogenannte „digitale Revolution“ in vollem Gange ist und ein großer Teil der Meinungsbildung in und mit digitalen Medien erfolgt (Tillmann 2017, 116), stellt sich daher die Frage, was der Literaturunterricht dazu beitragen kann, um die künftige Gesellschaft – in der nach Schrift, Sprache und Buchdruck vor allem digitale Medientechnologien eine zentrale Rolle spielen (Aßmann 2017, 137) – auf das Leben in dieser Welt vorzubereiten. In der zweiten Ausbildungsphase finden praktizierende Lehrkräfte unter anderem in schulinternen oder zentral angebotenen Fort- und Weiterbildungsmaßnahmen methodische und didaktische Unterstützung für den kompetenzorientierten Deutschunterricht mit digitaler Komponente. Im Rahmen der ersten Ausbildungsphase, dem Hochschulstudium, wird vor allem im Teilbereich der Mediendidaktik Grundlagenwissen über die Entwicklung, die Funktionsweise und die unterschiedlichen Einsatzmöglichkeiten von technischen Medien in Lehr-Lern-Szenarien vermittelt. Als Inspiration und Orientierungshilfe für Lehrer:innen in allen Ausbildungsphasen dienen Publikationen und praxisorientierte Handreichungen für Lehrkräfte zur Auseinandersetzung mit digitalen literalen Medien im Deutschunterricht.² Durchgeplante Unterrichtseinheiten finden sich beispielsweise zum Einsatz didaktisch aufbereiteter Computerspiele, zu unterschiedlichen Formen medialen Präsentierens (zum Beispiel mit Textverarbeitungsprogrammen oder Präsentationstools) sowie zum Recherchieren und Kommunizieren im Internet. Der Einsatz von E-Books in Verbindung mit dem Zugriff auf Online-Enzyklopädien und anderen mediengebundenen Funktionen wird als Form der digitalen Auseinandersetzung mit Digitalisaten vorgestellt. Hierbei werden die inhaltliche Erschließung und ein wiederholendes Lernen durch das Markieren signifikanter Inhalte oder Schlüsselbegriffe begleitet (Brand 2015, 135 f.). Weitere didaktisch fundierte Ideen für den digitalen Deutschunterricht stellt auch der Sachbuchautor und Lehrer Philippe Wampfler vor. Sie bestehen in der Erarbeitung eigener Wikipedia Einträge, der Arbeit mit Google-Docs als kollaborative Form der Textarbeit, dem Umschreiben von literarischen Texten als E-Mail, SMS oder X-Roman, der Erarbeitung von Unterrichtsblogs oder dem Verfassen von Gedichten mittels Messengerdiensten (Wampfler 2017). Darüber hinaus ist im Zuge der Fördermaßnahmen des DigitalPakt Schule (2019–2024 und 2025–2030) neben

² Die im Deutschunterricht vorkommenden digitalen Medien können nach Leubner (2014) unterschieden werden in literale Medien, auditive Medien, audiovisuelle Medien, Schreib-, Präsentations- und Publikationsmedien, Informations- und Kommunikationsmedien und Interaktions- und Handlungsmedien. Für diesen Artikel stellen digitale literale Medien die relevante Bezugsgröße dar. Die anderen Medienformen klammere ich aus. Digitale literale Medien unterscheiden sich von multimedialen digitalen Produktionen und von schriftsprachlichen Produktionen mit digitaler Druckvorlage, die als Druckversion publiziert sind. Digitale literale Medien umfassen sämtliche Medienproduktionen „[...], die auf Schriftsprache beschränkt oder zumindest im Wesentlichen konzentriert sind und in digitaler Form gespeichert bzw. publiziert vorliegen“ (Leubner 2014, 186). Digitale literale Medienproduktionen lassen sich wiederum in die zwei Haupttypen digital publizierter Texte und digitaler Literatur unterscheiden.

dem Ausbau der technischen Infrastruktur auch eine länderübergreifende Infrastruktur zur Bereitstellung und Bewertung digitaler Bildungsmedien erarbeitet worden, auf die Lehrkräfte zurückgreifen können.³ Dazu gehören vor allem die offene und frei zugängliche Bildungsmediathek *MUNDO*, die lizenzerrechtlich und qualitativ geprüfte Materialien für den Schulunterricht zur freien Verfügung stellt. Einen fortlaufend aktualisierten Überblick über Unterrichtsideen zur Auseinandersetzung mit digitalen literalen Medien im Deutschunterricht (und allen weiteren Unterrichtsfächern) bietet außerdem das Portal *digital.learning.lab*. Der kurSORische Streifzug durch Publikationen zum digitalaffinen Unterricht zeigt also eine Vielzahl an Angeboten und Unterrichtskonzepten für den Literaturunterricht. Eine Suchanfrage für die nach KMK-Kompetenzen gegliederten Angebote des *digital.learning.lab* zeigt allerdings, dass für den Kompetenzbereich „Problemlösen und Handeln“, der eine Auseinandersetzung mit Algorithmen vorsieht, nur wenige Ideen für den Literaturunterricht zur Verfügung stehen. Die Suchabfrage ergibt nur zwei Treffer. Auf der zentralen Plattform für Open Educational Resources (OER) *MUNDO* werden für Suchbegriffe wie „digitaler Literaturunterricht“, „digitale Textanalyse“ oder „Korpusanalyse“ gar keine Treffer angezeigt. Hinsichtlich des Einsatzes statistischer Verfahren im Umgang mit digitalen Texten im Deutschunterricht herrscht also eine gewisse Ratlosigkeit vor: „Da ihre Verfahren sowohl an die Programm- wie an die Deutungsseite hohe Anforderungen stellen, ist nicht auf Anhieb klar, wie diese Ansätze für die Schule fruchtbar gemacht werden können.“ (Wampfler 2017, 146) Ansätze aus dem Bereich *Distant Reading* im Schulunterricht sind selten zu finden. Überlegungen für eine erzähltheoretische Analyse von Kleists *Marquise von O.* mittels Suchfunktionen und Tag-Clouds (Wampfler 2017) oder die Analyse literarischer Texte mit dem *Distant Reading-Tool Voyant* (Kühnert 2017; Flüh 2024a [2019]) sind als Ausnahmen aufzufassen. Gleches gilt für Unterrichtsideen zur Sentimentanalyse mit *SentText* (Urs 2021), zur manuellen digitalen Annotation mit dem Textanalysetool *CATMA* (Flüh 2024 [2020]) oder zur Netzwerkanalyse (Flüh 2024b [2019]). Bis hierhin kann also festgehalten werden, dass es ein breites Spektrum und vielseitige frei verfügbare Unterrichtsideen für den Einsatz digitaler Medien im Deutschunterricht gibt. Weder *Distant-Reading*-Verfahren noch digitale *Close-Reading*-Verfahren, wie digitale manuelle Annotation (s. Kapitel „Digitale Literaturwissenschaft und Didaktik“), haben bisher Einzug in den Deutschunterricht gefunden. Das mag vor allem daran liegen, dass neben den ohnehin involvierten Bezugswissenschaften Literaturwissenschaft und Literaturdidaktik auch die Informatik als nahezu völlig fremde beteiligte Komponente eine Rolle spielt.

³ Der DigitalPakt Schule ist eine durch das Bundesministerium für Bildung und Forschung koordinierte Fördermaßnahme für den Ausbau der technischen Infrastruktur an Schulen in ganz Deutschland.

1.2 Digitale Literaturwissenschaft und Didaktik

Im universitären Bereich haben sich die Digital Humanities (DH) als Fachbereich, Methodenrepertoire und Forschungsgemeinschaft herausgebildet. Trotz der andauernden Diskussion darüber, was genau die Digital Humanities nun sind (Sahle 2015), lässt sich festhalten, dass es sich um einen Wissenschaftsbereich handelt, in dem geisteswissenschaftliche Forschungsfragen computergestützt bearbeitet werden (Limpinsel 2016) und zwar bereits seit den späten 1940er Jahren: Als eine Art Gründungsmythos der DH gilt die Erstellung des „Corpus Thomisticum“, eines ersten digitalen Indexes der Werke Thomas von Aquins auf Lochkarten, die der Jesuitenpater Roberto Busa in Zusammenarbeit mit IBM entwickelte (Lauer 2013, 103 f.). Bis heute ist die Möglichkeit der statistischen Betrachtung von geisteswissenschaftlichen (in Busas Fall theologischen) Werken ein zentraler Arbeitsbereich der Digital Humanities. Ein weiterer Teilbereich der DH ist die digitale Literaturwissenschaft. Ein zentraler methodischer Schwerpunkt der Computational Literary Studies (CLS) sind Ansätze, die unter dem Begriff *Distant Reading* zusammengefasst werden können. Der von dem Literaturwissenschaftler Franco Moretti eingeführte Begriff (Moretti 2016) beschreibt heute Analysen, in denen mithilfe von Computern und statistischen Verfahren große Textmengen untersucht werden. Digitale Literaturanalysen, in denen *Distant-Reading*-Verfahren eingesetzt werden, machen oft methodische Anleihen bei der Computer- und Korpuslinguistik (Jannidis 2010, 109). Anders als beim *Close Reading*, also der genauen, textnahen, auf ausgewählte Texte begrenzten Lektüre, bilden bei vielen *Distant-Reading*-Verfahren große Textkorpora den Untersuchungsgegenstand. Der quantitative Zugriff auf Literatur zielt in der Regel darauf ab, Verteilungsmuster zu erkennen und auf diese Weise die Ergebnisse von Einzelfallstudien in einen breiteren Kontext zu setzen. Dabei lassen sich je nach Zusammensetzung des Textkorpus und der Methode bspw. stilistische Besonderheiten (Schumacher 2022), thematische Schwerpunktsetzungen (Weitin & Herget 2017), emotionale Trends (Kröncke et al. 2023), Aspekte von Intertextualität (Yang et al. 2024), Genderreferenzen (Schumacher & Flüh 2023) oder – in einer populärwissenschaftlichen Aufbereitung der CLS – Eigenschaften von zeitgenössischen Bestsellern (Archer & Jockers 2017) für ganze Epochen, Œuvres oder unterschiedliche Gattungen analysieren. Doch in welcher Verbindung steht die digitale Literaturwissenschaft zur Didaktik und ferner zur Ausbildung angehender Lehrkräfte? Seitens der digitalen Literaturwissenschaft gibt es zwar Überlegungen und Ansätze für die Hochschullehre (Hirsch 2015; das Hauptangebot der Online-Plattform *forTEXT*; Mischke, Trilcke & Sluyter-Gäthje 2022; Bläß et al. 2022; Battershill & Ross 2022; Risam 2019; Jackaki & Faull 2016; Mauro 2016; Mahony & Pierazzo 2015), der Transfer in den schulischen Bereich und eine mögliche Integration in die Ausbildung angehender Lehrkräfte werden bisher aber eher wenig berücksichtigt.

Unter Rückbezug auf die eingangs skizzierte Rolle der Digitalisierung im bildungspolitischen und schulischen Kontext zeigt sich auf der einen Seite die Forderung nach einem Deutschunterricht, der an den KMK-Kompetenzen orientiert ist und Schüler:innen zur reflektierten Teilhabe an der digitalisierten Welt befähigt. Darunter fallen im Teilbereich „Problemlösen und Handeln“ die Anforderungen, „Digitale Medien und Werkzeuge zum Lernen, Arbeiten und Problemlösen nutzen“ und „Algorithmen kennen und formulieren“ zu können (KMK 2016, 19). Konkrete Ideen für den Literaturunterricht, die diese Aspekte adressieren, sind rar. Der größte Sprung in Richtung einer praktischen Auseinandersetzung mit Algorithmen scheint nicht Ergebnis einer gezielten bildungspolitischen Entscheidung, sondern primär durch wirtschaftliche Entwicklungen motiviert zu sein; schließlich hat erst die Publikation des Chatbots *ChatGPT* durch das amerikanische Unternehmen *OpenAI* die Auseinandersetzung mit generativer Künstlicher Intelligenz gesamtgesellschaftlich beinahe unausweichlich gemacht. Das zeigt sich auch in der Stellungnahme der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz (SWK 2024) zu Sprachmodellen und ihren Potenzialen im Bildungssystem. Darin vorgesehen sind unter anderem eine engere Zusammenarbeit zwischen Praxis und Wissenschaft zur Entwicklung fachspezifischer, altersgerechter Nutzungs- und Aufgabenszenarien. Das Positionspapier betont außerdem, dass Lehrende und Lernende eine umfassende Kompetenz zum lernförderlichen Umgang mit großen Sprachmodellen (Large Language Model/LLM) benötigen, die in allen Ausbildungsphasen vermittelt werden sollte. Während in der Grundschule und dem Beginn der Sekundarstufe I die Ausbildung von Lese- und Schreibkompetenzen im Vordergrund stehen, sollen im Laufe der Sekundarstufe I zunehmend auch digitale Tools zur Erstellung von Texten eingesetzt werden. Ab der achten Klasse – so das Positionspapier – ist der regelmäßige Einsatz von LLMs als Schreibunterstützung in allen schriftbasierten Fächern sinnvoll (SWK 2024, 18 f.). Im Hinblick auf den Umgang mit LLMs zeigt sich also eine technikbejahende Position, die ähnlich wie im Fall der KMK-Strategie aus dem Jahr 2016 eine Integration in den Schulunterricht vorsieht. Dafür müssen zunächst die Lehrkräfte lernen, wie Sprachmodelle funktionieren, für welche Aufgabenbereiche sie eingesetzt werden können und für welche nicht. Darüber hinaus bedarf es konkreter Szenarien für den Schulunterricht. Dass der regelmäßige Einsatz vor allem für die schriftbasierten Fächer vorgesehen wird, legt nahe, dass die Auseinandersetzung mit der Funktionsweise von Sprachmodellen angehende und praktizierende Deutschlehrer:innen in besonderem Maße betrifft.

Die digitale Literaturwissenschaft stellt ein umfangreiches, etabliertes und kritisch reflektiertes Methodenrepertoire zur digitalen Textanalyse bereit. Künstliche Intelligenz (KI) und große Sprachmodelle werden in den Geisteswissenschaften bereits seit den Anfängen von KI in den 1950er Jahren theoretisch perspektiviert und auch für literaturwissenschaftliche Fragestellungen eingesetzt (Cardoso et al. 2009; Shanahan & Clarke 2023). Eine Verbindung beider Bereiche ist eher als Ausnahmeerscheinung zu bewerten.

Das im folgenden Teil vorgestellte Lehrkonzept verknüpft beide Bereiche: Schule, hier bezogen auf die Ausbildung angehender Lehrkräfte für das Fach Deutsch, und Digital Humanities, hier bezogen auf die Computational Literary Studies (CLS) als literaturwissenschaftlicher Teilbereich der Digital Humanities.

2 Konzeptuelle Säulen des Seminars „Digitale Literaturwissenschaft und pädagogische Praxis“

Data-Literacy-Education für angehende Deutschlehrer:innen wird in zwei konzeptuellen Säulen aufgegriffen, die an den Kompetenzen orientiert sind, die die Studierenden durch die Seminarteilnahme erwerben sollen (s. Abb. 1).

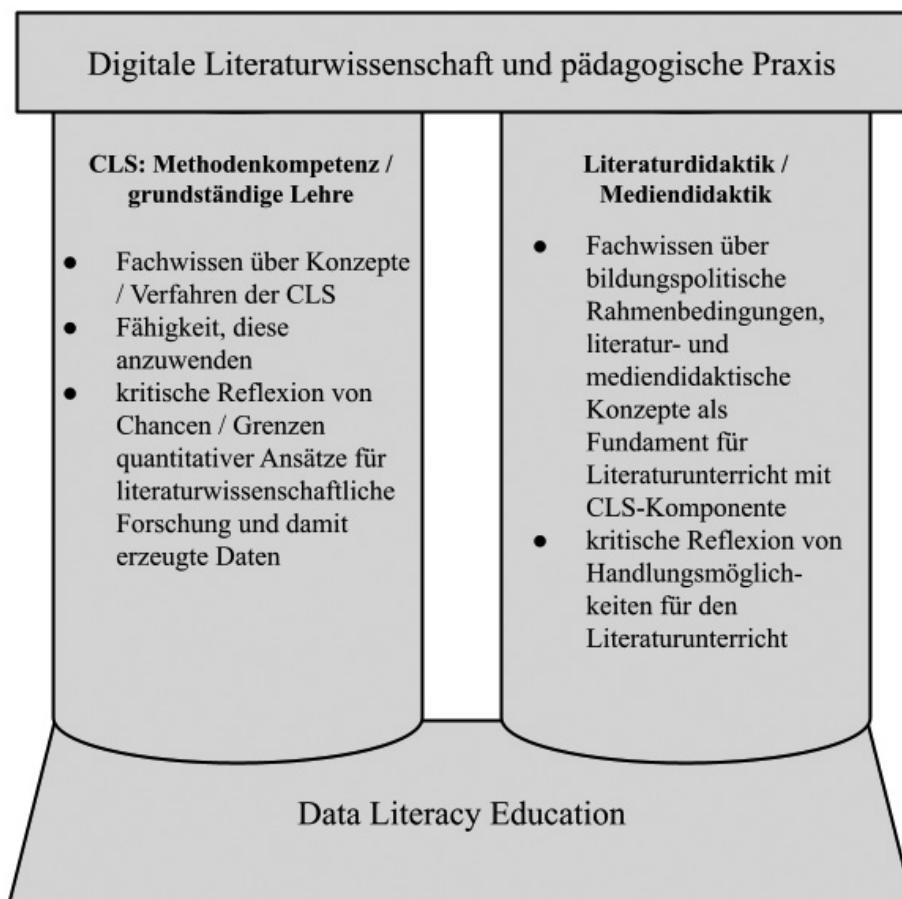


Abbildung 1. Zweisäuliges Konzept für das Seminar „Digitale Literaturwissenschaft und pädagogische Praxis“

2.1 Computational Literary Studies (CLS): Methodenkompetenz / grundständige Lehre: Vermittlung etablierter digitaler Verfahren und Methoden der Textanalyse

Im Bereich „CLS: Methodenkompetenz / grundständige Lehre“ werden drei Teilbereiche von Data Literacy angesprochen: 1. die Vermittlung von Fachwissen über Konzepte und Methoden der digitalen Literaturwissenschaft; 2. die Fähigkeit, diese anzuwenden und 3. die Reflexion über Chancen und Grenzen quantitativer Ansätze für die literaturwissenschaftliche Forschung. Der Fokus dieses Themenkomplexes liegt auf der praktischen Anwendung: Die Studierenden erwerben Methodenkompetenz im Umgang mit digitalen Verfahren der Textanalyse, indem diese mithilfe exemplarisch ausgewählter Tools selbst ausprobiert werden. Zu den wichtigsten unüberwachten Lernverfahren zählen beispielsweise digitale Stilometrie oder Topic Modeling. Verfahren des überwachten maschinellen Lernens werden in Form von Handwritten Text Recognition, Named Entity Recognition oder dem Modelltraining eines Classifiers zur automatischen Annotation von Genderrollen einbezogen.⁴ Ein weiterer Bestandteil sind digitale Varianten klassischer geisteswissenschaftlicher Methoden (beispielsweise digitale manuelle (kollaborative) Annotation oder digitale Netzwerkanalyse). Neben der praktischen Anwendung stellt die kritische Reflexion der Methoden und der damit erzeugten Daten einen wichtigen Bestandteil dar.

2.2 Didaktische Kompetenz und Beurteilungskompetenz: Rückkopplung an didaktische Konzepte und bildungspolitische Rahmenbedingungen

In diesem Teilbereich des Seminars werden zwei Aspekte von Data Literacy angesprochen: 1. die Vermittlung von Fachwissen über bestehende (medien)didaktische Konzepte als didaktisches Fundament für einen Literaturunterricht, in dem digitale Verfahren der Textdatenanalyse zum Einsatz kommen und 2. (da die Literaturdidaktik an diesem Punkt noch entwicklungsbedürftig ist) die Reflexion darüber, wie sich digitale Verfahren der Textanalyse auf bestehende didaktische Modelle auswirken und wie ein Transfer in den schulischen Bereich gestaltet werden kann. Zentral ist die Frage danach, welche Handlungsmöglichkeiten für den Literaturunterricht durch den Einsatz digitaler Verfahren der Textanalyse entstehen und wie Unterrichtskonzepte aussehen können, in denen unter Berücksichtigung des Primats der Pädagogik literarische Texte und Textkorpora digital analysiert werden. Zum didaktischen Grundstock zählen bspw. das TPACK- und das SAMR-Modell (Mishra & Koehler 2006; Zierer 2018; s. Kapitel „Verknüpfung von Wissen auf verschiedenen Ebenen“)⁵ und die Auseinandersetzung mit bildungspoliti-

⁴ Für eine Definition überwachter und unüberwachter Verfahren des maschinellen Lernens und deren Einsatz in den Geisteswissenschaften vgl. Andresen (2024).

⁵ Das TPACK-Modell (Kurzform für *Technological Pedagogical Content Knowledge*) zeigt in orientierungsstiftender Funktion, welche Wissensbereiche bei dem Einbezug von Technologien in den Schulunterricht beteiligt sind. Das Modell wird im folgenden Kapitel genauer vorgestellt. Das

schen Rahmenbedingungen in Form der KMK-Kompetenzbereiche, ausgewählte Positionspapiere und bestehende Ressourcen für einen digitalen Deutschunterricht, die im Seminar unter Rückbezug auf die CLS-Perspektive diskutiert werden.

3 Zielgruppe und Transferkonzept

Das Lehrangebot richtet sich an Lehramtsstudierende für das Unterrichtsfach Deutsch in der Sekundarstufe II, die sich im Masterstudium befinden. In diesem Stadium ist davon auszugehen, dass auf literaturwissenschaftliche und didaktische Grundkenntnisse aufgebaut werden kann, die wichtig sind, um die digitalen Methoden in eine literaturwissenschaftliche/didaktische Traditionslinie einordnen und kritisch hinterfragen zu können. Zugleich erfordert das Lehrangebot keine technischen Vorkenntnisse oder Erfahrungen mit den Methoden der CLS. Die Sitzungen mit DH-Anteil ermöglichen einen niedrigschwlligen Einstieg. Auf diese Weise soll die Hemmschwelle abgebaut werden, technische und quantitative Verfahren in die Literaturanalyse und die Unterrichtsplanung einzubeziehen. Die im Seminar verwendeten Tools sind so ausgewählt, dass Sie Mindestbedingungen für den Einsatz im Schulunterricht erfüllen. Dazu zählen aus meiner Perspektive, dass die Tools möglichst kostenfrei und unter einer Open Source Lizenz abrufbar sind, eine hohe Benutzerfreundlichkeit aufweisen (also über eine grafische Benutzeroberfläche verfügen und intuitiv bedienbar sind) und eine transparente Datenverarbeitung und -speicherung gewährleisten.

Durch den Wissenstransfer zwischen Literaturdidaktik und digitaler Literaturwissenschaft sowie die Synchronisierung von digitalen literaturwissenschaftlichen Grundkompetenzen mit einer fachspezifischen, didaktischen Kompetenzentwicklung erbringt das Seminar entscheidende Transferleistungen. Während die Studierenden im Seminar ei-

SAMR-Modell (Kurzform für *Substitution, Augmentation, Modifocation* und *Redefinition*) ist ein Arbeitsmodell, auf das sich Klaus Zierer (2018, 73 f.) bezieht. Es stammt von C. Punendedura und beschreibt vier Ebenen des Einbeugs von Technik im Schulunterricht: 1. *Substitution* (Ersetzung: Technik ist direkter Ersatz für Arbeitsmittel, es erfolgt keine funktionale Änderung), 2. *Augmentation* (Erweiterung: Technik ist ein direkter Ersatz des Arbeitsmittels, dabei erfolgt eine funktionale Verbesserung), 3. *Modification* (Änderung: Technik ermöglicht eine beachtliche Neugestaltung von Aufgaben) und 4. *Redefinition* (Neubelegung: Technik ermöglicht die Erstellung neuartiger Aufgaben, die bisher unvorstellbar waren). Ein Lerneffekt tritt im Vergleich mit traditionellen Medien vor allem auf der dritten und vierten Ebene ein: Je besser es Lehrpersonen gelingt, „neue Medien so einzusetzen, dass sie bisherige Aufgaben im Hinblick auf Anforderungsniveau und Kommunikation ändern und neu belegen, desto größer wird der Einfluss auf die Lernleistung“ (Zierer 2018, 75). Das Pramat der Pädagogik / Pramat des Pädagogischen bezeichnet im Hinblick auf den Einsatz von Technik im Schulunterricht den Grundsatz, dass der Einbezug neuer digitaler Medien und Technologien in den Schulunterricht nicht ohne didaktisches und pädagogisches Fundament erfolgen sollte. Dem Grundsatz ‘Pädagogik vor Technik’ folgend soll eine neue Technologie nicht unmittelbar nach ihrem Erscheinen in den Schulunterricht integriert werden, sondern nur dann, wenn sie dazu beiträgt, den Erziehungs- und Bildungsauftrag besser zu erfüllen.

nen Grundstock an etablierten Verfahren und Tools kennenlernen, ist es im Rahmen einer nachhaltigen Professionalisierung wichtig, sie auch über den universitären Kontext hinaus dazu zu befähigen, aus dem breiten Spektrum an Tools und Methoden eigenständig eine fundierte sowie lerngruppenorientierte Auswahl treffen und beurteilen zu können, welche Analyseverfahren/ Tools sich für den Einsatz im Schulunterricht eignen und welche nicht. Das Seminar soll diese Transferleistung unterstützen, indem ein Grundverständnis für die Funktionsweisen, die Chancen und die Grenzen quantitativer Verfahren der Textanalyse vermittelt und eine erste Annäherung an darauf abgestimmte Unterrichtskonzepte unternommen wird. Transfer ist außerdem als Brückenschlag zwischen etablierten Bezugswissenschaften (nicht-digitale Literaturwissenschaft/Literatur- und Mediendidaktik) und neuen Bezugswissenschaften (digitale Literaturwissenschaft/Literatur- und Mediendidaktik) gedacht. Leitend ist dabei die dem zweiseitigen Seminar-Konzept zugrundeliegende und im Seminar stetig verhandelte Frage, wie bestehende didaktische Konzepte und Ansätze der CLS zusammenpassen. Der Aufbau und die Vernetzung von fachlichem, pädagogischem und technologischem Wissen bildet eine wichtige Grundlage für einen Literaturunterricht, der Technologien nicht isoliert betrachtet, sondern Inhalte der Literaturdidaktik und Verfahren der DH sinnvoll kombiniert und durch eine didaktisch fundierte Integration digitaler Methoden in den Literaturunterricht einen verstehensorientierten Zugang zu und Umgang mit Daten vermittelt. Das Handwerkszeug dazu sollte im Rahmen des hier beantragten Lehrkonzepts erarbeitet, ausprobiert, evaluiert und für Dritte bereitgestellt werden.

4 Verknüpfung von Wissen auf unterschiedlichen Ebenen

Das TPACK-Modell (Mishra & Koehler 2006) ist sowohl ein Seminarinhalt, der gemeinsam mit den Studierenden und im Hinblick auf die Planung eigener Unterrichtseinheiten behandelt wird, als auch eine theoretische Grundlage und Orientierungshilfe für die Konzeptionierung und Planung der einzelnen Sitzungen. Um die Verwendung des Modells bei der Planung des Seminars soll es im folgenden Teil gehen. Das Modell zeigt, welche Wissensbereiche Lehrer:innen verknüpfen müssen, wenn digitale Technologien im Schulunterricht zum Thema gemacht werden. Um effektiv mit digitalen Technologien unterrichten zu können, so die Grundidee, benötigen Lehrer:innen eine Kombination von fachdidaktischen, pädagogischen und technologischen Wissensbeständen. Eine isolierte Technikbetrachtung ist nach dem TPACK-Modell deshalb nicht zielführend (Schmid, Krannich & Petko 2020). Das TPACK-Modell stellt eine Erweiterung des Shulman'schen Wissensmodells (Shulman 1987) dar und umfasst sieben Wissensbereiche (s. Abb. 2), von denen vier technologische Komponenten aufweisen. Die technikbezogenen Wissensbereiche stellen bei dem Transfer digitaler Verfahren der Textanalyse in den schulischen Bereich eine große Rolle und werden deshalb im folgenden Teil näher beschrieben. Dabei soll deutlich werden, welche Seminarinhalte welche Wissensbereiche aktivieren. Der Fokus liegt dabei auf Korpusanalyse.

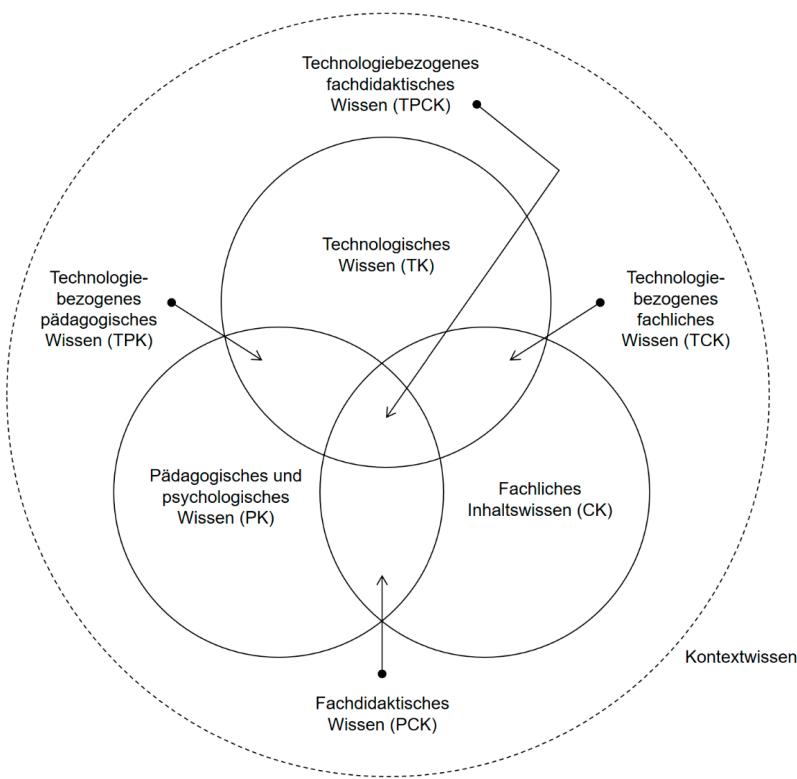


Abbildung 2. TPACK-Modell nach Mishra und Koehler (2006)

Technologisches Wissen (Technological Knowledge, TK): Technologiebezogenes Wissen umfasst Anwendungswissen und konzeptionelles Wissen über aktuelle und relevante Medien und Technologien. Das Anwendungswissen für Notebooks umfasst bspw. Kenntnisse über deren Bedienung, die Installation von Programmen, zur Nutzung oder Erstellung von Inhalten oder über das Beheben von Problemen. Konzeptionelles Wissen umfasst grundlegendes Wissen zur Funktionsweise digitaler Technologien (Schmid, Krannich & Petko 2020, 118). Der im Seminar entscheidende Medientyp sind sicherlich Digitalmedien, die für die computergestützten Literaturanalysen in besonderer Weise verwendet werden. Die Vermittlung der Methodenkompetenz, also die praktische Anwendung der Tools, folgt einem *Bring-your-own-device*-Ansatz.⁶ Die Arbeit am eigenen

⁶ Darauf wird bereits in der Seminarbeschreibung hingewiesen, damit die Studierenden noch vor der Anmeldephase wissen, dass für die Teilnahme ein Laptop benötigt wird und, falls nötig, Leihgeräte organisiert werden können. Alternativ kann das Seminar auch in einem Computerraum der Universität durchgeführt werden. Dadurch wird ausgeschlossen, dass Studierende ohne eigenes Gerät nicht am Seminar teilnehmen. Gleichzeitig entfallen die Vorteile, die durch die Arbeit am eigenen Gerät entstehen.

Gerät zielt darauf ab, technologiebezogenes Anwendungswissen direkt am Beispiel der Geräte zu vermitteln, die die Studierenden auch außerhalb des Seminars verwenden. Die Arbeit am eigenen Notebook und der Verzicht auf externe Hardware erhöhen außerdem die Einsteiger:innenfreundlichkeit, da die Studierenden mit einigen Grundfunktionen der Geräte bereits vertraut sind. Die gemeinsame Einarbeitung am eigenen Gerät in Präsenz ist wichtig, um im Zuge einer kleinschrittig erfolgenden Annäherung an die neuen Methoden eventuelle Berührungsängste mit digitalen technischen Verfahren abzubauen. Darüber hinaus ist der Blick über die Schulter auf den Bildschirm der Studierenden gerade bei der Installation und Anwendung der Tools wichtig, um eventuelle Fehler frühzeitig zu erkennen und gemeinsam beheben zu können. Das sog. *Trouble Shooting*, also die Behebung von Fehlermeldungen, die bereits bei der Installation oder später bei der Anwendung der meisten CLS-Methoden auftreten, wird zum Seminargegenstand und zur Herausforderung, die gemeinschaftlich gemeistert wird. Im Hinblick auf diverse Betriebssysteme und Kenntnisstände über Funktionsweisen und Ordnungsstrukturen der Computer sowie unterschiedliche Nutzungsrouterinen der Studierenden erweist sich die Vermittlung von technologiebezogenem Wissen als besonders elementar. Um bspw. mit eigenen Trainingsdaten ein Modelltraining durchzuführen und ein Textkorpus automatisch annotieren zu lassen, muss die Ordnerstruktur des Computers bekannt sein. Das mag trivial klingen, weicht in vielen Fällen aber von den alltäglichen Gebrauchsszenarien der Studierenden ab und sollte deshalb als Seminarinhalt berücksichtigt werden. Dabei gilt es – so meine Perspektive – transparent zu machen, dass *Trouble Shooting* ein mehr oder weniger präsenter Bestandteil der CLS ist und vermeintlich triviale Angelegenheiten – wie ein fehlendes Leerzeichen im Eingabefeld – über das Gelingen oder Misserfolg einer Korpusanalyse entscheiden kann.

Technologiebezogenes fachliches Wissen (*Technological Content Knowledge*, TCK): „Damit ist Wissen gemeint, das technologisches und schulfachbezogenes Wissen verbindet. Fast jedes Fachgebiet erlebt aktuell Veränderungen im Zusammenhang mit neuen Technologien. [...] Lehrpersonen müssen deshalb wissen, wie Technologien in den Hintergrunddisziplinen ihrer Schulfächer zur Wissensgenerierung, Wissensrepräsentation, Wissensnutzung und Wissenskommunikation eingesetzt werden.“ (Schmid, Krannich & Petko 2020, 119) So auch der Deutschunterricht, als dessen ‘Hintergrunddisziplin’ ich hier die *Computational Literary Studies* stark machen möchte. Die TCK-Komponente soll das Seminar besonders in der ersten Säule aufgreifen, indem ein Grundstock an digitalen Verfahren der Textanalyse und Verfahren des überwachten und unüberwachten Lernens zur Korpusanalyse theoretisch und praktisch unterrichtet werden und zu einem reflektierten Umgang mit digitalen Verfahren der Textanalyse angeleitet wird. Zum technologiebezogenen Wissen für angehende Deutschlehrer:innen zähle ich eine Einführung

in die zentralen Begrifflichkeiten der CLS und ein Kennenlernen der DH als Fachdisziplin sowie eine Einführung in die Korpuskonstituierung. Darüber hinaus gehört die digitale manuelle Annotation (mit dem Textanalysetool *CATMA*, Gies et al. 2024) als digitales *Close-Reading*-Verfahren zu den Seminarinhalten. Im Themenkomplex *Distant Reading* werden Topic Modeling (*Orange Data Mining*), Sentimentanalyse (*SentText*), gKI (*UHGP* und *FelloFish*) und überwachtes maschinelles Lernen (*StanfordNER*-Toolkit; Finkel, Grenager & Manning 2005) vermittelt.

Technologiebezogenes pädagogisches Wissen (*Technological Pedagogical Knowledge*, TPK): Mit diesem Wissensbereich ist im deutschsprachigen Raum in der Regel das mediendidaktische (Lernen mit Medien), medienerzieherische oder medienbildnerische (Lernen über Medien) Wissen gemeint (Schmid, Krannich & Petko 2020, 119). Diesen Aspekt greift das Seminar vor allem in der zweiten konzeptuellen Säule auf. Zum einen soll eine Grundlage geschaffen werden, die bestehende mediendidaktische Modelle und bildungspolitische Rahmenbedingungen, die die Unterrichtsplanung und den Schulalltag mitbestimmen, beinhaltet. Zum anderen geht es um eine Rückbindung zwischen Tools und Medien, die im Rahmen der CLS verwendet werden, und Medienmodellen und -typologien, die seitens der Mediendidaktik zur Verfügung stehen. Diesem Wissensbereich nähert sich das Seminar durch die Auseinandersetzung mit bildungspolitischen Rahmenbedingungen und Medientypologien, durch die Auswertung von Studien zum Einsatz digitaler Medien in der Schule und durch die Auseinandersetzung mit medienwissenschaftlichen Modellen (TPACK, SAMR-Modell).

Technologiebezogenes fachdidaktisches Wissen (*Technological Pedagogical Content Knowledge*, TPCK): In diesem Wissensbereich fallen die Wissensbereiche und Schnittbereiche ohne Technologie-Bezug, fachliches Wissen (CK), pädagogisches und psychologisches Wissen (PK), mit dem technologischen Wissen (TK) zusammen. Dieses Wissen „[...] bezieht sich darauf, wie eine bestimmte Technologie in einem bestimmten Fachgebiet didaktisch sinnvoll eingesetzt werden kann. Dieses hochspezifische Wissen bildet den Kern des TPACK-Modells, da erst in der Kombination sinnvolles Planungs- und Handlungswissen entsteht“ (Schmid, Krannich & Petko 2020, 119). Der globale Seminarverlauf ist an den Teilbereichen der Säulen orientiert. In einer Seminarsitzung steht jeweils die Einarbeitung in ein ausgewähltes Verfahren der digitalen Literaturwissenschaft im Vordergrund, in der darauffolgenden Sitzung ein mediendidaktisches oder literaturdidaktisches Konzept. Beide Säulen werden in möglichst jeder Sitzung am ‘Runden Tisch’ zusammengebracht, um eine isolierte Betrachtungsweise und den Aufbau von Inselwissen in einem Teilbereich zu vermeiden und die Aktivierung von TPCK zu begünstigen. Bei methodisch-praktischen Sitzungen werden Gelingensbedingungen und didaktische Grundlagen für einen Transfer in den schulischen Bereich reflektiert.

Bei Sitzungen zu mediendidaktischen Modellen wird überprüft, zu welchen bereits behandelten DH-Methoden dieses Modell passen könnte.

Die übrigen nicht-technologiebezogenen Komponenten des Wissensmodells tauchen im Seminar stetig auf, aber eher als Querschnittsthema und nicht klar durch die Taktung nach Sitzungen abgegrenzt. Die Bereiche pädagogisches und psychologisches Wissen sowie fachdidaktisches Wissen werden in einzelnen Sitzungen und spätestens bei der Anfertigung der Hausarbeiten aufgegriffen. Hier gilt es, vertiefend darzustellen und auszuarbeiten, welche literaturdidaktischen Ansätze auf die CLS-Analyse übertragbar sind, und eigene Ideen für den Unterricht zu entwickeln. Fachliches Inhaltswissen wird durch die Bearbeitung möglichst konkreter Forschungsfragen, die mittels digitaler Verfahren der Textanalyse bearbeitet werden, in jeder Sitzung mit CLS-Schwerpunkt mitbehandelt. Um zu zeigen, mit welchen Methoden welche Fragestellungen beantwortet werden können, eignet sich der Einbezug (im Impulsvortrag oder als vorbereitend zu lesende Seminarlektüre) von Fallstudien und / oder Konferenzabstracts, in denen Analysen in der Regel besonders kurz und bündig vorgestellt werden. Im Sinne der interdisziplinären Brückebildung erweist es sich außerdem als wichtig, die literaturwissenschaftliche Traditionslinie der digitalen Verfahren sichtbar zu machen (die die Studierenden in vielen Fällen in entsprechenden literaturwissenschaftlichen Fachseminaren bereits kennengelernt haben). Sowohl die manuelle Annotation von Textphänomenen als auch die Analyse epochentypischer Themen und emotionstragender Textstrukturen oder die Frage nach Autor:innenschaft und Schreibstilen gehören zum ‘traditionellen’ literaturwissenschaftlichen Forschungsdiskurs. Der digitale Zugriff auf Einzeltexte oder Textkorpora bietet eine andere Perspektive auf ‘klassische’ literaturwissenschaftliche Forschungsfragen. Er kann neue Erkenntnisse hervorbringen, alte Erkenntnisse in Frage stellen oder diese bestätigen. Dabei bleibt der Rückbezug auf bestehende Forschungsdiskurse und -ergebnisse bei der Auswertung und Interpretation der Ergebnisse digitaler Textanalysen ein genauso elementarer Bestandteil des wissenschaftlichen Arbeitens wie in literaturwissenschaftlichen Analysen, die ohne digitale Komponente auskommen. Dennoch unterscheiden sich Forschungsarbeiten der digitalen Literaturwissenschaft nicht nur inhaltlich, sondern auch formal und im Aufbau, bspw. durch ihr experimentelles Studiendesign, die Zitierweise für digitale Texte und Tools, den Einbezug von Visualisierungen und Datensätzen, von Forschungsbeiträgen aus der nicht-digitalen Literaturwissenschaft. Auch im Hinblick auf die Wissenschaftssprache für Korpusanalysen erweist sich der Einbezug von Fallstudien als Orientierungshilfe. Darüber hinaus können auf das Seminarthema abgestimmte Templates und Leitfäden eine hilfreiche Unterstützung für das Verfassen von Hausarbeiten sein.

4.1 Digitale Textkorpora und Korpusanalysen im Seminar „Digitale Literaturwissenschaft und pädagogische Praxis“

Computerlesbare Textkorpora bilden die Grundlage digitaler Textanalysen und sind ein Querschnittsthema der CLS. Dabei heben Herrmann und Lauer (2018) hervor, dass es trotz möglicher Rückbezüge auf die literaturwissenschaftliche Fachtradition und analog zum sprachwissenschaftlichen Teilbereich der Korpuslinguistik, die Teildisziplin der Korpusliteraturwissenschaft noch nicht gibt (Herrmann & Lauer 2018, 129). Das sei unter anderem darauf zurückzuführen, dass die Literaturwissenschaft und literaturwissenschaftliche Editionen stark am Kanon besonderer Werke ausgerichtet seien (Herrmann & Lauer 2018, 127). Im Fokus vieler literaturwissenschaftlicher Arbeiten stehe also eher das Besondere und nicht das Typische: „Besondere Texte werden hier vor allem als exemplarische Belege eingeschaltet oder auch in genaueren Einzelanalysen untersucht, nicht aber in größerem Umfang und zur Beschreibung genereller Trends oder allgemeiner Phänomene kompiliert.“ (Herrmann & Lauer 2018, 127) Nun schafft die Digitalisierung großer Bestände die Grundlage für eine Korpusliteraturwissenschaft, die abseits ausgewählter Einzeltexte oder Werkausgaben kanonisierter Autor:innen arbeitet (Herrmann & Lauer 2018, 129) und den als *the great unread* bezeichneten nicht kanonisierten, zahlenmäßig aber weit überlegeneren Bestandteil der Literaturproduktion, in den Blick zu nehmen (Moretti 2000).⁷ Das wohl bekannteste und umfangreichste Digitalisierungsprojekt ist sicherlich Google Books mit über 40 Millionen Büchern. Darüber hinaus gibt es historische und gattungsspezifische digitale Sammlungen, bspw. das gattungsübergreifend aufgebaute Deutsche Textarchiv (DTA) mit über 4000 deutschsprachigen Druckwerken aus der Zeit 1600 bis 1900 (Geyken et al. 2018), das Textgrid Repository mit über 600 deutschsprachigen literarischen Texten aus dem 15. bis 20. Jahrhundert (Wegstein, Rapp & Jannidis 2015) oder die Dramensammlung DraCor mit über 700 deutschsprachigen Dramen aus der Zeit zwischen 1510 bis 1947 und weiteren Dramenbeständen aus dem französischen, russischen, spanischen, englischen, niederländischen, italienischen, ungarischen, schwedischen, polnischen, ukrainischen, griechischen oder amerikanischen Sprachraum (Fischer et al. 2019). Unter einer Textsammlung oder einem Textkorporus ist in diesem Beitrag also eine bestehende größere Sammlung von kanonisierten und nicht-kanonisierten Texten gemeint, die meistens nach Epoche, Sprache, Textsorte oder

⁷ Moretti zufolge fokussiert die Literaturwissenschaft weniger als 1 % der gesamten literarischen Produktion und beschäftigt sich vorrangig mit immer denselben Texten und Autor:innen. Hermann und Lauer beschreiben diesen vergleichsweise kleinen Teil der gesamten Literaturproduktion als „die besonderen Werke“ (Herrmann & Lauer 2018, 127). Mit dem Fokus auf die wenigen bekannt gewordenen Werke und Autor:innen ist die Forschung auf die absoluten Ausnahmeherscheinungen ausgerichtet und blendet einen Großteil aller literarischen Werke aus. Herrmann und Lauer beschreiben den großen Anteil der unberücksichtigten Werke als das eigentlich Typische der Literatur (Herrmann & Lauer 2018, 127), Moretti in Anlehnung an Margaret Cohen (1999, 23) als *the great unread* (Moretti 2000).

Autor:in zusammengestellt sind. Im Hinblick auf konkrete Untersuchungsfragen, die die Studierenden bspw. in Hausarbeiten ausformulieren, wird aus der Textsammlung in der Regel eine Stichprobe (auch als Teilkorpus bezeichnet) zusammengestellt, anhand dessen der Untersuchungsfrage nachgegangen wird.⁸

Im Seminarkontext spielen Textsammlungen außerdem eine gleichbleibend große Rolle, wird doch für jedes Verfahren und jede Beispielanalyse ein Textkorpus benötigt, mit dem geübt und gearbeitet werden kann.

In einer einführenden Sitzung zum Thema Korpuskonstituierung werden zunächst etablierte, wissenschaftlich valide Datenquellen vorgestellt und unter quellenkritischer sowie didaktischer Perspektive diskutiert. Für diese Sitzung lesen die Studierenden vorbereitend einen Fachtext, der wissenschaftlich fundiert und niedrigschwellig in die literaturwissenschaftliche Korpusbildung einführt und via *forTEXT* – einer digitalen Plattform mit einsteiger:innenfreundlich aufbereiteten Einführungstexten zu Methoden der digitalen Textanalyse – zur Verfügung gestellt wird (Bläß 2024 [2020]). Im Seminar selbst folgt auf eine kurze Diskussion des Fachtextes eine theoretische Einführung in die Korpuskonstituierung mitsamt Kriterien zum Aufbau von Datensammlungen (nach Schöch 2017). Anschließend arbeiten die Studierenden in Kleingruppen mit unterschiedlichen digitalen Textsammlungen aus dem Bereich der digitalen Literaturwissenschaft. Dabei halten sie fest, welcher Sammelschwerpunkt dem Repertorium zugrunde liegt, in welchen Formaten die Texte angeboten werden und welche Filterfunktionen, implementierten Analysetools und Kriterien zur Qualitätskontrolle vorliegen. Darüber hinaus skizziert jede Kleingruppe ein Unterrichtsszenario, das sich mit dem von ihnen geprüften digitalen Repertorium verbinden ließe. Im Rahmen der Unterrichtsplanung üben sie außerdem den Download und das strukturierte Anlegen größerer thematisch kuratierter Textkorpora. Das mag trivial klingen, ist für viele Studierende aber absolutes Neuland. Die Beschäftigung mit Korpusbildung und der Umgang mit digitalen Quellen sind in der Literaturwissenschaft relativ neue Gegenstandsbereiche, sodass fundiertes Vorwissen nicht vorausgesetzt werden kann. Im Plenum werden die unterschiedlichen digitalen Sammlungen kurz vorgestellt und die Unterrichtsideen diskutiert. Die Sitzung zu digitalen Repertorien zielt vor allem darauf ab, digitale Textkorpora vorzustellen, mit denen die Studierenden im weiteren Seminarverlauf immer wieder arbeiten werden und auf die sie auch im Rahmen ihrer Arbeit an Schulen zurückgreifen können. Die Arbeit mit Textkorpora spielt im weiteren Seminarverlauf bei der Auseinandersetzung mit ganz unterschiedlichen Seminarinhalten eine wiederkehrende Rolle (s. Tabelle 1).

⁸ Für die adäquate Auswahl einer Stichprobe eignet sich der Beitrag von Christof Schöch (2017). Darin werden drei Strategien (repräsentative Stichprobe, balancierte Sammlung oder opportunistische Auswahl) zur Zusammenstellung einer Textsammlung, bei Schöch als Datensatz bezeichnet, vorgestellt und (statistische) Herausforderungen erläutert.

	Textkorpus	Seminarinhalt
1	‘Korpus bildungspolitische Schlüsseltexte’: KMK-Strategie „Bildung in der digitalen Welt“, Erweiterung der KMK-Strategie, Dagstuhl Erklärung, Erweiterung der Dagstuhl-Erklärung, Strategiepapier der Gesellschaft für Fachdidaktik (GFD) „Fachliche Bildung in der digitalen Welt“, DigitalPakt Schule	bildungspolitische Rahmenbedingungen
2	d-Prose 1870–1920 (2511 literarische Prosatexte)	Distant Reading mit <i>Voyant</i>
3	d-RoRo (115 Romane aus der Romantik)	Stilometrische Analyse mit <i>Stylo</i>
4	TextGrid Repository, Fantasykorpus	Sentimentanalysis mit <i>SentText</i>
5	Deutsches Textarchiv und / oder Andersen-Korpus	Topic Modeling
6	Materialplattform Grundschullehramt Deutsch	Prompt-a-thon mit <i>UHHGPT</i> ⁹ und <i>FelloFish</i>

Tabelle 1. Textkorpora und ausgewählte Sitzungsinhalte im Überblick

Nachdem im vorangegangenen Teil die Sitzungsinhalte vorgestellt wurden, in denen Textkorpora eine besonders wichtige Rolle spielen, soll im folgenden Teil unter Einbezug des TPACK-Modells erläutert werden, welche Wissensbereiche bei der Verbindung von Korpusliteraturwissenschaft und Schule aktiviert werden.

Technologiebezogenes Wissen umfasst nicht nur das Wissen über die Anwendung digitaler Verfahren der Textanalyse, die für Korpusanalysen eingesetzt werden können. Neben einem Überblick über bestehende Korpora, die nachgenutzt werden können, gehören auch Kenntnisse über den Download digitaler Textsammlungen, grundlegende Formatierungsarbeiten sowie Schritte des *Preprocessing* zum technologiebezogenen Wissen für Korpusanalysen. Konzeptionelles Wissen beinhaltet grundlegendes Wissen über Fachbegriffe aus dem Bereich der Korpuskonstituierung und Korpusanalyse, Strategien zum Aufbau von Datensammlungen und eine Auseinandersetzung mit Bias-Problematiken, die mit dem Aufbau von Datensammlung und deren Analyse einhergehen.

Im Bereich des technologiebezogenen fachlichen Wissens steht der Transfer digitaler Verfahren in den Literaturunterricht im Vordergrund. Ein Einbezug des „schulfachbezogenen Wissens“ bedeutet im Falle einer Korpusanalyse ein Abgleichen mit und

⁹ Das Format Prompt-a-thon wurde an der Universität Hamburg entwickelt und ist angelehnt an den in der Informatik gängigen Hack-a-thon, bei dem kollaborativ und in einer vorgegebenen Zeit an einer Lösung zu einer Herausforderung oder einer Idee gearbeitet wird. Der Prompt-a-thon ist also eine niedrigschwellige, praktische Einführung in der Arbeit mit gKI. Nach einer kurzen theoretischen Einführung arbeiten die Teilnehmenden kollaborativ an sogenannten *Challenges* / Herausforderungen. Dabei handelt es sich um vordefinierte spezifische Aufgaben, die in den Kleingruppen mithilfe von gKI gelöst werden. Die Universität Hamburg ermöglicht Mitarbeitenden und Studierenden seit April 2024 durch die universitätseigenen KI *UHHGPT* einen datenschutzkonformen Zugang zu *ChatGPT*.

Andocken an fachliche und jahrgangsspezifische Anforderungen (Lehrpläne für das Unterrichtsfach Deutsch, Bildungsstandards und KMK-Kompetenzen). Dabei geben die fachlichen Anforderungen einen Rahmen für die Korpuskonstituierung, indem sie bspw. die Auseinandersetzung mit ausgewählten Epochen und Gattungen vorschreiben und dadurch eine entsprechende Korpuskonstituierung nahelegen. Wenn bspw. im Textgrid-Repository eine Sammlung mit allen Märchen von Hans Christian Andersen im TXT-Format heruntergeladen wurde, müsste die geplante digitale Märchenanalyse im Lehrplan verortet werden und – je nach Methode der Korpusanalyse – eine KMK-Kompetenz herausgestellt werden, die im Rahmen der Unterrichtsstunde oder -einheit geschärft werden soll. Soll die Funktionsweise von Algorithmen diskutiert werden, bietet sich als Methode zur Analyse des Märchenkorpus ein Distant-Reading-Verfahren an. Im Hinblick auf eine fiktive Lerngruppe gilt es dann, einen passenden Unterrichtsentwurf zu entwickeln, der konkrete Lernziele und Unterrichtsmaterialien (Arbeitsblätter, Anleitung durch die Lehrkraft etc.) beinhaltet. Dazu muss die Analyse zunächst einmal selbst durchgeführt werden und dann für den Schulunterricht quasi in Einzelteile zerlegt werden, die für Schüler:innen umsetzbar sind. Die Verortung von Korpusanalysen im TPACK-Modell offenbart die unterschiedlichen Wissensbereiche, die im Hinblick auf Korpusanalysen für angehende Lehrkräfte von Bedeutung sind. Technologiebezogenes fachdidaktisches Wissen bedeutet im Hinblick auf Korpusanalysen in den meisten Fällen eine Reduktion der Komplexität. Nachdem die Studierenden Funktionsweise und Einsatz digitaler Verfahren der Korpusanalyse erlernt haben, müssen sie identifizieren, welche Komponenten unmodifiziert in den Schulunterricht übertragen werden können und welche Komponenten im Hinblick auf die (fiktive) Lerngruppe und das schulfachbezogene Wissen modifiziert werden sollten. In dieser Transferleistung liegt eine große Herausforderung. Sie kann im Seminar durch Diskussionen von Unterrichtskonzepten angebahnt werden und im Rahmen von Hausarbeiten weiter ausgearbeitet werden.

5 CLS-Korpusmethodik in der Lehre: Seminarresultate und abschließende Überlegungen

Wie die Verortung der am Brückenschlag zwischen CLS und Didaktik beteiligten Wissensbereiche gezeigt hat, erfordert das Seminar vor allem Transferleistungen. Die Verfahren der CLS stehen zwar in einer langen literaturwissenschaftlichen Traditionslinie und dürften in Grundzügen bekannt sein, sie erfordern neben der reinen Anwendungskompetenz aber auch einen neuen Blick auf die Analyse literarischer Texte. Von der Auswahl der Texte über die Definition des Phänomens bis zur Durchführung der Analyse und der Auswertung der Ergebnisse sind alle Arbeitsschritte an den Prämissen quantitativer Analysen ausgerichtet. Ein quantitativer Zugriff auf literarische Texte ist für die Studierenden in der Regel absolutes Neuland. Darüber hinaus ist auch die Übertragung in den schulischen Bereich herausfordernd, da die ‘frisch’ erlernten CLS-Verfahren mit den

Anforderungen einer fiktiven Lerngruppe abgeglichen und an die Kompetenzbereiche und Bildungsstandards angedockt werden müssen. Damit diese Transferleistungen gelingen, braucht es – neben der reinen Vermittlung von Fachwissen aus dem Bereich CLS und Didaktik – vor allem praktische Übung, Zeit und intensiven diskursiven Austausch. In einer zweiten Durchführung des Seminars im Wintersemester 2024/25 wurde ein Teil des Seminars im Block durchgeführt. Gerade im Hinblick auf komplexere Verfahren, in diesem Fall das Machine-Learning-Training eines Classifiers zur automatischen Annotation und eine damit vorgesehene Korpusanalyse, erweist sich das Blockseminar als Format der Wahl. Für das Modelltraining, die Anwendung des Classifiers, die Analyse und deren Auswertung braucht es – noch dringender als bei den übrigen Verfahren – Zeit am Stück für die technische Einarbeitung.

Im gesamten Seminar erweisen sich digitale Repositorien als wichtige Grundlage. Korpusanalysen stellen ein Querschnittsthema dar, wobei unterschiedliche Textkorpora jeweils im Rahmen von unterschiedlichen Analyseverfahren wie Stilometrie, Topic Modeling, Sentimentanalyse oder mithilfe automatischer Annotation untersucht werden.

Angesichts der vorgesehenen Kompetenzorientierung äußern sich die Studierenden oft kritisch bis ratlos. Im Gespräch wird immer wieder der Wunsch nach praktischen Anwendungsszenarien und konkreten Beispielen artikuliert, die zeigen, wie genau Literaturunterricht mit digitaler Komponente aussehen kann. Dass in Form der CLS eine Verbindung zwischen Algorithmen und Literatur besteht, wird in der Regel beinahe erleichtert, neugierig und gleichzeitig kritisch-abwartend wahrgenommen. Kann Literatur wirklich digital erforscht werden? Werden digitale Verfahren der Spezifik literarischer Texte gerecht? Welche Fachinhalte des Literaturunterrichts passen zu welcher CLS-Methode? Welche Phänomene erfasst der Computer anders als der Mensch und: Geht eine digitale Korpusanalyse wirklich immer schneller? Diese und viele weitere Fragen sind wiederkehrende Diskussionsinhalte des Seminars.

Der Blick auf die bereits eingegangenen Abschlussarbeiten zeigt, dass die Studierenden kreative und kluge Ideen für den Literaturunterricht mit digitaler Komponente entwickeln. So wurden bspw. Unterrichtsentwürfe zum Thema digitale Annotation mit dem Textanalysetool *CATMA* am Beispiel des Romans *Ellbogen* von Fatma Aydemir oder „Eine didaktische Annäherung am Beispiel von Wolfgang Borcherts Kurzgeschichte *Das Brot*“ ausgearbeitet. Weitere Hausarbeiten befassen sich mit der Anwendung von Sentimentanalyse oder dem Distant-Reading-Tool *Voyant* im Literaturunterricht – und sogar eine Bachelorarbeit zum Thema „Digital Humanities im Deutschunterricht – Voraussetzungen und Strategien für den Transfer in den schulischen Bereich“ ist aus dem Seminar hervorgegangen. Im Hinblick auf die Hausarbeiten fällt auf, dass die meisten Studierenden Thema und Untersuchungsfrage so wählen, dass eine digitale Analyse ei-

nes Einzeltextes durchgeführt werden kann. Unter methodischer Perspektive war die digitale manuelle Annotation mit dem Textanalysetool *CATMA* eindeutig am beliebtesten, wobei das Tool zur Analyse unterschiedlicher Phänomene und Texte herangezogen wurde und unterschiedliche Szenarien für den Schulunterricht entwickelt wurden.

Abschließend möchte ich noch einmal auf die im ersten Teil skizzierte und aktuell sehr lebendige Debatte über gKI in der Lehre zurückkommen. Die Auseinandersetzung mit gängigen Verfahren der digitalen Literaturwissenschaft, wie manuelle Annotation und Machine-Learning-Verfahren, bilden eine Brücke hin zum Verständnis von komplexeren Verfahren und Zusammenhängen, wie der Funktionsweise von LLMs. Die Seminarinhalte sind also bewusst so angeordnet, dass sie markante Entwicklungsstufen der CLS wiedergeben und der Komplexitätsgrad langsam steigt. Nun legt die aktuellebildungspolitische Debatte einen Fokus auf die Auseinandersetzung mit LLMs nahe. Die Auseinandersetzung mit Sprachmodellen zu priorisieren und isoliert von den Grundlagen der CLS zu betrachten, hieße in gewisser Weise, das Pferd von hinten aufzäumen. CLS-Grundlagenwissen bildet den Ausgangspunkt für ein tiefergehendes und kritisch-reflektiertes Verständnis von Tools und Verfahren, die mit gKI arbeiten. Die Auseinandersetzung mit gKI und deren Funktionsweisen baut auf Grundlagenwissen der CLS auf. Um die Empfehlungen der Ständigen Wissenschaftlichen Kommission nachkommen zu können und einen hinreichenden Wissensaufbau zu ermöglichen, bedarf es aus meiner Perspektive eines sukzessiven Wissensaufbaus (ausgehend von den CLS-Grundlagen) und keines Ad-hoc-Einstiegs beim komplexesten Lerninhalt (Funktionsweise und Anwendung generativer KI).

In einer der wenigen deutschsprachigen Einführungen in die Digital Humanities werden vier Bereiche benannt, die in der Literaturwissenschaft im Kontext der Digitalisierung und Vernetzung Veränderungen erfahren haben. Beides, sowohl die Digitalisierung als auch eine dadurch erhöhte Vernetzung mit anderen Fachbereichen, betreffen den Gegenstandsbereich der Literaturwissenschaft, wozu hier vor allem Texte gezählt werden (1). Eine Veränderung erfahren außerdem die Kommunikation über Texte (2), die Methoden zu deren Aufbereitung und die Analyseverfahren (3) und die Zusammenarbeit mit Archiven und Bibliotheken (4) (Jannidis 2022, 1 f.). Abschließend möchte ich die Lehre, also die Vermittlung von literaturwissenschaftlichem Wissen und literaturwissenschaftlichen Fertigkeiten, als fünften und weiteren Teilbereich der Literaturwissenschaft stark machen. Verändern sich die Fachinhalte, so verändern sich auch die Lehrinhalte und die Art und Weise, sie zu vermitteln. Einem Verständnis wissenschaftlicher Verwandtschaftsverhältnisse folgend, das Deutschdidaktik und nicht-digitale Literaturwissenschaft als (dialogbedürftige) Bezugswissenschaften betrachtet (Boyken 2016, 34), begreife ich die digitale Literaturwissenschaft auch als Bezugswissenschaft der Literaturdidaktik. Die zielgruppenorientierte Vermittlung einer literaturwissenschaftlichen

Grundausbildung oder Basiskompetenz fasse ich als weiteren Teilbereich auf, der durch die Digitalisierung und Vernetzung eine Veränderung erfährt. Zielgruppenorientierung bedeutet in diesem Fall, dass angehende Lehrkräfte einen Grundstock an Verfahren der literaturwissenschaftlichen Korpusanalyse kennenlernen und dazu angeregt werden, den Transfer in den Schulunterricht bereits in der ersten Ausbildungsphase mitzudenken. Da Technologien schnelllebig sind, ist es außerdem wichtig, Studierende über den universitären Kontext hinaus dazu zu befähigen, aus dem breiten Spektrum eigenständig eine fundierte und lerngruppenorientierte Auswahl treffen zu können. Neben der Einarbeitung in die Methoden wird insbesondere die Frage der Komplexitätsreduktion und des schulischen Anwendungsbezuges immer wieder durchdacht und neu verhandelt werden müssen. Um begründet beurteilen zu können, welche Analyseverfahren und Tools sich für den Einsatz im Schulunterricht eignen und welche nicht, müssen die unterschiedlichen Wissensbereiche ‘angezapft’ und miteinander in Verbindung gesetzt werden. Auf diese Transferleistung möchte der vorliegende Beitrag aufmerksam machen und in Form des oben skizzierten Seminarkonzepts einen Vorschlag zur Diskussion stellen, der die Vermittlung und Vernetzung von Fachwissen aus unterschiedlichen Bereichen vorsieht.

References

- Albrecht, Christian. 2014. A6 Fachspezifische mediendidaktische Konzeptionen. *Digitale Medien im Deutschunterricht*. Deutschunterricht in Theorie und Praxis. Handbuch zur Didaktik der deutschen Sprache und Literatur in elf Bänden. Band 8. Volker Frederking, Axel Krommer & Thomas Möbius, Hg. Baltmannsweiler: Schneider Verlag Hohengehren. 134–149.
- Andresen, Melanie. 2024. *Computerlinguistische Methoden für die Digital Humanities. Eine Einführung für Geisteswissenschaftler:innen*. Tübingen: Narr Francke.
- Archer, Jodie & Matthew L. Jockers. 2017. *Der Bestseller Code. Was uns ein bahnbrechender Algorithmus über Bücher, Storys und das Lesen verrät*. New York: St. Martin’s Press.
- Aßmann, Sandra. 2017. Medienpolitische Positionen, Forderungen und Strategien. *Medienkompetenz. Herausforderung für Politik, politische Bildung und Medienbildung*. Harald Gapski, Monika Oberle & Walter Staufer, Hg. Bonn: Bundeszentrale für politische Bildung. 136–145.
- Battershill, Claire & Shawna Ross. 2022. *Using Digital Humanities in the Classroom: A practical introduction for teachers, lecturers, and students*. London: Bloomsbury Academic.
- Bläß, Sandra. 2024 [2020]. Methodenbeitrag: Korpusbildung. *forTEXT* 1 (2). <https://doi.org/10.48694/fortext.3808>
- Bläß, Sandra, Marie Flüh, Dominik Gerstorfer, Evelyn Gius, Malter Meister, Julia Nantke & Mareike Schumacher. 2022. Forschendes Lernen digital. *DHd 2022 Kultu*

ren des digitalen Gedächtnisses. 8. Tagung des Verbands Digital Humanities im deutschsprachigen Raum (DHd 2022), Potsdam. <https://doi.org/10.5281/ZENO-DO.6327933>

- Boyken, Thomas. 2016. Über wissenschaftliche Verwandtschaftsverhältnisse. Versuch einer Einordnung der aktuellen Entwicklungen innerhalb der Deutschdidaktik aus literaturwissenschaftlicher Sicht. *Interdisziplinäre Forschung in der Deutschdidaktik. „Fremde Schwestern“ im Dialog. Positionen der Deutschdidaktik. Theorie und Empirie.* Band 2. Iris Winkler & Frederike Schmidt, Hg. Frankfurt am Main: Peter Lang. 23–41.
- Brand, Tilman von. 2015. Umgang mit Medien. *Methoden im Deutschunterricht. Exemplarische Lernwege für die Sekundarstufe I und II.* Tilman von Brand, Jürgen Baurmann, Wolfgang Menzel & Kaspar H. Spinner, Hg. Seelze: Kallmeyer Klett. 107–136.
- Cardoso, Amílcar, Tony Veale & Geraint Anthony Wiggins. 2009. Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI magazine* 30 (3), 15–22. <https://doi.org/10.1609/aimag.v30i3.2252>
- Cohen, Margaret. 1999. *The Sentimental Education of the Novel.* Princeton (NJ): Princeton University Press.
- digital.learning.lab.* <https://digitallearninglab.de/>. Zugriff: 7. Februar 2025.
- FelloFish* – Feedack für alle. <https://www.fellofish.com/>. Zugriff: 24. April 2025.
- Finkel, Jenny Rose, Trond Grenager & Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, 363–370. Available at: <http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf>. Zugriff: 24. April 2025.
- Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling & Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019: „Complexities“.* Utrecht University. <https://doi.org/10.5281/zenodo.4284002>
- Flüh, Marie. 2024 [2020]. Schulunterricht: Textanalyse mit CATMA unterrichten. *forTEXT* 1 (4). <https://doi.org/10.48694/fortext.3755>
- Flüh, Marie 2024a [2019]. Schulunterricht: Textvisualisierung mit Voyant unterrichten. *forTEXT* 1 (5). <https://doi.org/10.48694/fortext.3774>
- Flüh, Marie 2024b [2019]. Lehrmodul: Netzwerkanalyse mit Ezlinavis unterrichten. *forTEXT* 1 (6). <https://doi.org/10.48694/fortext.3781>
- Frederking, Volker & Axel Krommer. 2014. A7 Deutschunterricht und mediale Bildung im Zeichen der Digitalisierung. *Digitale Medien im Deutschunterricht. Deutschunterricht in Theorie und Praxis.* Handbuch zur Didaktik der deutschen Sprache

- und Literatur in elf Bänden. Band 8. Volker Frederking, Axel Krommer & Thomas Möbius, Hg. Baltmannsweiler: Schneider Verlag Hohengehren. 150–182.
- Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas & Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. *Digitale Infrastrukturen für die germanistische Forschung*. Henning Lobin, Roman Schneider & Andreas Witt, Hg. Berlin, Boston: de Gruyter. 219–248.
- Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Dominik Gerstorfer & Mari Akazawa. 2024. *CATMA 7* (Version 7.1). Zenodo. <https://doi.org/10.5281/zenodo.1470118>
- Herrmann, Berenike & Gerhard Lauer. 2018. Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. *Korpuslinguistik*. Joachim Gessinger, Angelika Redder & Ulrich Schmitz, Hg. Bielefeld: Account. 127–155.
- Hirsch, Brett D. 2015. *Digital humanities pedagogy: practices, principles and politics*. Cambridge: OpenBook Publishers.
- Jackaki, Diane & Katherine Faull. 2016. Doing DH in the classroom: transforming the humanities curriculum through digital engagement. *Doing Digital Humanities. Practice, Training, Research*. Constance Crompton, Richard Lane & Ray Siemens, eds. New York: Routledge. 385–373.
- Jannidis, Fotis. 2010. Methoden der computergestützten Textanalyse. *Methoden der literatur- und kulturwissenschaftlichen Textanalyse. Ansätze – Grundlagen – Modellanalyse*. Vera Nünning & Ansgar Nünning, Hg. Stuttgart, Weimar: Metzler. 109–132.
- Jannidis, Fotis. 2022. Digitale Literaturwissenschaft. Zur Einführung. *Digitale Literaturwissenschaft*. Germanistische Symposien. Fotis Jannidis, Hg. Stuttgart: Metzler. 1–19. https://doi.org/10.1007/978-3-476-05886-7_1
- Kröncke, Merten, Leonard Konle, Simone Winko, Simone & Fotis Jannidis. 2023. Gattungen und Emotionen in der Lyrik des Realismus und der frühen Moderne. *Book of Abstracts DHd 2023 Open Humanities Open Culture. 9. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2023)*. Trier, Luxemburg. <https://doi.org/10.5281/zenodo.7715402>
- KMK 2016 = Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). 2016. *Strategie „Bildung in der digitalen Welt“*. Berlin 2016. Available at: <https://www.kmk.org/aktuelles/artikelansicht/strategie-bildung-in-der-digitalen-welt.html>.
- Kühnert, Janina. 2017. Fachdidaktisches Essay: Beispielhafte Konzeption einer Literaturunterrichtseinheit mit Voyant. *Skriptorium* 6 (2017), 41–59.
- Lauer, Gerhard. 2013. Die Vermessung der Kultur. Geisteswissenschaften als Digital Humanities. *Big Data. Das neue Versprechen der Allwissenheit*. Heinrich Geiseler & Tobias Moorstedt, Hg. Frankfurt am Main: Suhrkamp. 99–116.

- Lecke, Bodo. 2008. Medienpädagogik, Literaturdidaktik und Deutschunterricht. *Deutsch-Didaktik. Leitfaden für die Sekundarstufe I und II*. Michael Kämpfer-van-den Boogaart, Hg. Berlin. 46–57.
- Leubner, Martin. 2014. B1 Digitale literale Medien im Deutschunterricht. *Digitale Medien im Deutschunterricht*. Deutschunterricht in Theorie und Praxis. Handbuch zur Didaktik der deutschen Sprache und Literatur in elf Bänden. Band 8. Volker Frederking, Axel Krommer & Thomas Möbius, Hg. Baltmannsweiler: Schneider Verlag Hohengehren. 185–212.
- Limpinsel, Mirco. 2016. Was bedeutet die Digitalisierung für den Gegenstand der Literaturwissenschaft? *Zeitschrift für digitale Geisteswissenschaften* 1 (2016). http://dx.doi.org/10.17175/2016_009
- Mahony, Simon & Elena Pierazzo. 2015. Teaching Skills or Teaching Methodology? *Digital humanities pedagogy: practices, principles and politics*. Brett D. Hirsch, ed. Cambridge: OpenBook Publishers. 215–225. Available at: <https://books.openedition.org/obp/1639>.
- Mauro, Aaron. 2016. Digital liberal arts and project-based pedagogies. *Doing Digital Humanities. Practice, Training, Research*. Constance Crompton, Richard Lane & Ray Siemens, eds. New York: Routledge. 373–384.
- Mischke, Dennis, Peer Trilcke & Henny Sluyter-Gäthje. 2022. Hackathons als kollektiv-kreative Bildungsergebnisse. Ein Konzept zur Gestaltung offener Lehrveranstaltungen in den Digital Humanities. *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum“ (DHd 2022), Potsdam*. <https://doi.org/10.5281/zenodo.6328099>
- Mishra, Punya & Matthew J. Koehler. 2006. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record* 108 (6), 1017–1054.
- Moretti, Franco. 2000. Conjectures on World Literature. *New Left Review* 1 (2000), 54–68. Available at: <https://tinyurl.com/moretti2000conjwl>. Zugriff: 24. April 2025.
- Moretti, Franco. 2016. *Distant Reading*. Konstanz: Konstanz University Press.
- MUNDO. <https://mundo.schule/>. Zugriff: 7. Februar 2025.
- Organe Data Mining. <https://orangedatamining.com/>. Zugriff: 7. Februar 2025.
- Risam, Roopika. 2019. *New digital worlds: postcolonial digital humanities in theory, praxis, and pedagogy*. Illinois: Northwestern University Press.
- Sahle, Patrick. 2015. Digital Humanities? Gibt's doch gar nicht! *Grenzen und Möglichkeiten der Digital Humanities. Sonderband der Zeitschrift für digitale Geisteswissenschaften*. Constanze Baum & Thomas Stäcker, Hg. text/html Format. https://doi.org/10.17175/sb001_004
- Schmid, Mirjam, Maike Krannich & Dominik Petko. 2022. Technological Pedagogical Content Knowledge. Entwicklungen und Implikationen. *Journal für LehrerInnenbildung* 20 (2020) 1, 116–124. https://doi.org/10.35468/jlb-01-2020_10

- Schöch, Christof. 2017. Aufbau von Datensammlungen. *Digital Humanities: Eine Einführung*. Fotis Jannidis, Hubertus Kohle & Malte Rehbein, Hg. Stuttgart: Metzler. 223–232.
- Schumacher, Mareike. 2022. Wie ›Der Mann auf dem Hochrad‹ den Protagonisten des ›Schlangenbaums‹ auf Abwege führte. Hypothesenleitete stilometrische Untersuchung zweier Romane Uwe Timms. *Zeitschrift für digitale Geisteswissenschaften*. Wolfenbüttel 2022. https://doi.org/10.17175/2022_004
- Schumacher, Mareike & Marie Flüh. 2023. Made to Be a Woman: A case study on the categorization of gender using an individuation-based approach in the analysis of literary texts. *Digital Humanities Quarterly – Special Issue: Categories in Digital Humanities* 17 (3). Dominik Gerstorfer, Evelyn Gius & Janina Jacke, Hg. https://doi.org/10.17175/2022_004
- SentText*. <https://thomasschmidtur.pythonanywhere.com/>. Zugriff: 7. Februar 2025.
- Shanahan, Murray & Catherine Clarke. 2023. Evaluating large language model creativity from a literary perspective. *ArXiv*. <https://doi.org/10.48550/arXiv.2312.03746>
- Shulman, Lee. 1987. Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review* (57), 1–22.
- Stanford NER-Toolkit*. <https://nlp.stanford.edu/software/CRF-NER.html>. Zugriff: 5. Februar 2025.
- Stylo*. <https://github.com/computationalstylistics/stylo>. Zugriff: 24. April 2025.
- SWK 2024 = Ständige Wissenschaftliche Kommission (SWK) der KMK. 2024. *Large Language Models (LLM) und ihre Potenziale im Bildungssystem*. Available at: https://www.kmk.org/fileadmin/Dateien/pdf/KMK/SWK/2024/SWK-2024-Im-pulspapier_LargeLanguageModels.pdf. Zugriff: 5. Februar 2025.
- Tillmann, Angela. 2017. Informationsverhalten von Kindern und Jugendlichen in digital-vernetzten Welten. *Medienkompetenz. Herausforderung für Politik, politische Bildung und Medienbildung*. Harald Gapski, Monika Oberle & Walter Staufer, Hg. Bonn: Bundeszentrale für politische Bildung. 116–125.
- UHHGPT*. <https://uhhgpt.uni-hamburg.de/login.php>. Zugriff: 24. April 2025.
- Urs, Henning. 2021. *Sentimentanalyse mit SentText*. Available at: <https://web2-unterricht.ch/digitale-transformation/sentimentanalyse-mit-senttext/>. Zugriff: 7. Februar 2025.
- Voyant*. <https://voyant-tools.org>. Zugriff: 7. Februar 2025.
- Wampfler, Philippe. 2017. *Digitaler Deutschunterricht: Neue Medien produktiv einsetzen*. Göttingen: Vandenhoeck & Ruprecht.
- Wegstein, Werner, Andrea Rapp & Fotis Jannidis. 2015. Textgrid – eine Geschichte. *TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungs-umgebung für die Geisteswissenschaften*. Heike Neuroth, Andrea Rapp & Sibylle Söring, Hg. Glückstadt: Hülbusch. 23–35.

- Weitin, Thomas & Katharina Herget. 2017. Falkentopics: Über einige Probleme beim Topic Modeling literarischer Texte. *Zeitschrift für Literaturwissenschaft* 47 (1). <https://doi.org/10.1007/s41244-017-0049-3>
- Yang Zhenyuan, Zhengliang Liu, Jing Zhang, Cen Lu, Jiaxin Tai, Tianyang Zhong, Yawei Li, Siyan Zhao, Teng Yao, Qing Liu, Jinlin Yang, Qixin Liu, Zhaowei Li, Kexin Wang, Longjun Ma, Dajiang Zhu, Yudan Ren, Bao Ge, Wei Zhang, Ning Qiang, Tuo Zhang & Tianming Liu. 2024. *Analyzing Nobel Prize Literature with Large Language Models*. <https://doi.org/10.48550/arXiv.2410.18142>
- Zierer, Klaus. 2018. *Lernen 4.0. Pädagogik vor Technik. Möglichkeiten und Grenzen einer Digitalisierung im Bildungsbereich*. Baltmannsweiler: Schneider Verlag Hohengehren.

Korpusbasierte Übungsmaterialien für die Vermittlung von Korpuskompetenzen im DaF-Unterricht

Carla Sökefeld

Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6, Postfach #15
20146 Hamburg, Deutschland
E-Mail: carla.soekefeld@uni-hamburg.de
ORCID iD: 0000-0003-4352-9732

Anastasiia Stulen

Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6, Postfach #15
20146 Hamburg, Deutschland
E-Mail: anastasiia.stulen@studium.uni-hamburg.de
ORCID iD: 0009-0009-8594-5727

Abstract. Während sich die Diskussion über den Einsatz von Korpora im Sprachunterricht lange Zeit auf Englisch als Fremdsprache konzentrierte, gibt es in letzter Zeit zunehmend Veröffentlichungen, die sich mit den Vorteilen des Einsatzes von Korpora auch für Deutsch als Fremdsprache befassen. Korpusbasierte Arbeit hat sich als vorteilhaft für den Sprachunterricht erwiesen. Aus Sicht philologischer Studiengänge ist es darüber hinaus vorteilhaft, die während des gesamten Studiums benötigten Korpuskenntnisse der Studierenden zu stärken, indem korpusbasierte Arbeit bereits in den Sprachunterricht integriert wird. Dieser Beitrag stellt eine Reihe von Unterrichtsmaterialien vor: Korpusbasierte Übungen zu häufigen Wortkombinationen und das DWDS-Wortprofil sowie korpuslinguistische „Spickzettel“, d. h. Glossare, mit Erklärungen gängiger korpuslinguistischer Begriffe. Die Materialien sind so konzipiert, dass sie die Hemmschwelle von Studierenden (und Lehrenden) gegenüber der Integration von Korpusarbeit in den Deutschunterricht als Fremdsprache senken sollen. Die Unterrichtsmaterialien wurden im Rahmen der vom DAAD geförderten germanistischen Institutspartnerschaft zwischen der Universität Vilnius und der Universität Hamburg entwickelt und zielen darauf ab, drei für das Germanistikstudium im Ausland wichtige Lernbereiche zu verbinden: Deutsch als Fremdsprache, linguistisches Wissen und Korpuskompetenz. Es gibt verschiedene Fallstricke bei der erfolgreichen Umsetzung der Korpusarbeit im Deutschunterricht, z. B. die Veränderlichkeit von Online-Korpora und Korpuswerkzeugen, die in diesem Beitrag behandelt werden. Eine erste Pilotphase der Materialien zeigt, dass insbesonde-

Submitted: 18/06/2025. Accepted: 18/09/2025

Copyright © 2025 Carla Sökefeld, Anastasiia Stulen. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

re die unterschiedliche Terminologie von Korpuswerkzeugen und Lehrbüchern für die Studierenden eine Herausforderung darstellt. Dennoch argumentiert dieser Beitrag, dass die Einbindung korpusbasierter Übungen in den Sprachunterricht vorteilhaft ist, um eine Grundlage für die Korpuskompetenz der Studierenden zu schaffen, die sie während ihres gesamten philologischen Studiums erweitern können.

Schlüsselwörter: Korpuslinguistik, Korpuskompetenzen, Deutsch als Fremdsprache, Data-Driven Learning, Kollokation, Kookkurrenz

Corpus-based teaching materials for incorporating corpus literacy into German as a foreign language classes

Abstract. While the discussion of using corpora in language learning was long focused on English as a foreign language, recently, there has been an increase in publications dealing with the benefits of employing corpora for German as a foreign language as well. Corpus-based work has been shown to benefit language learning. From the perspective of philological study programs, it is moreover beneficial to strengthen students' corpus literacy needed throughout their studies by already integrating corpus work into language courses. This paper presents a set of teaching materials: Corpus-based exercises on frequent word combinations and the DWDS-Wortprofil as well as corpus-linguistic "cheat-sheets", i.e., glossaries with explanations of common corpus-linguistic terms. The way the materials are designed is intended to alleviate students' (and teachers') inhibitions when it comes to integrating corpus work into German as a foreign language classes. The teaching materials were developed in the context of the "German Studies Institute Partnership" funded by the DAAD between Vilnius University and the University of Hamburg and strive to combine three learning areas vital for German Studies abroad: German as a foreign language skills, linguistic knowledge, and corpus literacy. There are several pitfalls concerning the successful implementation of corpus work into German classes, e.g., the mutability of online corpora and corpus tools, which are addressed in this paper. An initial piloting of the materials shows that especially differing uses in terminology between corpus tools and textbooks are challenging to students. Nevertheless, this paper argues that incorporating corpus-based exercises into language classes is beneficial to building a foundation for students' corpus literacy to be expanded throughout their philological studies.

Keywords: corpus linguistics, corpus literacy, German as a foreign language, Data-Driven Learning, collocation, cooccurrence

1 Einleitung

Das grundlegende Ziel des Fremdsprachenunterrichts ist es, umfassende Kompetenzen in Wort und Schrift in der Zielsprache zu vermitteln. In den letzten Jahren nehmen zudem damit verbundene Kompetenzen, die sich auf den Umgang mit digitalen Texten und Daten im Sinne einer Digital Literacy und/oder Data Literacy beziehen, immer mehr an Bedeutung zu (vgl. Kerschhofer-Puhalo 2021, 282–283; Frank et al. 2016, 5). Auch der Begriff Corpus Literacy (vgl. Mukherjee 2002) lässt sich in diese Konzepte einreihen, besonders wenn eine Fremdsprache im Kontext eines philologischen Studiums erlernt wird, in dem die Arbeit mit Korpora in linguistischen und literaturwissenschaftlichen Lehrveranstaltungen erwartbar ist. Mukherjee (2002) beschreibt Corpus Literacy als „in einem handwerklich-methodischen Sinne eine auf den Umgang mit Korpusdaten abzielende Weiterentwicklung der *computer literacy*, andererseits aber auch eine von der Computernutzung unabhängige inhaltliche und korpusanalytische Kompetenz“ (Mukherjee 2002, 179, Hervorhebung i. O.). Neben den notwendigen technischen Fähigkeiten, Abfragen in Korpora vornehmen und mit Korpustools wie beispielsweise AntConc (Anthony 2024) oder Sketch Engine (Kilgarriff et al. 2014) Frequenzlisten und Kookkurrenzen erstellen bzw. abrufen zu können, spielt demnach auch die „dateninterpretatorische Kompetenz“ (Mukherjee 2002, 180) eine wichtige Rolle für den Umgang und die Analyse mit Korpora. Callies (2019) beschreibt Corpus Literacy ebenfalls als ein „multi-componential set of complex skills“ (Callies 2019, 247), das sowohl grundlegende Konzepte der Korpuslinguistik als auch die Fähigkeiten der Durchführung von (komplexen) Korpusabfragen und die Interpretation von Korpusdaten umfasst. Als deutschsprachiges Äquivalent zu Corpus Literacy betrachten wir Korpuskompetenz als die technische und dateninterpretatorische Fähigkeit, Korpora und Korpuswerkzeuge für (linguistische)¹ Analysen einsetzen zu können.

Das Ziel der exemplarischen korpusbasierten Übungsmaterialien zu Kollokationen und Kookkurrenzen, die dieser Beitrag vorstellt, ist es, das Erlernen einer Sprache im Fremdsprachenunterricht – in unserem Fall dem Deutschen – mit dem Erlernen solcher Korpuskompetenzen zu verbinden. Im Rahmen der Germanistischen Institutspartnerschaft (GIP) „Korpusdidaktik für (formelhafte) Fachsprache“ (KoDi-FS) zwischen der Universität Vilnius und der Universität Hamburg, die seit 2022 vom DAAD gefördert wird (vgl. Zinsmeister et al. 2025), wurden zu diesem Zweck korpusbasierte Übungsmaterialien für den DaF-Unterricht an der Universität Vilnius erstellt, die kein korpuslinguistisches Vorwissen bei den Studierenden voraussetzen.² Studierende ohne Deutschkenntnisse

¹ Während Korpuskompetenzen auch für andere geisteswissenschaftliche Fächer, beispielsweise die digitalen Literaturwissenschaften als Teilbereich der Digital Humanities, eine Rolle spielen, liegt unser Fokus auf der Linguistik.

² Neben den Autorinnen waren auch Skaistė Volungevičienė und Heike Zinsmeister an der Arbeitsgruppe beteiligt.

können an der Universität Vilnius ein Germanistik-Studium beginnen und Sprachkenntnisse des Niveaus A2/B1 bis zum Ende ihres ersten Studienjahres erlangen, um anschließend an linguistischen und literaturwissenschaftlichen Lehrveranstaltungen mit der Unterrichtssprache Deutsch teilzunehmen. Die Vermittlung von Sprachkenntnissen nimmt deshalb für viele Studierende einen großen Teil des ersten Studienjahres ein. Über die grundlegenden Kompetenzen der Digital und Data Literacy hinaus verfolgt der Einsatz korpusbasierter Übungen das Ziel, Studierende bereits während des DaF-Unterrichts an linguistische Konzepte und korpuslinguistische Methoden heranzuführen. Idealerweise kann es so gelingen, die drei Aspekte des Fremdsprachenlernens, der linguistischen Wissensvermittlung und der Vermittlung von Korpuskompetenzen miteinander zu verknüpfen und dadurch ein forschungsorientiertes Lernen zu fördern.

Im Folgenden wird zunächst der theoretische Rahmen zu korpusdidaktischen Ansätzen in der Auslandsgermanistik gesteckt, wobei Kapitel 2.1 auf die Rolle eingeht, die Korpora aktuell im DaF-Unterricht spielen, während Kapitel 2.2 Herausforderungen und Chancen korpusdidaktischer Ansätze darlegt. In Kapitel 3 werden die exemplarischen Übungsmaterialien vorgestellt, die im Rahmen von KoDi-FS erarbeitet wurden. Dabei handelt es sich um Wortschatz- und Grammatikübungen sowie einen „Spickzettel“ als korpusdidaktisches Hilfsmittel, das für Studierende und Lehrende jeweils wichtige Grundbegriffe und Konzepte aufgreift und erklärt. Diese Übungsmaterialien sollen als „Fahrplan“ gelten und können von Lehrenden entweder direkt genutzt, oder nach den eigenen Vorstellungen und Bedürfnissen der Studierenden angepasst werden. In Kapitel 4 wird der Einsatz eines Auszugs der Materialien im Rahmen eines studentischen Tutorats an der Universität Vilnius beschrieben. Kapitel 5 zieht ein Fazit über den Nutzen und die Herausforderungen beim Einsatz der korpusbasierten Übungen im DaF-Unterricht.

2 Korpuskompetenz und Korpusdidaktik für die Auslandsgermanistik

Während die Nutzung von Korpora und eine damit verbundene Vermittlung von Korpuskompetenzen lange vornehmlich im Fremdsprachenunterricht mit der Zielsprache Englisch eine Rolle spielte (vgl. Lüdeling & Walter 2009, 2), zeigt sich in den vergangenen Jahren eine vermehrte Publikationstätigkeit auch für den DaF-Bereich. Besonders hervorzuheben ist hierbei zum einen die Zeitschrift *Korpora Deutsch als Fremdsprache (KorDaF)*, die 2021 von Carolina Flinz und Britta Hufeisen mit dem Ziel gegründet wurde, „den Einsatz von Korpora für DaF/DaZ/DaM [...] zu beleben und dadurch den Nutzen von Korpora und korpuslinguistischer Forschung in diesem Feld einer interessierten und breiten Öffentlichkeit zugänglich zu machen“ (Flinz et al. 2021, 2). Im selben Jahr erschien außerdem das Themenheft „KORPORA in DaF und DaZ: Theorie und Praxis“, herausgegeben von Carolina Flinz, in der *Zeitschrift für Interkulturellen Mehrsprachigkeitsunterricht*, dessen Beiträge verschiedenste Aspekte der Arbeit mit Korpora im DaF-Unterricht beleuchten und diskutieren.

Mithilfe einer Korpusdidaktik, wie sie im Projekt KoDi-FS diskutiert wird³, sollen die oben beschriebenen Korpuskompetenzen an Studierende vermittelt werden. Im Sinne eines Data-Driven Learning (vgl. Flinz 2021, 3) für den DaF-Unterricht wurden im Projekt Übungsmaterialien erarbeitet, die Korpustools und Korpusdaten direkt in gängige Aktivitäten im Unterricht einbeziehen, wodurch die bei Lehrenden wie Studierenden vorhandene mögliche „Skepsis gegenüber diesem Ansatz gemindert werden kann“ (Flinz 2021, 3). Lehrende können Korpora also sowohl dazu benutzen, „um mit authentischen sprachlichen Daten im Unterricht induktiv (Regeln aus Daten) sowie auch deduktiv (Exemplifizierung von Regeln auf der Basis von Daten) zu arbeiten“ (Flinz 2021, 4). Im Folgenden wird das Potential des Einsatzes von Korpora im DaF-Unterricht kurz umrissen, bevor auf Herausforderungen einer Korpusdidaktik für die Auslandsgermanistik eingegangen wird.

2.1 Korpora im DaF-Unterricht: Chancen und Potentiale

Neben Publikationen, die einen Überblick über vorhandene Korpusressourcen und deren Einsatzmöglichkeiten für den DaF-Unterricht liefern (z. B. Flinz et al. 2023), und solchen, die sich auf die Arbeit mit Lernerkorpora beziehen (z. B. Hirschmann et al. 2022), stellen aktuelle Arbeiten anhand konkreter Unterrichtsszenarien bestimmte Ressourcen vor und reflektieren deren Einsatz. So untersucht beispielsweise Gredel (2021), inwiefern digitale Korpusinfrastrukturen, die Daten aus dem Bereich der internetbasierten Kommunikation bereitstellen, im Fremdsprachenunterricht die Digital Literacy von Studierenden fördern können. Dafür stellt sie fünf didaktische Szenarien anhand verschiedener Korpora von Wikipedia-Daten vor, die im Deutschen Referenzkorpus (DeReKo) über die Schnittstelle COSMAS II zur Verfügung stehen, z. B. die Untersuchung von Sozialdeixis im Kontext der Nettiquette auf den Diskussionsseiten der Wikipedia oder die Untersuchung plattformspezifischer Sprachmuster mithilfe der in COSMAS II integrierten Kookkurrenzanalyse. Die vorgestellten didaktischen Szenarien verschränken Digital Literacy im Sinne z. B. der „Einhaltung von Netiquetten“ (Gredel 2021, 128) mit Aspekten der Corpus Literacy. Abschließend attestiert Gredel CMC-Korpora „ein großes Potenzial für den DaF/DaZ-Unterricht“ (Gredel 2021, 129).

Krekeler (2021) beschreibt den Einsatz von Korpora für den Wirtschaftsdeutschunterricht und stellt konkrete Übungsaufgaben vor. Im Erfahrungsbericht wird betont, dass „die Fähigkeit zum autonomen Umgang mit Korpora über einen längeren Zeitraum entwickelt werden muss“ (Krekeler 2021, 190), im Grundstufenzbereich jedoch auch schon mit vorbereiteten Korpusabfragen oder ausgewählten Konkordanzen gearbeitet werden kann. Eine Anpassung des Unterrichtszenarios und der spezifischen Korpusarbeiten an die jeweilige Studierendengruppe ist demnach unerlässlich.

³ Vgl. <https://www.korpuslab.uni-hamburg.de/kodi-fs/korpusdidaktik.html>, letzter Zugriff 16.09.2025.

Mollica und Steyer (2024) stellen die Entwicklung korpusbasierter Übungsmaterialien für ihren Anwendungsfall des Erwerbs von Phrasemen als „echtes Desideratum“ heraus, da „die Integration von korpusbasierten Beispielen und Übungen in den Unterricht [...] das Lernen von Phrasemen nicht nur effektiver, sondern ungleich interessanter [macht]“ (Mollica & Steyer 2024, 3). Über die Aktivierung und Motivation der Studierenden durch die Arbeit mit authentischen Sprachdaten hinaus bietet die Arbeit mit Korpora zudem den Vorteil, dass sie den Zugriff auf Strukturen und Verwendungsmuster erlaubt, die in den genutzten Lehrbüchern möglicherweise gar nicht vorhanden sind (vgl. Gilquin 2022, 112). Zudem bereiten authentische Beispiele Lernende „langfristig und behutsam darauf vor [...], daß authentische Daten sich eben häufig nicht mit rigiden, schulgrammatischen Regeln erklären lassen“, ein Umstand, der „für Fremdsprachenlerner häufig Probleme [bereitet], da natürlich zunächst von vereinfachenden Regeln auszugehen ist, diese aber oftmals in Annäherung an die muttersprachliche Realität irgendwann erweitert oder modifiziert werden müssen und deren Gültigkeit insofern relativiert werden muß“ (Mukherjee 2002, 180).

Studierende können sich anhand korpusbasierter Übungseinheiten im DaF-Unterricht also einerseits Wortschatz- und Grammatikwissen an authentisch produzierten Äußerungen erarbeiten. Andererseits erlangen sie damit bereits im Zuge des Sprachunterrichts Korpuskompetenzen, die für ihr weiteres Germanistikstudium essenziell sind. Insgesamt hat Data-Driven Learning (DDL, vgl. Johns & King 1991) vielfältige Vorteile:

DDL is alleged to enhance cognitive and metacognitive skills, increase sensitivity to authentic language use, provide an interactive approach to constructivist discovery learning, foster motivation especially through individualisation, promote reusable and transferable skills, favour autonomy for life-long learning, and correspond largely to current theories of second language acquisition. (Boulton & Tyne 2013, 99)

Empirische Studien belegen beispielsweise die Förderung der schriftlichen Sprachproduktion durch die Nutzung von Korpora (vgl. Boulton & Tyne 2013, 100). Ein korpusbasierter DaF-Unterricht bietet also vielfältige Potentiale, setzt allerdings eine Vermittlung von Korpuskompetenzen voraus, die bei Studierenden nicht als gegeben angenommen werden können. Dabei ergeben sich verschiedene Herausforderungen für die Korpusdidaktik, die im Folgenden diskutiert werden.

2.2 Korpusdidaktische Herausforderungen

Verschiedene Aspekte stellen für eine „angemessene Korpusdidaktik, die ein sinnvolles Repertoire an Korpuskompetenzen identifiziert und konkrete Methoden für deren Vermittlung vorschlägt“ (Zinsmeister et al. 2025, 46) eine Herausforderung dar. Gilquin (2022) merkt beispielsweise an, dass authentische Beispiele aus Korpora zu komplex

für Lernende sein können oder dass „the more implicit and inductive approach that is typically associated with data-driven learning does not suit every student’s learning style“ (Gilquin 2022, 115). Im Folgenden werden verschiedene Aspekte ausgeführt, die für die Erstellung der korpusbasierten Übungsmaterialien eine besondere Rolle spielten.

Zum einen können die korpusbasierten Übungsaufgaben immer nur zu einem spezifischen Korpus oder Tool und darüber hinaus zum aktuellen Stand dieses Korpus oder Tools entwickelt werden. Das führt dazu, dass Materialien schnell veralten, weil sich sowohl die Nutzeroberfläche als auch die zugrundeliegenden Daten stetig weiterentwickeln können. Mit dieser Problematik ist man im Rahmen korpusgestützter Arbeit grundsätzlich konfrontiert, weshalb der Umgang mit sich wandelnden Ressourcen eine grundlegende Korpuskompetenz darstellt und von Beginn an vermittelt werden sollte. Die Veränderlichkeit von Korpustools sollte deshalb explizit thematisiert werden. Grundsätzlich ist es das Ziel, den Studierenden korpuslinguistische Grundkompetenzen zu vermitteln, sodass sie in der Lage sind, nicht nur mit veränderten Nutzeroberflächen umzugehen, sondern sich perspektivisch auch in neue Tools einzuarbeiten. Gleichzeitig ist es entscheidend, Übungsmaterialien, die vornehmlich dem Spracherwerb im DaF-Unterricht dienen, nicht mit korpuslinguistischen Ausführungen zu überfrachten: Es muss „stets eine Balance zwischen der Thematisierung solcher problematischen Aspekte und den Übungen selbst [gefunden werden], sodass nicht zu viel Zeit für das Erlernen der Sprachkompetenzen verloren geht“ (Zinsmeister et al. 2025, 55).

Eine weitere Herausforderung ist der Umgang mit den Tools selbst. Um sich mit der Darstellung und den Annotationsebenen in Korpustools zurechtzufinden, müssen nicht nur Korpuskompetenzen erworben werden, sondern auch die notwendige linguistische Terminologie vorhanden sein. Die in Korpora verwendeten Fachbegriffe können sich von denen unterscheiden, die in Lehrbüchern verwendet werden und „insgesamt in Kontrast zu den eher intuitiven Zugängen zu grammatischen Strukturen besonders in Lehrwerken für niedrige Sprachniveaus“ (Zinsmeister et al. 2025, 54) stehen. Solche Diskrepanzen müssen durch die Lehrenden aufgefangen und idealerweise bereits bei der Erstellung von Übungsmaterialien reflektiert werden.

Damit ist bereits die Rolle der Lehrenden angeschnitten, die für den erfolgreichen Einsatz korpusbasierter Übungsmaterialien nicht unterschätzt werden darf. Um eine gelungene Vermittlung von Korpuskompetenzen im DaF-Unterricht zu gewährleisten, müssen diese Kompetenzen zunächst bei den Lehrpersonen ausgebildet werden (vgl. Mollica & Steyer 2024, 6). Idealerweise führen korpusbasierte Materialien niedrigschwellig an die Arbeit mit Korpora heran, sodass auch Lehrende mit weniger Vorwissen die Möglichkeit bekommen, sich anhand der Materialien einzuarbeiten. Schließlich kann auch die Materialität selbst zur Hürde werden, wenn etwa nicht alle Studierenden einen

eigenen Computer zur Verfügung haben und es keinen Zugang zu Computerräumen oder andere Möglichkeiten gibt, die Lernenden mit eigenen Geräten für die Korpusarbeit auszustatten. Eine „weiche“ Art der Heranführung ist allerdings trotzdem über die Verwendung gedruckter Handouts möglich (vgl. Gilquin 2022, 115).

Ma (2025) gibt verschiedene Empfehlungen für eine gelungene Vermittlung von Korpuskompetenzen, unter anderem die Priorisierung essenzieller Fähigkeiten und die Thematisierung von Einschränkungen bei der Arbeit mit Korpora. Die bisherige Forschung zeigt, dass die Vorteile eines Einsatzes korpusbasierter Übungen im DaF-Unterricht die Auseinandersetzung mit den beschriebenen Herausforderungen definitiv lohnenswert machen: „Korpora können als Hilfsmittel für die Produktion und die Rezeption benutzt werden; sie erweitern den Wortschatz, sie geben Hinweise zu einem präzisen und normkonformen Sprachgebrauch; sie fördern die Sprachenbewusstheit, sie unterstützen den Lernprozess und helfen, Formulierungsschwierigkeiten zu überwinden.“ (Flinz et al. 2021, 1). Studierende profitieren demnach vom Einsatz korpusbasierter Übungsmaterialien und der dafür erforderlichen gelungenen Vermittlung von Korpuskompetenzen. Im Folgenden wird die Konzeption der exemplarischen Übungsmaterialien beschrieben.

3 Exemplarische Übungsmaterialien

Die vorgestellten exemplarischen Materialien wurden für den Kurs „Gegenwartsdeutsch“ im B.A.-Studiengang „Germanistik“ an der Universität Vilnius konzipiert. Mit dem Thema „Feste feiern“ wurde ein Rahmen gewählt, der die Wortschatz- und Korpusarbeit für die Studierenden möglichst zugänglich machen sollte. In den Übungen geht es um konventionalisierte Wortverbindungen, die als „Schlüssel zu einer kulturell angemessenen Kommunikation“ (Mollica & Steyer 2024, 1) gesehen werden können und damit wichtig für den Sprachunterricht sind. Als „integrale[r] Bestandteil des Wortschatzes“ (Hallsteindóttir & Chrissou 2024, 8) spielen sie einerseits eine große Rolle für die Wortschatzerweiterung und -vertiefung und berühren andererseits auch grammatische Phänomene, da beispielsweise je nach Wortart verschiedene syntaktische Relationen zwischen Kollokationen bestehen. Als Schnittstelle für die Korpusarbeit wurde das DWDS-Wortprofil⁴ (vgl. Geyken 2011) gewählt, weil es öffentlich zugänglich und damit niedrigschwellig verwendbar ist. Darüber hinaus ist das DWDS als Wörterbuch bei einigen Studierenden möglicherweise schon bekannt, sodass der Wiedererkennungsfaktor idealerweise Hemmungen gegenüber der Korpusarbeit abschwächt. Alle Ausführungen beziehen sich auf das Wortprofil in der Version von September 2024. Zu diesem Zeitpunkt fand das Tutorat an der Universität Vilnius statt, in dem ein Teil der korpusbasierten Materialien eingesetzt wurde (siehe Kapitel 4).

⁴ <https://www.dwds.de/d/wortprofil>, letzter Zugriff: 16.09.2025.

Die Zielgruppe der Studierenden hat ein Sprachniveau ab A2/B1, also das Niveau, ab dem sich die Arbeit mit Korpora laut Blauth-Henke und Heinz (2009, 100) für Lernende anbietet. Krekeler (2021) hält Deutschkenntnisse „über dem Niveau B1“ (Krekeler 2021, 164) als Voraussetzung dafür, dass Lernende von korpusbezogenen Unterrichtsszenarien profitieren können. Allerdings wird eingeräumt, dass im Grundstufenbereich „vorbereitete Korpusabfragen eingesetzt und ausgewählte Konkordanzen präsentiert werden“ (Krekeler 2021, 190) können. Die in diesem Beitrag vorgestellten Übungen stellen hauptsächlich bereits abgefragte Suchergebnisse zur Verfügung und integrieren sie in Wortschatz- und Grammatikübungen. Darüber hinaus stehen für die einzelnen Übungen verschiedene Schwierigkeitsgrade zur Verfügung, sodass die Lehrkraft die Aufgaben angepasst an das jeweilige Sprachniveau der Studierendengruppe auswählen kann. Bei der Auswahl wurde die sprachliche Komplexität der Übungssätze berücksichtigt (vgl. Krekeler 2021, 190), allerdings wurden die Sätze nicht systematisch darauf geprüft, inwiefern ihr Wortschatz den verschiedenen Niveaustufen des Gemeinsamen Europäischen Referenzrahmen (GER) entspricht.

Die Zielsetzung bei der Materialerstellung war es, die folgenden drei Lernbereiche miteinander zu verbinden: 1) Sprachkompetenz (z. B. Wortschatz), 2) Linguistische Kompetenz (z. B. Wissen über Grammatik, Terminologie, semantische Relationen), 3) Korpuskompetenz. Die Grenzen zwischen diesen Bereichen verlaufen nicht unbedingt strikt, manche Phänomene können mehreren Lernbereichen zugeordnet werden. Zum Beispiel setzt eine Zuordnung von Satzgliedern (s. Abschnitt 3.2) sowohl Sprachkompetenz voraus, um den Satz überhaupt zu verstehen, als auch linguistische Kompetenz im Sinne des Wissens über die verschiedenen syntaktischen Funktionen und deren korrekte Zuordnung. Trotzdem eignet sich die Abgrenzung der drei Bereiche, um sich bewusst zu machen, welche Kompetenzen durch welche Aufgaben gefördert werden.

Die Übungsmaterialien⁵ wurden zunächst in Form von druckbaren Handouts konzipiert. Dies gewährleistet einen niedrigschwierigen Zugang, da nicht vorausgesetzt wird, dass Studierende eigene Computer in die Lehrveranstaltung mitbringen müssen oder dass es Zugang zu einem Computerraum gibt (wobei die parallele Arbeit mit der Schnittstelle des DWDS-Wortprofils natürlich zuträglich ist).

In Kapitel 3.1 wird zunächst eine theoretische Einordnung zur Terminologie, die für die Übungsmaterialien wichtig ist, gegeben. In Kapitel 3.2 und 3.3 werden dann die Materialien selbst vorgestellt. In der in Kapitel 4 beschriebenen Unterrichtssituation wurde im Rahmen eines studentischen Tutorats mit einer modifizierten Version der Materialien gearbeitet, um sich den Gegebenheiten vor Ort anzupassen. Im Folgenden wird die Grundkonzeption der Materialien beschrieben.

⁵ Alle Materialien (Handout und Spickzettel) sind unter folgenden Link abrufbar: <https://www.fdr.uni-hamburg.de/record/17926>, letzter Zugriff: 16.09.2025.

3.1 Theoretische Einordnung: Kollokationen und Kookkurrenzen

Zunächst ist es notwendig, eine Abgrenzung der Begriffe *Kollokation* und *Kookkurrenz* vorzunehmen, die eine gewisse Überschneidung aufweisen. Der Begriff *Kollokation* wurde von John R. Firth in den 1930ern geprägt, um „habitual‘ word combinations“ (Evert 2005, 15) zu beschreiben. Diese Definition wurde in der nachfolgenden Forschung unterschiedlich ausgelegt, sodass zwei grundsätzliche Unterscheidungen zwischen a) „semi-compositional word pairs, with one ‚free‘ element (the *base*) and the other element lexically determined (the *collocate*)“ (Evert 2005, 16) und b) statistisch signifikanten wiederkehrenden Wortkombinationen vorliegen.

Ein Beispiel für eine Kollokation nach a) wäre *eine Leistung erbringen*. Hier ist das Verb *erbringen* lexikalisch durch die Basis *Leistung* bestimmt, ein anderes Verb an dieser Stelle (z. B. **Leistung durchführen*) widerspricht dem Sprachgefühl. Kollokationen nach b) werden meist Kookkurrenzen genannt (vgl. Evert 2005, 17), um den Aspekt des statistisch signifikant häufigen gemeinsamen Auftretens zu fokussieren. Wenn z. B. in einem Korpus das Adjektiv *dreitätig* statistisch gesehen häufig mit dem Nomen *Fest* auftritt, handelt es sich dabei nicht um eine Kollokation im engeren Sinne, aber eben durchaus um eine Kookkurrenz.

Neben der konzeptuellen Definition muss zudem erläutert werden, was „gemeinsames häufiges Auftreten“ bedeutet. Innerhalb welcher Spanne von Wörtern wird ein Auftreten als „gemeinsam“ definiert? Nach welchem Maß wird die statistische Signifikanz des gemeinsamen Auftretens berechnet? Im DWDS-Wortprofil werden zunächst nach dem Assoziationsmaß logDice (Rychlý 2008) statistisch signifikante Wortverbindungen, also Kookkurrenzen, angezeigt. So wird auch in der Kurzbeschreibung des Wortprofils der Begriff Kookkurrenz verwendet und darauf hingewiesen, dass „[v]iele dieser Kookkurrenzen [...] sprachlich nicht weiter relevant [sind]“ und es „daher notwendig [ist], sprachlich bedeutsame Wortverbindungen, sogenannte Kollokationen, herauszufiltern“ (DWDS-Wortprofil). Gleichzeitig wird jedoch auf der Ergebnisseite des Wortprofils der Hinweis gegeben: „Klicken Sie auf eine Kollokation, um Belege angezeigt zu bekommen“ (DWDS-Wortprofil).⁶ Damit wird der Begriff *Kollokation*, der zuvor vom oberflächenbasierten Konzept der *Kookkurrenz* abgegrenzt wurde, in der Bedeutung ‚Kookkurrenz‘ verwendet. Diese terminologischen Feinheiten können potenziell bei Studierenden Verwirrung stiften und müssen bei der Erstellung und Verwendung der Übungsmaterialien berücksichtigt werden. Der Terminologie im Wortprofil folgend,

⁶ In der aktuellen Version des Wortprofils (Juni 2025) hat sich die Terminologie geändert, nun wird auch auf der Ergebnisseite der Begriff *Kookkurrenz* verwendet. Diese Änderung verdeutlicht die Schnelllebigkeit von digitalen Tools, die fortlaufend verbessert werden. Dies ist einerseits zu begrüßen, andererseits veralten Lernmaterialien dadurch schnell (vgl. Kapitel 2.2).

wird im Folgenden sowie in den Übungsmaterialien der Begriff *Kollokation* verwendet, auch wenn es sich bei einigen Beispielen streng genommen eher um Kookkurrenzen handelt.

3.2 Wortschatz- und Grammatikübungen mit dem DWDS-Wortprofil

Das Handout mit den Wortschatz- und Grammatikübungen ist so aufgebaut, dass sich erklärende Texte, Übungsaufgaben und Hinweisboxen abwechseln. Insgesamt enthält es sechs Übungsaufgaben, die zum Teil in verschiedenen Schwierigkeitsgraden zur Verfügung stehen. Es sind außerdem Lösungsvorschläge vorhanden, die in grauer Schriftfarbe hervorgehoben sind und von der Lehrkraft vor der Verteilung des Handouts entfernt werden sollten. Korpuslinguistische Fachbegriffe sind fettgedruckt, um darauf hinzuweisen, dass sie im Glossar der „Spickzettel“ (siehe Kapitel 3.3) nachgeschlagen werden können.

Ein kurzer Einführungstext zu Beginn des Handouts stellt die Übungseinheit vor, indem genannt wird, welche thematischen Bereiche die verschiedenen Aufgaben berühren und welche Korpuskompetenzen aktiviert werden. Zudem wird der Hinweis gegeben, dass sich alle Links und Informationen auf den Stand des Wortprofils von September 2024 beziehen.

An die Arbeit mit dem DWDS wird anhand des Wörterbucheintrags zu *feiern* herangeführt. Durch die vertraute Textsorte des Wörterbucheintrags soll den Studierenden, die noch nicht mit Korpora gearbeitet haben, der Einstieg erleichtert werden. Ein kurzer Begleittext im Handout thematisiert die verschiedenen Lesarten des Verbs *feiern* und gibt den Hinweis, dass es sich bei der Bedeutung 1 ‚ein feierliches, frohes Ereignis (in größerem Kreise) festlich begehen‘ um die zentrale und für die folgenden Aufgaben relevante Lesart handelt. So soll im Unterrichtsgespräch die Grundlage für das Verständnis der Bedeutung von *feiern* geschaffen werden. Über die Abbildung der „typischen Verbindungen“ aus dem Wörterbucheintrag, die als Screenshot ins Handout eingefügt ist, wird außerdem zum Konzept der Kookkurrenz hingeführt. Im Beschreibungstext wird sowohl thematisiert, dass Verbindungen mit unterschiedlichen Wortarten und grammatischen Relationen bestehen, als auch auf die beiden Termini Kollokation und Kookkurrenz hingewiesen. Diese Aspekte sollten im Unterrichtsgespräch aufgegriffen werden, der Spickzettel (siehe Kapitel 3.3) bietet dafür eine weitere Hilfestellung. Hinweisboxen wie in Abbildung 1 geben zudem Hintergrundinformationen und vermitteln an passenden Stellen korpuslinguistisches Wissen, beispielsweise zur Veränderlichkeit der Datengrundlage oder zu möglichen Fehlern bei der automatisierten Verarbeitung der Daten.



Hinweis 1: Datengrundlage des Wortprofils

Da das DWDS und seine Anwendungen stetig weiterentwickelt werden, kann es sein, dass sich auch die Grundlage des Wortprofils verändert. Deshalb ist es wichtig, immer zu dokumentieren, mit welcher Version oder an welchem Datum eine Abfrage durchgeführt wurde. Spätere Abfragen können andere Ergebnisse liefern, wenn sich die Korpusgrundlage in der Zwischenzeit verändert hat. Auch die Textsorten in einem Korpus haben einen großen Einfluss. Je nachdem, welche Textsorten in einem Korpus vorhanden sind, können sich die Ergebnisse stark unterscheiden. Das DWDS enthält hauptsächlich Zeitungstexte. Ein Korpus aus z.B. Fachtexten könnte ganz andere Kookkurrenzen zeigen.

Abbildung 1. Hinweisbox zur Datengrundlage des DWDS-Wortprofils

In der ersten Aufgabe (siehe Abbildung 2. Aufgabe 1 der Übungsmaterialien, in der es darum geht, mit der Darstellung im Wortprofil vertraut zu werden und die Bedeutung der Kollokationen zu reflektieren) werden die Studierenden anhand von Screenshots mit der Ansicht des Wortprofils vertraut gemacht. Zunächst sollen die Studierenden nur die Tabelle „Überblick“ analysieren und überlegen, welche Gemeinsamkeiten die Kollokationen haben. Aufgabe 1 berührt alle drei Lernbereiche: Sprachkompetenz wird benötigt und gefördert, um die Bedeutungen der Kollokationen zu aktivieren oder neu zu erlernen (entweder durch Fragen an die Lehrkraft, Diskussion der Studierenden untereinander oder Nachschlagen im Wörterbuch). Des Weiteren benötigen die Studierenden linguistische Kompetenz, um zum Beispiel zu erkennen, dass alle Kollokationen in der Tabelle „Überblick“ der gleichen Wortart (Nomen) angehören. Schließlich werden durch die Analyse der extrahierten Kollokationen auch Korpuskompetenzen gefördert, nicht zuletzt durch die Konfrontation mit Fachbegriffen wie *logDice* und *Frequenz*, die in der Tabelle auftauchen. Diese grundlegenden korpuslinguistischen Konzepte der Frequenz und des Assoziationsmaßes *logDice* werden in einem begleitenden Text im Anschluss an Aufgabe 1 dargelegt.

Der Begleittext weist außerdem auf die Möglichkeit hin, durch einen Klick auf eine der Kollokationen in der Tabelle zu Beispielsätzen aus dem Korpus zu gelangen. Eine weitere Hinweisbox verweist auf mögliche Fehler bei der automatisierten Verarbeitung, beispielsweise in der Zuweisung von Wortarten, die zu fehlerhaften Darstellungen führen kann. Auf diesen eher textlastigen Abschnitt des Handouts folgt die nächste Übungsaufgabe, bei der die Studierenden selbst eine Korpusabfrage durchführen sollen (siehe Abbildung 3).

Aufgabe 1

Seht euch die Liste der Kollokationen in der „Überblick“-Tabelle genauer an. Fallen euch Gemeinsamkeiten bei den Wörtern auf? Lassen sich vielleicht Gruppen bilden?

Lösungsvorschlag

Bei allen Wörtern handelt es sich um Nomen. Diese Nomen lassen sich in zwei Gruppen einteilen: Zum einen tauchen Ereignisse auf (*der Geburtstag, das Bestehen, das Jubiläum*) und zum anderen Personen (*der Popstar, der Rockstar, der Retter*). Das Verb *feiern* scheint also am häufigsten mit Personen oder Ereignissen, die gefeiert werden, aufzutreten.

Überblick ↴ ↵	logDice ↴	Freq. ↴	hat Verbzusatz ↴ ↵	logDice ↴	Freq. ↴
1. Geburtstag ⓘ	11.7	58706	1. bis	7.7	59
2. Bestehen ⓘ	10.8	25633	2. in	6.6	70
3. Richtfest ⓘ	10.6	2470	3. hinein	6.5	205
4. Jubiläum ⓘ	10.2	15561	4. mit	6.1	2185
5. Popstar ⓘ	10.1	388	5. von	5.7	21
6. Jubiläum ⓘ	9.7	2872	6. rein	5.7	48
7. Comeback ⓘ	9.4	9397	7. af	5.4	9
8. Premiere ⓘ	9.4	9721	8. nach	4.5	277
9. Rockstar ⓘ	9.3	101	9. weiter	4.4	191
10. Sieg ⓘ	9.1	21588	10. durch	4.3	211
11. Erfolg ⓘ	9.1	27592	11. zusammen	4.3	390
...

Abbildung 2: Screenshots aus dem DWDS-Wortprofil zu *feiern*: Der Überblick über alle grammatischen Funktionen (links) und die Tabelle der Verbzusätze (rechts). Abgerufen am 20.02.2024.

Abbildung 2. Aufgabe 1 der Übungsmaterialien, in der es darum geht, mit der Darstellung im Wortprofil vertraut zu werden und die Bedeutung der Kollokationen zu reflektieren

Aufgabe 2

Durchsuche die Korpora des DWDS nach eigenen Beispielsätzen für Partikelverben mit *feiern*. Gehe dazu über folgenden Link zur Korpusabfrage: <https://www.dwds.de/r>. Wähle als Korpus „Gegenwartskorpora mit freiem Zugang“ aus. Wähle bei Anzeige „KWIC“ aus. Gib dein Suchwort in der Infinitiv-Form ein, zum Beispiel *durchfeiern* oder *hineinfeiern*. In den Ergebnissen werden Belege in ganz unterschiedlichen Flexionsformen angezeigt. Suche ein paar Beispiele aus, die sich deiner Meinung nach gut eignen, um die Bedeutung deines Suchworts zu illustrieren.

Abbildung 3. In Aufgabe 2 werden die Studierenden dazu angeleitet, eine eigene einfache Korpusabfrage durchzuführen

Die Aufgabe soll niedrigschwellig mit dem Prozess der Korpussuche und korpuslinguistischen Fachbegriffen (KWIC) vertraut machen und bietet Gelegenheit, zu thematisieren, dass bei einer einfachen Wortformsuche verschiedene Flexionsformen gefunden

werden. Durch die Aufgabenstellung, passende Beispielsätze für das Suchwort zu finden, soll die Sprachkompetenz der Studierenden aktiviert werden.

Die darauffolgende Aufgabe 3 beschäftigt sich weiterhin mit Verbzusätzen und liegt in drei Schwierigkeitsgraden vor. Aus Platzgründen wird in Abbildung 4 jeweils nur der erste Lückensatz sowie die Lösung dargestellt. Die Sätze beziehen sich auf die zuvor thematisierte Tabelle aus dem Wortprofil zu *feiern* (siehe Abbildung 2).

Aufgabe 3 berührt vor allem die Lernbereiche der Sprachkompetenz und der linguistischen Kompetenz. In allen Schwierigkeitsgraden geht es darum, die Verwendung von *feiern* mit verschiedenen Verbzusätzen zu üben. Im ersten Schwierigkeitsgrad wird das Verb vorgegeben und die Studierenden haben die Aufgabe, es richtig zu konjugieren und in die vorgegebenen Lücken einzusetzen.

Aufgabe 3

In den folgenden Sätzen fehlen die Verbzusätze *mit, bis, hinein/rein, nach* und *durch*, die in Abbildung 2 aufgelistet sind. Die Sätze stammen direkt aus dem Korpus, das für die Erstellung des Wortprofils genutzt wurde.

Schwierigkeitsgrad 1: Setze die konjugierte Form von *feiern* und den Verbzusatz in die richtigen Lücken ein.

Auswahl: *mitfeiern, hineinfeiern/reinfeiern, nachfeiern* und *durchfeiern*

1. Viele _____ von Freitag bis Sonntag in Berliner Clubs _____. (durchfeiern)
...

Schwierigkeitsgrad 2: Finde den passenden Verbzusatz und setze ihn in die Lücken ein.

Auswahl: *mit, hinein/rein, nach* und *durch*

1. Viele feiern von Freitag bis Sonntag in Berliner Clubs _____.
...

Schwierigkeitsgrad 3: Finde den passenden Verbzusatz. Setze die konjugierte Form von *feiern* und den passenden Verbzusatz in die richtigen Lücken ein.

Auswahl: *mitfeiern, hineinfeiern/reinfeiern, nachfeiern* und *durchfeiern*

1. Viele _____ von Freitag bis Sonntag in Berliner Clubs _____.
...

Lösung

1. Viele feiern von Freitag bis Sonntag in Berliner Clubs durch.
...

Abbildung 4. Lückensätze aus Aufgabe 3 in verschiedenen Schwierigkeitsgraden

In der zweiten Variante ist der Verbzusatz nicht mehr vorgegeben: Hier sollen die Studierenden den passenden Verbzusatz identifizieren und in die Lücke einsetzen. Dafür kann einerseits das Wissen aus dem Unterrichtsgespräch zur zuvor besprochenen

Tabelle genutzt werden. Andererseits können die Studierenden auch dazu angeregt werden, selbst im Wortprofil zu recherchieren, um den passenden Verbzusatz zu finden. Im dritten Schwierigkeitsgrad werden die ersten beiden Versionen kombiniert; hier sollen die Studierenden nicht nur den passenden Verbzusatz finden, sondern zudem das Verb konjugieren und korrekt in die Lücken einsetzen. Die Lehrenden haben so die Möglichkeit, den passenden Schwierigkeitsgrad auszuwählen und die Übungsmaterialien an den Wissensstand und die Bedürfnisse ihrer Studierenden anzupassen.

Im Anschluss an Aufgabe 3 wird das zuvor besprochene Wortprofil von *feiern* mit dem von *Fest* kontrastiert (Tabelle 1).

<i>feiern</i>	<i>das Fest</i>
hat Adverbialbestimmung Bsp.: <i>etwas gebührend feiern</i>	hat Adjektivattribut Bsp.: <i>ein frohes Fest</i>
hat Akkusativ-Objekt Bsp.: <i>ein Jubiläum feiern</i>	ist Akkusativ-/Dativ-Objekt von Bsp.: <i>ein Fest organisieren</i>
hat Präpositionalgruppe Bsp.: <i>mit Gästen feiern</i>	ist in Präpositionalgruppe Bsp.: <i>auf das Fest freuen</i>
hat Verbzusatz Bsp.: <i>sie feiern den Geburtstag nach</i>	
	hat Genitivattribut Bsp.: <i>das Fest der Liebe</i>
	ist Genitivattribut von Bsp.: <i>der Organisator des Festes</i>
in Koordination mit Bsp.: <i>feiern und jubeln</i>	in Koordination mit Bsp.: <i>Feste und Bräuche</i>
hat Prädikativ Bsp.: <i>sie wird als Helden gefeiert</i>	hat Prädikativ Bsp.: <i>das Fest war legendär</i>
	ist Prädikativ von Bsp.: <i>das Leben ist ein Fest</i>
hat vergleichende Wortgruppe Bsp.: <i>er wird wie ein Popstar gefeiert</i>	hat vergleichende Wortgruppe Bsp.: <i>ein Fest wie das Museumsfest</i>
	ist in vergleichender Wortgruppe Bsp.: <i>Veranstaltungen wie Feste, Wettkämpfe und Feiern</i>
hat Subjekt Bsp.: <i>die Fans feiern</i>	ist Subjekt von Bsp.: <i>das Fest fällt aus</i>
hat Passivsubjekt Bsp.: <i>die Hochzeit wird gefeiert</i>	
	hat Präpositionalgruppe Bsp.: <i>ein Fest für die ganze Familie</i>

Tabelle 1. Vergleichende Übersicht der grammatischen Funktionen der Kollokationen von *feiern* und *Fest* im DWDS-Wortprofil

Die Studierenden werden darauf aufmerksam gemacht, dass hier Kollokationen mit unterschiedlichen grammatischen Relationen vorliegen, die jeweils für die Wortarten Verb (*feiern*) und Nomen (*Fest*) relevant sind. Die grammatischen Relationen werden so gegenübergestellt, dass Gemeinsamkeiten (sowohl *feiern* als auch *Fest* treten mit vergleichenden Wortgruppen auf) sowie Unterschiede (*feiern* kann ein Objekt haben; *Fest* kann als Objekt auftreten) deutlich werden. Die darauf folgende Aufgabe 4 (siehe Abbildung 5) liegt in zwei verschiedenen Schwierigkeitsgraden vor. Für den niedrigeren Schwierigkeitsgrad wurden einfache Hauptsätze ausgewählt, während im höheren Schwierigkeitsgrad auch komplexe Sätze zu analysieren sind.

Aufgabe 4

Schwierigkeitsgrad 1: Grammatische Funktionen erkennen. In den folgenden Sätzen sind das Zielwort *Fest* und der jeweilige Kollokator unterstrichen. In welcher syntaktischen Funktion steht das Wort *Fest* im jeweiligen Satz?

(Auswahl: Subjekt, Objekt, Prädikativ)

1. Da wollte der Berliner Senat ein großes Fest der Demokratie ausrichten.
2. Ausklingen soll das Fest gegen 18 Uhr.

...

Lösung

1. Da wollte der Berliner Senat ein großes Fest der Demokratie ausrichten. (Objekt)
2. Ausklingen soll das Fest gegen 18 Uhr. (Subjekt)

...

Schwierigkeitsgrad 2: Grammatische Relationen erkennen. In den folgenden Sätzen sind das Zielwort *Fest* und der jeweilige Kollokator unterstrichen. In welcher grammatischen Relation stehen das Wort *Fest* und sein Kollokator im Satz zueinander?

(Auswahl: hat Präpositionalgruppe, hat Adjektivattribut, ist Genitivattribut von, ist in Koordination mit)

1. Die ausgebildete Konzertsängerin spielte auch Klavier und trat auf großen Festen und Familienfeiern auf.
2. Auf dem Höhepunkt des Festes verkündet der Kaiser seine Verlobung mit Sissy.

...

Lösung

1. Die ausgebildete Konzertsängerin spielte auch Klavier und trat auf großen Festen und Familienfeiern auf. (ist in Koordination mit)
2. Auf dem Höhepunkt des Festes verkündet der Kaiser seine Verlobung mit Sissy. (ist Genitivattribut von)

...

Abbildung 5. Auszug aus Aufgabe 4, in der die grammatische Funktion des Zielworts *Fest* bzw. die grammatische Relation zwischen dem Zielwort *Fest* und dem jeweiligen Kollokator erkannt werden soll

In beiden Versionen sind in den Sätzen das Zielwort *Fest* sowie der Kollokator unterstrichen, sodass den Studierenden die Kollokationen (z. B. *ein Fest ausrichten*) bewusst

gemacht werden. Bei Schwierigkeitsgrad 1 haben die Studierenden die Aufgabe, die syntaktische Funktion zu erkennen, die das Zielwort *Fest* im jeweiligen Satz hat. Bei Schwierigkeitsgrad 2 hingegen geht es darum, die grammatische Relation, in der Zielwort und Kollokator zueinander stehen, zu erkennen. Dazu sollen die Studierenden sich an der zuvor besprochenen Übersichtstabelle (Tabelle 1) orientieren.

Auch Aufgabe 4 berührt vor allem Sprachkompetenz und linguistische Kompetenz. Durch die Rückkopplung an die Kollokationstabellen des DWDS-Wortprofils werden jedoch auch Korpuskompetenzen gestärkt.

Anschließend wird die Möglichkeit des Wortprofils, Kollokationen von zwei Wörtern miteinander zu vergleichen, eingeführt. Auch auf die Ansicht als Wortwolke (im Gegensatz zur bisherigen Darstellung als Tabelle) wird anhand der beiden synonymen Wörter *Fest* und *Party* eingegangen. Der Begleittext thematisiert semantische Unterschiede zwischen den beiden Nomen, die anhand der Kollokationen ersichtlich werden. Beispielsweise tritt *Fest* eher in Kollokationen auf, die auf traditionelle und religiöse Kontexte schließen lassen (z. B. *Fastenbrechen*, *christlich*), während *Party* häufig mit den Anglizismen *Dinner* und *Disco* auftritt, was auf informellere Kontexte schließen lässt. Der Begleittext beschreibt eine beispielhafte Analyse von Kollokationen und stärkt so die Korpuskompetenz der Analyse und Interpretation von Korpusdaten. Aufgabe 5 (siehe Abbildung 6) arbeitet wieder mit Lückensätzen, in denen anhand der zuvor gegebenen Informationen über die Kollokationen das passende Nomen eingesetzt werden soll.

Aufgabe 5

Bei *Fest* und *Party* handelt es sich um Synonyme, die sich allerdings in Bedeutungsnuancen voneinander unterscheiden, wie die vergleichende Betrachtung des Wortprofils zeigt. Die folgenden Lückensätze stammen aus dem Korpus, das für die Erstellung des Wortprofils genutzt wurde. Entscheide für jeden Satz, ob *Fest* oder *Party* thematisch und semantisch besser in die Lücke passt. Setze das Wort in der passenden Flexionsform und gegebenenfalls mit dem richtigen Artikel in die Lücke ein. Die jeweilige Kollokation im Satz ist unterstrichen. Auswahl: (das/ein) *Fest*, (die/eine) *Party*

1. Laut Polizei war der Junge auf _____ aufgetaucht, hatte zwei Flaschen Wodka in einem Zug getrunken.
2. Aus diesem Grund begehen die katholische und die evangelische Kirche _____ gemeinsam.

...

Lösung

1. Laut Polizei war der Junge auf der Party aufgetaucht, hatte zwei Flaschen Wodka in einem Zug getrunken.
2. Aus diesem Grund begehen die katholische und die evangelische Kirche das Fest gemeinsam.

...

Abbildung 6. Auszug aus Aufgabe 5, bei der die Studierenden das passende Nomen (*Party/Fest*) in die Lückensätze eintragen sollen

Auf Aufgabe 5 folgend wird die Frage „Wieso lohnt es sich, sich mit Kollokationen zu beschäftigen?“ in einem graphisch abgesetzten Kasten thematisiert, wobei vor allem auf die Aspekte der Wortschatzerweiterung und Wortschatzvertiefung eingegangen wird. Die Passage ist zur Motivation der Studierenden gedacht und kann von der Lehrkraft auch als Ausgangspunkt für eine Gruppendiskussion genutzt werden, um so die Studierenden zur eigenen Reflexion anzuregen.

Aufgabe 6 schließt das Handout ab. Es handelt sich um eine Textproduktionsaufgabe, in der die Studierenden den gelernten Wortschatz anwenden sollen, indem sie einen kurzen Text darüber schreiben, wie sie ihren nächsten Geburtstag feiern möchten. Dafür sollen Sätze mit dem Verb *feiern* und den Nomen *Fest* und *Party* gebildet und die Kollokationen aus dem Wortprofil genutzt werden. Die abschließende Frage danach, welche Kollokationen die Studierenden am nützlichsten für den eigenen Wortschatz finden, regt wiederum zur Reflexion des Gelernten an.

3.3 Korpusdidaktische Hilfestellung: „Spickzettel“ für Studierende und Lehrende

Zusätzlich zu den korpusbasierten Übungsaufgaben wird ein Glossar mit wichtigen korpusbezogenen Begriffen und kurzen Definitionen, das die Arbeit mit korpuslinguistischen Werkzeugen und Ressourcen im Unterricht erleichtern soll, als sogenannter „Spickzettel“ zur Verfügung gestellt. Um Korpuskompetenzen an Studierende vermitteln zu können, müssen Lehrende zuerst selbst eine ausreichende Korpuskompetenz erlangen. Damit der Einsatz korpusbasierter Übungen nicht nur für die Lernenden, sondern auch für ihre Dozierenden erleichtert wird, existiert der „Spickzettel“ in zwei Versionen: Zum einen mit der Zielgruppe der Studierenden, die darauf im Unterricht oder bei der Nachbereitung eine prägnante, niedrigschwellige Definition korpusbezogener Fachbegriffe nachschlagen können. Zum anderen wird eine erweiterte Version zur Verfügung gestellt, die sich an Lehrende richtet und weiterführende Hintergrundinformationen zu den thematisierten Begriffen und Konzepten anbietet. Die Bereitstellung von zwei verschiedenen Versionen bietet den Vorteil, dass der „Spickzettel“ für die Studierenden übersichtlich bleibt und nicht mit (interessanten, aber sprachlich und wissenschaftlich komplexen) Informationen überfrachtet wird. Der „Spickzettel“ für die Lehrenden bietet genau solche Informationen, die es erleichtern sollen, beispielsweise auf tiefer gehende Nachfragen von Studierenden reagieren zu können.

Die Grundidee der „Spickzettel“ ist es, den Studierenden und ihren Lehrenden ein weiteres Hilfsmittel an die Hand zu geben, das parallel zu den Übungsmaterialien genutzt werden kann. Dadurch, dass Informationen gebündelt zur Verfügung gestellt werden, müssen Studierende nicht selbst recherchieren, was zu einer Überforderung bei der Konfrontation mit unbekannten Konzepten führen könnte. Die „Spickzettel“ sollen dazu

beitragen, die Berührungsängste zu mindern, die möglicherweise zur Arbeit mit Korpora bestehen, da es sich bei der Korpuslinguistik um ein großes, zum Teil unübersichtliches Feld handelt, in dem es schwierig sein kann, sich alleine zurecht zu finden.

Spickzettel für Studierende	Spickzettel für Lehrkräfte
Kollokation (f, -en) – häufig auftretende Wortverbindung, bei der ein Element das andere bedingt (z.B. <i>Antrag stellen</i> , <i>Zähne putzen, hin und her</i>). Die Begriffe <i>Kookkurrenz</i> und <i>Kollokation</i> werden häufig auch synonym verwendet, sollten aber voneinander abgegrenzt werden (vgl. Evert 2008).	Die Begriffe Konkordanz , Kookkurrenz und Kollokation werden in der Korpuslinguistik verwendet, um das gemeinsame Auftreten von Token zu beschreiben. Konkordanz nennt man die Darstellungsform von Korpusbelegen, bei der neben dem Treffer-Token links und rechts der Kotext, z. B. jeweils zehn Wörter, angezeigt wird. Ein anderer Begriff für Konkordanz ist KWIC (<i>Keyword in Context</i>).
Kollokator (m, -en) – Wort, das in einer Kollokation mit einem anderen Wort verbunden ist (z.B.: <i>stellen</i> ist Kollokator von <i>Antrag</i>).	Der Begriff Kookkurrenz beschreibt das gemeinsame Vorkommen von Lexemen. Für verschiedene Fragestellungen ist es interessant, zu untersuchen, mit welchen anderen Lexemen das untersuchte Wort besonders häufig zusammen auftritt. Um das herausfinden zu können, muss zunächst festgelegt werden, wie sich „gemeinsames Auftreten“ definiert, und ab wann dieses als besonders häufig gilt.
Konkordanz (f, -en) – Darstellung eines Korpusbelegs im jeweiligen Kontext des verwendeten Korpus (z. B. zehn Wörter links und zehn Wörter rechts des Treffers), auch <i>Keyword in Context</i> (KWIC) genannt.	Eine sehr strenge Definition für ein gemeinsames Auftreten wäre, nur Wörter zu betrachten, die direkt nebeneinanderstehen. Oft wird allerdings mit einem Suchfenster von fünf Wörtern (also fünf Wörter vor und fünf Wörter nach dem Zielwort) oder auch dem ganzen Satz gearbeitet. Um die Häufigkeit zu berechnen, reicht es nicht, sich die reinen Frequenzen anzusehen, weil manche Lexeme (z. B. Artikel oder Konjunktionen) viel frequenter sind als andere. Stattdessen geht es darum, herauszufinden, ob zwei Wörter häufiger gemeinsam auftreten, als man es aufgrund ihrer Einzelfrequenzen erwarten würde. Ein in der Korpuslinguistik oft dafür verwendetes Assoziationsmaß ist logDice .
Die Unterscheidung zwischen Kookkurrenz und Kollokation: Die Kookkurrenz <i>ausgelassen</i> zu <i>feiern</i> im Wortprofil ist	

Abbildung 7. Ausschnitt aus dem „Spickzettel“ für Lehrkräfte. Die weiterführenden Informationen werden parallel zu den Informationen aus der Version für die Studierenden gegeben, sodass sie leicht auffindbar sind

4 Einsatz der Materialien

Im Folgenden wird der Einsatz eines Ausschnittes aus den vorgestellten korpusbasierten Übungsmaterialien im DaF-Unterricht im Studiengang „Germanistik“ an der Universität Vilnius beschrieben und reflektiert. Dabei wird in Kapitel 4.1 auf die Unterrichtssituation eingegangen. In Kapitel 4.2 wird die Evaluation vorgestellt und reflektiert: Zum einen wurde anhand eines kurzen Tests das vermittelte Wissen objektiv überprüft, zum anderen wurden anhand eines anonymen Fragebogens die subjektiven Erfahrungen der Studierenden mit den korpusbasierten Übungsmaterialien gesammelt.

4.1 Beschreibung der Unterrichtssituation

Die Materialien wurden zu Beginn des Herbstsemesters 2024 im Rahmen eines studentischen Tutorats erprobt. Zielgruppe des Unterrichts war eine mehrsprachige Lerngruppe mit Studierenden unterschiedlicher Erstsprachen, die gerade das zweite Studienjahr (drittes Semester) an der Universität Vilnius angefangen hatten und von Lehrkräften als fortgeschritten bezeichnet wurden. Zu den Muttersprachen der Studierenden zählten

Litauisch, Russisch und Ukrainisch. Im Kurs „Gegenwartsdeutsch“ wurde hauptsächlich mit dem Lehrwerk „Vielfalt“ B2.2 gearbeitet. Dennoch lernt jeder anders und das gleich hohe B2.1-Niveau kann nicht zwangsläufig vorausgesetzt werden. Die Vorerfahrungen mit Korpora und mit der Schnittstelle des DWDS waren unterschiedlich. Hauptsächlich kannten die Studierenden das DWDS-Wörterbuch als ein Nachschlagewerk für Wortbedeutungen. Die Korpuswerkzeuge der Plattform waren größtenteils unbekannt.

Für die Durchführung des Unterrichts mit dem korpusbasierten Material, der Bearbeitung der Tests und der abschließenden Diskussion mit Ausfüllen des Fragebogens wurden insgesamt drei Sitzungen eingeräumt. Die erste und die dritte Sitzung fanden im normalen Seminarraum statt und es wurde nur indirekt mit Korpora gearbeitet. Die zweite Sitzung fand in einem Computerraum statt, wo die Studierenden einen digitalen Test in Gruppen bearbeiten und Korpusabfragen direkt im DWDS vornehmen konnten. Alle Sitzungen wurden auf Deutsch gehalten. Im Gespräch untereinander griffen die Studierenden aber auch auf ihre Alltagssprache Litauisch und auf die Lingua Franca Englisch zurück. Da die Unterrichtseinheiten an unterschiedlichen Wochentagen stattfanden und keine Anwesenheitspflicht an der Universität Vilnius besteht, nahm eine unterschiedliche Anzahl der Studierenden an jeder Einheit teil. Im Durchschnitt waren neun Studierende in jeder Lerneinheit präsent, allerdings nahmen nur sechs an allen drei Einheiten teil.

Die in Kapitel 3 vorgestellten Materialien wurden gekürzt, um sie innerhalb der ersten Sitzung einsetzen zu können. Es wurde nicht zwischen den Schwierigkeitsgraden innerhalb einer Aufgabe unterschieden. Grundsätzlich kann es aber lohnenswert sein, eine Binnendifferenzierung innerhalb einer Lerneinheit zu ermöglichen, damit das Material für alle Teilnehmenden interessant und herausfordernd genug bleibt: „Damit sich das Lernen mit Erfahrungen von Selbstwirksamkeit verknüpft, sollten individuelle Arbeitsschwerpunkte im Anschluss an vorhandene, explizit benannte Ressourcen und die damit verbundene individuelle Lernausgangslage [...] herausgearbeitet werden“ (Junghans & Feindt 2020, 234).

Die im Unterricht genutzten Materialien können auf der Website der Germanistischen Institutspartnerschaft KoDi-FS abgerufen werden.⁷ Aufgabe 1 leistete den niederschwelligen Einstieg und machte die Studierenden mit der Schnittstelle vertraut. Die Schwierigkeitsgrade 1 und 2 von Aufgabe 4 der in Kapitel 3.2 vorgestellten Materialien wurden als selbständige Aufgaben eingesetzt, um gezielter mit den grammatischen Funktionen und Relationen zu arbeiten. Dabei musste die vom DaF-Lehrwerk „Vielfalt“ abweichenende grammatische Begrifflichkeit des DWDS-Wortprofils mehrfach thematisiert werden

⁷ <https://www.korpuslab.uni-hamburg.de/kodi-fs/archiv-aktivitaeten-2022-2024/uebungen-tutorat-vilnius-2024.pdf>, letzter Zugriff 16.09.2025.

(siehe Kapitel 2.2, z. B. *Verbzusatz*⁸ auf DWDS-Wortprofil gegenüber *Partikelverben*, *trennbare Verben* oder *Präverfügung* in DaF-Grammatiken, vgl. Nied Cursio 2020, 193). Im Unterricht muss eine solche Abweichung durch die Lehrkraft thematisiert und erklärt werden. Die Verständnisschwierigkeiten bei korpusbezogenen Terminen, z. B. der Unterschied zwischen Frequenz und logDice in den Ergebnistabellen des Wortprofils, oder bei der Semantik einzelner Wörter, wurden von den Studierenden angesprochen und konnten durch Erklärungen der Tutorin behoben werden.

Normalerweise stammt das im Unterricht verwendete Material aus dem Lehrbuch und wird mit einer Präsentation begleitet. Um den gewohnten Lernstil beizubehalten, wurden die korpusbasierten Übungen in Form des Handouts und einer begleitenden Präsentation eingesetzt. Die Schnittstellen des DWDS und des Wortprofils wurden auf der Leinwand gezeigt. Die längeren Textpassagen im in Kapitel 3.2 vorgestellten Handout, die der Erklärung dienen, wurden auf das für den Unterricht Essenzielle zusammengefasst und durch mündliche Erklärungen der Tutorin begleitet.

4.2 Eindrücke und Rückmeldungen der Studierenden

Bereits der Einstieg in die Sitzung (die Bedeutungsangaben zu *feiern*) förderte die Aktivierung des Vorwissens der Studierenden sowie ihren Sprachgebrauch und gab ihnen die Möglichkeit, den Unterricht interaktiv mitzugestalten. Obwohl *feiern* zum A1-Niveau-Wortschatz gehört und den Studierenden dadurch schon bekannt war, zeigten sie Interesse an den Lerninhalten, besonders in Bezug auf feine Details, die bei den authentischen Beispielsätzen und der Verwendung mit unterschiedlichen Verbzusätzen im Wortprofil zur Geltung kommen.

Im Laufe der Unterrichtseinheiten erwiesen sich die Studierenden oft als eher zurückhaltend, sodass es die Einschätzung erschwerte, ob die Lerneinheit möglicherweise als über- oder unterfordernd empfunden wurde. Um die Zufriedenheit und die Lernerfolge der Studierenden mit den Materialien messen zu können, wurde eine Überprüfung in Form eines Fragebogens und eines digitalen Tests durchgeführt, der ähnliche Aufgaben wie das Material der Lerneinheit enthielt und auf die direkte Arbeit mit dem DWDS (Wörterbuch und Wortprofil) abzielte. Der Test wurde in drei Gruppen bestehend aus jeweils drei Studierenden bearbeitet, die dazu alle verfügbaren Hilfsmittel nutzen durften.

Die Ergebnisse der Tests geben ein wichtiges Feedback für die Verbesserung der künftigen Unterrichtsgestaltung. Ein Problem zeigte sich jedoch darin, dass nicht alle Studierenden, die den Test bearbeiteten, tatsächlich auch bei der korpusbasierten Lernein-

⁸ Mittlerweile wurde die grammatische Funktion „hat Verbzusatz“ aus dem DWDS-Wortprofil von *feiern* entfernt (Stand: 17.06.2025).

heit anwesend gewesen waren. Obwohl die abwesenden Studierenden die Lerneinheit zuhause nachbereiten sollten, hatten sie größere Schwierigkeiten bei der Bearbeitung des Tests als diejenigen, die sich mit den Übungsmaterialien im Unterricht beschäftigt hatten. Dieser Umstand zeigt, wie wichtig die Begleitung der korpusbasierten Einheit durch eine Lehrkraft ist, um Verständnisprobleme auffangen zu können.

Dennoch zeigten auch die anderen Studierenden gewisse Verständnisprobleme, insbesondere in Bezug auf den Begriff der Kollokation, der für Aufgabe 3 (siehe Abbildung 8) eine wichtige Rolle spielt. Diese Aufgabe diente der Überprüfung der im Unterricht vermittelten Korpuskompetenzen: Für eine erfolgreiche Bearbeitung ist die Fähigkeit, eine eigenständige Suchabfrage im DWDS-Wortprofil durchzuführen, ebenso notwendig wie die Aktivierung des theoretischen Wissens über den korpuslinguistischen Begriff *Kollokation* und die damit verbundene linguistische Kompetenz, Kollokationen erkennen und passende Beispielsätze (Korpusbelege) finden zu können.

Aufgabe 3: Schreibt fünf Kollokationen und fünf Beispielsätze pro Teilnehmer*in in eurer Gruppe zu einem oder mehreren Wörtern eurer Wahl. Nutzt dafür das DWDS-Wortprofil <https://www.dwds.de/wp>. Eure Wörter dürfen mit den obigen Wörtern nicht identisch sein. Vergesst nicht, die Quelle anzugeben.

Abbildung 8. Formulierung der Aufgabestellung von Aufgabe 3 im digitalen Test

Im Kapitel 3.1 wird das Problem der Bedeutungsüberschneidung von *Kookkurrenz* und *Kollokation* im DWDS-Wortprofil beschrieben. Diese führte zu einer Verwirrung der Studierenden bei der Bearbeitung von Aufgabe 3 im digitalen Test: Die Studierenden sollten nach Kollokationen suchen, wobei sowohl Kookkurrenzen als auch Kollokationen auf der Webseite als Kollokationen bezeichnet wurden.⁹ Das führte dazu, dass die Studierenden auch „bloße“ Kookkurrenzen aus dem DWDS-Wortprofil als Kollokationen angaben, wie Abbildung 9 zeigt. Bei manchen Studierenden ist das Sprachgefühl weiter entwickelt als bei den anderen. Zusätzlich zu den Korpuskompetenzen ist – ganz abgesehen von den terminologischen Überschneidungen – auch linguistische Analysekompetenz (vgl. Abschnitt 1) notwendig, um einschätzen zu können, ob es sich bei einer Wortverbindung tatsächlich um eine Kollokation handelt.

Zuletzt wurde die subjektive Erfahrung der Studierenden mit den korpusbasierten Übungen anhand eines anonymen Fragebogens erhoben. Der Fragebogen bestand aus überwiegend geschlossenen, einerseits personenbezogenen Fragen und andererseits korpuslinguistischen Wissensfragen. Die Ergebnisse des Fragebogens können ausschließlich

⁹ Mittlerweile wird im DWDS-Wortprofil einheitlich der Begriff Kookkurrenz verwendet. Aus den vorhandenen Kookkurrenzen müssen Kollokationen noch herausgefiltert werden.

qualitativ betrachtet werden, da die Anzahl der Teilnehmenden für eine quantitative Analyse zu gering war. Laut Fragebogen unterscheiden sich die Meinungen zum Einsatz der korpusbasierten Materialien stark. Manche Studierenden fanden die neuen Zugänge interessant und nützlich, andere nahmen sie als verwirrend oder gar nutzlos wahr. Um den Mehrwert der Vermittlung von Korpuskompetenzen im DaF-Unterricht spürbar zu machen, sollte der Einsatz der korpusbasierten Aufgaben im Sprachunterricht zur Routine werden. Ein rein punktueller Einsatz scheint nicht immer zielführend zu sein und bei manchen Studierenden sogar eine Ablehnungshaltung hervorrufen. Durch eine Verfestigung im DaF-Unterricht können unterschiedliche thematische Zugänge zur Korpusarbeit gewählt werden (z. B. auch in Bezug auf Syntax, Orthographie, Morphologie, etc.), wodurch das Verständnis der Studierenden für die Anwendungsmöglichkeiten der Korpuslinguistik wachsen kann.

1. **Gehen zurück.** Bsp.

Zum Essen **geht** sie **zurück** in ihr Wohnzimmer.

2. **Gehen in Richtung.** Bsp.

Aber jemand muss entscheiden, **in** welche **Richtung** es **gehen** soll.

3. **Gehen also.** Bsp.

Also geht es weiter, immer weiter.

Quelle: <https://www.dwds.de/wp/gehen>

1. **Das Gesetz verabschieden**

Das dänische Parlament hat am Donnerstag ein **Gesetz verabschiedet**, das den unangemessenen Umgang mit Schriften von religiöser Bedeutung verbietet.

2. **Das Gesetz erlassen**

Es war richtig, diese **Gesetze zu erlassen**; das sagen selbst Autofahrerinnen und Autofahrer, Raucherinnen und Raucher.

3. **Das Gesetz verstößen**

Die ermittelnden Behörden müssen klären, ob dabei gegen **Gesetze verstößen** wurde.

DWDS-Wortprofil für „Gesetz“, erstellt durch das Digitale Wörterbuch der deutschen Sprache, <<https://www.dwds.de/wp/?q=Gesetz>>, abgerufen am 01.10.2024.

Abbildung 9. Auszug aus den Antworten von verschiedenen Studierenden auf Aufgabe 3. Bei den Wortverbindungen mit *gehen* handelt es sich um Kookkurrenzen, während es sich bei den Wortverbindungen mit *Gesetz* tatsächlich um Kollokationen handelt

5 Fazit

Lüdeling und Walter „plädieren [...] für ein spezifisches Korpustraining für die Lernenden“ (Lüdeling & Walter 2009, 8), um den sinnvollen Einsatz von Korpora zu gewährleisten. Oft ist es Lehrenden aus zeitlichen Gründen (vgl. Krekeler 2021, 189) nicht möglich, spezielle Arbeitseinheiten zur Nutzung von Korpora in den DaF-Unterricht

zu integrieren. Auch für Lernende ist die direkte und indirekte Arbeit mit Korpora im Fremdsprachenunterricht zeitaufwendig und kann deshalb zunächst als mühsam wahrgenommen werden. Im Hinblick auf die Weiterentwicklung digitaler Tools und der dafür notwendigen Kompetenzen in Studium und Beruf ist es jedoch empfehlenswert, sich dieser Herausforderung auch in der Sprachlehre zu stellen.

Der Einsatz korpusbasierter Materialien, wie sie in Form der Übungsaufgaben und dem „Spickzettel“ in diesem Beitrag vorgestellt wurden, soll eine Verknüpfung des Sprachunterrichts mit einer niedrigschwierigen Vermittlung von Korpuskompetenzen erleichtern. Es empfiehlt sich, die Korpusarbeit in die Unterrichtsaktivität so zu integrieren, dass sie genauso wie die Arbeit mit einem Lehrbuch stattfinden kann. Ziel ist es, durch die regelmäßige Arbeit mit Korpora deren Nutzung zunehmend als selbstverständlichen Bestandteil des Unterrichts zu etablieren. Dadurch können Berührungsängste abgebaut und der didaktische Mehrwert der Tools für Lernende erschlossen werden.

Im Gegensatz zu Lehrwerken, die häufig nur in längeren Intervallen aktualisiert werden, erlauben Korpora den Zugang zu sprachlichen Beispielen, die die tatsächliche Gegenwartssprache widerspiegeln. Die vorgestellten Übungsaufgaben können flexibel an andere Themenbereiche, beispielsweise jahreszeitliche Ereignisse oder aktuelle Trends, angepasst werden und so das Interesse der Studierenden wecken.

Für den Einsatz der vorgestellten Materialien hat sich gezeigt, dass besonders die korpuslinguistische Terminologie und Diskrepanz in der Verwendung von *Kollokation* und *Kookkurrenz* eine Herausforderung für die Studierenden darstellte. Das Erlernen von Korpuskompetenzen sollte allerdings nicht aufgrund korpuslinguistisch komplexer Konzepte vermieden werden. Vielmehr kann deren thematische Einbindung den kritischen Sprachumgang und das linguistische Wissen der Lernenden fördern. Allerdings stellt gerade die Entscheidung, solche Konzepte entweder auszuschließen oder gezielt zu thematisieren, für viele Lehrende eine Herausforderung dar, da sie ein hohes Maß an fachlichem Vorwissen und didaktischer Sicherheit erfordert. Ein Korpustraining der Lehrenden im DaF-Bereich darf also nicht außer Acht gelassen werden, wenn eine Integration der Vermittlung von Korpuskompetenzen in den Sprachunterricht gelingen soll.

Literatur

- Anthony, Laurence. 2024. *AntConc (Version 4.3.1)*. [Computer Software]. Tokyo, Japan: Waseda University. <https://www.laurenceanthony.net/software/AntConc>. Letzter Zugriff: 17.06.2025.
- Blauth-Henke, Christine & Matthias Heinz. 2009. Korpora und Fremdsprachendidaktik: Neue Perspektiven für Lehrer und Lerner. *Zeitschrift für romanische Sprachen und ihre Didaktik* 3 (1), 85–108.

- Boulton, Alex & Henry Tyne. 2013. Corpus linguistics and data-driven learning: A critical overview. *Bulletin suisse de Linguistique appliquée*. Neuchâtel: Institut de linguistique de l'Université. Rôle des pratiques langagières dans la constitution des espaces sociaux pluriels d'aujourd'hui Actes du colloque VALS-ASLA 2012 (Lausanne, 1-3 février 2012). 97–118.
- Callies, Marcus. 2019. Integrating corpus literacy into language teacher education: The case of learner corpora. *Learner Corpora and Language Teaching*. Sandra Götz & Joybrato Mukherjee, Hrsg. Amsterdam, Philadelphia: John Benjamins. 245–265.
- Evert, Stefan. 2005. *The Statistics of Word Cooccurrences. Word Pairs and Collocations*. Universität Stuttgart.
- Flinz, Carolina, Britta Hufeisen, R. Luppi & Ruth Mell. 2021. Zeit für eine neue Zeitschrift? – Die elektronische Zeitschrift KorDaF – Korpora Deutsch als Fremdsprache stellt sich vor. *Korpora Deutsch als Fremdsprache* 1 (1), 1–5.
- Flinz, Carolina. 2021. KORPORA in DaF und DaZ: Theorie und Praxis. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 1–43.
- Flinz, Carolina, Ruth M. Mell, Christine Möhrs & Tassja Weber. 2023. Korpora für Deutsch als Fremdsprache – Potenziale und Perspektiven. *Korpusgestützte Sprachanalyse: Grundlagen, Anwendungen und Analysen*. Michael Beißwenger, Eva Gredel, Lothar Lemnitzer & Roman Schneider, Hrsg. Tübingen: Narr Francke Attempto. 393–408.
- Frank, Mark, Johanna Walker, Julie Attard & Alan Tygel. 2016. Data Literacy: what is it and how can we make it happen? *The Journal of Community Informatics* 12 (3), 4–8.
- Geyken, Alexander. 2011. Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. *Korpora in Lehre und Forschung*. Andrea Abel & Renata Zanin, Hrsg. Bozen, Italien: Bozen University Press. 115–137.
- Gilquin, Gaëtanelle. 2022. Cognitive corpus linguistics and pedagogy. From rationale to applications. *Pedagogical Linguistics* 3 (2), 109–142.
- Gredel, Eva. 2021. CMC-Korpora und Digital Literacy in der Fremdsprachendidaktik: Relevanz, Potentiale und didaktische Szenarien. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 109–135.
- Hallsteindóttir, Erla & Marios Chrissou. 2024. Korpuslinguistische Ansätze der Phraseologie – und was nun? Phraseodidaktik und die Potenziale neuerer Zugänge der Sprachdidaktik. *Korpora Deutsch als Fremdsprache* 4 (1), 8–27.
- Hirschmann, Hagen, Anke Lüdeling, Anna Shadrova, Dominique Bobeck, Martin Klotz, Roodabeh Akbari, Sarah Schneider & Shujun Wan. 2022. FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. *Korpora Deutsch als Fremdsprache* 2 (2), 139–148. <https://doi.org/10.48694/kordaf.3552>
- Johns, Tim & Philip King, Hrsg. 1991. *Classroom Concordancing. English Language Research Journal* 4. Birmingham: University of Birmingham.

- Junghans, Carola & Andreas Feindt. 2020. Nachdenken und Sprechen über Unterricht und mich selbst: Reflexionsmethoden für den phasenübergreifenden Einsatz in der Lehrer_innenbildung. *HLZ – Herausforderung Lehrer*innenbildung* 3 (2). Bielefeld, Germany. 232–253. <https://doi.org/10.4119/hlz-2711>
- Kerschhofer-Puhalo, Nadja. 2021. Literacy – Dimensionen eines Begriffs und ein Modell zur kritischen Textarbeit. *Die Wirklichkeit lesen: Political Literacy und politische Bildung in der Migrationsgesellschaft*. Lisa Gensluckner, Michaela Ralser, Oscar Thomas-Olalde & Erol Yıldız, Hrsg. Bielefeld: transcript Verlag. 269–294.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1, 7–36.
- Krekeler, Christian. 2021. Korpustraining und datengestütztes Lernen im Wirtschafts-deutsch-Unterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 161–195.
- Lüdeling, Anke & Maik Walter. 2009. *Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung*. <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiterinnen/anke/pdf/LuedelingWalterDaF.pdf/view>. letzter Zugriff: 17.06.2025.
- Ma, Qing. 2025. Corpus Literacy and Data-Driven Learning. *The Palgrave Encyclopedia of Computer-Assisted Language Learning*. Lee McCallum & Dara Tafazoli, Hrsg. London: Palgrave Macmillan. 1–7.
- Mollica, Fabio & Kathrin Steyer. 2024. Korpusdaten und Phraseodidaktik: Probleme und Perspektiven. *Korpora Deutsch als Fremdsprache* 4 (1), 1–7.
- Mukherjee, Joybrato. 2002. *Korpuslinguistik und Englischunterricht: Eine Einführung*. Berlin u. a.: Peter Lang.
- Nied Cursio, Martina. 2020. Die Anwendung grammatischer Terminologie im Bereich des Deutschen als Fremdsprache. *Grammatik und Terminologie. Beiträge zur ars grammatica 2017*. Christian Lang, Roman Schneider, Horst Schwinn, Karolina Suchowolec & Angelika Wöllstein, Hrsg. Tübingen: Narr Francke Attempto. 189–212.
- Rychlý, Pavel. 2008. A Lexicographer-Friendly Association Score. *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 6–9.
- Zinsmeister, Heike, Carla Sökefeld & Skaistė Volungevičienė. 2025. Korpusdidaktik für formelhafte (Fach-)Sprache: Erfahrungen aus der Institutspartnerschaft zwischen Hamburg und Vilnius. *Info DaF* 52 (1), 45–59.

Building and annotating a bachelor paper abstract corpus: First findings about Latvian student abstracts

Baiba Egle

Liepaja Academy
Riga Technical University
Lielā iela 14
LV-3401 Liepāja, Latvia
E-mail: baiba.egle@rtu.lv
ORCID iD: <https://orcid.org/0000-0002-6207-7789>
<https://ror.org/00twb6c09>

Agnese Dubova

Liepaja Academy
Riga Technical University
Lielā iela 14
LV-3401 Liepāja, Latvia
E-mail: agnese.dubova@rtu.lv
ORCID iD: <https://orcid.org/0000-0002-7428-1824>
<https://ror.org/00twb6c09>

Dzintra Lele-Rozentāle

Liepaja Academy
Riga Technical University
Lielā iela 14
LV-3401 Liepāja, Latvia
E-mail: dzintra.lele-rozentale@rtu.lv
ORCID iD: <https://orcid.org/0000-0003-3181-6929>
<https://ror.org/00twb6c09>

Gints Jēkabsons

Institute of Applied Computer Systems
Riga Technical University
Zunda krastmala 10
LV-1048 Rīga, Latvia
E-mail: gints.jekabsons@rtu.lv
ORCID iD: <https://orcid.org/0000-0002-9575-2488>
<https://ror.org/00twb6c09>

Abstract. The paper discusses the problems identified in the early stages of an interdisciplinary project that focuses on the creation of a bilingual abstract corpus of bachelor's theses from a linguistic point of view. The aim of this paper is to summarize the identified problem cases and to show the perspective of linguistic analysis of the corpus to be formed based on the research of special literature on the abstract types, the methodology of abstract analysis of bachelor's theses and problems of text creation, as well as the results of the pilot study.

From the annotated corpus of texts, 25 Latvian abstracts and their corresponding 25 English translations were randomly selected. The texts were obtained from the Registry of Final Theses of Riga Technical University (RTU) (2023–2024). When looking at the abstracts in correlation with the RTU methodological instructions, it can be established that the text type 'abstract' is mentioned and briefly described, but the descriptions and scope

Submitted: 24/02/2025. **Accepted:** 30/09/2025

Copyright © 2025 Baiba Egle, Dzintra Lele-Rozentāle, Agnese Dubova, Gints Jēkabsons. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

of their structure differ between faculties and study programs. In the pilot study, based on the modified models by Swales and Feak (2009) and Hyland (2000/2004) concerning the structure of abstracts consisting of different moves and steps, the abstracts' text-internal sequence as well as quantitative indicators, such as detailed breakdowns of moves used, the length of an abstract, etc., were determined. Moreover, the relationship between moves forming patterns and the model of moves was adapted to the needs of researching texts written by Latvian students, and problems encountered during the intentional and deliberate annotation of the corpus were identified. These problems are mainly related to the lack of in-depth academic writing courses and the often-overgeneralized style of methodological instructions. The sequence of the moves and steps in the corpus is diverse. As abstracts constitute an internationally standardized text type, it does not seem purposeful to interpret differences in an intercultural context. This study has also found that the editing of abstract translations should be taught to students due to the way these translations are performed. The results of the pilot study show the need for modern academic writing support, which is the focus of further research.

Keywords: academic writing, undergraduate abstracts, bilingual text corpus, translation, Latvian, English

1 Introduction

The study is part of the project *Towards AI-Based Thesis Abstract Writing Aid: Bilingual Text Corpus Preparation, Analysis, and Model Development*, which was launched in late 2024 by RTU Liepaja Academy. The aim of the project is to prepare and analyse a bilingual abstract text corpus formed by the abstracts of the bachelor's theses of Riga Technical University (RTU) in Latvian and English over the last two years (2023–2024) in order to further develop a general abstract text model and its variants as a basis for the development of an academic writing tool to support the writing of the bachelor's thesis abstract.

In parallel to the work with the corpus, we study whether the structure of the abstracts complies with the instructions provided in the methodological materials of the RTU faculties. The project is currently still in its early stages, and therefore this paper will focus on the problems found in the early stages of a pilot study on 50 randomly selected abstracts, while also taking into account the research results of undergraduate student writing skills in other countries and languages. The corpus for the research project consists of over 1000 abstracts written by RTU students in 2023 and 2024. To gain insight into the corpus and what to expect in further research, we randomly selected 25 abstracts (each abstract has a Latvian and English version) to ensure that true randomness is achieved, and that there is no bias regarding the study programme or department. The

years 2023 and 2024 were selected to show the latest trends in abstract writing at RTU. Although abstracts are a standardized text type, they are far from a simple text to write, as students write an abstract usually for the first time writing their bachelor's thesis since they do not typically write abstracts for other assignments during their studies. The aim of this paper is to summarize the identified problem cases and to show the perspective of linguistic analysis of the corpus to be formed based on the research of special literature on abstract types, the methodology of abstract analysis of a bachelor's thesis along with problems of text creation, as well as on the results of the pilot study.

2 Theoretical framework

The analysis of abstracts as a text type is a current issue in academic writing within many linguistic cultures, whereas the number of publications devoted to this genre of text has become almost uncountable, which can be explained by the process of globalization and, in connection with this, the internationalization of studies and research. Studies on non-English abstracts of different study or academic career levels are relatively rare, although it is possible that these studies have not been published in English (for research on abstracts in the humanities in Lithuanian, English, and Russian, see Gobekci 2023, 33). In Latvian, abstracts have been studied as a type of secondary text in comparison with texts in German (Dubova 2009, 83–100). No research on abstracts of student papers has been published so far.¹

This text type in Latvian and English has become an obligatory requirement for all final works such as bachelor's and master's theses. This aspect of academic writing in Latvia has not been looked at in research previously, and RTU, as one of Latvia's largest universities, has a sufficient amount of final works that can be used as an example of the bigger picture in specific academic writing scenarios.

The theoretical basis of this article consists of abstract standards based on ISO and ANSI that give the international definitions of what an abstract should be as well as previous research, which:

1) are attributable to the abstract as text type and its subtypes, as well as 2) research that focuses on bachelor's theses, 3) describes the methodology of abstract analysis of bachelor's theses, and 4) identifies text formation and writing problems at the macro- and microstructure level of the text. These aspects are especially important in the first stage of corpus formation, annotation, and analysis.

¹ At the 29th international scientific conference *The Word: Aspects of Research* (2024), Laiveniece and Helviga presented the topic "Abstract of the study paper as a text genre: Main issues of structure and content".

2.1 Abstracts as a standardized text type

The International Standard (ISO) for abstracts, which is also recommended for the description of a thesis that could be written by an undergraduate student, is defined as a term for a text that “signifies an abbreviated, accurate representation of the contents of a document, without added interpretation or criticism and without distinction as to who wrote the abstract” (ISO, 1). According to the ISO Standard, 3 abstract types are described: 1) informative, which “present(s) as much as possible of the quantitative and/or qualitative information contained in the document”, 2) indicative or descriptive, which can be a guide to the type of document the principal subjects covered, and the way the facts are treated, and 3) informative-indicative, “when limitations on the length of the abstract or the type and style [...] make it necessary to confine informative statements to the primary elements of the document and to release other aspects to indicative statements” (*ibid.*). The choice of a certain type is therefore influenced by the type of the main text.

The distinction between informative and indicative abstracts is also made by the American national standard developed by the National Information Standards Organization (ANSI/NISO) describing informative abstracts as “generally used for documents pertaining to experimental investigations, inquiries, or surveys” and containing “the purpose, methodology, results, and conclusions presented in the original document” (ANSI/NISO). Indicative abstracts, in turn, “are best used for less-structured documents, such as editorials, essays, opinions or descriptions; or for lengthy documents [...]” (*ibid.*) and present the “purpose or scope of the discussion or descriptions” as well as “essential background material, the approaches used, and/or arguments presented in the text” (*ibid.*).

In addition to these three types, there is an extended list, which is not relevant in the context of an undergraduate bchelor’s thesis as this classification refers to another specific document or a different text type context, such as critical abstract and slanted abstract (ANSI/NISO, 18), or highlight abstract (Asikuzzaman 2024).

Abstracts can also be distinguished by form as either paragraphing or structured abstracts (ANSI/NISO, 5).

The description of the text type abstract was detailed by Busch-Lauer (2012). Her research focuses on the communicative aspects of this text type and authorship, content and positioning. Busch-Lauer based her abstract categorization on ISO, ANSI, as well the German Committee for Terminology and Language of the German Documentation Association, and the German Institute for Standards definitions for informative, indicative, informative-indicative, as well as structured abstracts.

Busch-Lauer describes 6 types from the communicative point of view:

1. depending on the time of writing the text – retrospective (written after the primary text) and prospective abstracts (for example, conference abstracts);
2. by the author – as the sole author of the abstract (*Autorenabstract*), an abstract created by someone else (*Fremdabstract*), and machine created abstracts (*maschinell erstelltes Abstract*);
3. by content – informative, indicative and mixed form abstract;
4. by place in the text, for example, as a subtext between the title and the body text, as an abstract in a foreign language at the end of a journal or collection, and as an autonomous text in an *Abstracting Journal*;
5. by form and layout – text, structured and *Schlagwortabstract* (based on *keywords*) (*Text-, Struktur- and Schlagwortabstracts*); and
6. abstract in the language of the document and in a foreign language (Busch-Lauer 2012, 7).

Following this classification, the bachelor's thesis abstracts analysed in our study are characterized by the following traits: retrospective, they are *Autorenabstracts* – written by the authors of the theses, content – informative, indicative or informative-indicative, they are paratexts in the full version of the bachelor's thesis, created in text or in a structured form in the language of the document (bachelor's thesis) as well as in a foreign language (English)².

2.2 Bachelor's thesis abstracts in research

Research devotes less attention to bachelor's theses than to master's theses, doctoral theses and journal articles. Without denying the fact that published article abstract research can provide theoretical support, for example, to creation of study materials for bachelor's students, however, it is necessary to emphasize the different prerequisites that distinguish bachelor's students from other target groups. Students have relatively little experience in research, discourse and text formation of research questions; as also, a limited (and short) period of time is devoted to the acquisition of the basics of the discipline, which is usually 3 years. For this reason, the creation of the corpus planned in the project is focused on the bachelor's level of education, considering the gradualness in the acquisition of text formation.

In part, this may be due to the different requirements of universities in different countries. For example, in Indonesia, as well as in Latvia, students must write abstracts to bachelor's

² Exceptions are certain works written by foreign students in English. Their abstracts are translated into Latvian.

theses in two languages, specifically, in their mother tongue, i.e., Indonesian, as well as in English. Therefore, the question of the equivalence of abstracts in both languages arises. The structure of native language texts translated into English is not necessarily equivalent to the target language abstract genre (Suryani & Rismiyanto 2019, 193). When explaining the lack of research on bachelor's thesis abstracts, Suryani and Rismiyanto emphasize that "the students are still considered new to the academic community and are still guided in conducting research. That can be the reason why few, even might be none [sic!], studies are found on bachelor's thesis abstract" (Suryani & Rismiyanto 2019, 192).

Swales and Feak's book *Abstracts and the Writing of Abstracts* (2009) highlights the 'pedagogical consequences' (Swales & Feak 2009, xi) by pointing to the global increase in the role of the English language and the research literature devoted to it. Their book is addressed to "graduate students and junior researchers" (ibid., xiii), but it is also relevant for undergraduate students, especially if the text type of abstracts is included in the study program.

Abstract analysis is based on the structure of 5 rhetorical moves described by Swales and Feak (2009, 5). A move is explained here as "a stretch of text that does a particular job. It is a functional, not a grammatical term. A move can vary in length from a phrase to a paragraph" (ibid.). These 5 moves, corresponding to the IMRaD (Introduction, Methods, Results, and Discussion) structure, are as follows:

- Move 1: Background, introduction, situation;
- Move 2: Present research, purpose;
- Move 3: Methods, materials, subjects, procedures;
- Move 4: Results, findings;
- Move 5: Discussion, conclusion, implications (ibid.).

This subdivision coincides with the functional moves, as listed by Hyland (2000): Introduction, Purpose, Method, Product, and Conclusion, which underlies a large number of abstract analyses (Darabad 2016; Pratiwi, Hermawan & Muniroh 2021; Putri, Hermawan & Muniroh 2021; Ramadhini et al. 2021, etc.).

We use a slightly modified division in our annotation (see description of methodology in Section 4.2).

Šulovská's (2022) paper, dedicated to the study of academic writing, including the abstract as a genre in English, describes mainly Slovak undergraduate students' abstract writing in ESP classes at the Faculty of Arts, Comenius University, Bratislava. She describes informative abstracts as 'complete', and indicative abstracts as 'limited'. The

informative abstract structure consists of 5 moves – background, purpose, methods, results, and conclusions – but, in the abbreviated case of the ‘reduced abstract’, there are only 3 moves: purpose, methods, and results. Abstracts consist of one paragraph, and the word limit is between 100–500. According to Šulovská, the language of abstracts is characterized by an impersonal style, i.e., passive forms are used, while avoiding pronouns of persons, which, however, depends on the discipline. The abstract typically uses the formal academic style, and, for the acquisition of this style, its typical vocabulary is offered: key verbs, nouns, adverbs, and adjectives (Šulovská 2022, 76).

Descriptive abstracts, which are often written in philosophy, are described as consisting of the background, purpose and focus of the paper or article, without specifying the methods, results and conclusions (Šulovská 2022, 77–86). Šulovská associates moves with the function expressed in a certain style by questions as well as with the use of their characteristic tense. In informative abstracts, the move Background (What?) is implemented with the Present Simple tense form, the Purpose (Why?) uses the Present Simple, Present Perfect and Past Simple tenses, the Methods (How?) use the Past Simple tense, the Results (What?) employ the Past Simple (and Past Perfect), and the Conclusion is presented in the Present Simple tense; tentative verbs, adverbs, adjectives and modals. Descriptive abstracts characterize the Background (What?) with Present Simple, the Purpose (Why?) with Present Simple, and the Focus (What?) is implemented with the Present Simple, tentative verbs, adverbs, adjectives and modals (Šulovská 2022, 87–88).

Frydrychova Klimova wrote about the acquisition of the abstract genre in English to “demonstrate how to teach formal writing, particularly the writing of abstracts in English” (Frydrychova Klimova 2015, 908), by offering guidelines and reflecting the typical mistakes of Czech students and academics when writing English abstract texts.

Student thesis abstracts from the Department of English Education in UIN Syarif Hidayatullah, Jakarta were analysed by Luthfiyah, Alek and Fahriany at the level of text cohesion and moves, by rating the use of cohesive devices based on their gradual technique (high, medium, low) as medium (Luthfiyah, Alek & Fahriany 2015, 148). Moves patterns mostly do not follow the pattern outlined by Swales and Feak (2009), and errors are detected in the tense and passive voice usage, which is explained by the students’ “lack of knowledge about the abstract features” (*ibid.*, 157).

Undergraduate thesis abstracts are also viewed interlingually and intralingually. When emphasizing that “analysis of rhetorical moves of abstracts written in two languages is still scarce”, Putri, Hermawan and Muniroh (2021), based on Hyland’s 5 moves patterns, studied the abstracts of undergraduate as well as graduate theses and dissertations created

by students at 3 different educational levels in Indonesian, as well as their English translations. The number of moves in Indonesian and English was slightly different (284 moves in Indonesian and 281 in English). Apart from move 3 (method) and move 4 (results), all other moves showed differences in different levels of education. At the first two study levels, less attention is paid to conclusions (Putri, Hermawan & Muniroh 2021, 164–166). These results show that the usage of moves in Indonesian and English is relatively similar.

The results of research on scientific articles are also considered important for the acquisition of English writing proficiency. A number of articles focus on abstracts of scientific articles, emphasizing that the results of the analysis will also be useful for students when learning text creation in English without differentiating the level of study: “Finally, the results suggest some key implications for teachers, learners, and all practitioners working in the field of Discourse Analysis (DA), English for Specific Purposes (ESP), and English for Occupational Purposes (EOP).” (Darabad 2016, 137)

The approach to writing bilingual abstracts can be different, and the emphasis is most often on writing English texts. Thus, for example, Frydrychova Klimova analyses the errors of 3rd year Czech part-time students of the University of Hradec Kralove, Czech Republic bachelor’s work in English abstract texts. She found influence of the Czech language arising in the translation process and concluded that the errors stem “from the linguistic-stylistic point of view”, and that Czech students have difficulties in English “word order and objectivity of one’s abstract”, as well as a variety of common grammatical and language use issues that are typical in texts by foreign language learners (Frydrychova Klimova 2013, 514). As a result, “the methodological message for the teachers is to make students first summarize any English text before they start writing any abstract which might be based on the Czech text” (*ibid.*, 516).

The writing process and contents of bachelor’s paper abstracts should be seen as a separate text type as the requirements, writer experience, and context are different from master’s, doctoral and scientific research. A bachelor’s paper abstract is a type of scientific text that a student has to produce for their final thesis as a required part for all students of Riga Technical University, which is a requirement for all study programmes and fields. This might be the first time the student encounters the need for an abstract to be written. While, throughout their studies, students read scientific texts, they might not pay attention to the abstracts of those texts, and especially to the finer structural aspects of an abstract. The goal of a bachelor’s paper abstract can be seen as significantly different from an abstract of a scientific paper written by an experienced researcher – as the bachelor’s student does not need to ‘sell’ the idea of the research to entice the readers. Commonly, bachelor’s thesis abstracts reflect the student’s research, which is, in most

cases, more practical than theoretical – for example, a student might describe a plan to increase employee motivation in one specific company without claiming (or attempting) to have found a ground-breaking theory or application. Another factor to consider is the limited readership. While the abstracts are made publicly available in Latvia, it is not legally required to make the bachelor's thesis itself publicly available. Thus, the initial readership of the bachelor's thesis abstract consists of the advisor, the reviewer and the bachelor's defence committee. Later, successfully defended bachelor's theses might be of interest for other students of the same study program, but, usually, a bachelor's thesis abstract might not have the larger potential reader public like a peer-reviewed scientific article published in a journal.

Instead, the Bachelor's paper abstract could be treated as a type of a stepping stone towards academic writing – where the writer needs more guidance and assistance to learn the skills they need. Bachelor's level students may have some experience of writing essays and other homework tasks set by their teachers, but, most often, homework does not require the writing of an abstract. The guidelines provided by their university are often the main, and possibly only, document which the students consult when writing their bachelor's paper abstract. That is why some institutions will give a template for an abstract to aid students in their writing, but even the templates are not given with extensive descriptions and tips on how to write the abstract.

Good guidelines could be a way to improve the writing quality of the abstracts that BA students submit. The writers of the guidelines should take into account the amount of experience the students have, how the material they can access is laid out, and what requirements are set for the students, so that abstracts match the requirements and expectations of the institution and the study program.

3 Abstract in bachelor's theses in Latvian universities: The example of RTU

The Latvian education system requires that bilingual abstracts are written for the bachelor's thesis.

RTU is one of the largest universities in Latvia, and its main fields of study are natural sciences, engineering and technology, social sciences and humanities, and art sciences. Consequently, our corpus consists mainly of the abstracts of bachelor's theses in the field of engineering, which is expected for a university that traditionally used to focus on a large variety of engineering-related fields and added a strong Humanities branch only in 2024. In accordance with the requirements in force in Latvia, abstracts are included in the full text of the bachelor's thesis and are available in universities' repositories. bachelor's theses and abstracts have been collected in the Registry of Final Theses of the RTU since 2010.

While RTU does have scientific writing courses for students regardless of their study program, there are no specific lectures for the study of abstract writing, and the only sources of information on abstract writings are the methodological instruction offered by the faculty, materials available on the internet, information obtained in consultations, and, of course, the exchange of experience with course mates.

RTU methodological guidelines for the development of final papers have been developed both for individual faculties, such as the Faculty of Computer Science and Information Technology (DITF 2023), as well as for specific study programs, such as the Professional Bachelor's Study Program "Heat, Gas and Water Technology" (BMF PBSP 2024) of the Faculty of Construction and Mechanical Engineering. The instructions stipulate that an abstract in Latvian and its English translation are a mandatory part of the bachelor's thesis. DITF (2023, 5) describes an abstract as consisting of four parts: keywords, a brief description of the content of the work with an introduction to the study, the purpose and results, and data on the scope of the work. The abstract specified in BMF (2024, 9) must contain four parts: the name of the author of the work, as well as the title; then, the topic, content, main results and conclusions; moreover, the language of the work must be indicated, and, finally, the length of the work must be stated. Keywords must be specified after the text of the abstract. The length of the abstract in both instructions differs: DITF (2023, 5) restricts it to no more than one page, while BMF (2024, 9) limits the text to no more than three-quarters of a page. Differences thus appear both in the structure and length of the abstract. Therefore, presumably, abstract texts will be different, which may also be determined by the specifics of the field of science. All the methodological instructions of RTU on the development of the final papers studied indicate that the abstract in Latvian, followed by an analogue abstract in English. Therefore, we presume that the English abstract is a subsequent translation of the Latvian abstract.

4 Methodology

4.1 Description of the material

50 abstracts (25 in Latvian and 25 in English) from the hitherto annotated corpus (2023–2024) were randomly selected with the following aims:

- 1) determine the number of words in Latvian and English abstracts;
- 2) determine the implementation of moves and steps based on the Moves and Steps models by Swales and Feak (2009), Hyland (2000/2004), etc., and our modified model (see below);
- 3) explore the patterns forming moves, and explain their relationship with the role of external factors in text formation;
- 4) infer perspectives for further research.

Metadata is excluded from the further analysis. This metadata incorporates keywords, as well as the description of the length of the work, number of pages, number of chapters, number of attachments, number of literature sources, number of images, number of tables, etc., which are marked with [[len]] when annotating the corpus. Also, information about the author, supervisor, title, type of work, language, etc., which obtains a marker in the annotation of the corpus [[meta]], is excluded from this analysis.

The abstracts in both languages were annotated; however, the Latvian abstract can be considered the original source material, even though abstracts in both Latvian and English were analysed.

4.2 Description of the analysis

The following steps are taken during the analysis:

- 1) anonymizing and annotating the selected texts, which, in this pilot study, was carried out manually in a group of 3 annotators. This made it possible to immediately discuss the encountered problems and agree on a solution;
- 2) tabular compilation of detected moves and steps;
- 3) preparation of texts for qualitative and quantitative analysis with *Sketch Engine*;
- 4) linguistic (genre) analysis of texts with the aim of analysing how coherence is implemented;
- 5) verifying the equivalence of the translated abstracts.

The annotation of the whole corpus takes place, based on the 5-moves-model of Hyland, and adapted to the specifics of this research project, for example, by having move 2 split into 3 different steps so that to enhance the clarity of the annotation process and to ensure that the annotation process matches the guidelines that the students receive:

- Move 1: Background [[back]] – description of the current situation, state-of-the-art, background, introduction to the topic;
- Move 2: Aim/purpose, which divides into 3 steps: motivation/gap [[motiv]], aim [[aim]], hypothesis/question [[hyp]];
- Move 3: Methodology/structure divided into 2 steps: thesis structure [[struct]] and methods/materials/subjects/procedures [[meth]];
- Move 4: Results/findings [[res]];
- Move 5: Discussion/conclusion/significance [[conc]].

An example of an annotated abstract in Latvian and English is given in Table 1.

Latvian version

Katrā valstī un nācijā ir neredzīgi un vājredzīgi cilvēki, taču šai iedzīvotāju grupai informācijas uztveršana un piekļūstamība internetā ir ierobežota. Pirms mājaslapu uzlabošanas, tajās ir jānoskaidro piekļūstamības problēmvietas. To var izdarīt, testējot attiecīgās mājaslapas. Mājaslapu piekļūstamības iespējas var testēt gan manuāli, gan izmantojot kādu automatizētu rīku. [[back]]

Bakalaura darba mērķis ir salīdzināt pieejamos tīmekļa piekļūstamības testēšanas rīkus, pamatojoties uz to spēju identificēt mājaslapu neatbilstības WCAG vadlīnijām, darboties dažādās tehnoloģijās un vidēs, kā rezultātā izstrādāt vadlīnijas šādu rīku izvēlei. [[aim]]

Bakalaura darbā tika izvirzīti automatizēto piekļūstamības testēšanas rīku salīdzināšanas kritēriji, balstoties uz literatūras analīzi. Vadoties pēc izvēlētajiem kritērijiem, tika veikti eksperimenti, kas ietver testēšanas rīku ātrdarbību, precīzitāti, spēju noteikt neatbilstības WCAG vadlīniju pamatprincipiem un darboties dažādās vidēs, interneta pārlūkos u.c. [[meth]] Pēc eksperimentu pabeigšanas darba autors izstrādā vadlīnijas piekļūstamības testēšanas rīka izvēlei. [[res]] Nemot vērā uz bakalaura darba autora veikto eksperimentu rezultātiem izstrādātās vadlīnijas, to lietotājiem būs iespēja izdarīt pamatošu piekļūstamības testēšanas rīka izvēli atbilstoši savām vajadzībām un prasībām. [[conc]]

English version

Blind and visually impaired people are part of every country and nation, but this group has limited access to information and the Internet. Before improving websites, problem areas need to be identified. This can be done by testing the websites. Website Accessibility can be tested either manually and or using automated tools. [[back]]

The aim of this thesis is to compare available web Accessibility testing tools based on their ability to identify non-compliance of websites with WCAG guidelines, to work in different technologies and environments, and to develop guidelines for the selection of such tools. [[aim]]

Based on a literature analysis the author sets out criteria for comparing automated Accessibility testing tools. Based on the selected criteria, experiments were carried out, which included the speed of the tools, the accuracy of the tools, the ability of the tools to detect inconsistencies with the basic principles of the WCAG guidelines, the ability of the tools to operate in different environments, web browsers, etc. [[meth]] *After the completion of the experiment, the author of the work puts forward guidelines for the selection of Accessibility testing tools.* [[res]] *Following the results of the experiment carried out by the author, the user of the guidelines can choose the appropriate tool for his situation.* [[conc]]

Table 1. Example of an annotated abstract in Latvian and English

The pilot study was carried out in parallel with the initiated annotation of the corpus, which takes place in several stages, and which has not yet been completed. Its preliminary results are also partially reflected in this article. The pilot study offers a detailed and broader review of the texts with the aim of recording the problems expected in a more extensive corpus analysis in the future.

5 Results of the pilot study

5.1 Quantitative results

The average number of words in Latvian abstracts is 191 words, with the longest abstract comprising 298 words, and the shortest abstract comprising 101 words. The number of abstract words translated into English is higher than in the original Latvian abstracts. On average, English abstracts consist of 243 words, with a maximum of 406 and a minimum of 128 words (see Figure 1). The length of the analysed abstracts corresponds to the length specified by ISO 214.

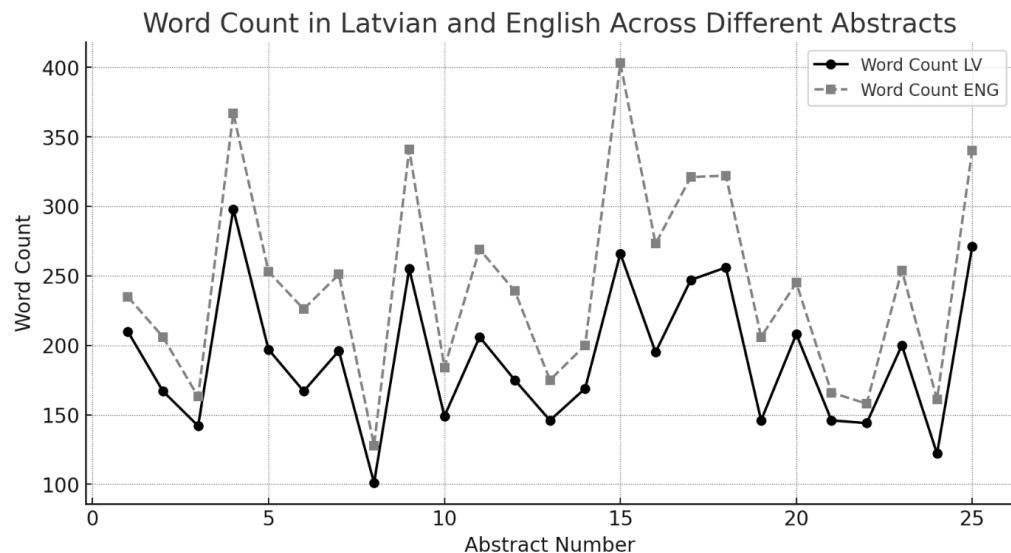


Figure 1. Number of words in abstracts in Latvian and English

RTU general requirements for a bachelor's thesis and other final theses contain an indication that the requirements for the structure and scope of the thesis are determined by the faculties (Guidelines for RTU 2014). This could explain the different abstract lengths in the Registry of Final Theses. The aim of this research is to give insight into various engineering-related abstracts without specifying the exact fields or subfields of science.

Regarding the annotation of moves, no differences were found between the Latvian and English texts as they are direct translations, however, in the larger corpus of abstracts that is not part of this study, some differences may exist. The results of this process are summarized in Table 2 for each selected text with the aim of visualizing the structure of the texts. Moves are arranged in the order in which they are found in the text. In some cases, one move breaks down if it has another move in the middle. This usage of atypical, repeated moves has been described in Luthfiyah, Alek and Fahriany (2015),

where they call it ‘move cycling’, while other authors have called it a ‘hybrid move’, but we use ‘split’ for conciseness and to match the moves found in the Latvian student writing, see Table 2 below.

#	Moves						
1	M2		M3		M4		
2	M3						
3	M5						
4	M3			M5			
5	M2	M3	M2	M1	M4		
6	M3	M4	M2	M3			
7	M1	M2	M3	M5			
8	M2			M3			
9	M2		M3		M5		
10	M2			M4			
11	M2			M3			
12	M2	M4	M5	M4	M5		
13	M2		M3		M4		
14	M2		M3		M4		
15	M5		M2		M4		
16	M2	M1	M3	M4	M3		
17	M3	M5	M3	M4	M3		
18	M1	M4	M2	M4	M3		
19	M1	M2	M3	M4	M5		
20	M2		meta		M3		
21	other	M2	M3	M4	M5		
22	M1	M2	M3	M4	M3		
23	M2			M3			
24	meta	M2	other	M2	M3		
25	M2		M3		other		

Table 2. Move patterns

Move patterns show great diversity in the structure of texts, and this applies not only to the representation of moves, but also to the order in which they appear in the text. Only one of the 25 texts contained the traditional order of moves from 1 to 5 (see No. 19, Table 2), and only 2 more contained all five moves, but in a mixed or divided order. Surprising is a group with 1 move – in two cases, M5 [[conc]], whereas in one – M3 [[meth]]. The dominant structure is with 2 moves: there are 7 texts containing 2 moves; 5 with M2M3, excluding a split with [[meta]]. There were 8 texts with 3 moves, of which, only 2 are with identical structures, specifically, M2M3M4. It is obvious that one cannot speak of a certain textual

tradition, and the question that arises is whether this affects the coherence of the text. The variety of models and divisions of moves suggests potential difficulties in text creation and serve as an incentive for the creation of teaching materials.

A different result from the requirements of the writing guidelines that demand all moves to be present in the abstract is observed when summarizing the number of moves in the analysed texts and calculating their percentage (see Table 3).

Moves	Absolute frequency (split moves)	Percentage
Move 1	6	24%
Move 2	20	80%
Move 3	20 (5)	80%
Move 4	13 (4)	52%
Move 5	13 (1)	52%

Table 3. Moves frequency relative to the number of texts

None of the five moves are represented ideally in all 25 (50) texts. The dominant moves are Move 2 and Move 3, so it can be said that, most often, attention is paid to the goal (also motivation) and methods (also structure). Surprisingly few abstracts contain background information. Since none of the moves reaches at least 90% in all texts, it is not possible to talk about any obligatory category based on Hüttner's breakdown, according to which, 90–100% means that the category is Obligatory, 50–89% stands for Core, 30–49% is within the range of Ambiguous, while 1–29% is perceived as Optional (Hüttner 2010, 205). In the selected texts, M2 and M3 are to be interpreted as the core, M4 and M5 as ambiguous, and M1 as optional. When compiling this assessment with the requirements of the methodological instructions, a discrepancy can be observed as the abstracts do not match the requirements. It is possible that incorporating a larger number of texts would change these percentages.

Texts differ in the number of moves represented in them. The percentage is shown in Table 4.

Number of moves (from all possible 5 moves)	Number of texts	Percentage
5	3	12%
4	4	16%
3	8	32%
2	7	28%
1	3	12%

Table 4. Number of moves represented in texts

As the proportion reflected in the table shows, the corpus of the pilot study is dominated by texts containing 2 and 3 moves, which are present in 28% and 32% of texts, respectively. 4 texts contain 4 moves (16%). 3 texts represent all 5 moves; also, three texts contain 1 move, which, in the latter case, stands at 12%. This number should be viewed together with the diverse layout of moves in the text shown in Table 2.

Steps form a subcategory of moves. In the corpus, steps are divided in M2 and M3, and the quantitative results are presented in percentages in Table 5. The numbers in brackets denote steps that have been split, where, besides the original function, they have signs of another step.

Steps	Step: motiv	Step: aim	Step: struct	Step: meth
Number of texts	5	18	18 (4)	13 (1)
Percentage	20%	72%	72% (16%)	52 (4%)

Table 5. Steps of Move 2 and 3. The number in brackets represents the corresponding step splits

Step [[motiv]], found in 5 texts, in four cases is considered to be an addition to the goal, but in one case it replaces the wording of the goal, which usually appears explicitly with the signal word *goal*. The formulation of the goal was found only in 18 texts, although, in the methodological instructions, it is emphasized as an essential component of the abstract. The relatively high proportion of step [[struct]] – at 72% – is due to two reasons. Traditionally, in texts of abstracts, as in the introductions to bachelor's theses, tasks are formulated to which the structure of the work is subject, and therefore they coincide. The second reason is the use of the indicative abstract, which includes a description of the textual structure. Step [[meth]], one of the three most important constituent parts of abstracts next to aim and results, appears in only 52% of texts. Again, the reason can be found in the fact that these abstracts are of the indicative type, which allows an indication of the use of methods, but does not specify them, thus resulting in the label [[struct]].

5.2 Equivalence of translations

Translations are often studied in connection with scientific article abstracts, emphasizing the usefulness of the acquired knowledge, also “when it comes to designing material for students [...] with a view to helping new entrants into the academic discourse community who face difficulty with producing clear and coherent abstracts” (Pezzini 2003, 97).

Previous research also expresses the opinion that students who are not native speakers of English and are not studying English professionally should not abandon translation but

instead create an abstract by using the 5-move model, as developed by Hyland, so that to ensure that they are “following the conventional English rhetorical moves” (Suryani & Rismiyanto 2019, 197).

All students at RTU are required to submit an abstract in English, and therefore students have to find a way to create the English text. Most of the time, students focus on the Latvian version of the abstract text and then use the Latvian version as the basis for the English text. As the English text is a requirement but not a distinct writing priority, students are likely to choose various machine translation tools to produce the English text, which is not forbidden as per the guidelines. Some of the abstracts analysed in this paper had a reasonably acceptable English version. Despite some awkward textual choices, the text could overall be seen as acceptable in terms of lexis, grammar, and overall content. In this context, it should be noted that Latvian students focus more on what was required in the university guidelines and not on the English language tradition of writing an abstract; therefore, the abstracts may fit the requirements set by the institution, but not the overall concept of what an abstract and its writing process is like in English.

Overall, the Latvian and English language abstracts were similar in length, and all of them contained the same information in both language versions. While most English versions of the abstracts seem to have at least some post-machine translation editing, as the majority of the texts are reasonably coherent in English, there were some linguistic issues that the authors might have failed to notice, perhaps due to time constraints. Sometimes, the wrong choice of verbs/grammatical constructions might cause confusion. For example, the sentence *the construction of the smart private house wiring was created* implies to the reader that the wiring in this project was physically completed, but, in the Latvian version of the abstract, it is only a draft of potential wiring that could be implemented.

Misuse or potential false friends were also observed, most strikingly in this example: *The graphic part consists of 5 pages*. With this, the student means five pages of drawings and graphs, and not explicit content. In two separate English translations, where the author means to describe the content of the thesis, there is a translation mishap calling it *the job* which is a literal translation from the Latvian word *darbs* used in the original *paper=work*. Similarly, typically used Latvian *noskaidrot* which means ‘to find out’, was most often translated to *clarified* in various tenses. While it is not a significant mistake overall, it appears several times in texts by various authors.

The quality of the translation also heavily depends on the quality of the original text in Latvian. Confusing original sentences hardly ever get clearer after translation, especially if the author chose not to re-read and edit the text. It can be found, for instance, in the following example of language use:

More extensive consideration is given to mobile verifiers, as they are based on the work itself. This example shows that the author was unable to explain their idea in the abstract – from this line, it appears as if there are mobile verifiers that are based on the findings of the bachelor’s paper research, when, in fact, the author’s research focused on evaluating the already existing mobile verifying apps.

Another issue that machine translation cannot solve and which solely depends on the human editor’s efforts are spelling errors that can occur in the original text. For example, one of the abstracts annotated for the corpus had a misspelling of the Latvian – *tīrgus* ‘market’, the misspelling *tigrus* instead of *tīrgus* turning into *tiger* in English. Some students might have too much trust in machine translation capabilities, without double-checking the results and the original text. In traditional translation study theory, it is often advised that a translation should be reviewed not only by the translator themselves, but also by a native speaker of the target language for correctness. The reviewing and editing process for the translation is likely quite short, or perhaps even non-existent, depending on the student’s perception of the importance of the English abstract of their bachelor’s thesis. Students know that their work will be definitely read by the reviewer and their advisor, but they likely do not treat the English version of their abstract as a significant text that represents their work. For graduate studies, especially at the doctorate level, abstract writing skills are important, and these skills should theoretically consistently improve, but, for bachelor’s students, this might be the only time they are tasked with translating an abstract to English.

This highlights the need to have a more in-depth look at the English language versions of the abstracts that will be annotated for this research project beyond the selected 25 texts discussed in this paper. While the abstracts are almost equivalent, with move-for-move and step-for-step matching in the Latvian and English versions, the language quality and editing importance could be another factor to include in the prospective AI tool to help the students write, especially while taking into account their level of experience with abstracts and their potentially limited experience with post-editing the scientific text machine translation output.

6 Conclusions and suggestions regarding corpus annotation and future work

Abstracts are an important, but so far undervalued text type in the programs of Latvian universities, and the annotation and more detailed analysis of the corpus in the pilot study makes it possible to draw several conclusions divided into two groups. First of all, there are some possible solutions to improve student academic writing: a detailed methodological instruction, including familiarization with the types of abstracts, and the lack of special in-depth academic writing courses at the undergraduate level. Since

the abstract of the bachelor's thesis is the first contact with this text type, and as the writing of the text most likely takes place within a limited period of time, the great diversity of moves patterns, multiple split cases and observed moves inconsistencies for signal words such as *aim*, *results*, *conclusion* are inevitable. As a result, in the process of annotating texts, it was necessary to focus on the content of the text, and, if necessary, to adapt to the fact that students used an atypical order of the moves at times. As a result, the annotation process is time-consuming, which is also reinforced by the fact that most abstracts consist of one paragraph and are relatively rarely structured. The issue of annotating informative and indicative abstracts according to the same criteria is debatable. For informative abstracts, it is possible to evaluate all moves and steps, whereas, for indicative abstracts, instead of methods, results and conclusions, there may be an indication of the structure, which is therefore the dominant move in the analysed corpus. Looking at the results of the analysis of the Latvian abstracts, it seems that there is no need to raise such a question of differences between an indicative and informative abstract in the language combination Latvian – English because:

- 1) there is no established genre tradition in Latvian;
- 2) often, in writing guidelines/instructions there is a requirement to translate close to the original text, and thus it results in awkward abstract texts in two languages, and neither of those actually meets the guidelines;
- 3) the type of text is standardized, and the orientation toward ISO requirements takes place regardless of the language used.

Differences between Latvian and English are possible at the microstructure level by selecting the appropriate phrases for each language.

The second group of conclusions summarizes the possibilities in regard to future research, both in terms of external factors and in terms of working with the corpus. The following should be added to the external factors in this context:

- 1) systemic and systematic error analysis;
- 2) development of methodological tools that would allow students to learn this process independently;
- 3) in the first stages of higher education, when writing semester papers and bachelor's theses, the choice should fall in favour of the structured abstract form, as they have an easier-to-understand content which students can reproduce more easily and learn the basics of the text type;
- 4) teaching students the meaning of the type of text: focusing on learning logical text formation and how a publicly available text, such as an abstract, can be useful in their further careers;

- 5) a special set of academic writing classes with exercises in Latvian and English, including abstract types, signal words and related key phrases, verb tenses and order usage moves.

The question of dividing abstracts into study levels is debatable. Since students at the bachelor's and master's/doctoral levels have different prerequisites for abstract writing, it is desirable to separate the bachelor's level in text formation from the other two. This does not preclude the use of the experience gained from the scientific article abstract research and adaptation to the level of undergraduate studies.

Acknowledgments

This article has been supported by research and development Grant No. C4835. ZPD.PI.0024P1 under the EU Recovery and Resilience Facility funded Project No. 5.2.1.1.i.0/2/24/I/CFLA/003 “Implementation of Consolidation and Management Changes at Riga Technical University, Liepaja University, Rezekne Academy of Technology, Latvian Maritime Academy and Liepaja Maritime College for the Progress towards Excellence in Higher Education, Science, and Innovation”.

Data Sources

Registry of Final Theses of RTU – Rīgas Tehniskā universitāte. [Riga Technical University]. 2010–2025. *Noslēgumu darbu publiskā datu bāze*. [Public database of final theses]. Available at: <https://www.rtu.lv/lv/studijas/bakalaura-limena-studijas/noslēgumu-darbu-registrs>. Accessed: 10 February 2025.

References

- ANSI/NISO – ANSI/NISO Z39.14-1997 (R2015). *An American National Standard. Developed by the National Information Standards Organization*. Published by the National Information Standards Organization Baltimore, Maryland, USA. <https://doi.org/10.3789/ansi.niso.z39.14-1997R2015>
- Asikuzzaman, Md. 2024. *What is an Abstract? Definition, Purpose, and Types Explained*. Available at: <https://www.lisedunetwork.com/what-is-an-abstract-definition-purpose-and-types-explained/>. Accessed: 10 February 2025.
- BMF PBSP – RTU Būvniecības un mašīnzinību fakultātes (BMF) profesionālās bakalaura studiju programma “Siltuma, gāzes un ūdens tehnoloģija”. [RTU Faculty of Construction and Machinery, study programme “Heating, gas and water technologies”]. 2024. *Metodiskie norādījumi bakalaura darba ar projekta daļu izstrādāšanai un aizstāvēšanai*. [Methodological guidelines for the development and defense of bachelor papers with a project part]. Rīga: RTU Izdevniecība.

- Available at: <https://ebooks.rtu.lv/wp-content/uploads/sites/32/2024/09/Metodiskie-noradijumi-bakalaura-darba-ar-projekta-dalu-izstradei-un-aizstavesanai.pdf>. Accessed: 10 February 2025.
- Busch-Lauer, Ines. 2012. Abstracts – eine facettenreiche Textsorte der Wissenschaft. *Linguistik online* 52 (2), 5–22. <https://doi.org/10.13092/lo.52.293>
- Darabad, Ali Mohammadi. 2016. Move Analysis of Research Article Abstracts: A Cross-Disciplinary Study. *International Journal of Linguistics* 8 (2), 125–140. <https://doi.org/10.5296/ijl.v8i2.9379>
- DITF – RTU Datorzinātnes un informācijas tehnoloģijas fakultāte. [Faculty of Computer Science and Information Technology]. 2023. *Norādījumi studiju noslēgumu darbu noformēšanai*. [Guidelines for formatting of final student works]. Rīga: RTU Izdevniecība. Available at: https://ebooks.rtu.lv/wp-content/uploads/sites/32/2023/03/9789934226960-DITF_metodiskie_norad-2021-LV.pdf. Accessed: 10 February 2025.
- Dubova, Agnese. 2009. Sekundārie teksti vācu un latviešu valodā. [Secondary texts in Latvian and German]. *Zinātniskā komunikācija starpkultūru kontekstā*. [Scientific communication in an intercultural context]. Agnese Dubova, Māra Leitāne & Dzintra Lele-Rozentāle, eds. Ventspils: Ventspils Augstskola. 83–100.
- Frydrychova Klimova, Blanka. 2013. Common Mistakes in Writing Abstracts in English. 3rd World Conference on Learning, Teaching and Educational Leadership (WCLTA 2012). *Procedia – Social and Behavioral Sciences* 93 (1), 512–516. <https://doi.org/10.1016/j.sbspro.2013.09.230>
- Frydrychova Klimova, Blanka. 2015. Teaching English Abstract Writing Effectively. 5th World Conference on Learning, Teaching and Educational Leadership, WCLTA 2014. *Procedia – Social and Behavioral Sciences* 186, 908–912. <https://doi.org/10.1016/j.sbspro.2015.04.113>
- Gobekci, Erika. 2023. Rhetorical structure and linguistic features of research article abstracts in the humanities: the case of Lithuanian, English, and Russian. *Taikomojī kalbotyra* 19, 33–56. <https://doi.org/10.15388/Taikalbot.2023.19.4>
- Guidelines for RTU – Rīgas Tehniskā universitāte. Studiju departaments. [Riga Technical University]. 2014. *Norādījumi studiju noslēgumu darbu noformēšanai*. [Guidelines for the formatting of final theses]. Rīga: Rīgas Tehniskā universitāte. Available at: https://www.rtu.lv/writable/public_files/RTU_nordjumi_studiju_noslēgumu_darbu_noformanai.pdf. Accessed: 10 February 2025.
- Hüttner, Julia. 2010. The potential of purpose-built corpora in the analysis of student academic writing in English. *Journal of Writing Research* 2 (2), 197–218. <https://doi.org/10.17239/jowr-2010.02.02.6>
- Hyland, Ken. 2000/2004. *Disciplinary discourses: Social interactions in academic writing*. Michigan Classics Edition. Ann Arbor, MI: University of Michigan Press.

- ISO – *International Organization for Standardization. Documentation – Abstracts for publications and documentation. Documentation – Analyse pour les publications et la documentation.* ISO 214–1976 (E), 1976.
- Luthfiyah, Alek & Fahriany. 2015. An Investigation of Cohesion and Rhetorical Moves in Thesis Abstracts. *IJEE (Indonesian Journal of English Education)* 2 (2), 145–159. <https://doi.org/10.15408/ijee.v2i2.3086>
- Pezzini, Ornella Inês. 2003. Genre Analysis and Translation – an Investigation of Abstracts of Research Articles in Two Languages. *Cadernos de Tradução* 2 (12), 75–108.
- Pratiwi, Dian, Budi Hermawan & Rd. Dian Muniroh. 2021. Rhetorical Move Analysis in Humanities and Hard Science Students' Undergraduate Thesis Abstracts. *Advances in Social Science, Education and Humanities Research* 546. Proceedings of the Thirteenth Conference on Applied Linguistics (CONAPLIN 2020), 121–128. <https://doi.org/10.2991/assehr.k.210427.019>
- Putri, Fanny, Budi Hermawan & Rd. Dian Muniroh. 2021. Rhetorical Move Analysis in Students' Abstracts Across Degrees. *Advances in Social Science, Education and Humanities Research* 546. Proceedings of the Thirteenth Conference on Applied Linguistics (CONAPLIN 2020), 162–167. <https://doi.org/10.2991/assehr.k.210427.025>
- Ramadhini, Tasya Maharani, Isti Tri Wahyuni, Nida Tsania Ramadhani, Eri Kurniawan, Wawan Gunawan & R. Dian Dia-an Muniroh. 2021. The Rhetorical Moves of Abstracts Written by the Authors in the Field of Hard Sciences. *Advances in Social Science, Education and Humanities Research* 546. Proceedings of the Thirteenth Conference on Applied Linguistics (CONAPLIN 2020), 587–592. <https://doi.org/10.2991/assehr.k.210427.089>
- Suryani, Fitri Budi & Rismiyanto. 2019. Move Analysis of the English Bachelor Thesis Abstracts Written by Indonesians. *Prominent Journal of English Studies* 2 (2), 192–199.
- Swales, John M. & Christine B. Feak. 2009. *Abstracts and the Writing of Abstracts. Vol. 1 of the Revised and Expanded Edition of English in Today's Research World.* Ann Arbor, MI: University of Michigan Press.
- Šulovská, Denisa. 2022. *Selected topics from academic writing.* Comenius University in Bratislava, Faculty of Arts, Bratislava: STIMUL.

How to annotate a corpus for research on frames in the language of mathematics

Bernhard Fisseni

Department for Linguistics
Institute for German Studies
Faculty of Humanities
University of Duisburg-Essen
Universitätsstraße 12
D-45141 Essen, Germany
E-Mail: bernhard.fisseni@uni-due.de
ORCID iD: <https://orcid.org/0000-0003-3434-8964>
<https://ror.org/04mz5ra38>

Bernhard Schröder

Department for Linguistics
Institute for German Studies
Faculty of Humanities
University of Duisburg-Essen
Universitätsstraße 12
D-45141 Essen, Germany
E-Mail: bernhard.schroeder@uni-due.de
ORCID iD: <https://orcid.org/0000-0002-7847-1393>
<https://ror.org/04mz5ra38>

Deniz Sarikaya

Institute for Medical Engineering
Ethical Innovation Hub
Universität zu Lübeck
Ratzeburger Allee 160
D-23562 Lübeck, Germany
E-Mail: d.sarikaya@uni-luebeck.de
ORCID iD: <https://orcid.org/0000-0001-8951-8161>
<https://ror.org/00t3r8h32>

Department of History, Archaeology, Arts,
Philosophy, and Ethics (HARP)
Centre for Logic and Philosophy of Science
Vrije Universiteit Brussels
Pleinlaan 2
BE-1050 Brussels, Belgium
<https://ror.org/006e5kg04>

Submitted: 01/11/2025. **Accepted:** 16/12/2025

Copyright © 2025 Bernhard Fisseni, Deniz Sarikaya, Bernhard Schröder. Published by Vilnius University Press
This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract. This paper outlines a plan for designing a corpus-linguistic project from which an analysis of frames specific to mathematical proof texts shall be derived. Previous work has already developed instances of frames for mathematical texts. We have argued that these frames are well suited to model how background knowledge enriches explicitly given information. To this end, a collection of mathematical texts needs to be annotated.

We describe the idea behind the corpus annotation, frame semantics and how we adapted frames for mathematical texts in particular by distinguishing between structural proof frames and ontological ones. Ontological frames for instance correspond to proof techniques, while structural frames model domain knowledge like exact definitions of mathematical structures. We describe annotation principles, explain qualifications annotators should have and how such annotations can be evaluated.

We explain potential linguistic research questions: The corpus will allow us to study the linguistic means by which frames are introduced and signaled. We plan to generalize beyond previous case-based analyses and to provide a foundation for broader empirical research on mathematical language and practice. We argue that these can be used in deeper semantic parsing and the development of interactive theorem-proving software. We furthermore assume that this perspective on mathematical text will also have implications on the philosophy of mathematics and its practice.

Keywords: frames, mathematical language, special languages, corpus annotation, mathematical proofs, philosophy of mathematical practice

1 Introduction

This paper is a whitepaper for a corpus-linguistic project. We aim to apply the theoretical concept of frames to mathematical texts. The project builds on earlier work within the *Naproche* project¹ and first applications of the concept of frames to proofs involving the mathematical concept of induction. Frames allow to model how explicitly given information is combined with expectations deriving from background knowledge (see section 2 for more context).

In the context of mathematical proofs, the frame concept can be applied to model how the expectations that mathematical readers have due to their mathematical training allow them to interpret a mathematical text and to complement it with additional relevant information. This is particularly important in the case of mathematical proofs, which notoriously contain gaps (see section 3 for a discussion of relevant research questions).

¹ Available at: <http://www.naproche.net>. Accessed: 5 December 2025.

Until now, research on frames in mathematical texts has been based on case studies. To be able to generalize our findings, we want to develop a corpus of mathematical texts annotated with structural and ontological frames, especially with regard to the introduction of new frames and the linguistic means of signaling frame use. For now, we will focus on proofs from mathematical textbooks involving various basic proof strategies.

We claim that the frame project can be helpful for natural language processing in general: The frame approach for mathematical text could be helpful in mathematical text parsing and in the usage of interactive and automated theorem provers. We have also already argued that the frame approach has philosophically interesting consequences for debates about mathematics (see section 3.1).

From a cognitive perspective we consider frames a useful tool for modelling learning and understanding of mathematical structures.

In section 2, we give a more detailed introduction to the frame concept and the research context. In section 3, we discuss how the corpus may be used in research of (the philosophy of) mathematical practice and language with respect to frame acquisition and usage.

This corpus annotation goes hand-in-hand with the creation of a library of structural and ontological frames. This library will contain structural information about the logical, conceptual and linguistic relationships between the frames in the library, comparable to a tag set or a set of grammatical relations in the background of traditional linguistic corpora. We will discuss the annotations guidelines to be further developed in this corpus linguistic project in section 4 and section 5 and discuss the evaluation of the annotation procedure and its results in section 6. Finally, after the conclusion in section 7, we return to the bigger picture in section 8.

2 The concept of frames²

In the context of mathematical proofs, readers of proofs have expectations due to their mathematical training. These allow them to interpret a mathematical text and to complement it with additional relevant information. Originally developed in linguistics, cognitive science and artificial intelligence, the concept of frame – nowadays very polysemous – is employed in models of how understanding combines explicitly given information with expectations derived from background knowledge.

2.1 Related framework

Minsky (1974) introduces frames as a general “data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child’s birthday

² The text of this section is closely based on the corresponding section 2 of Carl et al. (2021).

party" (Minsky 1974, 1); frames are part of a frame-system. Despite the reference to "situations", frames can be used to model concepts in the widest sense, already evidenced by the examples in Minsky (1974), from vision to story understanding.³

Each frame contains slots (or features), which can be (sub-)frames again. Slots can have default values or carry constraints. They are filled by concrete values. Some formal systems such as the Frame Representation Language (FRL, cf. Roberts & Goldstein 1977) permit the 'lazy' calculation of values from others by attaching procedures to slots. For instance, a circle in Euclidean geometry, the centre and the radius suffice to describe the circle. Thence, we may calculate the circle diameter, if needed. Related to feature structures (see, e.g., Carpenter 1992), later formalizations of the frame concept use an inheritance hierarchy of types to constrain both the proliferation of features and their values.

FrameNet⁴ is an important linguistic project using frames. Its hierarchical model of (mainly) verb semantics defines frames whose main constituents are the participants and their semantic roles. Frames are evoked by verbs; they provide roles that can be taken on by entities in the discourse universe. These roles can be either core or non-core roles, which captures the salience and optionality of the roles within a frame. For example, the FrameNet frame *Commerce_buy* has the core roles BUYER and GOODS, while SELLER, MONEY and MEANS (e.g., cash vs. check) and many others are non-core roles. Differences in role assignment capture semantic (and pragmatic) differences between verbs, as in the following example. Here, the explicitly realized frame elements are annotated with their roles:

- (1) [Buyer *John*] *bought* [Goods *a beautiful medieval book*] [Time *yesterday*]. [Seller *Peter*] *sold* [Goods *a beautiful medieval book*] to [Buyer *John*] for [Money *twenty Euros*].

The frames representing buying and selling can be represented in a feature-value matrix as follows. The example in Figure 1 shows the *buy* frame, also illustrating subframes (in the TIME field). An exclamation mark indicates core roles, and the semantics of the expression is described by double brackets. *Point-in-time*, *person*, *money*, *purpose* are type labels constraining potential fillers. We abbreviate the frame, indicating this with the ellipsis dots, thus omitting slots that have not been realised explicitly. By convention, a description in this way is generally partial (and thus ellipsis dots are not really needed).

³ Schank & Abelson (1977) develop the related concept of scripts, which adds a temporal dimension. Our use of frames also resembles scripts if one reads the constituents of certain proof structural frames (as introduced below) as a plan for linear text organisation.

⁴ See, e.g., Ruppenhofer et al. (2006) and the project's website at <https://framenet.icsi.berkeley.edu>. Accessed: 1 December 2025.

$\left[\begin{array}{ll} \text{buy} & \\ \text{BUYER!} & \llbracket \text{John} \rrbracket \\ \text{GOODS!} & \llbracket \text{a beautiful medieval book} \rrbracket \\ \text{TIME} & \llbracket \text{yesterday} \rrbracket \\ \text{SELLER} & \text{person} \\ \text{MONEY} & \text{money} \\ \text{PURPOSE} & \text{purpose} \\ \dots & \end{array} \right]$	$=$	$\left[\begin{array}{ll} \text{buy} & \\ \text{BUYER!} & j \\ \text{GOODS!} & b \\ \text{TIME} & \left[\begin{array}{ll} \text{point-in-time} & \\ \text{YEAR} & 2018 \\ \text{MONTH} & 02 \\ \text{DAY} & 28 \\ \text{HOUR} & \{1, \dots, 24\} \\ \text{MINUTE} & \{0, \dots, 60\} \\ \dots & \end{array} \right] \\ \text{SELLER} & \text{person} \\ \text{MONEY} & \text{money} \\ \text{PURPOSE} & \text{purpose} \\ \dots & \end{array} \right]$
$\left[\begin{array}{ll} \text{sell} & \\ \text{SELLER!} & \llbracket \text{Peter} \rrbracket \\ \text{BUYER!} & \llbracket \text{John} \rrbracket \\ \text{GOODS!} & \llbracket \text{a beautiful medieval book} \rrbracket \\ \text{TIME} & \text{point-in-time} \\ \text{MONEY} & \llbracket \text{twenty Euros} \rrbracket \\ \text{PURPOSE} & \text{purpose} \\ \dots & \end{array} \right]$	$=$	$\left[\begin{array}{ll} \text{sell} & \\ \text{SELLER!} & p \\ \text{BUYER!} & j \\ \text{GOODS!} & b \\ \text{TIME} & \text{point-in-time} \\ \text{MONEY} & 20\text{€} \\ \text{PURPOSE} & \text{purpose} \\ \dots & \end{array} \right]$

Figure 1. A frame representation buying and selling a book as in the sentences of example (1). ! marks core roles

While both sentences evoke similar frames, the SELLER slot need not be filled in the first case. However, it is present and could be filled, just as the TIME slot specifying when the transfer occurs is present (but usually non-core) with most verbs and can be filled with more or less specific values.⁵

The concept of frame has been developed further in linguistic and philosophic projects recently, most notably in Düsseldorf in the SFB 991: *Die Struktur von Repräsentationen in Sprache, Kognition und Wissenschaft* (see, e.g., Gamerschlag et al. 2014; Gamerschlag et al. 2015), exploring how frames are connected to the semantic category of functional concepts (see Löbner 2015), and also investigating the history of scientific language and even connecting to discourse analysis (see e.g. Ziem 2008; Ziem 2014). With respect to the formal representation of frames, Petersen (2015) develops a model using feature structures (closely related to those explored by Carpenter 1992)

⁵ Even core roles can be omitted sometimes, as in *John finally sold his car*. Salience and optionality must hence be considered to be gradual.

and highlights the connection between frames and functional concepts. In the realm of mathematics the frame concept was first used in the context of didactics by Davis (1984). From a constructivist perspective Davis (1984) models the possibly erroneous and over-generalized individual knowledge of learners by frames. In this approach frames are a tool for the modelling of individual conditions for success or difficulties in learning processes. Many of the frames causing errors are overgeneralizations of mathematical knowledge acquired up to a certain stage, as, e.g., a frame of binary operations equipped with features of addition which is transferred to multiplication and leads to invalid calculations like $4 \times 4 = 8$ (Davis 1984, 111–113). In other cases, Davis (1984) argues that explanations for certain mathematical phenomena given by students depend on very basic pre-mathematical frames.

Modelling inductive proofs, Fissen et al. (2019) see frames as guiding the processing of proofs based on usual mathematical practice. Carl et al. (2021) relate the frames approach to concepts of understanding and especially Avigad's (2008) ability-based account of understanding mathematical proofs. Fissen, Sarikaya & Schröder (2023) discuss how the concept of frames can be used to explore and explain the notion of innovation in mathematical practice.

The approach to applying the frame concept to mathematical proofs taken in this paper builds on the three aforementioned papers. It hypothesizes, as stated in the introduction, that (at least) two kinds of frames play a crucial role. Structural frames on the one hand schematically model the structure of proofs and definitions, similar to what Engel (1999) presents as proof techniques. Ontological frames on the other hand model domain knowledge: mathematical structures and typical patterns of expressions referring to them and their elements. Even if this distinction between structural and ontological frames is assumed to be heuristic and depends on the approach and field of the proof, it is useful for the investigation of mathematical texts and therefore will underlie our annotation.

The mathematical subfield and the type of text determine expectations regarding presence and explicitness of proof elements, as attested to by the differences between, e.g., textbooks articles in a mathematical journal. Frames, formalizing these expectations, bridge the gap between the text and more formal representations. Handbook articles on specific areas often present proof techniques; these can be understood as building blocks for frames that can also be used in innovative ways.

2.2 Interaction of frames, introducing the frame of induction

In frame systems, frames can interact slots of one frame can be filled from neighbouring, sub- or superordinate frames. For instance, the form of an induction depends on the un-

derlying inductive type. The general structure in Figure 2⁶ gives the typical parts of the induction proof such as the BASE-CASE, the INDUCTION-STEP and the INDUCTION-VARIABLE. As the form of an induction on some type is also dependent on the structure of said type, the mentioned slots are complemented by those for the INDUCTION-SIGNATURE and the INDUCTION-DOMAIN. The non-recursive BASE-CONSTRUCTORS of the latter provide the cases in the BASE-CASE of the induction, and its RECURSIVE-CONSTRUCTORS derive the ‘successors’ of the base values, thus informing the INDUCTION-STEP.

This explicit frame structure (defined in the appendix of Fisseni et al. 2019) allows multiple BASE-CONSTRUCTORS and STEP-CONSTRUCTORS. Types with one base constructor and one recursive constructor allow for a simpler, more specialized frame.

Structurally, a typical induction on natural numbers will contain one BASE-CASE and one INDUCTION-STEP. Often BASE-CASE concerns the element 0 or 1 and INDUCTION-STEP refers to the successor function, $\lambda x.x + 1$, which given a natural number returns the next one. When an induction is done on all even numbers, the step function can return the next even number, $\lambda x.x + 2$.⁷

An induction on complex formulas contains CASE-PROOFS for each kind of atomic formulas in the PROOF of the BASE-CASE, and one element in the CASE-PROOFS for each connective of the formal language in the PROOF of the INDUCTION-STEP. Ontologically, the form of hypotheses of the BASE-CASE and INDUCTION-STEP are constrained by the BASE-CONSTRUCTORS and RECURSIVE-CONSTRUCTORS: The former by default concerns the value of one of the BASE-CONSTRUCTORS and the latter applications of the RECURSIVE-CONSTRUCTORS.

In the view adopted in Fisseni et al. (2019) and Carl et al. (2021), frames – both ontological and structural ones – have a conceptual and a form dimension. The latter consists in text-structuring elements, notational conventions and linguistic triggers. An example of a linguistic trigger are certain plural constructions like “ L_1 and L_2 are parallel”, which presuppose a symmetric relation (here: *parallel*, meaning that L_1 is parallel to L_2 and – due to the symmetry – also that the inverse holds: L_2 is parallel to L_1) and thus trigger the ontological frame of symmetric relations (see, e.g., Cramer & Schröder 2012).

In this bilateral sign-like conception frames deviate from Fillmore’s original understanding of frames as a mere representation of meaning. We chose this approach because structural frames in mathematics do not just define slots and their semantic relation, but

⁶ The structure is adjusted to multiple BASE-CONSTRUCTORS and STEP-CONSTRUCTORS, taken from the appendix of Fisseni et al. (2019); a simpler version can be defined for types with one base constructor and one recursive constructor.

⁷ Alternative formalizations are possible, see section 6.

also on different levels abstraction influence how proofs are “told”. This comprises, e.g., the choice of variables and notation at lower levels of abstraction, and the conventional order of proof parts at higher levels. The latter reminds of event sequences in narrative texts and is clearly related to aspects of cognitive organization.

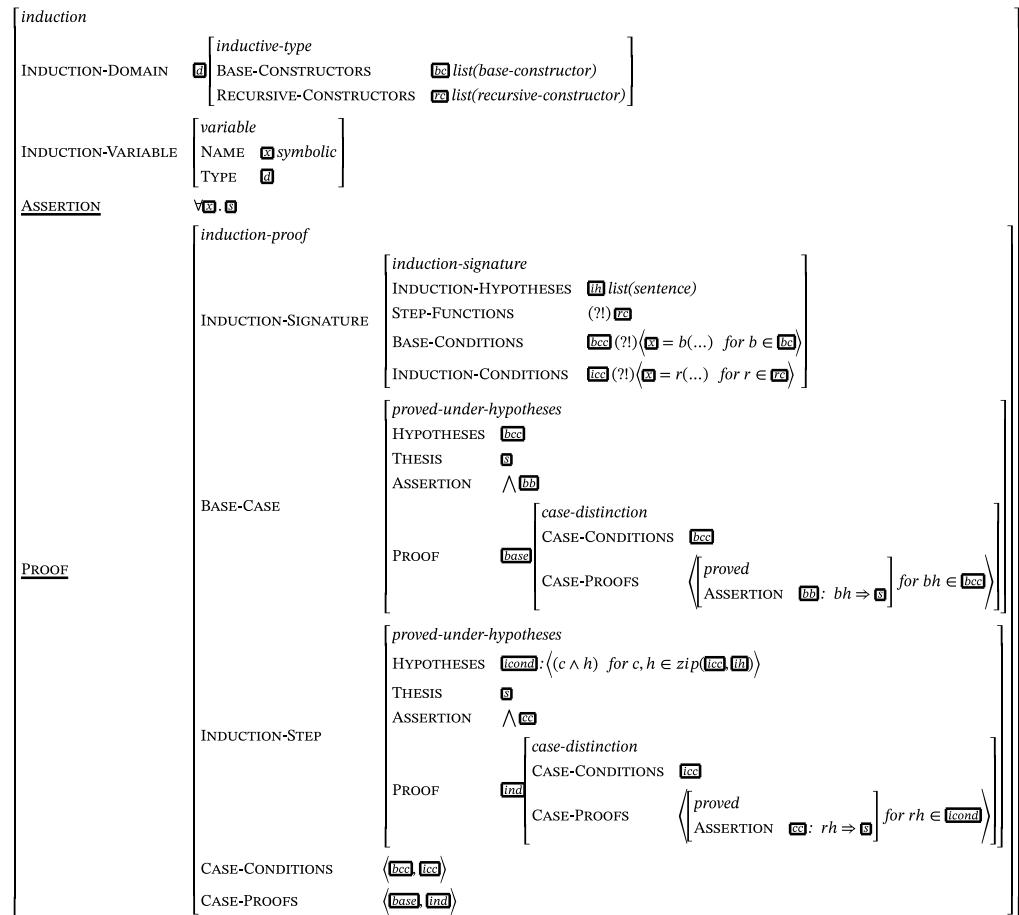


Figure 2. Induction frame. (?) marks default fillers of a slot. The whole representation describes the proof, but INDUCTION-DOMAIN and INDUCTION-VARIABLE are displayed as sub-frames and connected to different places in the PROOF sub-frame by structure sharing (indicated by boxes)

In this respect our frames resemble linguistic constructions in the sense of construction grammar as advocated by Goldberg (2006) and others: Like constructions, frames are organized in type hierarchies of more general and more specialized frames. Similar to the acquisition of linguistic constructions in constructivist language acquisition approaches

(cf. Tomasello 2005), more general frames are usually acquired by generalization from more specific frames. Despite these parallel conceptualizations, frames differ from usual concepts of linguistic constructions as the restrictions on the form side are less strict. Most frames can be evoked by a huge range of formal realizations and hints, such as certain argumentation patterns. The form-concept relation is less transparent and more abstract compared with constructions usually discussed in linguistics.

3 Research questions

In this section, we present some research questions we want to address, and which would profit from a corpus such as the one we are designing.

3.1 Philosophical research questions

Since mathematical texts play a role in what mathematicians do, it can be expected to make a contribution for the study of mathematical practice. This means that we can expect results relevant to the (relatively) newly formed discipline of the philosophy of mathematical practice, which is currently stressing the need to incorporate the work of mathematicians into philosophical considerations on mathematics. Or in the words of a founding member of the Association for Philosophy of Mathematical Practice created in 2009, P. Mancosu:

[A]nyone familiar with contemporary philosophy of mathematics will be aware of the need for new approaches that pay closer attention to mathematical practice. (Mancosu 2008, preface)

While this community is still a minority, it is constantly growing. If we begin to focus more on the practice of working mathematicians, rather than focusing on purely metaphysical questions of the status of numbers, it becomes interesting to use existing frameworks from philosophy of science. We can note here that the frame approach has already been used in philosophy of science in general (Kornmesser 2018). This opens up the possibility to discuss the issues raised in the SFB 991 at the University of Düsseldorf⁸ for the philosophy of mathematics.

Text is a crucial medium for the transfer of mathematical ideas, agendas and results in the scientific community and in the context of education. This makes the focus on mathematical texts a natural and important part of the philosophical study of mathematics. Moreover, it opens up the possibility to apply a huge corpus of knowledge available from the study of texts in other disciplines to problems in the philosophy of mathematics. Big data studies of the corpus of all texts in the ArXiV are reported by Sørensen et al.

⁸ See <https://frames.phil.uni-duesseldorf.de/> and the publications cited in the previous section.

(2024) and Johansen et al. (2022). But these texts are of course not meaningfully annotated, even though there is some structure gained from those articles that also upload their LaTeX code, like the proof environment (`\begin{proof} ... \end{proof}`) often annotating which part of the text corresponds to proofs.

In earlier articles we already sketched parts of the philosophical impact of the frame approach. Carl et al. (2021) argued that many aspects of a prominent operationalization of understanding can be explained in terms of frames, for instance frame-identification as the ability to give a high-level outline. Fisseni, Sarikaya and Schröder (2023) talked about aspects of novelty and the connection of far-apart fields of mathematics. Here frames were used to explain, among others, how different perspectives on an object can make different questions salient and how notational systems and metaphors can motivate new steps that can then later be properly embedded into a theory. We suspect that many philosophical debates might benefit from the concept of frames, including questions on explainability, notations, questions of (national) styles of mathematics. We also sketched some future case studies, like the development of the *forcing* technique in set theory.

Finally, the frame approach may motivate the semantic view of theories. This view states that a theory can be identified with a collection of its models, rather than with a set of its true statements. Some of those models have another status than other, they are intended models. This can be read in straight forward Tarskian style, i.e. we read model as defined in the mathematical discipline of model theory. This relates to ontological frames, as ontological frames do not only capture the definition of an object, but we could develop a routine that ‘guesses’ what would be expected fillers. For instance, when dealing with a topological space, we may assume that in addition this space is likely a Hausdorff space. A reader could simply then fill in a metric space or even \mathbb{R} as their paradigmatic example for a space. This would be wrong, at least in full generality, but it could work in many cases. Anyhow it might simply show that the frames of the author (or a community at large) are non-ideal, and hence do not actually capture the theory fixed by the axioms, for instance when non-topologists most of the time (wrongly) implicitly assume that topological space is a metric space.

This also shows in another notion related to structuralism, namely theory-nets. As frames are organized in an inheritance hierarchy, we can map out whole theories of mathematical structures. This is highly parallel to thoughts of specialization discussed for empirical theories.

Specialization introduces another type of relation among theory-elements to account for the inner structure of theories. A large number of scientific theories, in the ordinary sense of the term, have been reconstructed in the form of a tree-like structure with a basic theory-element at the top and several branches of more special theory-elements.

The underlying idea is that any intended application of any specialized theory-element T is also an intended application of the more basic theory-elements being higher up in the hierarchy. Through specialization, the substantial laws of different theory-elements can be superimposed (Andreas & Zenker 2014).

To take an example take the theory-net of algebraic structures. We do not start with the group axioms per se, but with different models or examples. This mimics how we learn and also the historical developments. We study the integral numbers long before the algebraic structure of a ring and before the group structure of $(\mathbb{Z}, +)$. Frames are learned by examples, and we keep paradigmatic examples in our mind. This semantic view of theories can be associated with structuralist thinking, where a theory is not just seen as a set of truths (or a set of axioms and its deductive closure) but putting models forefront and the relation with each other. So each theory comes with paradigmatic examples. Groups are introduced after and relying on the study of natural numbers, whole numbers and symmetries. Often, we observe additional structure (namely the ring structure), but time shows that there are non-ring groups that are interesting as well. We can even study algebraic structures which are weaker than groups, but those have been less successful, i.e. are not much used, in current practice.

A few further directions we also mentioned as possible future works in the quoted papers are questions of the identity of proofs. It is clearly not necessary for the actual text of two proof texts to be identical. Being subsumed under the same frame and having (in some sense) equivalent fillings of the slots may be a better identity criterion.

Frames can also help to think about the discussion around the derivation-indicator view (see, e.g., Azzouni 2004; Carl & Koepke 2010), as it describes a possible intermediate step. Whether this is really an intermediate step to a derivation is of course up to discussion.

Finally, it is to be stressed that frames might well relate to something cognitively *real* opening a window to many questions of the epistemology of mathematics.

3.2 Linguistic research questions

As mentioned in the introduction, we want to study the acquisition and usage of frames in mathematical texts.

One point of departure is the differentiation between ontological and structural frames sketched in section 2.1. As this distinction is heuristic, we can characterize annotation in the following way: High-level annotation mainly concerns the identification of structural frames, and determining potential triggers (for instance, *by induction*) and fillers of frame slots (for instance, the induction anchor).

Structural frames are high-level structures because these frames generally connect several phrases across the proof, which fill different slots one by one. Ontological frames tend to act as fillers, and their slots are mostly filled immediately in definitions or declarations.

We assume that slots follow certain structural patterns. For instance, a typical induction anchor contains fixing a variable, a typical induction step contains a predication about two discrete values a natural ‘step’ apart; these two patterns also indicate the higher-level frame of induction. We call such structurally marked constructions frame indicators; these give less clear indication of frames, but a combination of indicators with a suitable context may be sufficient to trigger a frame. (Instances of) ontological frames are used/presupposed as fillers and therefore referenced in formulaic or linguistic expressions. Structural frames, on the contrary, are assumed to be expressed across several sentences.

Frames model how explicit and implicit information are combined with background knowledge. This connects well with a typical feature of mathematical proof texts: They often contain many gaps that are left – to quote a classic phase ‘as an exercise to the reader’. Omissions may include non-trivial steps or even the full omission of (large parts of) proofs. The reader is expected to be able to ‘fill in the details’.

Another particularity of mathematical proofs is that they often contain explicit marking, often combined with numbering, for text structure such as definitions, theorems, and proofs, at times even of single formulas. Broader study of these patterns will allow us to shed some light on the question how pervasive these patterns are, and how they interact with the introduction and use of frames.

On the linguistic side of things, the outlined project aims to provide insights into the structuring of texts and the management of discourse referents by frames for mathematical (proof) texts. This goal already started early in the mentioned Naproche project (Schröder & Koepke 2003; Fissen 2003).

Unlike many other text types, the intended semantic meaning of a proof text can often be reconstructed quite unambiguously. This makes this text type an ideal research case for studying the interplay between text/process structuring and ontological frames. Therefore, we claim that this work will be a great test case for Natural Language processing in general. This application scenario allows for studying the feasibility of algorithmically constructing a (logically) largely unambiguous text interpretation (e.g., in the sense of Proof Representation Structures as described in Cramer et al. 2009; Cramer, Koepke & Schröder 2011) using frame-based techniques in a relatively ideal domain.

Basically, our hypothesis is that basic frames are introduced by various explicit indicators. For ontological frames, we expect introduction by definitions and the explicit introduction of notational means. In the case of structural frames, we expect initially the explicit marking of the structuring of the proof and explanations of this structure. It is an empirical question to which extent the order of elements follows a conventional order or is varied in mathematical text.

For frames that are already known to the reader, we want to study how they are signaled. We expect that this is done by a variety of linguistic means, including anaphoric references, conventions of function or variable names and potentially typical triggers or indicators. The corpus will allow us to evaluate the conditions under which these means are used, and whether they depend on context.

4 Corpus annotation: General considerations

The corpus as envisaged in this article is designed to allow linguistic and philosophical investigation of frame usage in mathematical texts.

The corpus will be annotated using standard formats to ensure interoperability and reusability.⁹ In our area, the relevant specifications are the TEI¹⁰ guidelines for text encoding, but also the MMT¹¹ framework for mathematical knowledge representation. We focus on TEI compatibility here (see section 4.2).

To start in as theory-neutral a way as possible, initial annotation will be surface-oriented. This means that we only annotate stretches of text with respect to the role they play in mathematical frames we assume they represent. Specifically, we do not annotate information that is only implicitly given or must be supplemented by semantic inference and type-shifting (see next section for examples).

We thus distinguish the following components of annotation process: Going bottom-up, identifying shallow frame constituents (structural patterns) that, by default, can be viewed as slot fillers (among these, indicators) and frame triggers. Due to the recursivity of frames, however, not all fillers can be shallow constituents, but sometimes even full frames will fill a slot.

Structural frames are not necessarily explicitly introduced by any fixed lexical items. In a proof by induction any overt reference to *induction*, *inductivity* etc. may be missing.

⁹ As is standard by now, software tools for annotation and management of the corpus will be released as open source wherever possible.

¹⁰ Available at: <https://tei-c.org/>. Accessed: 1 December 2025.

¹¹ Available at: <https://uniformal.github.io>. Accessed: 1 December 2025.

In some cases, the relevant frame can only be inferred by the logical properties and the relation of its arguments. In this respect frame annotation may resemble the proposal of a “‘loose’ layer” of inferred frames in addition to a “‘strict’ FrameNet-compatible lexical layer” (Remijnse & Minnema 2020, 13). The essential difference to these layered approaches is that our inferred structural frames do not overlay lexical frames. Contrary to the flat conception of lexical frames structural mathematical frames are recursively nested with a increasing tendency of lexical grounding of frames closer to the leaves of this nested structure.

Determining frames, which is a mixed process of top-down hypotheses and bottom-up pattern recognition (triggers, configurations of indicators), including assigning shallow and recursive constituents to slots. The latter includes identifying those fillers which are not explicit in the text.

In the example proof of Figure 3¹², the fact that it is a proof by induction is mentioned explicitly only in the very last sentence. For the expert reader the proof strategy becomes clear in the course of reading by the type of claim, by the subgoals of the proof, by the demand to repeat the argument, and finally the explicit mention of the trigger word *induction*. So, the structural frame of a proof by induction could assumed as a top-down hypothesis, and its slots can be filled by the respective proof parts.

Proof. First, the second statement is indeed more precise than the first: let $k \geq 1$ be such that $f^k = 0$ but $f^{k-1} \neq 0$; there exists $v \neq 0$ such that $f^{k-1}(v) \neq 0$, and we obtain $k \leq n$ by applying the second result to this vector v . We now prove the second claim. Assume therefore that $v \neq 0$ and that $f^k(v) = 0$ but $f^{k-1}(v) \neq 0$. Let t_0, \dots, t_{k-1} be elements of \mathbf{K} such that

$$t_1 v + \dots + t_{k-1} f^{k-1}[sic!](v) = 0.$$

Apply f^{k-1} to this relation; since $f^k(v) = \dots = f^{2k-2}[sic!](v) = 0$, we get

$$t_1 f^{k-1}(v) = t_1 f^{k-1}(v) + t_2 f^k(v) + \dots + t_{k-1} f^{2k-2}[sic!](v) = 0,$$

and therefore $t_1 f^{k-1}(v) = 0$. Since $f^{k-1}(v)$ was assumed to be non-zero, it follows that $t_1 = 0$. Now repeating this argument, but applying f^{k-2} to the linear relation (and using the fact that $t_1 = 0$), we get $t_2 = 0$.

Then similarly we derive by induction that $t_i = 0$ for all i , proving the linear independence stated.

(Kowalski 2016, 93); in the first equation, the exponent $k - 1$ has to be replaced by $k - 2$; in the line below and the second equation, $2k - 2$ by $2k - 3$.

Figure 3. Example proof without annotation

¹² The annotated version is shown in Figure 4.

[**Tag: Structure** Proof.] First, the second statement is indeed more precise than the first: let $[formula: sentence k \geq 1]$ be such that $[formula: sentence f^k = 0]$ but $[formula: sentence f^{k-1} \neq 0]$; there exists $[formula: sentence v \neq 0]$ such that $[formula: sentence f^{k-1}(v) \neq 0]$, and we obtain $[formula: sentence k \leq n]$ by applying the second result to this vector $[formula: referring v]$. We now prove the second claim. Assume therefore that $[formula: sentence v \neq 0]$ and that $[formula: sentence f^k(v) = 0]$ but $[formula: sentence f^{k-1}(v) \neq 0]$. Let $[formula: referring t_0, \dots, t_{k-1}]$ be elements of $[formula: referring K]$ such that

$$[formula: sentence t_1 v + \dots + t_{k-1} f^{k-1}[sic!](v) = 0.]$$

Frame: Base-Case | Proof Apply $[formula: referring f^{k-1}]$ to this relation; since $[formula: sentence f^k(v) = \dots = f^{2k-2}[sic!](v) = 0]$, we get

$$[formula: sentence t_1 f^{k-1}(v) = t_1 f^{k-1}(v) + t_2 f^k(v) + \dots + t_{k-1} f^{2k-2}[sic!](v) = 0,]$$

and therefore $[formula: sentence t_1 f^{k-1}(v) = 0]$. Since $[formula: sentence f^{k-1}(v) \neq 0]$ was assumed to be non-zero, it follows that $[formula: sentence t_1 = 0]$.

Frame: Induction-Step | Proof Now repeating this argument, but applying $[formula: referring f^{k-2}]$ to the linear relation (and using the fact that $[formula: sentence t_1 = 0]$), we get $[formula: sentence t_2 = 0]$.

Then similarly we derive $[trigger: induction$ by induction] that $[formula: sentence t_i = 0]$ for all $[formula: referring i]$, proving the linear independence stated.

(Kowalski 2016, 93); in the first equation, the exponent $k - 1$ has to be replaced by $k - 2$; in the line below and the second equation, $2k - 2$ by $2k - 3$.

Figure 4. Example proof from Figure 3 with annotations

To lay a good foundation for the high-level annotation, it is desirable to develop a surface-level annotation which provides categories that are candidates for triggers and ontological frames, but also treats mathematical formulas and uses linguistic preprocessing to guide annotators (see next section).

Based on previous work on frames and mathematical language, the following correspondences between linguistic structure and frame structure are expected; in these, we must take into account that mathematical text uses formulas both as referring and as sentence-valued expressions. Therefore, mathematical formulas will be annotated and tentatively tagged for both uses. A deep annotation of formulas is not intended at this stage. To a large extent the syntactic structure of formulas could be parsed automatically, but semantic ambiguities remain to be resolved. $a(b - 1)$ could mean a function a applied to the argument $b - 1$ or the product of a and $b - 1$. For our purposes the annotation of frame indicators (see below) within formulas and of parts of formulas which

are targets of slot references (e.g., “the sum on the left side of the formula”) could be relevant.¹³

First, a structural frame that is explicitly realized must be connected to a span of text, at least several sentences. Its slots can be of different size. Secondly, ontological frames are by default realized as referring expressions, i.e. noun phrases or variables. Of course, other realizations are possible (see Tanswell & Inglis 2022 for a discussion of some cases). Third, triggers (structural and ontological frames) must be connected to individual words or phrases. These, of course can be discontinuous, as German particle verbs (*umformen*, *formen*... *um* ‘transform an equation’) or analytic verb forms (*wird*... *gezeigt* ‘is shown’) which can be extended by adverbs (‘is presently transformed’). Fourth, indicators correspond to propositional units, i.e. at least one sentence or sentence-valued formula. Indicators may be discontinuous, as they may be separated by other indicators or fillers.

4.1 An example proof

The example proof from Kowalski (2016, 93) is shown in two versions: in Figure 3 without annotation, and in Figure 4, we illustrate the frame annotation without showing any XML; the annotation principles are sketched below. Chunking and pos-tagging are also omitted, but formulas are tagged as referring or sentence-valued. “Lexical units that do not play a role as triggers or indicators for the specific frames under discussion (e.g., *let*, *since*, *therefore*), have not been annotated in the example to keep it readable.

While the proof uses induction on natural numbers, it just mentions the induction frame (trigger!) late after giving an example for the step from the base case to the immediate successor. The clearest trigger thus is the phrase “by induction” in the last sentence. But there are several indicators before that like the step from $t_1=0$ to $t_2=0$ in the preceding sentence. That i is to be the INDUCTION-VARIABLE with $i=1$ as the BASE-CONDITION can be inferred from the step from t_1 to t_2 and the universal quantification. Only the ASSERTION is stated explicitly. The whole proof has to be reconstructed from the hint that in the step from $i=1$ to $i=2 f^{k-2}$ has to be applied instead of f^{k-1} , which leads to the general term $f^{k-(i+1)}$ for the step from i to $i+1$.

4.2 Annotation principles

Annotation will be TEI-compatible, using stand-off annotation, i.e. annotations will be linked to the text by pointers. This allows us to add annotation layers that are structurally very different from the original structure.

¹³ A reviewer of this paper asked about annotation of formulas “as they are read”. This would be a kind of “normalization” of mathematical texts, which could lead to lexical targets for FrameNet-frames. As we do not assume that there exists a canonical reading for formulas in general, the definition of special formal triggers and indicators seems more feasible at this stage.

The text will be recorded using standard-TEI segments and specific mathematical annotation, taking special care of numbering and labelling to allow for research on text structure. Mathematical text is usually written in the LaTeX¹⁴ system. As the TEI stylesheets¹⁵ do not allow conversion from (but only to) LaTeX format, we will have to build our own conversion, building on Pandoc's¹⁶ TEI Simple conversion.

The following annotation will be in-line, as we consider it part of the text, and no structural divergence can arise: Text structure will be annotated using standard TEI elements, formulas will be annotated as such, and there will be linguistic annotation on the level of chunking.

Stand-off annotation will contain the following layers: First, linguistic annotation on the word level. To achieve optimal alignment between linguistic structure and annotation, we will annotate the text linguistically using chunking and POS-tagging (for the rationale, see below). Secondly, formula annotation. Formulas will heuristically be classified as sentence-valued or referring expressions (see below, section 4.3). Thirdly, frame annotation. Based semantically on typed feature structures (see, e.g., Carpenter 1992), this annotation layer will be implemented using TEI feature structures to the extent possible.

Coreference will not be annotated in the first phase, as this task can be complicated in cases such as plural entities (e.g., y and $f(x)$). This layer of annotation may be added in a later phase, and may be necessary to resolve certain slot fillers.

4.3 Before manual annotation

The following automatic linguistic preprocessing will be applied and provides goals for pointers of stand-off annotation. In cases linguistic preprocessing leads to inadequate results, it will have to be corrected in manual annotation.

First, sentences will be annotated with parts of speech (POS) to help identify referring expressions and hence reference to ontological frames. Also, trigger candidates (e.g., *by induction*, *by contradiction*) shall be marked. As such expressions may be formulaic, formulas and mixed expressions must be annotated and heuristically classified as sentence-valued (e.g., equations as in *hence, $a+b=0$*) or referring expressions (e.g., simple variables or appositions like *the equation $a+b=0$*). Interpretation of formulas is outside the scope of this step; it will be taken up later in frame analysis.

¹⁴ Available at: <https://www.latex-project.org/>. Accessed: 1 December 2025.

¹⁵ Available at: <https://github.com/TEIC/Stylesheets>. Accessed: 1 December 2025.

¹⁶ Available at: <https://pandoc.org/>. Accessed: 1 December 2025.

5 Annotation guidelines

Writing annotation guidelines is not trivial. In our case, the task is complicated by the fact that the goal structure is not necessarily close to the surface structure of the text, but is rather oriented towards a higher-level representation of proofs.

Lemnitzer & Zinsmeister (2015) list the following criteria for good annotation guidelines:

- a) a list of all tag names (mostly descriptive abbreviations) together with their full names (category names),
- b) definitions of the categories,
- c) prototypical annotation examples for the categories,
- d) tests that help to decide whether a category applies,
- e) problematic examples with [correct] annotations,
- f) typical confusion categories (i.e. competing tags) with examples. (Lemnitzer & Zinsmeister 2015, 102, our translation)

We will strive to meet these criteria. For reasons of space and avoidance of boredom, we cannot provide a full set of annotation guidelines here. Instead, we will discuss some of the challenges that arise in writing such guidelines and sketch our approach to these challenges. We will not give examples of tests or confusion categories here yet, but we will discuss some problematic examples.

As explained earlier, the resulting frame structure will (by default, but not necessarily) be much more complete than the text – filling the gaps –, so that both sub-frame structure and fillers may have to be inferred. This is different from, for instance, the linguistic annotation in FrameNet, which focusses on non-recursive structures with the verb as the base (and explicit trigger) of the frame, while our frames are recursive and need not have a unique and unequivocal trigger: an induction need not contain any explicit reference to induction.

5.1 Annotators

Annotators must be able to build up the frame structure for themselves. As this is a hermeneutic process guided partly by the proof text, but also by mathematical culture, annotators must be educated enough in mathematics to understand the text and supplement the missing parts. These aspects of mathematical practice can be imparted in training workshops or self-paced learning.

It is questionable whether it is possible to find annotators with advanced competence. Hence, annotators will be instructed to stick to the units determined by preprocessing if possible.

5.2 How to start manual annotation

In the beginning, annotators will be presented with the text of the proof. Annotation spans will be aligned with certain linguistic categories.

Based on this, annotators will mark up triggers and indicators as well as (corresponding) structural frames, and will build up the frame structure by assigning fillers to slots. The frame structure will be built up according to the structural frames, referencing triggers and indicators. Ontological frames will be marked up as they are encountered, but only those structures that are explicitly realized in the text will be annotated.

5.3 The frame library and the challenge of new frames

The annotation guidelines will contain a description of frames. The frame library will not be expanded during the project to achieve consistency. If frames are missing, annotators will be instructed to mark these cases and skip annotation. Later, the respective proofs will be reinspected, and the frame library will be revised and expanded if necessary. This may lead to re-annotation of other proofs, as well.

An excerpt of the inheritance hierarchy of frames in the *proved* family is shown in Figure 5.

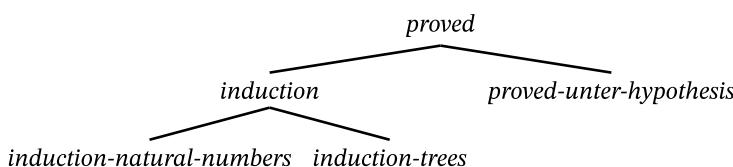


Figure 5. Inheritance hierarchy of frames in the *proved* family

5.4 Challenge avoided: Bridging structural and logical gaps

Logical gaps can occur on different levels, and are often covered by the classical statements *left to the reader, is obviously true* etc. Filling these gaps corresponds to completing the goal frame; however, there may be more than one way to do it.

Structural gaps occur in two cases. First, slots or sub-frames may be only partially realized, i.e. only a slot of a sub-frame is realized. This can be annotated by linking the realization to its slot, but it begs the question how to resolve sub-frame types etc. Secondly, slot fillers may differ in abstraction from the slot value they must provide. An example is the step function. We again distinguish two cases. The step statement is often given by an instance of its application, i.e. in the induction step; alternatively, the step state-

ment may be expressed by referring in a very abstract way (*the next element*, or similar). In either case, reconstructing the step function involves fine-grained interpretation and semantic typing.

How to bridge such gaps can be considered an empirical question. We have therefore decided not to annotate such gaps in the first phase. How to fill them best will be one of the first uses of the corpus.

5.5 Challenge avoided: Variable use

Certain values such as induction variables occur many times in a proof text. Annotating all instances manually is tedious but annotating them automatically may lead to incorrect results.

Complications arise from using variables of the same name in different scopes of the same text. For instance, a variable may be bound to a value in the induction anchor, but then used as a free variable in the induction step. Similarly, a variable may be fixed in a hypothesis, but then used as unbound in the conclusion.

We have therefore decided not to annotate variable use for now. Investigating how to best deal with variable use will be one of the first uses of the corpus.

5.6 Challenge: Resolving typing ambiguities

Annotation guidelines must give indications on how to deal with ambiguities. We now discuss examples of ambiguities, also sketching our approach to avoiding them or resolving them in context.

In annotating induction proofs, ambiguities arise between the step, the domain of induction and the function. This is based on an inner-theoretical choice: If the domain is constructed in a way that its constructor corresponds to the step function, one gets a proliferation of domains, i.e., induction on odd or even numbers, on numbers divisible by ten etc. If, on the other hand, one takes a number of basic types and treats the function as independent of the constructor, this results in many step functions. Logically, both approaches are equivalent, but they lead to different annotation strategies.

We have decided to follow the surface structure of the proof as much as possible, i.e., when an induction is performed on a certain type, we annotate this type as the domain of induction, and the step function is inferred from the text of the induction step. In case the proof is so sketchy that we cannot decide which approach is used, we follow the second approach, i.e., use a small number of basic types and infer the step function.

6 Evaluation

Whether annotation guidelines have the desired effect of leading to correct annotations and as few deviations as possible between different annotators should be evaluated in any corpus building project.

The evaluation can be based on metrics for inter-annotator comparisons. However, this cannot be done on the basis of kappa values such as Cohen's kappa or Fleiss' kappa alone, since it is not only a matter of categorizing data by assigning it to frames and slots, but also of annotating the textual extension on the form side of the frames and the relational assignment of slots to frames.

Since the requirements for a comparison of frame annotations are similar to those for the evaluation of parser output, the metrics of parser evaluation will be discussed here. We distinguish the mere comparison of two or more annotations from the evaluation relative to a gold standard as a reference annotation. Parser evaluation is usually done against a gold-standard. For our scenario we consider the greatest possible inter-annotator agreement based on the annotation guidelines to be developed and refined in an iterative process as the goal of the first step of the evaluation. If a high degree of agreement is attained, a gold-standard could be devised on this basis.

In our corpus annotation we will limit ourselves to structural frames and a flat annotation of ontological types at the structural leaves and annotation of frame triggers, as elaborated in the two preceding sections. This leads to non-cyclic tree graphs with possibly crossing edges.

If we disregard evaluation metrics which take specific error types of syntactic parsing into account, we essentially have four types of metrics (cf. Romanyshyn 2021).

In **leaf-ancestor evaluation** (cf. Sampson & Babarczy 2003), the paths from the root node to the leaves in the parse tree are compared for each word. The distances are measured by the minimum edit distance relative to the number of nodes in the parse tree of the gold standard. In this metrics errors near the tree root have a great impact because they affect more paths than errors lower in the tree.

Cross-bracketing (as e.g. defined in Caroll, Briscoe & Sanfilippo 1998) counts how many text parts are subsumed (bracketed) differently under frame slots by determining how many beginnings and ends of text parts do not match for the annotations to be compared. So, it can be used to determine whether the text parts that can be assigned to slots and frames match in different annotations. It is therefore a measure mostly for quality of the leaf-related parts of the annotation.

While the leaf-ancestor evaluation only determines the minimum editing distance for the paths from the terminal nodes to the root nodes, the minimum **tree edit distance** (TED)¹⁷ measures how many editing steps (adding, deleting and renaming nodes) are required to transform the given tree into the gold standard. Variants arise depending on the weighting of different editing operations. In the case of strongly divergent frame analyses, however, it can be difficult to find suitable editing strategies that lead to a minimization of the editing steps.

The **ParsEval measures** determine the number of constituents that have the same extension in the parse tree as those in the gold standard.¹⁸ The precision is defined as the ratio of the correctly determined constituents to the constituents in the parser output, the recall is the number of correctly determined constituents in the parser output in relation to the number of constituents in the gold standard. An F-score can be calculated on the basis of precision and recall as the harmonic mean of both values which could be weighted in favour of one or the other value. This measure only provides information about the extent of the text parts corresponding to the nodes in the parse tree and the number of nodes between root and leaves.

The classification of the constituents is not taken into account in the ParsEval base measure. In labelled precision and labelled recall, however, it is possible to determine how many matches there are in text extension and label.

In a more refined metrics of ParsEval, the errors in nodes can also be weighted depending on how far they are from the leaves. This is based on the assumption that faults closer to the root are more serious than faults closer to the leaves. In the evaluation of frames other forms of weighting could be taken into consideration.

In section 2, we mentioned the similarities and differences between our conception of structural frames and linguistic constructions. While both share their bilateral structure of form and meaning, we conceive frames as much less bound to the surface structure of the text. The textual borders of certain frames and their fillers may be more disputable than in the case of grammatical constructions. Therefore, essentially text extension-based metrics as cross-bracketing and ParsEval seem less suited for our purpose than metrics highlighting the tree structure as leaf-ancestor evaluation and TED.

Any metrics taking into consideration the node labels of the tree, i.e. frame types and slots, should regard that the types are differently distant from each other. A frame subtype is to be treated differently from a completely contradictory type assignment. Hier-

¹⁷ For a survey of problems related to computing the TED see Bille (2005).

¹⁸ For the original introduction of the ParsEval concept see Black et al. (1991).

archy-based measures for frame type distances could be a suitable approximation for the assessment of the quality of frame type matching. So, recognizing a proof correctly as an induction proof but being mistaken in the type of induction (e. g. *induction-natural-numbers* instead of *induction-trees* in the hierarchy fragment of figure 2 would nevertheless allow for a correct detection of slot fillers and would not make the annotation useless as a whole.

But it seems well motivated that any metrics should take the root distance into consideration. Mistakes higher up in the frame structure tree, i.e., principal misinterpretations of the proof structure, have more serious effects on the whole annotation than mistakes closer to leaves.

7 Conclusion

We have shown why the systematic creation of a corpus of mathematical texts annotated with structural and ontological frames can contribute to several research goals in linguistics, philosophy of mathematics and natural language processing and highlighted the challenges of designing annotation for this corpus.

8 Outlook

The corpus will allow us to study the linguistic means by which frames are introduced and signalled. We sketched the hypotheses above, and expect that using the corpus, we will be able to test and refine these hypotheses.

In section 5, we discussed the challenges of writing annotation guidelines for frame annotation, and we had to postpone some important aspects such as the filling-in of proof gaps and the study of ambiguities. The corpus will allow us to study these aspects in a broader context and refine the annotation and the guidelines for taking a next step in the hermeneutic spiral.

References

- Andreas, Holger & Frank Zenker. 2014. Basic concepts of structuralism. *Erkenntnis* 79 (S8), 1367–1372. <https://doi.org/10.1007/s10670-013-9572-y>
- Artstein, Ron & Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 (4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Avigad, Jeremy. 2008. Understanding Proofs. *The Philosophy of Mathematical Practice*. Paolo Mancosu, ed. Oxford: Oxford University Press. 317–353. <https://doi.org/10.1093/acprof:oso/9780199296453.003.0013>

- Azzouni, Jody. 2004. The derivation-indicator view of mathematical practice. *Philosophia Mathematica* 12 (2), 81–106. <https://doi.org/10.1093/philmat/12.2.81>
- Bille, Philip. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science* 337 (1–3), 217–239. <https://doi.org/10.1016/j.tcs.2004.12.030>
- Black, E., S. Abney, D. Flickenger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, et al. 1991. A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19–22, 1991*. Patti Price, ed. San Mateo, California: Morgan Kaufmann. Available at: <https://aclanthology.org/H91-1060/>. Accessed: 8 December 2025.
- Carl, Merlin, Marcos Cramer, Bernhard Fisseni, Deniz Sarikaya & Bernhard Schröder. 2021. How to Frame Understanding in Mathematics: A Case Study Using Extremal Proofs. *Axiomathes* 31 (5), 649–676. <https://doi.org/10.1007/s10516-021-09552-9>
- Carl, Merlin & Peter Koepke. 2010. Interpreting Naproche – an algorithmic approach to the derivation-indicator view. *Proceedings of the International Symposium on Mathematical Practice and Cognition*, 7–10.
- Caroll, John, Ted Briscoe & Antonio Sanfilippo. 1998. Parser evaluation: A survey and a new proposal. *International Conference on Language Resources and Evaluation*. European Language Resources Association. 447–454. Available at: <https://api.semanticscholar.org/CorpusID:7042755>. Accessed: 8 December 2025.
- Carpenter, Bob. 1992. *The Logic of Typed Feature Structures*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511530098>
- Cramer, Marcos, Bernhard Fisseni, Peter Koepke, Daniel Kühlwein, Bernhard Schröder & Jip Veldman. 2009. The Naproche project: Controlled natural language proof checking of mathematical texts. *Controlled Natural Language*. Norbert E. Fuchs, ed. Berlin & Heidelberg: Springer. 170–186. https://doi.org/10.1007/978-3-642-14418-9_11
- Cramer, Marcos, Peter Koepke & Bernhard Schröder. 2011. Parsing and Disambiguation of Symbolic Mathematics in the Naproche System. *Intelligent Computer Mathematics*. James H. Davenport, William M. Farmer, Josef Urban & Florian Rabe, eds. Berlin, Heidelberg: Springer. 180–195. https://doi.org/10.1007/978-3-642-22673-1_13
- Cramer, Marcos & Bernhard Schröder. 2012. Interpreting Plurals in the Naproche CNL. *Controlled Natural Language*. Michael Rosner & Norbert E. Fuchs, eds. Berlin, Heidelberg: Springer. 43–52. https://doi.org/10.1007/978-3-642-31175-8_3
- Davis, Robert B. 1984. *Learning mathematics. The Cognitive Science Approach to Mathematics Education*. Norwood, New Jersey: Ablex Publishing Corporation.
- Engel, Arthur. 1999. *Problem-solving strategies*. New York: Springer.

- Fisseni, Bernhard. 2003. *Die Entwicklung einer Annotationssprache für natürlichsprachlich formulierte mathematische Beweise*. Bonn: Rheinische Friedrich-Wilhelms-Universität, Philosophische Fakultät Magisterarbeit. https://naproche-net.github.io/downloads/2003-Magister_Fisseni.pdf. Accessed: 8 December 2025.
- Fisseni, Bernhard, Deniz Sarikaya, Martin Schmitt & Bernhard Schröder. 2019. How to Frame a Mathematician. Modelling the Cognitive Background of Proofs. *Reflections on the Foundations of Mathematics: Univalent Foundations, Set Theory and General Thoughts*. Stefania Centrone, Deborah Kant & Deniz Sarikaya, eds. Cham: Springer International Publishing. 417–436. https://doi.org/10.1007/978-3-030-15655-8_19
- Fisseni, Bernhard, Deniz Sarikaya & Bernhard Schröder. 2023. How to frame innovation in mathematics. *Synthese* 202 (4). <https://doi.org/10.1007/s11229-023-04310-3>
- Gamerschlag, Thomas, Doris Gerland, Rainer Osswald & Wiebke Petersen, eds. 2014. *Frames and Concept Types*. Heidelberg: Springer. <https://doi.org/10.1007/978-3-319-01541-5>
- Gamerschlag, Thomas, Doris Gerland, Rainer Osswald & Wiebke Petersen, eds. 2015. *Meaning, Frames, and Conceptual Representation*. Düsseldorf: Düsseldorf University Press. <https://doi.org/10.1515/9783110720129>
- Goldberg, Adele. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199268511.001.0001>
- Hovy, Eduard. 2015. Corpus Annotation. *Oxford Handbook of Computational Linguistics*. Ruslan Mitkov, ed. 2nd edn. <https://doi.org/10.1093/oxfordhb/9780199573691.013.011>
- Ide, Nancy & James Pustejovsky, eds. 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-024-0881-2>
- Johansen, Mikkel Willum & Josefine Lomholt Pallavicini. 2022. Entering the valley of formalism: Trends and changes in mathematicians' publication practice – 1885 to 2015. *Synthese* 200 (3), 239. <https://doi.org/10.1007/s11229-022-03741-8>
- Kornmesser, Stephan. 2018. Frames and concepts in the philosophy of science. *European Journal for Philosophy of Science* 8 (2), 225–251. <https://doi.org/10.1007/s13194-017-0183-3>
- Kowalski, Emmanuel. 2016. *Linear Algebra*. Lecture Notes, ETH Zürich. Available at: <https://people.math.ethz.ch/~kowalski/script-la.pdf>. Accessed: 8 December 2025.
- Lemnitzer, Lothar & Heike Zinsmeister. 2015. *Korpuslinguistik*. 3rd edn. Tübingen: Narr.
- Löbner, Sebastian. 2015. Functional concepts and frames. *Meaning, Frames, and Conceptual Representation. Studies in language and cognition*. Thomas Gamerschlag, Doris Gerland, Rainer Osswald & Wiebke Petersen, eds. Düsseldorf: Düsseldorf University Press. 15–42. <https://doi.org/10.1515/9783110720129>

- Mancosu, Paolo, ed. 2008. *The Philosophy of Mathematical Practice*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199296453.001.0001>
- Minsky, Marvin. 1974. *A Framework for Representing Knowledge*. MIT AI Laboratory Memo. Cambridge, MA, USA: Massachusetts Institute of Technology/MIT.
- Petersen, Wiebke. 2015. Representation of concepts as frames. *Meaning, Frames, and Conceptual Representation*. Vol. 2. Thomas Gamerschlag, Doris Gerland, Rainer Osswald & Wiebke Petersen, eds. Düsseldorf: Düsseldorf University Press. 43–67. <https://doi.org/10.1515/9783110720129>
- Remijnse, Levi & Gosse Minnema. 2020. Towards Reference-Aware FrameNet annotation. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*. Tiago T. Torrent, Collin F. Baker, Oliver Czulo, Kyoko Ohara & Miriam R. L. Petrucci, eds. Marseille, France: European Language Resources Association. 13–22. Available at: <https://aclanthology.org/2020.framenet-1.3/>. Accessed: 8 December 2025.
- Roberts, R. B. & Ira P. Goldstein. 1977. *The FRL Manual*. Cambridge, MA, USA: Massachusetts Institute of Technology/MIT.
- Romanyshyn, Mariana. 2021. The Dirty Little Secret of Constituency Parser Evaluation. Available at: <https://www.grammarly.com/blog/engineering/the-dirty-little-secret-of-constituency-parser-evaluation/>. Accessed: 8 December 2025.
- Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petrucci, Christopher R. Johnson & Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute.
- Sampson, Geoffrey & Anna Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Journal of Natural Language Engineering* 9 (4), 365–380. <https://doi.org/10.1017/S1351324903003243>
- Schank, Roger C. & Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: L. Erlbaum.
- Schröder, Bernhard & Peter Koepke. 2003. ProofML – eine Annotationssprache für natürliche Beweise. *Sprachtechnologie für die multilinguale Kommunikation – Textproduktion, Recherche, Übersetzung, Lokalisierung – Beiträge der GLDV-Frühjahrstagung 2003*. Uta Seewald-Heeg, ed. St. Augustin: Gardez! 428–441. <https://doi.org/10.21248/jlcl.18.2003.48>
- Sørensen, Henrik Kragh, Sophie Kjeldbjerg Mathiasen & Mikkel Willum Johansen. 2024. What is an experiment in mathematical practice? New evidence from mining the Mathematical Reviews. *Synthese* 203 (2), 1–21. <https://doi.org/10.1007/s11229-023-04475-x>
- Tanswell, Fenner Stanley & Matthew Inglis. 2022. The Language of Proofs: A Philosophical Corpus Linguistics Study of Instructions and Imperatives in Mathematical Texts. *Handbook of the History and Philosophy of Mathematical Practice*. Bharath Sriraman, ed. Cham: Springer International Publishing. 1–28. https://doi.org/10.1007/978-3-030-19071-2_50-1

- Tomasello, Michael. 2005. *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge, MA, USA: Harvard University Press.
- Ziem, Alexander. 2008. Frame-Semantik und Diskursanalyse – Skizze einer kognitions-wissenschaftlich inspirierten Methode zur Analyse gesellschaftlichen Wissens. *Methoden der Diskurslinguistik*. Ingo H. Warnke & Jürgen Spitzmüller, eds. Berlin & Boston: De Gruyter. 89–116. <https://doi.org/10.1515/9783110209372.2.89>
- Ziem, Alexander. 2014. Von der Kasusgrammatik zum FrameNet. *Grammatik als Netzwerk von Konstruktionen. Sprachwissen im Fokus der Konstruktionsgrammatik*. Alexander Lasch & Alexander Ziem, eds. Berlin: De Gruyter. 261–290. <https://doi.org/10.1515/9783110353693.263>

KI als Hilfsmittel für die Formulierung von Suchanfragen in Korpora

Skaistė Volungevičienė

Lehrstuhl für Deutsche Philologie

Institut für Sprachen und Kulturen des Ostseeraums

Philologische Fakultät

Universität Vilnius

Universiteto g. 5

LT-01131 Vilnius, Litauen

E-Mail: skaiste.zabarauskaite@flf.vu.lt

ORCID iD: <https://orcid.org/0009-0001-6206-1541>

<https://ror.org/03nad84>

Maximilian Arndt

Institut für Empirische Sprachwissenschaft

Johann-Wolfgang-Goethe-Universität Frankfurt am Main

Rostocker Straße 2

D-60629 Frankfurt am Main, Deutschland

E-Mail: maximilianarndt1998@gmail.com

<https://ror.org/04cvxnb49>

Abstract. Die Bedeutung von Korpusarbeit und korpuslinguistischen Methoden nimmt in der linguistischen Forschung und im DaF-Unterricht stetig zu, da sie eine umfassende und systematische Analyse sprachlicher Phänomene ermöglichen. Häufig arbeiten Forscherinnen und Forscher dabei mit einer Vielzahl von Korpora, die unterschiedliche Datensätze und Strukturen aufweisen. Eine zentrale Herausforderung besteht in der Vielfalt der jeweils eingesetzten Abfragesysteme. Diese sind oft komplex und heterogen, sodass ein erheblicher Zeitaufwand erforderlich ist, um ihre Funktionsweise zu verstehen und sie effektiv zu nutzen. Die damit verbundene Vielfalt kann die Suche nach spezifischen sprachlichen Mustern oder Phänomenen sowohl für erfahrene Nutzerinnen und Nutzer als auch für Anfängerinnen und Anfänger erschweren und frustrierend machen, da um-

Submitted: 07/11/2025. **Accepted:** 19/12/2025

Copyright © 2025 Skaistė Volungevičienė, Maximilian Arndt. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

fangreiche Einarbeitungszeiten in die jeweiligen Abfragesprachen notwendig sind. An- gesichts dieser Herausforderungen kann die Integration von KI als Werkzeug zur Formu- lierung von Suchanfragen in Korpora künftig eine entscheidende Rolle spielen.

Vor diesem Hintergrund untersucht der vorliegende empirische Beitrag das Potenzial KI-gestützter Systeme zur Formulierung Suchanfragen in COSMAS II. Grundlage der Untersuchung bilden 50 typische korpuslinguistische Suchaufgaben unterschiedlicher Art, die von unerfahrenen Nutzerinnen und Nutzern ohne computerlinguistische Vor- kenntnisse formuliert werden. Die KI-generierten Abfragen werden hinsichtlich ihrer formalen Korrektheit, Präzision und Funktionalität analysiert.

Methodisch basiert die Studie auf einer qualitativen Auswertung der Abfrageergebnisse. Die Ergebnisse zeigen, unter welchen Bedingungen KI als unterstützendes Werkzeug bei der Korpusabfrage eingesetzt werden kann und wo ihre Grenzen liegen. Darüber hinaus wird diskutiert, welche Informationen der KI zur Verfügung stehen müssen, um korpus- linguistische Suchprozesse effektiv zu unterstützen.

Schlüsselwörter: KI, Formulierung von Korpusabfragen, Korpora, korpuslinguistische Methoden, DaF, COSMAS II

Artificial intelligence as a support tool for corpus query formulation

Abstract. The importance of corpus work and corpus-linguistic methods is steadily increasing in linguistic research and in the field of German as a Foreign Language (GFL), as they enable a comprehensive and systematic analysis of linguistic phenomena. Researchers often work with a large number of corpora that differ in their datasets and structures. One challenge in this work, however, is the diversity of query systems used across individual corpora. These systems can be complex and heterogeneous, which often requires a considerable amount of time to understand their functionality and use them effectively. This diversity can make the search for specific linguistic patterns or phenomena difficult and frustrating for both experienced users and beginners, as a great deal of time must be spent familiarizing oneself with the particularities of individual query systems. In response to these challenges, the integration of AI as a tool for formulating search queries in corpora may play a decisive role in the future.

In this context, the present empirical contribution examines the potential of AI-support- ed systems for formulating complex search queries in COSMAS II. The study draws on 50 typical corpus-linguistic search tasks of different types, formulated by inexperienced users without prior knowledge of computational linguistics. The AI-generated queries are analyzed with regard to their formal correctness, precision, and functionality.

Methodologically, the study is based on a qualitative analysis of the query results. The results indicate the conditions under which AI can be used as a supportive tool in corpus querying and where its limitations lie. The contribution also discusses what information needs to be available to the AI in order to effectively support corpus-linguistic search processes.

Keywords: AI, corpus query formulation, corpora, corpus-linguistic methods, GFL, COSMAS II

1 Einleitendes

Die fortschreitende Digitalisierung prägt zunehmend alle Bereiche der Wissenschaft und verändert auch die Methoden und Fragestellungen innerhalb der Sprachwissenschaft grundlegend (McShane & Nirenburg 2021, 7). Insbesondere die computergestützte Korpuslinguistik hat sich zu einem zentralen Forschungsfeld entwickelt, da sie es ermöglicht, Sprache auf der Grundlage umfangreicher, authentischer Datensammlungen empirisch zu untersuchen. Korpora stellen damit eine essentielle Ressource für die Analyse sprachlicher Strukturen, Entwicklungen und Gebrauchsweisen dar.

Eine der größten Herausforderungen in der praktischen Arbeit mit Korpora besteht jedoch darin, adäquate Forschungsanfragen zu formulieren, um spezifische sprachliche Phänomene gezielt zu untersuchen. Das Erstellen solcher Korpusanfragen erfordert nicht nur Kenntnisse in Logik und Linguistik, sondern auch ein Verständnis der technischen Anfragesprachen, die von verschiedenen Korpusplattformen verwendet werden (Hunton 2006, 241). Diese Anforderungen erschweren insbesondere Einsteigerinnen und Einsteigern den Zugang zu den Daten und können den Lernprozess erheblich verlangsamen.

Vor diesem Hintergrund gewinnt die Frage an Bedeutung, inwiefern aktuelle Entwicklungen im Bereich der KI den Umgang mit Korpora vereinfachen und zugänglicher gestalten können. *Large Language Models* (LLMs) wie ChatGPT bieten durch ihre dialogische Struktur und ihr Verständnis natürlicher Sprache das Potenzial, niedrigschwellige Unterstützung bei der Formulierung und Interpretation von Korpusanfragen zu leisten. Ziel dieses Beitrags ist es daher, zu untersuchen, wie solche Modelle eingesetzt werden können, um Barrieren bei der Arbeit mit Korpusdaten zu verringern und den Zugang zu sprachwissenschaftlichen Ressourcen zu erleichtern.

Im Zentrum der Untersuchung steht das Korpusverwaltungssystem COSMAS II (Leibniz-Institut für Deutsche Sprache (IDS), *Corpus Search, Management and Analysis System*), das eine zentrale Rolle in der Korpusforschung im deutschsprachigen Raum einnimmt. Die Plattform bietet Zugang zu einer Vielzahl synchroner und diachroner Korpora und ermöglicht so die Analyse der deutschen Sprache in unterschiedlichen

historischen und gegenwärtigen Kontexten. Trotz dieser umfangreichen Funktionalität stellt die Komplexität der zugrunde liegenden Abfragesyntax und der begleitenden Dokumentation für viele Nutzende eine erhebliche Hürde dar. Die verfügbaren Anleitungen umfassen zahlreiche Webseiten mit detaillierten Erläuterungen und Beispielen¹, was insbesondere für unerfahrene Nutzende eine kognitive Überforderung darstellen kann, zumal die Materialien nicht immer in der Muttersprache der Anwenderinnen und Anwender verfasst sind. Der effiziente Umgang mit COSMAS II setzt daher nicht nur Zeit und Geduld, sondern auch eine ausgeprägte Fähigkeit zum logischen und strukturellen Denken voraus.

Aus diesen Beobachtungen ergibt sich ein klarer Forschungsbedarf: Es gilt zu klären, inwiefern KI-gestützte Systeme dazu beitragen können, die Nutzung von COSMAS II zu erleichtern, indem sie komplexe Strukturen verständlicher machen und die Formulierung linguistisch präziser Abfragen unterstützen.

Dieser Beitrag ist das Ergebnis eines zweimonatigen wissenschaftlichen Praktikums², das im Sommer 2024 vom Litauischen Wissenschaftsrat (*Lietuvos mokslo taryba*) unterstützt und an der Universität Vilnius erfolgreich durchgeführt wurde.

2 Kurzer Überblick über den aktuellen Forschungsstand

Die aktuelle Forschung betont zunehmend die Bedeutung digitaler Kompetenz als zentrale Voraussetzung für den erfolgreichen Einsatz von Künstlicher Intelligenz (KI) im Sprachunterricht und Sprachlernen. Mit der jüngsten Verbreitung benutzerfreundlicher generativer KI-Werkzeuge (z. B. ChatGPT) hat der Einsatz von KI und insbesondere Generativer KI (GenAI) im Bildungsbereich erheblich an Bedeutung gewonnen (vgl. Curry & McEnery 2025, 241). Unter digitaler Kompetenz wird dabei nicht nur die Fähigkeit verstanden, digitale Technologien zu nutzen, sondern auch, sie kritisch zu bewerten, reflektiert einzusetzen und gegebenenfalls selbst zu gestalten. Diese Kompetenz umfasst verschiedene Dimensionen, die in der Forschung als Kernbereiche beschrieben werden: das Erkennen und Verstehen digitaler Technologien, die Anwendung und Bewertung ihrer Funktionen, die ethisch verantwortungsvolle Navigation in digitalen Umgebungen sowie die Fähigkeit, selbst digitale Inhalte zu erstellen. Ein hohes Maß an digitaler Kompetenz gilt als Voraussetzung dafür, dass Lehrkräfte und Lernende KI-basierte Werkzeuge effektiv, sicher und didaktisch sinnvoll in den Unterricht integrieren können.

¹ Online Hilfe zu COSMAS II Suchanfragen ist unter <https://www2.ids-mannheim.de/cosmas2/win-app/hilfe/suchanfrage/> zu finden. Man kann eines der acht möglichen Themen auswählen: grafische Eingabe, Zeileneingabe, Morph-Assistent, Auswahllisten, Wortformlisten, Statistik, Sonderzeichen und reguläre Ausdrücke.

² Informationen zu finanzierten Sommerpraktika sind unter <https://lmt.lrv.lt/lt/veiklosritys/mokslo-finansavimas/karjeros-mobilumo-sklaidai-skatinimo-priemones/studentu-tyrimai-vasaros-metu/> zu finden.

Mehrere Studien zeigen, dass Lehrkräfte durch den gezielten Ausbau digitaler und technologischer Fähigkeiten ihre Unterrichtspraxis verbessern und Lernprozesse stärker personalisieren können. Dadurch lassen sich Lernumgebungen schaffen, die sowohl motivierend als auch adaptiv auf die individuellen Bedürfnisse der Lernenden eingehen (Kohnke et al. 2023; Pokrívčáková 2019). Besonders im Zusammenhang mit KI-Anwendungen eröffnet digitale Kompetenz neue Möglichkeiten der Unterrichtsgestaltung, etwa durch automatisiertes Feedback, adaptive Lernsysteme oder interaktive Kommunikationswerkzeuge. Lehrkräfte mit höherer digitaler Kompetenz sind besser in der Lage, die Funktionsweisen von KI-Werkzeugen zu verstehen und diese gezielt in sprachdidaktische Szenarien einzubinden.

Forschungsarbeiten zeigen jedoch auch, dass viele Lehrende und Lernende bislang nur über begrenzte Erfahrungen und Kenntnisse im Umgang mit KI-gestützten Anwendungen verfügen. So belegen empirische Studien, dass sowohl das Wissen über KI-Systeme als auch die praktische Anwendungskompetenz vieler Nutzerinnen und Nutzer noch unzureichend ausgeprägt sind. Trotz dieser Defizite zeigen Lehrkräfte eine grundsätzlich positive Einstellung gegenüber computer- und KI-gestütztem Lernen. Frühere Untersuchungen im Bereich des *Computer-Assisted Language Learning* (CALL) und des darauf aufbauenden *Intelligent CALL* (ICALL) belegen, dass Lehrkräfte neuen Technologien gegenüber offen sind und ihr Potenzial zur Förderung sprachlicher Lernprozesse anerkennen (Heift & Schulze 2007, 63).

3 Methodische Ansätze zur Integration von KI in die Korpusforschung

Am Anfang der vorliegenden Untersuchung stand die Hypothese, dass der Einsatz von KI-basierten Systemen wesentlich dazu beitragen kann, komplexe Abfrageprozesse zu vereinfachen, die Bedienbarkeit von Korpusplattformen zu verbessern und den Zugang zu linguistischen Daten für eine breitere Nutzengruppe zu erleichtern.

Die im Rahmen der Untersuchung identifizierten methodischen und technischen Herausforderungen verdeutlichen zugleich, dass die Integration von LLMs in komplexe Korpusverwaltungssysteme wie COSMAS II gegenwärtig noch erhebliche Entwicklungarbeit erfordert. Ein Vergleich mit bestehenden Lösungen aus anderen Sprachräumen zeigt jedoch, dass niedrigschwellige Zugänge grundsätzlich realisierbar sind. So entwickelten Milička und Šebestová (2024) eine Anwendung, die Korpusrecherchen im Tschechischen Nationalkorpus ermöglicht. Dieses Korpus ist im Gegensatz zu COSMAS II ohne vorherige Anmeldung zugänglich und bietet damit insbesondere für unerfahrene Nutzende einen deutlichen Vorteil. Die dokumentierten Ergebnisse und die erprobte Anwendung veranschaulichen, dass praktikable Ansätze existieren, um die Barrieren zwischen Nutzenden und Korpussystemen zu verringern (Milička & Šebestová 2024). Daraus

lässt sich ableiten, dass bei entsprechender technischer Anpassung auch für COSMAS II die Entwicklung eines benutzerfreundlicheren, KI-gestützten Zugangs denkbar ist.

In folgenden Unterkapiteln werden verschiedene mögliche Ansätze vorgestellt. Im Mittelpunkt steht die Frage, inwiefern diese Technologien den Forschungsprozess nicht nur effizienter, sondern auch inklusiver gestalten können, indem sie den Zugang für Personen ohne vertiefte technische Kenntnisse erleichtern.

Darüber hinaus werden die jeweiligen Vor- und Nachteile unterschiedlicher KI-basierter Methoden diskutiert und ihre Anwendbarkeit im akademischen Lehr- und Lernkontext kritisch bewertet. Ziel ist es, ein differenziertes Verständnis der Potenziale und Grenzen solcher Verfahren zu entwickeln.

3.1 Von der Idee zum ausgewerteten Ergebnis

Ein naheliegender Ansatz besteht darin, Forschungsanfragen in natürlicher Sprache zu formulieren und die entsprechenden Ergebnisse automatisiert durch COSMAS II ausgeben zu lassen. Eine derartige Lösung wäre ideal, da sie die Nutzenden von der Notwendigkeit entbindet, sich intensiv mit der spezifischen Korpusanfragesprache auseinanderzusetzen. So könnte beispielsweise eine Eingabe wie *Ich möchte alle Verwendungen des Suffixes -ung in Verbindung mit Negationspräfixen finden* automatisch in eine korrekte COSMAS II-Abfrage übersetzt und ausgeführt werden. Fragt man eine KI entsprechend, erhält man folgende Antwort: *Ja – ich kann dir die Suchintention in eine korrekte COSMAS-II-Abfrage übersetzen. Ausführen (also tatsächlich im DeReKo/COSMAS II laufen lassen) kann ich hier allerdings nicht, weil ich keinen direkten Zugriff auf dein COSMAS-II-Konto bzw. die Korpora habe.*

Wenn du orthografisch nach Wörtern suchst, die mit einem Negationspräfix beginnen und auf „...ung“ enden, ist `#REG()` dafür am passendsten: www2.ids-mann...

cosmas

Code kopieren

```
#REG( ^(un|in|im|il|ir|non)-?.*ung$ )
```

Was das macht

- `^(un|in|im|il|ir|non)` = Wortanfang mit einem der Präfixe
- `-?` = optionaler Bindestrich (falls mal „un-...“ geschrieben wird)
- `.*ung$` = endet auf „ung“ www2.ids-mann...

Tipp: Stell in COSMAS II (Optionen) am besten Groß-/Kleinschreibung ignorieren, damit du nicht extra `[Uu]` etc. bauen musst. (Die Wirkung von Groß-/Klein hängt bei `#REG()` von der Option ab.) www2.ids-mann...

Abbildung 1. Darstellung einer von ChatGPT generierten COSMAS-II-Abfrage

Die Abbildung 1 zeigt die vom ChatGPT empfohlene Verwendung des regulären Ausdrucks `#REG(^|un|in|im|il|ir|non)-?.*ung$)` sowie eine schrittweise Erläuterung der einzelnen Bestandteile der Regex-Abfrage (Wortanfang, Präfixauswahl, optionaler Bindestrich und Suffix). Ergänzend enthält die Antwort von ChatGPT einen Hinweis zur empfohlenen Deaktivierung der Groß- und Kleinschreibung in den COSMAS-II-Optionen, um die Abfrage zu vereinfachen.

Die praktische Umsetzung dieses Ansatzes erweist sich, wie aus der Antwort des Bots hervorgeht, als mit erheblichen Herausforderungen verbunden. Für eine direkte Interaktion mit COSMAS II wird eine *Application Programming Interface* (API) benötigt. Dies erfordert nicht nur Programmierkenntnisse, etwa in *Python*, sondern auch ein vertieftes Verständnis der Funktionsweise von ChatGPT und der Entwicklung entsprechender Anwendungen (OpenAI 2024). Hinzu kommt, dass COSMAS II nur mit einem aktiven Benutzer-Login zugänglich ist, was zusätzliche Anforderungen an Authentifizierung und Zugriffsrechte stellt.

Aufgrund dieser technischen und administrativen Hürden ist dieser Ansatz zwar konzeptionell überzeugend, in der praktischen Umsetzung jedoch nicht niedrigschwellig realisierbar und somit für die Mehrheit der Studierenden kaum geeignet.

3.2 Von der Idee zum fertigen Code

Ein alternativer Ansatz besteht darin, aus einer natürlichsprachlichen Anfrage automatisch den entsprechenden Code in der COSMAS II-Anfragesprache zu generieren. Beispielsweise könnte die Eingabe *Ich möchte alle Wörter mit Negationspräfix und der Endung -ung finden* in den formalen Ausdruck (*ver oder anti* oder un* oder non* oder des* oder ent**) und *ung* übersetzt werden. Dieser Ansatz hat den Vorteil, dass er die Komplexität der formalen Anfragesprache reduziert und den Nutzenden unterstützt, präzisere und effektivere Anfragen zu formulieren.

Gleichzeitig bestehen jedoch auch hier Herausforderungen. Natürlichsprachliche Eingaben sind häufig mehrdeutig oder unpräzise, was zu fehlerhaften oder unvollständigen Abfragen führen kann, insbesondere bei komplexeren linguistischen Fragestellungen mit mehreren Suchparametern oder Abhängigkeiten. Zwar liefert die kostenpflichtige Version von ChatGPT (zum Zeitpunkt der durchgeführten Untersuchung GPT-4.0) bei gezielten Anweisungen und Beispielen in der Eingabezeile deutlich bessere Ergebnisse, dennoch bleibt eine manuelle Überprüfung und gegebenenfalls Korrektur der generierten Abfragen unverzichtbar.

3.3 Präzisierung der Eingabeformulierung

Um die Wahrscheinlichkeit von Missverständnissen zu reduzieren, kann die Anfrage exakter und strukturell klarer formuliert werden. Beispielsweise ließe sich anstelle einer freien Eingabe der Satz *Suche nach den Vorsilben ver-, anti-, un-, non-, des- und ent- in Verbindung mit der Endsilbe -ung* verwenden. Eine derart präzise Formulierung trägt zur besseren Interpretierbarkeit bei und kann die Genauigkeit der resultierenden COSMAS II-Abfragen erhöhen.

Obwohl dieser Ansatz insbesondere bei einfacheren Suchanfragen zu einer Verbesserung führt, stößt er bei komplexeren linguistischen Strukturen an seine Grenzen. Die KI kann Schwierigkeiten haben, die erforderliche logische Struktur, Operatoren und Platzhalter korrekt zu kombinieren, insbesondere wenn abstraktere linguistische Konzepte einbezogen werden. Folglich bleiben auch in diesem Fall eine ergänzende Anleitung sowie eine kritische Nachkontrolle durch die Nutzenden erforderlich.

3.4 Verwendung einer halbcodierten Eingabesprache

Ein deutlich erfolgversprechenderer Ansatz besteht in der Verwendung einer halbcodierten Sprache, die Elemente der formalen COSMAS II-Anfragesprache in vereinfachter und informeller Weise abbildet. Eine entsprechende Eingabe könnte etwa lauten *Suche nach: (Vorsilben: ver-, anti-, un-, non-, des-, ent-) in Verbindung mit (Endung: -ung)*. Durch die Nutzung von Klammern und Interpunktionszeichen werden die Beziehungen zwischen den Elementen deutlicher und Missverständnisse reduziert.

Dieser Ansatz führt in der Regel zu zuverlässigeren Ergebnissen, da die KI durch die strukturelle Vorgabe klarere Hinweise auf benötigte Operatoren und Platzhalter erhält. Dennoch bleibt ein gewisses Fehlerpotenzial bestehen, insbesondere in Bezug auf die präzise Wahl der Operatoren und deren korrekte Positionierung. Eine nachträgliche menschliche Kontrolle und gegebenenfalls Anpassung der generierten Abfragen ist daher weiterhin notwendig.

3.5 Vom Code zur erklärten Abfrage

Ein besonders lernförderlicher Ansatz besteht darin, einen funktionierenden Code in der COSMAS II-Anfragesprache als Eingabe zu verwenden und durch die KI eine erläuternde Beschreibung seiner Funktionsweise generieren zu lassen. So könnte etwa der Ausdruck *(ver oder anti* oder un* oder non* oder des* oder ent*) und ung* von der KI folgendermaßen interpretiert werden: *Diese Anfrage sucht nach allen Wörtern, die mit den Präfixen ver-, anti-, un-, non-, des- oder ent- beginnen und mit dem Suffix -ung enden.*

Diese Methode bietet erhebliche Vorteile in Bezug auf Verständlichkeit und Genauigkeit, da sie auf bereits funktionierendem Code basiert. Insbesondere fortgeschrittene Versionen von ChatGPT, wie GPT-4.0 oder höhere, ermöglichen die Generierung präziser und nachvollziehbarer Erklärungen, die das Verständnis der Anfragesprache und der Funktionsweise von COSMAS II nachhaltig fördern können.

Darüber hinaus ist dieser Ansatz für Studierende mit grundlegenden Kenntnissen der COSMAS II-Syntax und einem Verständnis für linguistische Operatoren gut umsetzbar und bietet eine praxisnahe Möglichkeit, analytische und technische Kompetenzen zu verbinden.

3.6 Zwischenfazit

Die analysierten und oben beschriebenen Ansätze zeigen, dass der Einsatz von KI in der Korpusarbeit unterschiedliche Potenziale bietet, deren Nutzen jedoch stark von der Form der Interaktion abhängt. Während vollständig automatisierte Lösungen derzeit technisch und organisatorisch kaum umsetzbar sind, erweisen sich teilautomatisierte Verfahren, insbesondere die Nutzung halbcodierter Eingaben, als praktikabler und lernfördernder.

Zudem zeigt sich, dass die KI-gestützte Erklärung bestehender Codes eine effektive Unterstützung beim Verständnis der COSMAS-II-Syntax darstellt. Insgesamt sind die größten Fortschritte dort zu erwarten, wo KI-Systeme als ergänzende Hilfsmittel eingesetzt werden, nicht als vollständiger Ersatz menschlicher Analysekompetenz.

4 Ergebnisse der KI-gestützten Korpusarbeit

Im Rahmen der durchgeführten Untersuchung wurde die Leistungsfähigkeit von ChatGPT bei der Formulierung und Umsetzung von Korpusanfragen in COSMAS II empirisch getestet. Insgesamt wurden 50 Suchanfragen entwickelt und ausgewertet, um feststellen zu können, in welchem Maße das LLM in der Lage ist, die spezifische Syntax und Logik der Korpusprache korrekt zu interpretieren und anzuwenden.

Für alle Anfragen, die morphosyntaktische Annotationen erforderten, also die Verwendung des Operators *MORPH*, wurde das TAGGED-C-Archiv der morphosyntaktisch annotierten Korpora (CONNEXOR) von COSMAS II genutzt. In allen übrigen Fällen kam das W-Archiv der geschriebenen Sprache zum Einsatz.

Die Ergebnisse der Anfragen wurden anschließend inhaltlich und formal ausgewertet und in 2 Gruppen eingeteilt, je nachdem, inwieweit die vom LLM generierten Formulierungen der COSMAS-II-Syntax entsprachen. Weiter werden diese Gruppen ausführlicher vorgestellt.

4.1 Zum methodischen Vorgehen

Wie bereits zuvor angemerkt, wurde der vorliegenden Untersuchung die Annahme zugrunde gelegt, dass die Nutzerinnen und Nutzer kaum mit den Suchoperatoren von COSMAS II vertraut sind, über wenig oder keine Erfahrung in der Formulierung entsprechender Suchanfragen verfügen und zudem keine computerlinguistischen Kenntnisse besitzen. Zum Zeitpunkt der Durchführung der Untersuchung stand die Version GPT-4.0 von ChatGPT zur Verfügung, die im Rahmen der Analyse eingesetzt wurde.

Ein erstes praktisches Hindernis ergab sich aus der Anmeldungspflicht für COSMAS II. Da entsprechend der getroffenen Annahmen keine Programmierkenntnisse vorausgesetzt werden konnten, wurde dieses Problem auf einfache Weise umgangen: Die vom Sprachmodell generierten Suchanfragen wurden manuell in das Suchfeld von COSMAS II übertragen, anschließend wurden Probesuchen durchgeführt und die resultierenden Trefferlisten ausgewertet.

Wie bereits im Einleitungskapitel angedeutet, existieren zahlreiche Anleitungen mit detaillierten Erläuterungen und Beispielen zur Formulierung von COSMAS-II-Suchanfragen. Die detaillierte Einarbeitung in diese Materialien ist für Nutzerinnen und Nutzer, insbesondere ohne einschlägige Vorerfahrung, jedoch nur eingeschränkt realistisch, da sie mit einem beträchtlichen Zeitaufwand verbunden ist. Zudem ist die gezielte Nutzung der Hilfefunktionen für Personen ohne entsprechende Vorkenntnisse nur begrenzt praktikabel. Vor diesem Hintergrund wurde entschieden, das Sprachmodell schrittweise durch gezielte Prompts an die COSMAS-II-Syntax heranzuführen. Zu berücksichtigen ist dabei, dass GPT-4.0 zum Untersuchungszeitpunkt noch nicht in der Lage war, selbstständig automatisch auf frei zugängliche Online-Anleitungen zuzugreifen. Zum Zeitpunkt der Veröffentlichung dieses Beitrags ist dies mit der Version GPT-5.2 hingegen möglich, was sich unter anderem an den Verweisen auf die offizielle COSMAS-II-Hilfeseite am Ende einzelner Informationsabschnitte erkennen lässt (vgl. Abb. 1 in diesem Beitrag).

Da von keinerlei Vorerfahrung der Nutzerinnen und Nutzer mit den Suchoperatoren von COSMAS II ausgegangen wurde, wurde aus den in Kapitel 3 vorgestellten Ansätzen jener gewählt, der auf eine Interaktion mit dem Sprachmodell in natürlicher Sprache setzt. Zeigten sich die generierten COSMAS-II-Abfragen als unzureichend, wurden die natürlichsprachlichen Eingaben schrittweise modifiziert, beispielsweise durch die Verwendung von Operatoren, die aus der Schulmathematik oder der elementaren Logik bekannt sind. Darüber hinaus wurde das Sprachmodell in den Prompts aufgefordert, in natürlicher Sprache zu erläutern, welche Suchoperation durch den jeweils erzeugten Code im Korpus ausgeführt wird. Auf diese Weise konnte überprüft werden, ob die zugrunde liegende Suchintention korrekt verstanden und umgesetzt worden war.

In mehreren Fällen führten auch wiederholte Präzisierungen sowie Hinweise auf Fehlinterpretationen nicht zu zufriedenstellenden Ergebnissen. In solchen Situationen war es erforderlich, selbstständig in den COSMAS-II-Anleitungen nach geeigneten Suchoperatoren zu recherchieren und diese dem Sprachmodell als explizite Vorgabe bereitzustellen.

Die folgenden Unterkapitel 4.2 und 4.3 widmen sich einer detaillierten Beschreibung der im Rahmen dieses Vorgehens erzielten Ergebnisse. Zunächst werden Beispiele erfolgreicher Umsetzungen der COSMAS-II-Syntax durch das LLM vorgestellt. Anschließend werden fehlerhafte bzw. unzureichende Implementierungen beschrieben, um die Grenzen und Herausforderungen beim Einsatz LLMs zur Generierung komplexer Korpusabfragen festzustellen.

4.2 Erfolgreiche Umsetzungen der COSMAS-II-Syntax durch das LLM

Von den insgesamt 50 generierten Suchanfragen ließ sich eine Gruppe identifizieren, bei der das LLM die Syntax, Operatoren und semantischen Strukturen der COSMAS-II-Anfragesprache korrekt oder nahezu korrekt umsetzte. In diesen Fällen waren die Ergebnisse nachvollziehbar, logisch aufgebaut und führten zu sinnvollen linguistischen Analysen. Typisch für diese Gruppe war die korrekte Anwendung der Operatoren */+w*, */s0* und *MORPH*, das richtige Erkennen von Platzhaltern sowie die Fähigkeit, Kontextanalysen und Vergleiche zwischen Wortformen, Regionen und Zeiträumen vorzunehmen.

Die Anfrage *&Zeitenwende* (*Wie hat sich der Gebrauch des Wortes Zeitenwende nach 2022 verändert?*) wurde weitgehend korrekt umgesetzt. Das Modell erkannte, dass das Präfix *&* alle Wortformen abruft, interpretierte den Ausdruck *nach 2022* zunächst lokal statt temporal, konnte aber nach weiteren Hinweisen in natürlicher Sprache die Sortierung nach Jahr und relativer Häufigkeit vorschlagen.

Auch bei *lizensier** (*Finde alle Formen von lizensieren und vergleiche sie mit lizenziieren*) identifizierte das LLM den richtigen Platzhalteroperator und empfahl die Durchführung einer zweiten Vergleichssuche, um Häufigkeiten gegenüberzustellen.

Für die Anfragen *Velo* und *Fahrrad* lieferte das Modell korrekte Hinweise auf regionale Filterung und Berechnung relativer Häufigkeiten pro eine Million Wörter. In Kombination konnte es Synonyme vergleichen, benötigte jedoch für die Berechnung eine Beispielanleitung.

Bei der Anfrage *Fernsprech** (*Wie kann der Gebrauch von Fernsprecher, Fernsprechanschlage usw. im Laufe der Zeit untersucht werden?*) schlug das Modell zunächst eine falsche Syntax zur zeitlichen Einschränkung vor, korrigierte diese aber nach Hinweisen.

Eindeutig erfolgreich verlief die Anfrage *ich /s0 aß*, mit der Satzkollokationen gesucht wurden. Die Syntax */s0* wurde korrekt angewendet, um die beiden Wörter im selben Satz zu finden.

Ebenso zeigte das Modell bei *MORPH(A){2}* (*Suche eine Sequenz von zwei Adjektiven*) ein gutes Verständnis der morphologischen Operatoren und gab sowohl *MORPH(A) /+w1 MORPH(A)* als auch die vereinfachte Form *MORPH(A){2}* korrekt aus.

Auch Abstands- und Ignorierungsoperatoren wurden erfolgreich angewandt: Die Anfragen *weil /+w3:20,s0 MORPH(V)*, *\$Fake /+w1 \$News* und *unweit /p0 &liegen* wurden vollständig korrekt umgesetzt. Besonders bei der Letzteren zeigte das LLM, dass es den Abschnittsoperator */p0* richtig interpretieren und zusätzliche Hinweise zur Sortierung nach Jahrzehnten geben konnte.

Komplexere Abfragen mit logischen Operatoren wie *Propagand**, *Querdenker nicht Corona* und **tion** und **ung** wurden ebenfalls mit hoher Genauigkeit umgesetzt. Nach kleineren Korrekturen konnte das Modell den richtigen Negationsoperator sowie kombinierte Bedingungen korrekt darstellen.

Auch im Bereich der Mustererkennung zeigten sich solide Ergebnisse. So lieferte das LLM für *+?ente+* (*Suche nach Wörtern mit der Zeichenfolge ente*) nach präzisierenden Hinweisen die richtige Lösung *?ente?*, sowie für *Merkel** die nahezu korrekte Abfrage zur Suche nach Wörtern, die die Buchstabenfolge *Merkel* enthalten.

Anfragen, die auf semantische oder pragmatische Bezüge abzielten, wurden ebenfalls erfolgreich verarbeitet. So etwa *Wie kann ich untersuchen, in welchem Kontext Bier steht?*, bei der das LLM nach vorherigen Beispielen die Muster korrekt rekonstruierte, oder *Maus nicht Computer**, wo es die Negation nach einer Korrektur richtig interpretierte.

Auch Suchanfragen mit numerischen oder regulären Ausdrücken wurden weitgehend korrekt umgesetzt. Bei *viktorianisch* /s0 (#REG(18[0-9]{2}) | #REG(19[0-1][0-9]|1920))* wurde eine funktionierende Lösung generiert, die das Ziel (die Kombination von *viktorianisch* mit Jahreszahlen zwischen 1800 und 1920) erreichte.

Schließlich konnte das LLM bei Anfragen mit logischen Kombinationen mehrere Begriffe oder Phrasen korrekt verknüpfen. Beispiele hierfür sind (*Konrad /+w1 Adenauer*) ... (*Olaf /+w1 Scholz*) (*Suche nach allen Bundeskanzlern*) sowie (*zum /+w1 Beispiel*) oder (*z. /+w1 B.*) oder (*bspw.*), die nach Korrekturhinweisen vollständig korrekte Ausgaben lieferten. Auch Distanzsuchen wie *weit /w4:10 weg* oder Wortkombinationen wie *Haus* /w0 *Schuh* wurden fehlerfrei erkannt.

Diese Gruppe erfolgreicher Anfragen verdeutlicht, dass das LLM in der Lage ist, bei klar definierten, formal eindeutigen Aufgaben die COSMAS-II-Syntax präzise umzusetzen. Insbesondere bei linearen Operatorstrukturen, morphologischen Suchbefehlen und einfachen logischen Kombinationen erzielt das Modell konsistente Ergebnisse. Schwierigkeiten traten vor allem bei mehrdeutigen oder zu allgemein formulierten Eingaben sowie bei numerischen Bedingungen auf. Die Ergebnisse zeigen jedoch deutlich, dass LLMs in der Lage sind, strukturierte Abfragesprachen partiell korrekt zu reproduzieren, sofern sie durch geeignete Prompts und klare Instruktionen gesteuert werden.

4.3 Fehlerhafte oder unzureichende Umsetzungen der COSMAS-II-Syntax

Diese Gruppe umfasst alle Anfragen, bei denen das LLM deutliche Schwierigkeiten zeigte, die Syntax der COSMAS-II-Anfragesprache korrekt zu interpretieren oder anzuwenden. Fehler traten vor allem bei der Verwendung von Operatoren, Platzhaltern und logischen Verknüpfungen auf. Auch das Verständnis von Zeichenfolgen, Reihenfolgen und regulären Ausdrücken erwies sich in vielen Fällen als unzureichend. Häufig waren wiederholte Korrekturen und erläuternde Beispiele erforderlich, ohne dass eine formal korrekte Anfrage erzielt werden konnte.

Die Anfrage *richtig /+w1 ,und‘ /+w1 wichtig (Suche nach der Wortfolge richtig und wichtig)* zeigte grundlegende Probleme im Umgang mit Anführungszeichen und Operatoren. Das Modell behandelte *und* als logischen Operator anstelle eines festen Bestandteils der Wortfolge. Auch die Notwendigkeit, die Reihenfolge mittels */+w1* festzulegen, wurde zunächst nicht erkannt.

Eine komplexere Variante (*richtig /+w1 ,und‘ /+w1 wichtig) /p0 Antisemitismus* führte zu mehrfach redundanten und syntaktisch falschen Ergebnissen. Das Modell gab u. a. *Antisemitismus /p0 ,richtig und wichtig‘* aus, ohne die überflüssige Wiederholung zu vermeiden.

Bei der Anfrage *Standart nicht Standarte***, die den falschen Gebrauch von *Standard* untersuchen sollte, wurde der Ignorierungsoperator *\$* fälschlich als Negationsoperator interpretiert. Anschließend wurden auch die Zeichen *%*, *-* und *#* irrtümlich verwendet, bevor nach wiederholtem Hinweis der korrekte Operator *nicht* erkannt wurde.

Auch bei der Mustererkennung zeigten sich Defizite. So interpretierte das LLM in *Tele?+on* die Platzhalterzeichen *?* und *+* fehlerhaft, indem es mehr Zeichen als erlaubt zwischen *Tele-* und *-on* zuließ. Erst nach wiederholter Erklärung konnte das Modell *Telefon* korrekt als mögliches Ergebnis erkennen.

Bei *aß oder (habe /s0 gegessen)* und der erweiterten Anfrage (*ich /s0 aß) oder (habe /s0 ich /s0 gegessen)* oder (*ich /s0 habe /s0 gegessen*) traten wiederholte Fehler im Umgang

mit Platzhaltern und logischen Verknüpfungen auf. Die Formulierungen wurden zwar syntaktisch korrekt, aber semantisch redundant umgesetzt, da zwischen den Varianten *ich habe gegessen* und *habe ich gegessen* kein Unterschied besteht.

Die Anfragen zum Kasusgebrauch, etwa *wegen /+w1 (mir dir dem einem) (Dativ)* und *(wegen /+w1 (des eines)) oder meinewegen oder deinewegen (Genitiv)*, zeigten, dass das Modell Schwierigkeiten hatte, grammatische Rollen korrekt zu erfassen. Objektpronomen wurden irrtümlich als Genitivformen interpretiert, und die Klammersetzung blieb fehlerhaft.

Auch im Bereich der Nebensatzanalyse traten Fehler auf. Die Anfrage *weil /+w2,s0 MORPH(V)* sollte Verbletzstellung in Nebensätzen identifizieren, wurde jedoch nur als Suche nach *weil* und einem Verb im selben Satz mit bis zu zwei Wörtern Abstand interpretiert. Eine logische Differenzierung zwischen Verbendstellung und Verbzweitstellung war nicht möglich. Ähnliche Probleme zeigten sich bei *weil /+s0 MORPH(V) %+w1,s0 **, wo zwar die Bestandteile korrekt erklärt, jedoch keine funktionierende Gesamtsyntax generiert wurde.

Auch bei semantischen Kontextanfragen wie *&frisch (Untersuchung des Kontexts von frisch)* blieb das Verständnis oberflächlich: Das Modell schlug lediglich *frisch /+w1** vor, ohne die Möglichkeit einer Kontextanalyse in den Ergebnissen zu erkennen.

Bei *er*in* zeigte sich ein partiell korrektes Verständnis der Escape-Sequenz, jedoch Schwierigkeiten bei der praktischen Umsetzung. Erst nach Hinweisen konnte das Modell passende Beispiele wie *Arbeiterin** und *Arbeiterinnen** nennen.

In der Anfrage *Falsch nicht &falsch* (Suche nach Komposita mit Falsch- unter Ausschluss des Adjektivs falsch)* wurde erneut der falsche Operator *%* verwendet. Erst nach dem Hinweis, dass *nicht* der korrekte Negationsoperator ist, konnte eine korrekte Version erstellt werden.

Ein weiteres Beispiel für semantische Fehlinterpretationen stellt die Anfrage *treu #OV Be*ung* dar. Das Modell nannte *Betrug* oder *Belehrung* als wahrscheinlichste Ergebnisse und konnte den eigentlichen Zusammenhang, etwa *Betreuung*, auch nach mehreren Versuchen nicht erkennen. Das Verständnis von Zeichenfolgen innerhalb eines Wortes erwies sich somit als stark begrenzt.

Auch bei *ber?chtigen (Suche nach berichtigen und berechtigen)* nutzte das Modell zunächst reguläre Ausdrücke (#REG), statt die geforderten Platzhalter. Erst nach mehreren Korrekturen wurde *ber?chtigen* als korrekte Lösung ausgegeben. Ein ähnliches Problem trat bei *B?ot* auf, wo das LLM *Boot* und *Brot* nicht zuverlässig erkannte.

Bei der Anfrage (*wo /w0 <sa> /+s0 (?) /w0 <se>*), die Sätze mit *wo* am Anfang und einem Fragezeichen am Ende finden sollte, gelang erst nach mehrfacher Erklärung eine annähernd korrekte Formulierung. Das Modell verwechselte den Platzhalter *?* mit einem Operator und erkannte erst nach dem Hinweis auf die Notwendigkeit eines Backslashes (*!?*) die richtige Syntax.

Schließlich erwies sich die Kombination von Operatoren als besonders schwierig. Die Anfragen *raten #IN ver*ten* und *ente #IN st*en*, die nach eingebetteten Zeichenfolgen in zusammengesetzten Wörtern suchten, führten zu keinen sinnvollen Ergebnissen. Das Modell konnte die Funktionsweise des Operators *#IN* nicht korrekt umsetzen und keine passenden Wortbeispiele liefern.

Diese Gruppe zeigt deutlich die Grenzen der aktuellen Leistungsfähigkeit von LLMs im Umgang mit formalen Korpusanfragesprachen. Die meisten Fehler resultierten aus der Übertragung natürlichsprachlicher Logik auf regelbasierte Systeme. Schwierigkeiten traten insbesondere bei der korrekten Verwendung von Negations- und Kombinationsoperatoren, der Platzhalterlogik sowie bei verschachtelten Strukturen auf. Auch die semantische Interpretation komplexer Zeichenfolgen blieb oft unzureichend.

Die Ergebnisse verdeutlichen, dass LLMs zwar in der Lage sind, einfache Strukturen korrekt zu reproduzieren, bei formal strengen Abfragesprachen jedoch schnell an systemische Grenzen stoßen. Verbesserungen könnten durch den Einsatz spezialisierter Trainingsdatensätze, die gezielt auf linguistische Abfragesprachen zugeschnitten sind, sowie durch die Kombination mit regelbasierten Parsern erzielt werden. Erst durch solche hybriden Ansätze ließe sich die Zuverlässigkeit und Präzision im Umgang mit Korpusanfragen signifikant erhöhen.

4.4 Konzeptionelle und technische Herausforderungen der Untersuchung

Die Auswahl des Korpusverwaltungssystems COSMAS II und des *Large Language Models* ChatGPT ergab sich nach einer Sichtung der verfügbaren Plattformen als naheliegend. Bereits zu Beginn des Forschungsprojekts zeigte sich, dass diese Kombination sowohl aufgrund ihrer wissenschaftlichen Relevanz als auch der mit ihr verbundenen technischen und methodischen Herausforderungen besondere Eignung für die Untersuchung aufweist. COSMAS II, eine der zentralen Plattformen für die Korpusforschung im deutschsprachigen Raum, bietet umfangreiche sprachliche Ressourcen, die für die vorliegende Analyse von zentraler Bedeutung sind. ChatGPT wiederum stellt eines der bekanntesten und am weitesten verbreiteten *Large Language Models* dar und gilt damit als ein naheliegendes Werkzeug für die Erforschung KI-gestützter Ansätze in der Korpusarbeit.

Während die Auswahl der Untersuchungsobjekte vergleichsweise unkompliziert erfolgte, erwies sich die Entwicklung eines geeigneten methodischen Ansatzes als komplexe Aufgabe. Zentrale Herausforderung war die Konzeption einer Schnittstelle zwischen beiden Technologien, die eine niedrigschwellige Nutzung ermöglichen sollte. Von besonderer Bedeutung war hierbei die Frage, inwiefern Studierende ohne vertiefte technische Kenntnisse in die Lage versetzt werden können, präzise und relevante Korpusanfragen zu formulieren. In der praktischen Umsetzung zeigte sich jedoch, dass die Realisierung dieses Ziels mit erheblichen technischen und konzeptionellen Schwierigkeiten verbunden ist. Um die KI-gestützte Unterstützung valide gestalten zu können, war es notwendig, das Modell mit domänenspezifischem Wissen über COSMAS II zu versorgen und damit zu trainieren. Dieses Training setzt wiederum ein detailliertes Verständnis der Funktionsweise und der umfangreichen Operatoren von COSMAS II voraus.

Die Verwendung von LLMs wie ChatGPT zur Unterstützung in der Korpusforschung, insbesondere bei der Formulierung und Interpretation von Anfragen in COSMAS II, besitzt grundsätzlich das Potenzial, den Zugang zu komplexen sprachlichen Daten erheblich zu erleichtern. In der praktischen Anwendung traten jedoch mehrere Schwierigkeiten auf, die sich sowohl aus der spezifischen Struktur der Korpusanfragesprache als auch aus den Eigenschaften des LLM selbst ergaben. Die Ursachen hierfür liegen in der Diskrepanz zwischen der natürlichsprachlichen Trainingsbasis der Modelle und den formal-logischen Anforderungen linguistischer Korpusplattformen. Zukünftige Lösungsansätze könnten in der Entwicklung spezialisierter Trainingsdatensätze, der Integration regelbasierter Module oder der Kopplung von LLMs mit domänenspezifischen Assistenzsystemen liegen. Trotz dieser Einschränkungen bleibt das Potenzial von KI-basierten Systemen bestehen, insbesondere in der Funktion als Vermittlungs- oder Übersetzungsschicht zwischen natürlicher Sprache und Korpusanfragesprache, um so den Zugang zu sprachwissenschaftlichen Ressourcen langfristig zu erleichtern.

5 Zusammenfassung

Im Rahmen der Untersuchung wurde das Potenzial von KI, insbesondere des *Large Language Models* ChatGPT, zur Unterstützung bei der Formulierung und Durchführung von Korpusanfragen im Korpusverwaltungssystem COSMAS II analysiert. Ziel war es, zu ermitteln, inwieweit der Einsatz von LLMs dazu beitragen kann, die Zugänglichkeit und Effizienz der Korpusnutzung insbesondere für Studierende und Nutzerinnen und Nutzer ohne vertiefte Kenntnisse der Korpusanfragesprachen zu verbessern.

Die Analyse der Interaktionen zwischen dem LLM und der COSMAS II-Anfragesprache zeigte sowohl vielversprechende Ansätze als auch deutliche Einschränkungen. LLMs besitzen grundsätzlich das Potenzial, als niedrigschwellige Unterstützung bei der Kor-

pusarbeit zu fungieren. Durch ihre Fähigkeit, natürliche Sprache zu verarbeiten, können sie Benutzende bei der Formulierung von Abfragen unterstützen und somit den Zugang zu Korpora erleichtern. Auf diese Weise lassen sich Einstiegshürden in die Korpusforschung verringern und Lernprozesse in der sprachwissenschaftlichen Ausbildung effizienter gestalten.

Gleichzeitig wurde deutlich, dass die Arbeit mit der spezifischen Syntax von COSMAS II erhebliche Schwierigkeiten mit sich bringt. Das LLM hatte häufig Probleme, formale Operatoren und Zeichen korrekt zu interpretieren, wodurch fehlerhafte oder unvollständige Abfragen entstanden. Diese Beobachtungen verdeutlichen, dass große Sprachmodelle zwar über ausgeprägte Fähigkeiten in der Verarbeitung natürlicher Sprache verfügen, jedoch bei der Handhabung komplexer, formal definierter Syntaxstrukturen an ihre Grenzen stoßen. Eine präzise Anleitung und eine klar strukturierte Benutzerführung bleiben daher unerlässlich, um Missverständnisse und Fehlinterpretationen zu vermeiden.

Darüber hinaus zeigte sich, dass die Qualität der Ergebnisse stark von der Formulierung der Anfragen abhängt. Während bei einfacheren, eindeutig strukturierten Eingaben zufriedenstellende Ergebnisse erzielt wurden, führten komplexere Abfragen mit mehreren Bedingungen, Operatoren oder linguistischen Spezifikationen häufig zu Fehlinterpretationen.

Insgesamt lässt sich festhalten, dass LLMs wie ChatGPT ein hohes Potenzial als unterstützende Werkzeuge in der Korpusforschung besitzen. Sie können dazu beitragen, den Zugang zu sprachwissenschaftlichen Ressourcen zu erleichtern und den Lernprozess insbesondere für Einsteigerinnen und Einsteiger zu unterstützen. Um dieses Potenzial vollständig auszuschöpfen, sind jedoch gezielte Weiterentwicklungen erforderlich, insbesondere durch die Integration fachspezifischer Trainingsdaten, die Anpassung an formale Anfragesprachen sowie die Entwicklung benutzerfreundlicher Schnittstellen. Eine solche Verbindung von linguistischer Präzision und technischer Zugänglichkeit könnte langfristig entscheidend dazu beitragen, die Nutzung von Korpora im Forschungs- und Lehrkontext zu optimieren.

Author contributions

Skaistė Volungevičienė: conceptualization, methodology, formal analysis, writing – original draft, review & editing. **Maximilian Arndt:** methodology, formal analysis, investigation, writing – original draft, writing – review & editing.

Quellenverzeichnis

COSMAS. 1991–2024. *COSMAS I/II – Corpus Search, Management and Analysis System*. Mannheim: IDS. Available at: <https://www.ids-mannheim.de/cosmas2/>.

Literaturverzeichnis

- Curry, Niall & Tony McEnery. 2025. Corpus linguistics for language teaching and learning: A research agenda. *Language Teaching* 58 (2), 232–251. <https://doi.org/10.1017/S0261444824000430>
- Heift, Trude & Mathias Schulze. 2007. *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York: Routledge.
- Hunton, Susan. 2006. Corpus Linguistics. *Encyclopedia of Language & Linguistics*. Keith Brown, ed. 2. Ausgabe. Elsevier Science. 234–248.
- Kohnke, Lucas & Benjamin Luke Moorhouse & Di Zou. 2023. ChatGPT for language teaching and learning. *RELC Journal* 54 (2), 537–550. <https://doi.org/10.1177/00336882231162868>
- McShane, Marjorie & Sergej Nirenburg. 2021. *Linguistics for the Age of AI*. Cambridge: The MIT Press. <https://doi.org/10.7551/mitpress/13618.001.0001>
- Milička, Jiří & Denisa Šebestová. 2024. Query a corpus in near-natural language: A human-friendly corpus query language not only for linguists. *Crossing Boundaries through Corpora: Innovative Approaches to Corpus Linguistics*. Sara Buschfeld, Patricia Ronan, Theresa Neumaier, Andreas Weilinghoff & Lisa Westermayer, eds. Amsterdam: John Benjamins. 248–262. <https://doi.org/10.1075/scl.119.10mil>
- OpenAI. 2024. *GPTs vs Assistants*. Available at: <https://help.openai.com/en/articles/8673914-gpts-vs-assistants>. Accessed: 1 September 2024.
- Pokrivčáková, Silvia. 2019. Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education* 7 (3), 135–153. <https://doi.org/10.2478/jolace-2019-0025>

Vom Seminar zur Praxis: Korpusgestützt lehren lernen

Patrick Grommes

Institut für Germanistik
Fachbereich Sprache, Literatur und Medien I
Fakultät für Geisteswissenschaften
Universität Hamburg
Von-Melle-Park 6, Postfach #15
D-20146 Hamburg, Deutschland
E-Mail: patrick.grommes@uni-hamburg.de
<https://ror.org/00g30e956>

Anastasiia Stulen

Institut für Germanistik
Universität Hamburg
Von-Melle-Park 6, Postfach #15
D-20146 Hamburg, Deutschland
E-Mail: anastasiia.stulen@uni-hamburg.de
ORCID iD: 0009-0009-8594-5727
<https://ror.org/00g30e956>

Abstract. Die Umsetzung des Konzepts des datengestützten Lernens (DDL) im Fach Deutsch als Fremdsprache (DaF) in einer praxisorientierten Weise erfordert sowohl die Bereitschaft der Lehrkräfte, sich in Korpuslinguistik weiterzubilden, als auch die Motivation der Lernenden zur Teilnahme, da beide Gruppen diesem Ansatz oft skeptisch gegenüberstehen. Dieser Artikel skizziert drei Phasen der Korpusidentifizierung und -implementierung in DaF und stützt sich auf zwei Seminare an der Universität Hamburg, die sich mit der Entwicklung korpusbasierter Lehransätze für DaF-Lernende im Ausland befassten. Die Ergebnisse der Seminare werden zusammen mit Erkenntnissen aus der Literatur zum DDL im DaF herangezogen, um Kriterien für die Reflexion über einen korpusbasierten DaF-Kurs zu formulieren. Anhand einer Lerneinheit zu den kausalen

Submitted: 22/11/2025. **Accepted:** 03/12/2025

Copyright © 2025 Patrick Grommes, Anastasiia Stulen. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Konnektoren *denn*, *weil* und *da*, die während des zweiten Seminars an der Universität Hamburg entwickelt und an der Universität Vilnius getestet wurde, wird eine konkrete Unterrichtssituation analysiert und diskutiert. Die Auswertung der Lerneinheit zeigt, dass die Entwicklung eines korpusbasierten Sprachunterrichts einen schrittweisen und systematischen Prozess erfordert. Kontinuierliche Fortbildung und berufliche Weiterentwicklung sind unerlässlich, um Lehrkräfte besser auf diese Methoden vorzubereiten und ihre Akzeptanz in der Praxis zu verbessern.

Schlüsselwörter: Deutsch als Fremdsprache (DaF), *data-driven learning*, Korpuslinguistik, DaF-Lehre, DWDS, KWIC

From seminar to practice: Learning to teach using corpus-based methods

Abstract. Implementing the idea of data-driven learning (DDL) in German as a foreign language (DaF) in a way that works in practice requires both teachers' willingness to engage in further training in corpus linguistics and students' motivation to participate, as both groups often express skepticism toward this approach. This article outlines three phases of corpus identification and implementation in DaF and draws on two seminars at the University of Hamburg that focused on developing corpus-based teaching approaches for DaF students abroad. Findings from the seminars, together with insights from the literature on DDL in DaF, are used to formulate criteria for reflecting on a corpus-based DaF course. Using a learning unit on the causal connectors *denn* 'because', *weil* 'because', and *da* 'because', which was developed during the second seminar at the University of Hamburg and tested at the Vilnius University, a specific teaching situation is analysed and discussed. The evaluation of the learning unit shows that the development of corpus-based language teaching requires a gradual and systematic process. Continuous training and professional development are essential to better prepare teachers for these methods and to enhance their acceptance in practice.

Keywords: German as a foreign language, data-driven learning, corpus linguistics, teaching German as a foreign language, DWDS, KWIC

1 Einleitung

Wer sich mit *data-driven learning* (DDL) oder dem Einsatz von Korpora in der Fremdsprachenlehre in einem weiteren Sinne befasst, wird nicht überrascht sein, wie Studierende, aber auch Fremdsprachenlehrende, reagieren, wenn ihnen angekündigt wird, dass genau das von ihnen erwartet wird. „Die Arbeit mit Korpora ist komplex und intransparent“ und „ich gebe dem jetzt noch eine letzte Chance“. Das waren Reaktionen auf die Vorstellung des Seminarplans im Sommersemester 2022 im Seminar zu „Konnektivität:

Korpusanalysen aus dem DaF/DaZ-Erwerb“ an der Universität Hamburg. Bei einem Workshop im Rahmen einer Germanistischen Institutspartnerschaft (GIP) zwischen dem Institut für Germanistik der Universität Hamburg und dem Lehrstuhl für Deutsche Philologie der Universität Vilnius (vgl. hierzu Zinsmeister, Sökefeld & Volungevičienė 2025) äußerten litauische Lehrende für „Deutsch als Fremdsprache“ (DaF) bei einer Diskussion dieses Seminarkonzepts im Herbst 2022 ebenfalls Skepsis. Korpusdaten seien gerade für Studierende auf Anfängerniveau verwirrend und Lehrende seien auf die Arbeit mit Korpora nicht vorbereitet bzw. deren Nutzung sei wenig intuitiv.

Ausgehend von dieser Diskussion wurde im Rahmen des Workshops in Vilnius beschlossen, eine Neuauflage des Seminars im Sommersemester 2023 an der Universität Hamburg zu versuchen, an der sich Lehrende aus Vilnius beteiligen und die möglichst in eine Praxiserprobung der im Seminar erarbeiteten Konzepte und Verfahren in Vilnius münden sollte. Somit befanden sich alle Beteiligten in einer vielschichtigen Lehr-Lern-Konstellation. Am Seminar in Hamburg haben Studierende der germanistischen Linguistik, der Allgemeinen Sprachwissenschaft und verschiedener Lehramtsstudiengänge, die alle einen Studienschwerpunkt in der Vermittlung von DaF und DaZ¹ hatten, teilgenommen. Außerdem waren DaF-Lehrende aus Vilnius ebenfalls in der Rolle von Studierenden beteiligt. Bei der Erprobung eines korpusgestützten Lehrkonzepts in Vilnius wechselte dann eine Studierende des Studienschwerpunkts DaF in Hamburg aus dieser Rolle und agierte in Vilnius als Lehrende. Die Studierenden in Vilnius wiederum lernten Deutsch als Fremdsprache in ihrem Studium, dessen Schwerpunkt nicht notwendigerweise in der Germanistik lag. Die Mehrzahl dieser Studierenden belegte keine germanistischen Studiengänge.

In diesem Beitrag soll das Vorgehen im Seminar in Vilnius im Vergleich zum Hamburger Seminar sowie im Lichte der Diskussion um den Einbezug korpuslinguistischer Methoden und Ressourcen in die Fremdsprachenvermittlung reflektiert werden. Dazu wird zunächst ein kurzer Überblick über die entsprechende Forschung gegeben, bevor die fraglichen Seminare eingehender vorgestellt werden. Dabei werden wir in groben Zügen das Vorgehen und die Ergebnisse der im Seminar entstandenen Unterrichtsvorschläge darstellen und dann detaillierter auf die Konzeption und Durchführung einer korpusgestützten Unterrichtseinheit eingehen, die von der Zweitautorin dieses Beitrags entwickelt wurde. Erkenntnisse zu Potentialen und Beschränkungen des korpusgestützten Lehrens und Lernens, die in diesem Prozess gewonnen wurden, schließen unseren Beitrag ab.

¹ DaZ: Deutsch als Zweitsprache.

2 Zum Einsatz von Korpora in Deutsch als Fremdsprache

Seit dem Beginn der 2000er Jahre lassen sich unseres Erachtens drei Phasen der Auseinandersetzung mit Korpora in DaF ausmachen. Während die erste Phase sich recht klar von der zweiten und dritten abgrenzen lässt, gibt es zwischen den letzten zwei Phasen Überschneidungen. In den folgenden Abschnitten versuchen wir, diese Phasen zu charakterisieren.

2.1 Korpora in DaF in der ersten Dekade des 21. Jahrhunderts: Potentiale und Perspektiven

Als Kristallisierungspunkt für Beiträge aus der ersten Phase dürfte die inzwischen in der dritten Auflage vorliegende Einführung in die Korpuslinguistik von Lemnitzer und Zinsmeister (2006; 2015) gelten. Damit war die zuvor meist im englischsprachigen Raum geführte Diskussion über die Forschungs- und Vermittlungsoptionen, die Korpora eröffnen, im deutschsprachigen Raum angekommen. Außerdem wurden hier auch erste DaF-didaktische Zugänge diskutiert (Lemnitzer & Zinsmeister 2015, 190–192).

Fandrych und Tschirner (2007) zeigen dann recht detailliert mögliche Verbindungslien zwischen Erkenntnissen der Spracherwerbsforschung und korpuslinguistischen Fragestellungen. Im Ausblick ihres programmativen Beitrags heben sie hervor, „dass ein korpuslinguistisches Grundwissen elementarer Bestandteil der Ausbildung im Fach Deutsch als Fremdsprache sein sollte“ (Fandrych & Tschirner 2007, 203). Lüdeling und Walter (2009) stellen übersichtlich den damaligen Stand der Möglichkeiten des Korpus-einsatzes in der DaF-Vermittlung dar, oft ausgehend von Arbeiten aus dem englischsprachigen Raum. Sie zeigen die folgenden Hauptlinien auf, wobei sie zwischen einem qualitativen und einem quantitativen Vorgehen unterscheiden. Bei Ersterem nutzen die Lehrenden wie die Lernenden Korpusdaten, um sich anhand dieser Daten mit authentischem Sprachgebrauch auseinanderzusetzen. Somit können Lehrende insbesondere bei der Materialgestaltung entlastet werden, während Lernende u. a. für Variation in der Sprache sensibilisiert werden können (Lüdeling & Walter 2009, 4–9). Quantitative Verfahren können dagegen helfen, die Relevanz sprachlicher Ausdrücke bzw. Strukturen zu prüfen und ggf. daran angepasste Vermittlungsschwerpunkte zu setzen. Ferner kann überprüft werden, inwieweit vorhandene Lehrmaterialien den realen Sprachgebrauch reflektieren. Es wird allerdings auch betont, dass insbesondere Lernende mit grundlegenden korpuslinguistischen Methoden vertraut gemacht werden müssen, bevor sie in der Lage sind, sich eigenständig mit quantitativen Zugängen zu befassen (Lüdeling & Walter 2009, 9–13).

2.2 Korpora in DaF in der zweiten Dekade: Zwischen Desillusionierung und Machbarkeitstest

Die zweite Phase der Auseinandersetzung mit Korpora in der DaF-Vermittlung bestätigt die Aussagen der ersten Phase weitgehend. Die zweite Phase ist unseres Erachtens ab der späten Mitte der 2010er Jahre anzusetzen. In gewisser Weise wird hier schon auf die erste Phase zurückgeblickt und zumeist konstatiert, dass die hohen Erwartungen insbesondere an die Nutzung von Korpora in der Sprachvermittlung nicht eingelöst wurden. Nolting und Radtke (2019, 110) stellen bspw. fest, dass „[...] bisher [...] die Bedeutsamkeit der Arbeit mit Korpora im Deutschunterricht allenfalls in einigen theoretischen Betrachtungen ansatzweise herausgestellt worden [ist]“. Die Auswirkungen zeigen sich über diese zweite Phase hinaus. So merkt Link (2025) noch an, dass Korpora und DDL kaum im DaF-Kontext zum Tragen kommen. Sie macht das zum Teil an den in unserer Einleitung zitierten Aspekten fest. Zudem werden mangelnde Ausbildung der Lehrenden bezüglich Korpusmethoden und Ressourcenknappheit ins Feld geführt (Link 2025, 139). Diese Punkte werden in Krekeler (2021) sehr deutlich, einem Text, der als Scharnier zwischen den hier postulierten Phasen zwei und drei verstanden werden kann. Krekeler zeichnet minutiös alle Schritte nach, die gegangen werden müssen, um erfolgreich in einem DDL-Szenario zu arbeiten. Er stellt nachvollziehbar dar, dass DDL den erhofften didaktischen Mehrwert bringt, wenn eine Reihe von Voraussetzungen erfüllt sind. Dazu gehört *corpus literacy* auf Seiten der Lehrenden wie der Lernenden. Bei Letzteren kommt hinzu, dass eine Verunsicherung durch Vielfältigkeit und Unschärfen in authentischen Daten wohl frühestens ab dem B2-Niveau toleriert wird bzw. nicht ungebührlich mit dem Lernprozess interferiert (Krekeler 2021, 190–191).

Vor diesem Hintergrund verwundert es etwas, dass Bedarfe und Bedürfnisse der Korpusnutzenden eher selten untersucht werden. Dies wurde schon von Fandrych et al. (2016) festgestellt und wird noch in Farroni (2024, 96) gefordert. Erstere sehen eine kleine Zahl von Studien aus den frühen 2000er Jahren, die die Nutzenden schriftsprachlicher Korpora im Blick haben, und wenden sich selbst Nutzenden deutschsprachiger Korpora der gesprochenen Sprache zu, die bis dahin noch nicht untersucht wurden. Dabei springen einige Ergebnisse besonders ins Auge. Bspw. gibt es eine recht große Bandbreite an Nutzendentypen, bei denen gerade die eher nicht technisch oder computerlinguistisch Interessierten den größten Anteil ausmachen (Fandrych et al. 2016, 282). Zudem stellen sie fest, dass die Nutzenden nur sehr eingeschränkt willens oder in der Lage sind, sich mit den Anforderungen verschiedener Plattformen bspw. hinsichtlich der Abfragesprachen auseinanderzusetzen, wobei zugleich hohe Erwartungen an die Nutzerfreundlichkeit der Plattformen gestellt werden. Eine plattformübergreifende Dacharchitektur wird dabei als Lösungsansatz gesehen, auch wenn einige grundsätzlich unvereinbare Nutzungsansprüche und -szenarien nebeneinander bestehen bleiben dürften (Fandrych et al. 2016, 286).

Diese Befunde beziehen sich zwar auf die Nutzung von Korpora gesprochener Sprache, dürften aber mit großer Wahrscheinlichkeit auf andere Korpora übertragbar sein.

2.3 Korpora in DaF in der dritten Dekade: Praxisreife und Konsolidierung

Es ist insgesamt wenig überraschend, dass die gleichen Bedenken weiterhin bestehen. Sie werden auch in Studien der dritten, aktuellen Phase der Beschäftigung mit Korpora in DaF, die wir ab etwa 2021 beginnen lassen würden, vorgebracht. Ein wichtiger Kritikpunkt ist, dass vor jeder Korpusnutzung in DaF (noch) eine Einarbeitung in korpuslinguistische Methoden, und sei es auf noch so grundlegender Ebene, notwendig ist. In diese Richtung deuten Befunde, die Florou (2025) berichtet. Sie geht u. a. der Frage nach, wie man Einstellungen angehender Fremdsprachenlehrender zum Einsatz von Korpora im Unterricht zum Positiven verändern kann. Die Teilnehmenden in ihrer Studie sollten mittels *AntConc* (Anthony 2018) eigene kleine Korpora einrichten und daran anschließend Aufgaben für Englisch- bzw. Italienisch-Lernende entwickeln. Der spannende Befund ist, dass die anfänglich deutlich ausgeprägte Skepsis gegenüber korpusgestütztem Arbeiten im Anschluss an die Intervention in ihr Gegenteil verkehrt wurde (Florou 2025, 20). Ein noch ungelöstes, eher DaF-spezifisches Problem dabei ist, dass nach wie vor viele und zudem aktuelle Einführungen in die Korpuslinguistik mit einem Schwerpunkt auf Fremdsprachenvermittlung nur auf Englisch vorliegen. Dies kritisiert auch Krekeler (2023) in seiner Sammelrezension entsprechender Titel.

Dennoch deutet sich in dieser dritten Phase eine Einlösung der Versprechen aus der ersten Phase an. Dies mag damit zusammenhängen, dass mit dem eJournal *KorDaF – Korpora Deutsch als Fremdsprache*² seit 2021 ein Rahmen für entsprechende Arbeiten geschaffen wurde. Das hat mindestens zwei Vorteile. Für Beiträge aus einem zugleich DaF- und korpusdidaktischen Umfeld gibt es nun eine klare Anlaufstelle und einen naheliegenden Publikationsort. Außerdem ist es nun für DaF-Lehrende bzw. diejenigen, die diese ausbilden, einfacher, sich einen Überblick über Methoden und Unterrichtsszenarien zu verschaffen. Dies hat sich auch als vorteilhaft für das zweite hier noch zu diskutierende Seminar erwiesen.

Als hilfreich für diesen Beitrag hat sich der Aufsatz von Krekeler (2021) erwiesen, in dem er den gesamten Zyklus der Entwicklung und Durchführung einer Unterrichtseinheit beschreibt und kritisch reflektiert. Die von ihm diskutierten Schritte sollen hier im weiteren Verlauf auf die an der Universität Hamburg durchgeführten Lehrveranstaltungen und die Praxiserprobung in Vilnius bezogen werden. Zusätzlich liegt mit Link (2025) eine Untersuchung vor, in der die Erfahrungen DaF-Lernender mit einer DDL-Lerneinheit an einer italienischen Universität berichtet werden. Da eine ähnliche Evaluation auch im

² *KorDaF – Korpora Deutsch als Fremdsprache*, abrufbar unter: www.kordaf.de.

Anschluss an die Erprobung in Vilnius stattgefunden hat, sollen schließlich diese Ergebnisse vergleichend diskutiert werden.

3 Kriterien für die Reflexion einer korpusdidaktischen DaF-Lehrveranstaltung

Krekeler (2021, 163) unterscheidet auf der Basis der im Fachdiskurs diskutierten didaktischen Aspekte drei Bereiche, denen er insgesamt neun Themen zuordnet. Die drei unterschiedenen Bereiche sind Planung, Unterricht und Didaktik. Zur Planung gehört einerseits die Frage, ob mit Korpusanalysen und -tools sinnvolle zusätzliche Daten für den Unterricht gewonnen werden können, und andererseits die Auswahl zum Unterricht passender und für DDL geeigneter Korpora. Im weiteren Verlauf seines Beitrags macht Krekeler daher sehr deutlich, dass immer die (sprach-)unterrichtlichen Ziele die Richtung vorgeben und im Auge behalten werden müssen. Die korpuslinguistischen Anteile sollten diesen förderlich sein (Krekeler 2021, 164–165).

Unter dem Stichwort „Unterricht“ erfasst er dann vier Themen, die typische korpuslinguistisch geprägte Aufgabenstellungen im Fremdsprachenunterricht spiegeln. Er listet die folgenden Punkte auf: „Sprachliche Muster thematisieren“, „Mit Konkordanzen lernen“, „Fehlerkorrekturen durchführen“ und „Korrekttheit/Angemessenheit thematisieren“ (Krekeler 2021, 163). Die Fragen, die man sich hier zu stellen habe, sind u. a. die, ob die Auseinandersetzung bspw. mit sprachlichen Mustern oder Korrektheit und Angemessenheit im jeweiligen Unterricht zielführend ist oder ob nicht etwa der variationsreiche, in Korpora vorfindliche Sprachgebrauch überfordert. Analog ist bei Fehlerkorrekturen, die korpusunterstützt von Lernenden durchzuführen sind, zu fragen, wie weit Lernende ohne eine sprachwissenschaftliche Vorbildung diese sinnvoll bewältigen können. Genauso muss bei Konkordanzanalysen gefragt werden, was die Lernenden sich dadurch auf welche Weise aneignen sollen, und ob diese Aneignung wirklich stattfindet.

Schließlich macht Krekeler (2021, 163–164) noch drei didaktische Themen aus: „Zum Konzept des entdeckenden Lernens“, „Zur Rolle der Lehrperson“ und „Corpus Literacy vermitteln“. Die ersten beiden Themen verbindet er u. a. mit dem Begriff der Lernerautonomie. Diese wird einerseits mit dem Konzept des entdeckenden Lernens in Beziehung gesetzt, bei dem Krekeler aber bezweifelt, dass dieses als didaktische Grundlage allein tragfähig ist. Auf der anderen Seite ist davon auszugehen, dass Lehrpersonen anfangs steuernd eingreifen müssen. Dies dürfte nicht zuletzt mit der zu vermittelnden *corpus literacy* zusammenhängen. Zu dieser gehört es u. a. zu wissen, welche – dem Spracherwerb dienlichen – Informationen aus einem Korpus gewonnen und wie Funde interpretiert werden können. Gerade das kann aber bei den Lernenden in der Regel nicht vorausgesetzt werden.

Im insgesamt positiven Fazit zu seiner Unterrichtseinheit weist Krekeler explizit darauf hin, dass selbst bei den von ihm unterrichteten, eher fortgeschrittenen Lernenden mit vorbereiteten Abfrageergebnissen, die dann aber eigenständig interpretiert wurden, gearbeitet werden musste. Auf dieser Grundlage ließ sich dann aber das erhoffte stimulierende entdeckende Lernen realisieren (Krekeler 2021, 190–191). In eine ähnliche Richtung weisen die Ergebnisse, die Link (2025) aus zwei korpuslinguistischen Seminaren mit italienischen Deutsch-Studierenden an einer italienischen Universität generieren konnte. Hier konnten die Lernenden im Anschluss an eine intensive (korpus-)linguistische Vorbereitung in Kleingruppen ein eigenes Projekt zu Frequenzen und Verwendungskonzepten von Neologismen durchführen. In der Auswertung zeigte sich, dass die Studierenden vor allem auf der methodischen Ebene und zu linguistischen Fragestellungen dazugelernt hatten. Dagegen schätzen sie den Mehrwert für den Spracherwerb eher gering ein. Dabei ist zu berücksichtigen, dass unterschiedliche Niveaustufen zusammentrafen und Italienisch die Arbeitssprache war (Link 2025, 146).

Im Anschluss an Krekeler (2021) und Link (2025) sind unseres Erachtens nach diesem knappen Überblick die folgenden Kriterien von zentraler Bedeutung für das Gelingen korpusgestützten Lehrens und Lernens im DaF-Unterricht:

Die Lernenden brauchen

- a) mindestens grundlegende korpuslinguistische Kenntnisse oder einen ihrem Kenntnisstand entsprechend aufbereiteten Zugang zu Korpusdaten,
- b) eine Aufgabenstellung, die sich erkennbar auf den eigentlichen Unterrichtsgegenstand bezieht, vor allem, wenn dieser nicht in erster Linie linguistischer Natur ist,
- c) Resultate aus den Aufgabenstellungen, die den Mehrwert korpuslinguistischer Methoden erkennen lassen.

Die Lehrenden brauchen fachliche und didaktische Kompetenzen, um die Lernenden an genau den genannten Punkten angemessen einschätzen und unterstützen zu können. Die entsprechende Ausbildung der Lehrenden sollte aus naheliegenden Gründen bspw. während eines DaF-Studiums oder auch – und das ist wohl noch eher ein Desideratum – im Rahmen von Fortbildungen geschehen.

4 Skizze zweier Seminare zum korpusgestützten Lehren und Lernen an der Universität Hamburg

Wie eine korpusdidaktische Ausbildung im Rahmen eines germanistischen Studiengangs in Deutschland gelingen kann und welche Fallstricke dabei lauern, soll nun anhand zweier Seminare des Erstautors skizziert werden. Die Zweitautorin nahm am zwei-

ten dieser Seminare teil und entwickelte daraus eine Lerneinheit für ein DaF-Tutorat an der Universität Vilnius, das im weiteren Verlauf des Beitrags noch näher beleuchtet wird.

Das erste Seminar zum Thema „Konnektivität: Korpusanalysen aus dem DaF/DaZ-Erwerb“ fand im Sommersemester 2022 an der Universität Hamburg statt. Das Seminar war für alle linguistischen und Lehramts-Masterstudiengänge des Instituts geöffnet und es gab eine optionale Übung zum Thema „Erstellung und Annotation von Sprachkorpora – Texte junger Schreibender linguistisch untersuchen“, die allerdings nur von sechs der am Seminar Teilnehmenden besucht wurde. Schon zu Beginn des Seminars zeigte sich, dass sowohl zum linguistischen Thema des Seminars – Konnektivität – als auch zur Korpuslinguistik sehr unterschiedliche und zum Teil sehr rudimentäre bzw. wie oben schon erwähnt eher negativ konnotierte Vorerfahrungen vorlagen. Daher wurde ungefähr das erste Drittel des Seminars auf die Auseinandersetzung mit dem Thema „Konnektivität“ verwendet. Dabei wurde im Anschluss an Rehbein, Hohenstein und Pietsch (2007) zunächst ein weiter Konnektivitätsbegriff zugrundegelegt, der nicht nur sprachliche Mittel wie Konjunktionen, sondern auch semantische Relationen und Bezüge zur Produktions- bzw. Rezeptionskonstellation berücksichtigt. Den Studierenden blieb freigestellt, ob sie in diesem Rahmen arbeiten oder sich auf Konnektoren nach der recht strikten Definition des Handbuchs der Konnektoren (HdK-1) (Pasch, Brause, Breindl & Waßner 2003) fokussieren wollten. Für die Korpusuntersuchungen bot sich Letzteres an, da eindeutig war, nach welchen Ausdrücken im Korpus zu suchen sein würde, und so standen Konnektoren im Sinne des HdK im weiteren Seminarverlauf im Zentrum.

Als Datengrundlage diente ein Korpus aus ca. 180 Texten ein- und mehrsprachiger Schülerinnen und Schüler an Hamburger Schulen, die zu zwei ca. zwölf Monate auseinanderliegenden Messzeitpunkten erhoben wurden. In der seminarbegleitenden Übung wurden auf Basis dieses Korpus parallel erste Überlegungen zur Datenaufbereitung als Voraussetzung einer korpuslinguistischen Untersuchung angestellt. Die Analyse der Korpusdaten folgte bspw. Fragestellungen, die sich an Bachmann (2005) anschließen. Demzufolge markieren jüngere Schülerinnen und Schüler – hier des 8. Schuljahrs – Verknüpfungen im Text oft expliziter als solche aus dem 10. Schuljahr. Eine Überverwendung des Konnektors *da* illustriert der folgende Auszug aus einem Text eines Schülers aus dem 8. Schuljahr³. Weitere Beispiele und Befunde zu den Erwerbsfragen werden in Grommes (2024) diskutiert.

- (1) *Er hebt den Stift auf und wirft ihn mir ins Gesicht. Da war ich schon auf 180. Da hab ich mein Lineal genommen und zu ihm geworfen, aber ihn verfehlt. Er nimmt mein Lineal vom Boden und zerbricht es in 2 Teile. Da kamen meine Kollegen*

³ Orthographie und Interpunktions folgen dem handschriftlichen Original des Schülers.

und sagten zu mir mit diesen Worten: „Oha Alpi, wenn ich du wäre hätte ich ihm richtig auf die Fresse gehauen. Da hab ich erst eine Weile überlegt. Da hab ich mich doch entschieden ihn zu schlagen.

Ein weiteres Drittel des Seminars widmete sich korpuslinguistischen Grundlagen auf der Basis von Lemnitzer und Zinsmeister (2015) sowie Andresen und Zinsmeister (2019). Das Schlussdrittel galt den Projekten der Studierenden, von denen aber nur ein Teil korpuslinguistische Fragestellungen bearbeitete, während die Übrigen bspw. Lehrmaterialien in Bezug auf die Vermittlung von Konnektivität untersuchten. Hier zeigt sich unseres Erachtens recht eindeutig, dass der korpuslinguistische Anteil des Seminars von vielen Studierenden als nicht ausreichend dafür angesehen wurde, sich korpuslinguistische Methoden und Tools soweit anzueignen, dass eigenständig damit gearbeitet werden kann.

Dieser unbefriedigende Befund wurde auf einem Workshop im Rahmen der GIP der Universitäten Hamburg und Vilnius diskutiert, wobei deutlich wurde, dass auch erfahrene DaF-Lehrende sehr zurückhaltend auf korpusgestütztes Lehren reagieren. Dabei wurde ein Faktor unterstrichen, den auch die Hamburger Studierenden nannten: Viele Tools sind aus Sicht fachfremder Nutzenden nicht besonders anwenderfreundlich. Wobei im Seminar *AntConc* (Anthony 2018) verwendet wurde, das eher auf der positiven Seite steht, während etwa *Annis* (Krause & Zeldes 2016) eher verunsichert. Schließlich wurde vereinbart, im Sommer 2023 ein vergleichbares Seminar durchzuführen, an dem – via *Zoom* – Lehrende aus Vilnius teilnehmen sollten. In diesem Seminar sollte von vorneherein eine klare Anwendungsperspektive gegeben werden. Daher hieß es in der Seminarankündigung explizit, dass gefragt wird,

... wie die Erkenntnisse und Methoden der Korpuslinguistik für den Unterricht insbesondere von Deutsch als Fremdsprache zugänglich gemacht werden können. Dazu ist zunächst zu eruieren, was korpusbasiertes Arbeiten in der Sprachvermittlung überhaupt heißen kann, um dann zu überlegen, wie Sprachlerende auf die Arbeit mit Korpora und dazugehörenden Tools vorbereitet werden können.

Im zweiten Seminar waren wieder im Prinzip alle Masterstudiengänge vertreten, allerdings mit zehn Studierenden nur halb so viele Teilnehmende wie im Jahr davor. In diesem Durchgang wurde nur ein knappes Viertel der Zeit auf das Thema „Konnektivität“ verwendet. Außerdem hatten ungefähr die Hälfte der Teilnehmenden korpuslinguistische Kenntnisse aus früheren Seminaren. Daher konnte recht frühzeitig in Überlegungen zur Nutzung von Korpora in der Sprachvermittlung eingestiegen werden. Tabelle 1 skizziert den zeitlichen Ablauf sowie die institutionelle Anbindung und beteiligte Zielgruppen.

2022					... Oktober
April	Mai	Juni	Juli		
Seminar: Konnektivität: Korpusdatenanalysen aus dem DaZ/ DaF-Erwerb (Universität Hamburg) (Teilnehmende: Studierende verschiedener Master-Studiengänge (Germanistische Linguistik, Allgemeine Sprachwissenschaft, Lehramt Deutsch) der Universität Hamburg)					
Übung: Erstellung und Annotation von Sprachkorpora – Texte junger Schreibender linguistisch untersuchen (Universität Hamburg)					
2023					September
April	Mai	Juni	Juli	...	September
Seminar: Konnektivität: Korpusdatenanalysen aus dem DaZ/ DaF-Erwerb (Universität Hamburg, in Kooperation mit Vilnius im Rahmen der GIP) (Teilnehmende: Studierende verschiedener Master-Studiengänge (Germanistische Linguistik, Allgemeine Sprachwissenschaft, Lehramt Deutsch) der Universität Hamburg; außerdem: DaF-Lehrende der Universität Vilnius)					Tutorat Anastasiia Stulen an der Universität Vilnius im Rahmen der GIP
	virtuelle Hospitationen Hamburger Studierender in DaF-Kursen in Vilnius	Workshop „Kor- pusgestützte Fachphraseolo- gie“ an der Uni- versität Ham- burg im Rahmen der GIP			darin: Erprobung der korpusgestützten Lerneinheit zu Konnektoren (Zielgruppe: DaF- Lernende)
		darin: Zwischen- bericht aus dem Kooperationsse- minar			

Tabelle 1. Zeitlicher Verlauf und Beteiligte der Seminare, Workshops und der Lerneinheit

Einblicke in die Entwicklung und zum Einsatz von korpusgestütztem Lehrmaterial im DaF-Unterricht boten insbesondere Beiträge aus dem eJournal *KorDaF* (Gredel 2021; Moroni 2021; Salzmann 2021; Meier-Vieracker 2022; Wiesmann 2022). Durch (hybride) Hospitationen in Seminaren in Vilnius sollte ein Zugang zur unmittelbaren

DaF-Vermittlung ermöglicht werden. Als Ergebnis des Seminars in Hamburg sollten die Studierenden eigene Vorschläge für korpusgestützte Lerneinheiten entwickeln, die perspektivisch im Rahmen der zuvor genannten GIP in Form von Praktika oder Tutoraten in Vilnius erprobt werden konnten.

Insgesamt wurden im Seminar an der Universität Hamburg im Sommer 2023 sieben Konzepte für Lerneinheiten erarbeitet. Diese werden in der folgenden Übersicht kurz skizziert. Die ersten drei Konzepte nutzen Korpora zur gesprochenen Wissenschaftssprache⁴ (GeWiss). Das vierte Konzept stützt sich auf das DWDS, im fünften wird auf das Vorgehen in Meier-Vieracker (2021) Bezug genommen und im sechsten auf Wiesmann (2022).

- 1) Schriftsprachliche Merkmale der gesprochenen Wissenschaftssprache ermitteln nach vorheriger Einführung in korpuslinguistische Arbeitsweisen und Werkzeuge.
- 2) Übungen zur Verwendung von *also* in Präsentationen von Deutsch als L1-Sprechenden auf der Basis von aus dem Korpus gewonnenen Materialien.
- 3) Übungen zum Hörverstehen mit Fokus auf *also* auf Basis authentischer Daten.
- 4) Arbeit mit dem DWDS-Wortprofil in einer Lerneinheit zu Präpositionen.
- 5) Themen- und interessenspezifische Erkundung der syntaktischen Merkmale von Konnektoren auf Basis eines Korpus zur Sprache in Fußballkommentaren.
- 6) Überprüfung automatischer Übersetzungen von Rechtstexten mit Hilfe des GeLeCo-Korpus.

Außerdem ist als siebtes Konzept die Lerneinheit „Verbstellung in Kausalsätzen mit *denn, weil, da*“ entstanden, die im Rahmen eines Tutorats in Vilnius im Herbst 2023 erprobt und durch die dortigen Studierenden evaluiert wurde. Das Projekt wurde zunächst im Rahmen einer Hausarbeit dokumentiert. Nun soll es in diesem Beitrag in den Abschnitten 5 und 6 vorgestellt und diskutiert werden.

5 Lerneinheit zur Verbstellung in Kausalsätzen mit *denn, weil, da*

Die Lerneinheit zur Verbstellung in Kausalsätzen wurde für den Bachelorkurs „Wirtschaftsdeutsch“ an der Universität Vilnius konzipiert. Die Zielgruppe bestand aus 14 mehrsprachigen Studierenden des dritten, fünften und siebten Fachsemesters. Sie hatten unterschiedliche Sprachniveaus zwischen A2+ und B2+ und befanden sich in verschiedenen Studiengängen. Die Mehrzahl studierte Deutsche Philologie als Hauptfach, die anderen hatten Deutsche Sprache und Kultur als Nebenfach oder lernten Deutsch als eine zweite Fremdsprache in der Fachpädagogik (Englisch/Litauisch und Deutsch).

⁴ GeWiss – Gesprochene Wissenschaftssprache, abrufbar unter: gewiss.uni-leipzig.de.

Durch die Hospitationen im Rahmen des Tutorats in verschiedenen Seminaren, darunter auch in den Kursen „Gegenwartsdeutsch“ und „Wirtschaftsdeutsch“, konnten die Lernprozesse der Studierenden in authentischen Unterrichtssituationen systematisch beobachtet werden. Die dabei gewonnenen Erkenntnisse sowie die Analyse der Arbeit der Studierenden im Kurs „Wirtschaftsdeutsch“ beeinflussten den Aufbau und die Gestaltung der Aufgabentypen der Lerneinheit zu kausalen Konnektoren so, dass das Unterrichtsgeschehen interaktiv und eher nach gewohnten Mustern für die Studierenden abließ.

Im Fokus der Lerneinheit standen die Konnektoren *denn*, *weil* und *da*, für deren Auswahl es zwei Gründe gibt. Zum einen werden in den Lehrwerken „Begegnungen“, mit denen im DaF-Unterricht für die Niveaustufen A1+ bis B1+ an der Universität Vilnius wie dem Kurs „Gegenwartsdeutsch“ gearbeitet wird, ausschließlich *denn* und *weil* als kausale Konnektoren vermittelt (siehe Buscha & Szita 2013; 2013a; 2013b). *Da* war den Studierenden nur als temporaler Konnektor bekannt. Wie Breindl (2018, 30) zeigt, verwenden Lernende den Konnektor *da* gerade in argumentativen Texten wenig bis gar nicht im Gegensatz zu Deutsch als Erstsprache Sprechenden, die diesen Konnektoren sogar gegenüber *weil* bevorzugen. So weicht die Verwendungshäufigkeit dieser Konnektoren bei DaF-Lernenden auffällig von der bei Erstsprachensprechenden des Deutschen ab. Das ist problematisch, weil argumentative Texte, wie sie auch für Wirtschaftsdeutsch relevant sind, sich nach Breindl (2004, 432) durch eine hohe Frequenz an Konnektoren auszeichnen. Der zweite Grund für die Auswahl dieser drei Konnektoren besteht darin, dass ihre Verwendungsbeschränkungen nicht offensichtlich sind, sich aber durch Beispiele erschließen lassen. Somit hatte die Lerneinheit zum Ziel, den Lernenden eine größere Variationsbreite beim Konnektoreneinsatz zu ermöglichen und sie für die unterschiedlichen Gebrauchsbedingungen der drei Konnektoren zu sensibilisieren. Da *weil* und *da* als Subjunktionen Nebensätze mit Verbendstellung einleiten, während *denn* in Hauptsätzen mit Verbzweitstellung auftritt, wurde zudem die Notwendigkeit gesehen, diese syntaktischen Beschränkungen in die Lerneinheit aufzunehmen.

Vor diesem Hintergrund ergab sich eine vierstufige Aufgabenfolge, die in eine Unterrichtseinheit im Umfang von zweimal neunzig Minuten eingebettet war. Zunächst wurde das Vorwissen der Lernenden zu den Konnektoren aktiviert und darauf aufbauend wurden die Unterschiede zwischen den drei fokussierten Ausdrücken erarbeitet. In der ersten Aufgabe sollten die Lernenden Konnektoren in einem wirtschaftsjournalistischen Text auffinden und deren syntaktische Positionierung erfassen. Der Text entstammte dem Subkorpus „Wirtschaftsblatt 2000–2016“ des DeReKo (IDS 2022), aus dem für die Zwecke der Unterrichtseinheit wiederum ein Teilkorpus „Wirtschaftsblatt 2016“ gebildet wurde. Da das DeReKo (IDS 2022) über COSMAS II abgerufen und im Unterricht nur das Ergebnis der Abfrage, also der Text im Hand-Out, verwendet wurde, wird die

Schnittstelle weder im Unterrichtsmaterial noch hier erläutert. Es diente lediglich als Quelle des Textes in der Aufgabe 1 sowie der Sätze der Aufgaben 2 und 3.

In der zweiten Aufgabe sollten semantische und syntaktische Merkmale der Konnektoren erkannt und zugeordnet werden. Dabei wurde auf die Wortfolge und Verbstellung aufmerksam gemacht. Anschließend wurden in der Aufgabe 3 anhand des von Helbig und Buscha (1999, 448) verwendeten Positionsmodells des deutschen Satzes, das dem Feldermodell ähnlich ist und den Studierenden in der Begrifflichkeit vertraut war, aus dem Korpus vorausgewählte Sätze analysiert.

Aufgabe 3: Lest die Sätze. Markiert die Grenzen zwischen den Teilsätzen. Ergänzt die Tabelle.

Bemerkung: für jeden Teilsatz gibt es eine Zeile. Manche Felder können leer bleiben.

0. Das sei ihm allerdings schwergefallen, denn all seine Kühe kenne er beim Namen.
1. Doch das ist reine ökonomische Fantasie, denn der Aktivitätsgrad der Österreicher ist seit vielen Jahrzehnten konstant.
2. Rund 16 Prozent der Unternehmen setzen auf externe Anbieter, weil die IT-Abteilung nicht zeitnah liefern kann – 2015 waren es knapp 13 Prozent.
3. Da die österreichischen Bauern nur zwei Prozent zur gesamten EU-Milchmenge beitragen, würde ein theoretischer Milchliefer-Stopf nichts beeinflussen.

	Position 0	Position 1	Position 2	Position 3	Position (...n)	Satzende
0.		Das	sei	ihm	allerdings	schwergefallen,
	denn	all seine Kühe	kenne	er	beim Namen.	
1.						
2.						

Abbildung 1. Ausschnitt aus der Aufgabe 3 zur Wortstellung in Haupt- und Nebensätzen mit kausalen Konnektoren (Positionsmodell nach Helbig und Buscha 1999, 448)

Die Schritte, die in der dritten Aufgabe durchgegangen werden müssen, sollten die Studierenden in die Lage versetzen, die richtige Wortfolge und Verbstellung bei der Verwendung der unterschiedlichen Kausalkonnektoren einzuhalten und in der vierten Aufgabe eigenständig Konnektoren in Korpusdaten finden sowie ihre Eigenschaften anschließend ausgehend von den KWiC-Belegen im Plenum vorstellen zu können.

In der Aufgabe 4 wurde auf das DWDS zurückgegriffen, weil die Suche mit dieser Korpuschnittstelle nach einer kurzen Einführung schon gut durchführbar ist. Dementsprechend wurden im Unterricht nur einfache Begriffe wie *KWiC* bzw. *keyword in context* erläutert und die grundlegende Logik der Suchmaske wie der Ergebnisansichten erklärt (siehe Abbildung 2). So waren die Studierenden schnell in der Lage, selbstständig zu arbeiten und sich mit den Daten zu befassen.

Wortschatz zur Aufgabe 4:

1. **das Korpus, die Korpora:** ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Diese Daten sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert. Die Daten sind typischerweise über spezialisierte Suchmaschinen mit linguistischen Abfragewerkzeugen abfragbar;
2. **die Suchabfrage, -n:** ein Prozess, bei dem der Suchmaschine der Suchausdruck übermittelt wird;
3. **der Treffer, -:** ein Korpussegment, das auf das Muster eines Suchausdrucks passt;
4. **der Beleg, -e:** ein Korpussegment, das als Ergebnis eines erfolgreichen Matches zurückgegeben wird;
5. **KWiC (Keyword in Context)** oder **das Schlüsselwort:** der Beleg, der rechts und links mit dem Text umgebenden Zeichenkette;
6. **die Kollokation/Kookkurrenz, -en:** feste Wortverbindung, die im Text häufiger als andere vorkommt.

Mehr dazu Lemnitzer & Diewald 2022, 374–390.

Korpusbelege DWDS-Kernkorpus (1900–1999)

Belege

Treffer

Keyword (Schlüsselwort)

Reihenfolge	Jahr	Textausschnitt	Rechtskontext
1: b 1999	Degenhardt, Fra	zweier namhafter, der Familie verpflichteter Soziologen der hiesigen Universität	zufolge, und das überzeugte schließlich die entscheidenden Herren ...
2: b 1999	Engler, Wolfgang	als Oberschüler Gasthörer bei berühmten Philosophen an der Berliner Universität	, verschlug es ihn dann doch noch in den Krieg, zur Ostfront, von der...
3: b 1999	Engler, Wolfgang	andere gesetzt und eine kleine Widerstandsguppe an der Bonner Universität	gegründet.
4: b 1999	Engler, Wolfgang	Künstlerverbände, Hochschulen, Universitäten	und wissenschaftliche Akademien schlossen sich der allgemeinen W...
5: b 1999	Engler, Wolfgang	»Ich war damals in Greifswald an der Universität	und hatte meine beste DDR-Zeit, das heißt, ich war eigentlich nie so ...
6: b 1999	Engler, Wolfgang	der Umstände zufrieden mit mir wie Anfang der 60er Jahre dort an der Universität	<
7: b 1999	Engler, Wolfgang	der siebziger Jahre begegnete man den einen an Hochschulen und Universitäten	, wo sie diplomierten und promovierten.
8: g 1999	Hannover, Hein	des angesehene Wirtschaftswissenschaftlers Prof. Dr. Rudolf Hickel	Bremen) mit, das sich kritisch mit dem Konzentrationsprozeß im Ba...
9: g 1999	Hannover, Hein	25.4.1977 in den göttinger nachrichten, einer vom AstA der Göttinger Universität	herausgegebenen Zeitung, veröffentlichte.
10: g 1999	Hannover, Hein	aus Solidarität mit dem in die Schußlinie geratenen AstA der Universit...	der lagegespie

Abbildung 2. Arbeitsblatt mit Glossar und Einblick in Ergebnisse der Korpusabfrage im DWDS-Kernkorpus (Stand: September 2023)

Die Herangehensweise in dieser Unterrichtseinheit war bewusst sehr niedrigschwellig gewählt. Ein sehr kleines Teilkorpus mit wirtschaftsjournalistischen Texten wurde für

die Vorbereitung der Einheit gebildet und genutzt, um unmittelbar mit authentischen und für den Seminargegenstand relevanten Beispielen arbeiten zu können. Diese Beispiele waren kuratiert, um das Vorwissen der Studierenden zu berücksichtigen, wie Krekeler (2021, 190–191) es fordert. Sicherheitshalber wurde mit weiteren Lehrenden aus Vilnius ein Glossar zu den Texten erarbeitet, mit denen die Studierenden im Kurs „Wirtschaftsdeutsch“ ggf. vorhandene Wortschatzlücken füllen konnten. Die folgenden Beispiele repräsentieren einen Auszug aus dem Glossar:

- (2) **(stark) rudern** (umgangspr.): *sich bemühen oder Anstrengungen unternehmen.*
- (3) **der milchintensivste Monat:** *der Monat, in dem besonders viel Milch produziert wird.*
- (4) **kalben:** *der Prozess, bei dem eine Kuh ein Kalb zur Welt bringt.*
- (5) **der Strafabschlag, Strafabschläge:** *Geld, das von einem Betrag abgezogen wird, normalerweise als Strafe.*
- (6) **Kredite bedienen:** *Rückzahlungen leisten, normalerweise auf einen Kredit bezogen.*

Für die selbstständige Arbeit der Studierenden wurde mit dem DWDS ein leicht zugängliches Korpus gewählt. So musste ihnen zusätzlich zu den Unterrichtsinhalten nicht noch ein Korpusanalysetool erklärt werden und sie kamen relativ schnell zu Erfolgserlebnissen. Die Erwartung war an dieser Stelle, dass die Studierenden so zur eigenständigen Arbeit mit Korpora im weiteren Verlauf ihres Studiums motiviert werden können.

6 Auswertung und Diskussion der Lerneinheit

Wie im Kapitel 3 schon erwähnt, haben die Studierenden in der Untersuchung von Link (2025) vor allem auf methodischer und linguistischer Ebene dazugelernt, während sie den Zuwachs ihrer DaF-Kompetenz geringer einschätzten. Aus der fragebogenbasierten Evaluation der in Vilnius durchgeführten Unterrichtseinheit ergibt sich ein dazu komplementäres Bild.

In dem Fragebogen, der aus offenen und multiple-choice Fragen bestand, wurden die Studierenden um eine Einschätzung ihres Vorwissens zu Konjunktionen und zur Korpuslinguistik befragt und am Ende der Lerneinheit um eine Einschätzung ihres Lernzuwachses gebeten. Dabei zeigte sich, dass acht der vierzehn Teilnehmenden ihr Vorwissen zu den Kausalkonnektoren eher gut bis sehr gut einschätzten (siehe Abbildung 3). Das passt dazu, dass zwölf Teilnehmende mindestens im dritten Studienjahr waren und außerdem über die Hälfte von ihnen Deutsche Philologie im Hauptfach studierte.

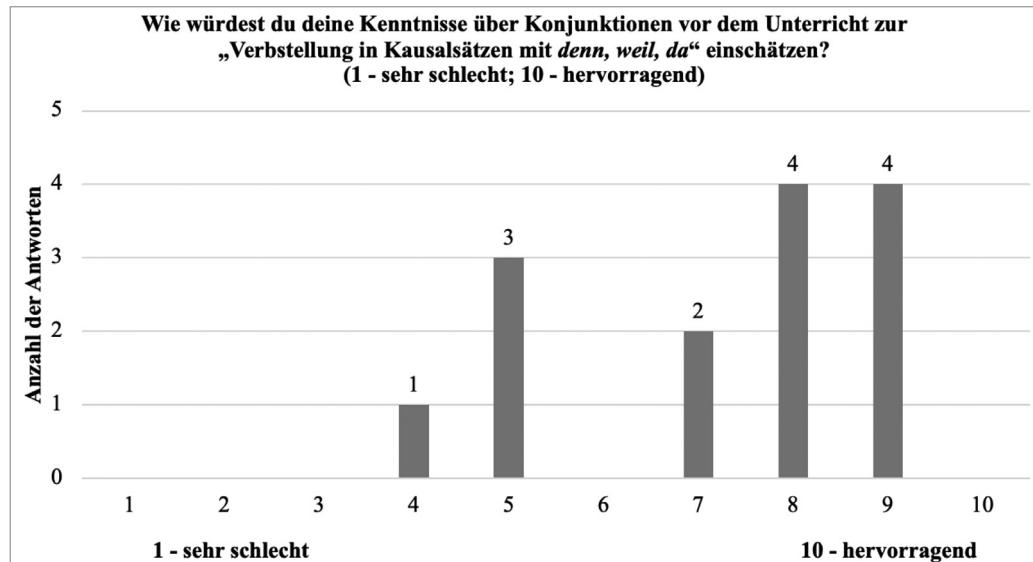


Abbildung 3. Selbsteinschätzung der Studierenden vor der Lerneinheit (n = 14)

Trotz dieses hohen Ausgangsniveaus verschoben sich die Einschätzungen nach der Einheit nochmal etwas nach oben (siehe Abbildung 4).

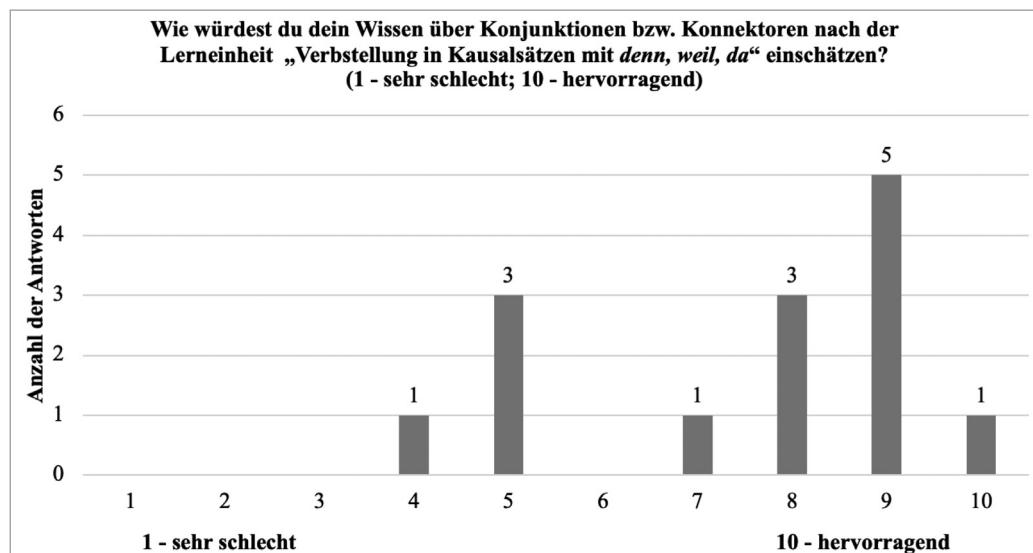


Abbildung 4. Selbsteinschätzung der Studierenden nach der Lerneinheit (n = 14)

Die Textvorlagen wurden insgesamt als sehr nützlich eingeschätzt, wobei sie aber von einigen Studierenden im Freitext-Kommentar als zu einfach beanstandet wurden. Wesentlich schwieriger schien die Auseinandersetzung mit dem Positionsmodell zu sein. Hier wurde aber hervorgehoben, dass dieses nützlich sei, um die Besonderheiten des Deutschen gegenüber dem Litauischen besser zu verstehen.

Zum korpuslinguistischen Anteil der Einheit sagten zwölf der vierzehn Teilnehmenden, dass die Erläuterung grundlegender korpuslinguistischer Begriffe und Konzepte für sie notwendig gewesen sei. Das ist erstaunlich, weil zugleich die Hälfte der Teilnehmenden angab, schon Erfahrungen mit korpuslinguistischem Arbeiten in DaF gemacht zu haben. Die eigentliche korpuslinguistische Aufgabe, die direkte Arbeit mit Korpora, wurde von acht Teilnehmenden als erfolgreich bzw. nützlich bezeichnet. Zwei Lernende sahen keinen Nutzen und vier äußerten sich nicht dazu. Als nützlich wurden einerseits der methodische Einblick und andererseits der Zugriff auf authentische Beispiele eingeschätzt.

Betrachtet man nun noch einmal die Kriterien für erfolgreiches korpusgestütztes Lehren und Lernen aus Abschnitt 3, so ist zunächst festzuhalten, dass Merkmal a), das u. a. grundlegende korpuslinguistische Kenntnisse der Lernenden fordert, als erfüllt betrachtet werden kann. Auch wenn viele der Studierenden schon über erste korpuslinguistische Erfahrungen verfügten, sahen doch nahezu alle die methodische Hinführung zur korpuslinguistischen Übung als notwendig an. Über die Hälfte konnten diese dann aber eigenständig durchführen. Im Übrigen wurde ein leichter Zugang durch die Vorauswahl der Texte bzw. den Rückgriff auf die Suchmaske des DWDS ermöglicht.

Merkmal b) fordert, dass Aufgabenstellungen einen klaren Bezug zum Unterrichtsgegenstand haben sollen. Auch dies wurde hier erfüllt, da einerseits zum Thema passende Texte zur Verfügung standen und andererseits sprachliche Ausdrücke und Strukturen behandelt wurden, die u. a. für diese Art Texte kennzeichnend sind (vgl. Abbildung 5).

Mit Merkmal c) wird schließlich verlangt, dass die Resultate aus den Aufgabenstellungen den Mehrwert korpuslinguistischer Methoden erkennen lassen. Dieses Merkmal wurde nur teilweise erfüllt. Auch ohne die korpuslinguistischen Übungen hätten die Studierenden auf Basis der vorausgewählten Texte in vergleichbarer Weise arbeiten und über die Konnektorenverwendung reflektieren können. Allerdings sahen zumindest einige der Studierenden die Auseinandersetzung mit den selbst gefundenen Beispielen als Gewinn an, ergänzend sahen einige in den methodischen Kompetenzen Potenzial für ihr weiteres Lernen. Auf der Seite der Tutorin zeigte sich, dass relativ unkompliziert authentische Beispiele generiert und Alternativen zu herkömmlichen Übungsformen in den Unterricht integriert werden können. Selbst wenn diese keinen übermäßigen Lernzuwachs im Vergleich zum Hergebrachten bewirken, sorgen sie für Abwechslung und damit Aktivierung und Motivierung der Lernenden. Die Bedeutung von Motivation als Antriebsfaktor

Konjunktion	Funktion/Bedeutung	Beispiel	Verbstellung
und	verbindet Sätze	Nun hat Sportminister Gerald Klug wieder rund vier Millionen € an Spitzensportförderung verteilt. <u>und</u> wieder brach das Jammern wegen angeblicher Ungerechtigkeiten <u>aus</u> .	Position 2
denn	leitet einen Hauptsatz ein, der vorher Gesagtes begründet	Das polnische Sozialversicherungsamt ZUS ist dieser Tage alarmiert. <u>denn</u> dem Sozialversicherungsfonds FUS <u>werden</u> bis 2020 gut 100 Milliarden € für Pensionen fehlen.	Position 2
weil	leitet einen kausalen Nebensatz ein; muss in der Antwort auf eine direkte Frage gebraucht werden	Warum gerade eine Milliarde €? Woher kommt diese Zahl? <u>Weil</u> es einen Einmaleffekt in dieser Höhe gab.	Verb am Ende
da	leitet einen kausalen Nebensatz ein; steht meistens am Satzanfang	<u>Da</u> sich das gesamte Gebäude auf einer Ebene <u>befindet</u> , lässt sich die Barrierefreiheit leichter umsetzen.	Verb am Ende

Abbildung 5. Tabellenübersicht zu Konjunktionen Lösung zur Aufgabe 2

wird u. a. in Kormos und Dörnyei (2004) belegt. Allerdings weist Kleppin (2004, 10) darauf hin, dass es nicht die eine anregende Lernumgebung oder Aufgabengestaltung für alle Lernenden einer Lerngruppe gibt. Die von ihr referierte Befundlage zeigt aber, dass wechselnde Lernumgebungen eher motivationsfördernd sind. Dies ist unseres Erachtens durchaus auch als ein Mehrwert des korpuslinguistischen Arbeitens zu verstehen.

7 Fazit

In diesem Beitrag haben wir die Entstehung einer korpusgestützten DaF-Lerneinheit aus einem Seminar heraus bis hin zu ihrer praktischen Erprobung vor dem Hintergrund des aktuellen Diskussionsstandes zur Einbindung korpuslinguistischer Methoden und Werkzeuge in die Vermittlung von DaF nachvollzogen. Dabei sollte nicht außer Acht gelassen werden, dass die hier in den Abschnitten 5 und 6 vorgestellte Lerneinheit nicht das einzige Produkt des vorhergehenden Seminars ist. Insgesamt sind sieben Unterrichtsvorschläge von unterschiedlichem Umfang entwickelt worden, von denen aus Ressourcengründen nur einer erprobt werden konnte. Dieses Ergebnis zeigt aber, dass die korpusdidaktische Diskussion nicht nur seit etwa 2021 wieder Fahrt aufgenommen hat. Da viele der Unterrichtsvorschläge von Beiträgen aus der Fachliteratur inspiriert sind, ist davon auszugehen, dass ein zunehmender Transfer in die Praxis gelingt. Damit kann an Krekeler (2021, 191) anschließend festgehalten werden, dass die „Korpusrevolution im Fremdsprachenunterricht“ vielleicht ausgeblieben ist, eine Evolution aber offensichtlich stattfindet.

Evolution ist allerdings ein Prozess der kleinen Schritte und größeren Zeiträume. Auf den ersten Blick ist der zeitliche Rahmen des hier Berichteten sogar relativ knapp gehalten. Blendet man den ersten Seminarversuch aus, hat alles innerhalb eines Semesters stattgefunden. Es ist aber fraglich, ob ohne das vorausgehende Seminar an der Universität Hamburg die Vorbereitung und Durchführung der Lerneinheit an der Universität Vilnius im gleichen zeitlichen Rahmen möglich gewesen wäre. Und nicht zuletzt hat unsere Diskussion der Evaluation gezeigt, dass zwar ein insgesamt positiver Eindruck bleibt, aber kein sprunghafter Fortschritt erzielt wurde. Auf der Haben-Seite steht sicherlich, dass es möglich ist, in diesem zeitlichen Rahmen zumindest auf Seiten der – zukünftigen – Lehrenden hinreichend *corpus literacy* aufzubauen, sodass die Arbeit mit Korpora und den dazugehörigen Methoden und Tools als sinnvolle Ergänzung der eigenen Lehrkompetenz begriffen wird und ein erster Praxistransfer möglich wird.

Dieser Zugewinn ist aber nur systematisch und nachhaltig zu haben, wenn entsprechende Angebote kontinuierlich und zuverlässig im Lehrangebot von germanistischen und DaF-Studiengängen erscheinen. Die beiden hier diskutierten Seminare haben sehr klar gezeigt, dass auch an stark empirisch arbeitenden Standorten noch nicht davon ausgegangen werden kann, dass alle interessierten Studierenden erreicht werden und dass unter diesen noch große Verunsicherung auszumachen ist. Dies setzt sich bei schon in der Praxis stehenden DaF-Lehrenden fort. Insofern wäre sowohl eine konsequente methodische Grundbildung als auch ein entsprechendes Fortbildungsangebot wünschenswert.

Author contributions

Patrick Grommes: conceptualization, methodology, formal analysis, investigation, writing – original draft, writing – review & editing. **Anastasiia Stulen:** methodology, formal analysis, investigation, writing – original draft, writing – review & editing, visualization.

Datenquellen

- DWDS. *DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftsysteem zur deutschen Sprache in Geschichte und Gegenwart*. Hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften. Verfügbar unter: <https://www.dwds.de/>.
- IDS. 2022. *DeReKo. Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2022-I*. Release vom 08.03.2022. Mannheim: Leibniz-Institut für Deutsche Sprache. PID: 00-04B6-B898-AD1A-8101-4.
- Meier-Vieracker, Simon, ed. 2021. *Korpora zur Fußballlinguistik (Release 2021-09)*. Professur für Angewandte Linguistik, TU Dresden. Verfügbar unter: www.fussballlinguistik.de/korpora.

Literatur

- Andresen, Melanie & Heike Zinsmeister. 2019. *Korpuslinguistik*. Tübingen: Narr Francke Attempto.
- Anthony, Lawrence. 2018. *AntConc (Version 3.57)*. Tokyo: Waseda University. Available at: www.laurenceanthony.net/software/.
- Bachmann, Thomas. 2005. Kohärenzfähigkeit und Schreibentwicklung. *Literale Textentwicklung*. Helmuth Feilke & Regula Schmidlin, eds. Frankfurt: Peter Lang. 155–183. <https://doi.org/10.3726/978-3-653-01158-6>
- Beißwenger, Michael, Lothar Lemnitzer & Carolin Müller-Spitzer, eds. 2022. *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium*. Paderborn: Brill & Fink. <https://doi.org/10.36198/9783838557113>
- Breindl, Eva. 2004. Konnektoren in Übungsgrammatiken. *Übungsgrammatiken Deutsch als Fremdsprache. Linguistische Analysen und didaktische Konzepte*. Peter Kühn, ed. Regensburg: FaDaF. 426–458.
- Breindl, Eva. 2018. Konnexion in argumentativen Texten von DaF-Lernern und Muttersprachlern. *Deutsche Sprache. Zeitschrift für Theorie, Praxis, Dokumentation*. Ludwig Eichinger & Angelika Linke, eds. IDS Mannheim, Mannheim: Erich Schmidt Verlag. 22–36.
- Buscha, Anne & Szilvia Szita. 2013. *Begegnungen. Integriertes Kurs- und Arbeitsbuch. Sprachniveau A1+*. 2., unveränderte Auflage. Leipzig: SCHUBERT-Verlag.
- Buscha, Anne & Szilvia Szita. 2013a. *Begegnungen. Integriertes Kurs- und Arbeitsbuch. Sprachniveau A2+*. 2., unveränderte Auflage. Leipzig: SCHUBERT-Verlag.
- Buscha, Anne & Szilvia Szita. 2013b. *Begegnungen. Integriertes Kurs- und Arbeitsbuch. Sprachniveau B1+*. 2., veränderte Auflage. Leipzig: SCHUBERT-Verlag.
- Fandrych, Christian, Elena Frick, Hanna Hedeland, Anna Iliash, Daniel Jettka, Cordula Meißner, Thomas Schmidt, Franziska Wallner, Kathrin Weigert & Swantje Westpfahl. 2016. User, who art thou? User Profiling for Oral Corpus Platforms. *LREC 2016, Tenth International Conference on Language Resources and Evaluation. May 23–28, 2016, Grand Hotel Bernardin Conference Center, Portorož, Slovenia*. Nicoletta Calzolari, ed. Paris: European Language Resources Association. 280–287. Verfügbar unter: <https://aclanthology.org/L16-1043/>. Letzter Zugriff: 12. November 2025.
- Fandrych, Christian & Erwin Tscherner. 2007. Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivwechsel. *Deutsch als Fremdsprache* 44 (4), 195–204. <https://doi.org/10.37307/j.2198-2430.2007.04.02>
- Farroni, Cristina. 2024. Korpora im DaF-Unterricht: Einsatz im Schulbereich. *Linguistik Online* 126 (2), 83–98. <https://doi.org/10.13092/lo.126.11044>
- Florou, Katerina. 2025. Harnessing the Power of Corpus Linguistics in Language Education: A Student-Centered Approach. *Journal of Contemporary Language Research* 4 (2), 16–22. <https://doi.org/10.58803/jclr.v4i2.132>

- Gredel, Eva. 2021. Wikipedistik in Vermittlungskontexten des DaF-/ DaZ-Unterrichts: Kollaborative Textproduktion mithilfe von CMC-Korpora analysieren und reflektieren. *Korpora Deutsch als Fremdsprache* 1 (2), 74–94. <https://doi.org/10.26083/tuprints-00020711>
- Grommes, Patrick. 2024. Herausforderungen bei der Aneignung konzessiver Konnektoren bei Schüler*innen mit Deutsch als Erst- oder Zweitsprache. *Spracherwerb in DaZ und DaF – Forschung, Didaktik, Praxis*. Sabrina Ballestracci & Silvia Introna, eds. Berlin: Frank & Timme. 113–134. https://doi.org/10.57088/978-3-7329-9016-0_5
- Helbig, Gerhard & Joachim Buscha. 1999. *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. 19. Auflage. Leipzig & Berlin: Verlag Enzyklopädie & Langenscheidt.
- Kleppin, Karin. 2004. „Bei dem Lehrer kann man ja nichts lernen.“ Zur Unterstützung von Motivation durch Sprachlernberatung. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 9 (2). Verfügbar unter: <https://zif.tujournals.ulb.tu-darmstadt.de/article/id/2756/>. Zugriff: 3. Dezember 2025.
- Kormos, Judit & Zoltán Dörnyei. 2004. The interaction of linguistic and motivational variables in second language task performance. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 9 (2). Verfügbar unter: <https://zif.tujournals.ulb.tu-darmstadt.de/article/id/2749/>. Zugriff: 3. Dezember 2025.
- Krekeler, Christian. 2021. Korpustraining und datengestütztes Lernen im Wirtschafts-deutsch-Unterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 161–195. Verfügbar unter: <https://zif.tujournals.ulb.tu-darmstadt.de/article/id/3326/>. Zugriff: 3. Dezember 2025.
- Krekeler, Christian. 2021. Korpustraining und datengestütztes Lernen im Wirtschafts-deutsch-Unterricht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 26 (1), 161–195. Verfügbar unter: <https://zif.tujournals.ulb.tu-darmstadt.de/article/id/3326/>. Zugriff: 3.12.2025.
- Krekeler, Christian. 2023. Korpuslinguistik und Fremdsprachenunterricht [Sammelrezension]. *Informationen Deutsch als Fremdsprache* 50 (2–3), 120–124. <https://doi.org/10.1515/infodaf-2023-0012>
- Krause, Thomas & Amir Zeldes. 2016. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities* 31 (1), 118–139. <https://doi.org/10.1093/llc/fqu057>
- Lemnitzer, Lothar & Nils Diewald. 2022. Abfrage und Analyse von Korpusbelegen. *Forschen in der Linguistik. Eine Methodeneinführung für das Germanistik-Studium*. Michael Beißwenger, Lothar Lemnitzer & Carolin Müller-Spitzer, eds. Paderborn: Brill & Fink. 374–390. <https://doi.org/10.36198/9783838557113>
- Lemnitzer, Lothar & Heike Zinsmeister. 2006. *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

- Lemnitzer, Lothar & Heike Zinsmeister. 2015. *Korpuslinguistik. Eine Einführung*. 3. Auflage. Tübingen: Narr.
- Link, Sabrina. 2025. Data driven learning with students of German: A practical example of a small-scale corpus-based research project. *Korpora Deutsch als Fremdsprache* 5 (1), 138–149. <https://doi.org/10.48694/kordaf.4113>
- Lüdeling, Anke & Maik Walter. 2009. Korpuslinguistik für Deutsch als Fremdsprache. Sprachvermittlung und Spracherwerbsforschung. Stark erweiterte Fassung von Anke Lüdeling & Maik Walter (2010) Korpuslinguistik. *Deutsch als Fremdsprache*. HSK 35. Mouton de Gruyter: Berlin. Verfügbar unter: <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/LuedelingWalterDaF.pdf>
- Meier-Vieracker, Simon. 2022. Fußballwortschatz digital. Korpuslinguistische Ressourcen für den Sprachunterricht. *Korpora Deutsch als Fremdsprache* 2 (1), 7–24. <https://doi.org/10.48694/kordaf-3485>
- Moroni, Manuela. 2021. Zum Einsatz von Daten aus FOLK und DeReKo für die Untersuchung und Didaktisierung der Formen *nach dem Motto* und *von wegen*. *Korpora Deutsch als Fremdsprache* 1 (1), 51–78. <https://doi.org/10.48694/tujournals-39>
- Nolting, Antje & Nadja Radtke. 2019. Korpusbasierte Lexikografie: Nutzung von Korpora und Analysewerkzeugen im Unterricht für Deutsch als Muttersprache und Fremdsprache. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 24 (1), 107–126. <https://doi.org/10.26083/tuprints-00013091>
- Pasch, Renate, Ursula Braußé, Eva Breindl & Ulrich Hermann Waßner. 2003. *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfung (Konjunktionen, Satzadverbien und Partikeln)*. Berlin u. a.: De Gruyter. (HdK-1)
- Rehbein, Jochen, Christiane Hohenstein & Lukas Pietsch. 2007. Connectivity as an object of linguistic research in multilingualism. *Connectivity in Grammar and Discourse*. Jochen Rehbein, Christiane Hohenstein & Lukas Pietsch, eds. Amsterdam & Philadelphia: John Benjamins. 1–18. <https://doi.org/10.1075/hsm.5.02reh>
- Salzmann, Katharina. 2021. Zum Einsatz des GeWiss-Korpus im Rahmen einer mehrsprachigkeitsdidaktischen Vermittlung der gesprochenen Wissenschaftssprache. *Korpora Deutsch als Fremdsprache* 1 (1), 79–96. <https://doi.org/10.48694/tujournals-40>
- Wiesmann, Eva. 2022. Rechtssprachenbezogene Wissensvermittlung. Zum Einsatz von GeLeCo in der Übersetzausbildung. *Korpora Deutsch als Fremdsprache* 2 (1), 42–60. <https://doi.org/10.48694/kordaf-56>
- Zinsmeister, Heike, Carla Sökefeld & Skaistė Volungevičienė. 2025. Korpusdidaktik für formelhafte (Fach-)Sprache: Erfahrungen aus der Institutspartnerschaft zwischen Hamburg und Vilnius. *Informationen Deutsch als Fremdsprache* 52 (1), 45–59. <https://doi.org/10.1515/infodaf-2025-0009>

Daiktavardžio *pyktis* valentingumas Dabartinės lietuvių kalbos tekstyne

Veslava Čižik-Prokaševa

Bendrinės kalbos tyrimų centras

Lietuvių kalbos institutas

P. Vileišio g. 5

10308 Vilnius, Lietuva

El. paštas: veslava.prokaseva@lki.lt

ORCID ID: <https://orcid.org/0000-0002-3005-0338>

<https://ror.org/011v22a28>

Anotacija. Straipsnyje nagrinėjamas daiktavardžio *pyktis* valentingumas. Remiantis Dabartinės lietuvių kalbos tekstyno medžiaga norėta nustatyti šio daiktavardžio junglu-mą trimis aspektais: sintaksiniu, semantiniu bei leksiniu. Bandyta išsiaiškinti, kas daro įtaką daiktavardžio *pyktis* valentingumui, kokias veiksmažodžio *pykti(s)* valentingumo savybes perima šis daiktavardis. Darbe taikyti interpretacijos, aprašomasis analitinis, gramatinės, semantinės, leksinės analizės metodai. Nustatyta, kad daiktavardis *pyktis* ne tik paveldi veiksmažodžių *pykti* ir *pyktis* valentingumą, bet ir įgyja papildomų savybių.

Raktažodžiai: daiktavardis *pyktis*, valentingumas, leksinis semantinis laukas

Valency of the noun *pyktis* ‘anger’ in the Corpus of Contemporary Lithuanian Language

Abstract. By using methods of interpretation, descriptive analysis, grammatical analysis, semantic and lexical analysis, 292 sentences containing different forms of the noun *pyktis* ‘anger’ drawn from the Corpus of Contemporary Lithuanian Language were analysed. The research showed that the syntactic valency of the noun *pyktis* ‘anger’ partly differs from that of the verb *pykti(s)* ‘to be angry (with)’. The findings indicate that the syntactic valency of the noun *pyktis* ‘anger’ is influenced by the semantics and grammar of the verb *pykti(s)* ‘to be angry’ and nouns within its minimal semantic field. Specifi-

Submitted: 19/06/2025. **Accepted:** 05/11/2025

Copyright © 2025 Veslava Čižik-Prokaševa. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

cally, the nominative case governed by the verb *pykti(s)* ‘to be angry’ is replaced by the genitive case governed by the noun *pyktis* ‘anger’. Similar to the verb *pykti(s)* ‘to be angry’, the noun *pyktis* ‘anger’ requires constructions such as *ant* ‘on’ + genitive, *dėl* ‘because of’ + genitive, *už* ‘for’ + accusative, *su* ‘with’ + instrumental and a subordinate clause. Furthermore, *pyktis* ‘anger’, like *priešiškas* ‘hostile’, *jausmas* ‘feeling’, *nusistatymas* ‘attitude’ and *piktumas* ‘malice’, requires a dative actant and can take a construction with *atžvilgiu* ‘in relation to’ instead of the dative. Similar to *nusistatymas* ‘attitude’, *pyktis* ‘anger’ can take a construction with *prieš* ‘against’ and *tarp* ‘among’.

The noun *pyktis* ‘anger’ has the same semantic valency as the verb *pykti(s)* ‘to be angry’, just there are more ways to express the contentive and the percipient. When a person, who is experiencing anger, is mentioned or emphasized in a sentence, the noun *pyktis* ‘anger’ agrees with the genitive or is used in a construction with *tarp* ‘among’ percipient. When the noun *pyktis* ‘anger’ agrees with a dative or is involved in constructions with *ant* ‘on’, *prieš* ‘against’ or *atžvilgiu* ‘in relation to’ contentive, it expresses a content of state. When the cause is actualized in a sentence, *pyktis* ‘anger’ agrees with the construction *dėl* ‘because of’ + genitive causative, whereas, when the reason is highlighted, it is used with the construction *už* ‘for’ + accusative motive. However, it can describe a peripheral person, in this case *pyktis* ‘anger’ agrees with the construction *su* ‘with’ + instrumental comitative.

The collected examples show that the semantic valency of *pyktis* ‘anger’ is more diverse than that of the verb. A genitive actant can indicate not only humans, but objects or abstract concepts too. An anger can be directed towards anything: humans, animate and inanimate objects, or abstract concepts. Nonetheless, a construction *su* ‘with’ + instrumental is not unique to humans, it can indicate abstract concepts too.

Hence the noun *pyktis* ‘anger’ inherits the valency of verbs *pykti(s)* ‘to be angry’, while also acquiring new properties from nouns within its minimal semantic field.

Keywords: noun *pyktis* ‘anger’, valency, lexical semantic field

1 Įvadas

Valentingumas – „tai žodžio reikšmės lemiamą ypatybę atverti tam tikrą skaičių laisvų vietų, kurias sakinyje arba žodžiu junginyje užima tam tikros reikšmės žodžiai bei jų formos arba jų atitikmenys“ (Sližienė 1994, 16; Vaičiulytė-Seménienė 2001, 19)¹. Valentingumas paprastai siejamas su trimis lygmenimis: 1) loginės semantikos, kur nustatomas

¹ Plačiau apie nevienodą valentingumo apimties supratimą žr. Vaičiulytė-Seménienė 2001, 9–10.

žodžio laisvų vietų skaičius ir jų galimų aktantų semantinės funkcijos; 2) sintaksės, kur nustatoma valentinių aktantų gramatinė forma ir jų būtinumas ar fakultatyvumas; 3) leksinės semantikos, kur nustatoma valentinių aktantų semantika (plačiau Sližienė 1994, 16; Vaičiulytė-Seménienė 2001, 19–45).

Šiuo metu Lietuvoje valentingumo tyrimai nesulaukia tokio dėmesio kaip antroje praejusio amžiaus pusėje. Kaip žinia, valentingumo teorijos pradininku laikomas Lucienas Tesniére (1959). Didelis ir Gerhardo Helbigo (1992) indėlis į valentingumo sampratą. Lietuvoje pirmąjį darbą apie veiksmažodžių sintaksines savybes paskelbė Emma Geniušienė (1971): veiksmažodžius ji suskirstė į nulinio valentingumo, monovalentinius, bivalentinius ir trivalentinius. Tačiau veiksmažodžių valentingumo tyrimai dažniausiai siejami su Nijolės Sližienės vardu: ji parašė trijų tomų veiksmažodžių junglumo žodyną (1994, 1998, 2004). Nuo L. Tesniére laikų valentingumo sąvoka buvo taikoma veiksmažodžiams, vėliau ji pritaikyta ir kitoms kalbos dalims. Kaip žinoma, valdyti gali įvairiausią dalių žodžiai: ne tik veiksmažodžiai ir būdvardžiai, bet ir daiktavardžiai², kai kurie prieveiksmiai, skaitvardžiai bei vienas kitas įvardis (žr. Labutis 1998, 34). Semantinis būdvardžių valentingumas apžvelgiamas Dalijos Tekorienės straipsniuose (1983; 1985). Išsamiai ir sistemingai būdvardžių (vyriškosios, moteriškosios ir bevardės giminės) valentingumą aprašė Loreta Vaičiulytė-Seménienė (2001).

Nagrinėjant būdvardžių ir daiktavardžių valentingumą, remiamasi tokiais pat kriterijais, kaip ir nagrinėjant veiksmažodžių valentingumą: nustatoma, kiek palydovų gali prijungti žodis, kokios tų palydovų reikšmės, formos ir semantinės funkcijos. Nustatyti palydovų reikšmes svarbu, nes daiktavardžio vartojimas su skirtingos reikšmėmis žodžiais gali rodyti, kad tas daiktavardis vartojamas skirtingomis reikšmėmis. Jeigu daiktavardis vartojamas tik su tam tikros reikšmės žodžiais, tai rodo to daiktavardžio specifinį junglumą. Skirtingos palydovų formos dažniausiai žymi ir tų žodžių skirtinges atliekamas sakinyje funkcijas, tačiau gali rodyti ir kalbos gebėjimą įvairiomis formomis reikšti tą pačią funkciją. Semantinių funkcijų nustatymas rodo, kaip semantiškai interpretuojama daiktavardžiu ir jo palydovais nusakoma tam tikra situacija. Nors daiktavardžių valentingumo tyrimų esama labai mažai (van Durme, ed. 1997; Rathert, Alexiadou, eds. 2010; Alexiadou, Rathert, eds. 2010; Spevak, ed. 2014), pastebėta, kad daiktavardžių valentingumo tyrimams trūksta sąvokų ir tik jiems skirtos teorijos. Pvz., ką daryti, jei būtinasis papildinys nėra išreikštas? Juk, skirtingai nuo veiksmažodžio, daiktavardis gali būti pavartotas be būtinojo palydovo nepažeidžiant sakinio sintaksinės struktūros (pvz., *Namo statyba užtruko / Statyba užtruko*). Ar daiktavardžių reikalaujami žodžiai skirtystini į būtinuosius, fakultatyviuosius ir laisvuosius, kaip daroma su veiksmažodžių palydovais, kol kas irgi neaišku, tai parodys išsamūs daiktavardžių tyrimai. Be to, daiktavardžių va-

² Paminėtina, kad požiūris į daiktavardžių valentingumą skiriasi: yra tokį, kurie mano, jog valentingumas būdingas tik veiksmažodiniams daiktavardžiams (pvz., Grimshaw 1990), ir tokį, kurie visiškai neigia daiktavardžio valentingumą (pvz., Mackenzie 1997).

lentingumas yra „lankstesnis“: dažnai yra galimybė rinktis iš sintaksinių variantų (pvz., *pyktis sau / ant savęs*). Taigi yra nemažai klausimų, iš kuriuos reikėtų atsakyti.

Lietuvių kalboje daiktavardžių valentingumas nėra sistemingai aprašytas. Iki šiol aprašyti tik kelių daiktavardžių valentingumo savybės (Vaičiulytė-Seménienė 2020 (*atitiktis, atitikimas*), 2023a (*prieraišumas*), 2023b (*atviravimas*); Čižik-Prokaševa 2024 (*neapykanta, priešišumas, nemeilė, netolerancija*)). Nustatyta, kad tirtų daiktavardžių objektų raiškai įtaką daro predikatai iš jų minimalaus leksinio semantinio lauko; kad daiktavardžių prijungiamų papildymų skaičius didesnis nei atitinkamų veiksmažodžių; kad daiktavardžiai turi sintaksinio valentingumo variantų; kad daiktavardžių dažniausiai prijungiami papildymai gali neatitikti kalbos normų. Šie tyrimai parodė daiktavardžių valentingumo tyrimų svarbą bei reikalingumą.

Žinoma, kad vienų veiksmažodinių daiktavardžių sintaksinis valentingumas sutampa su atitinkamų veiksmažodžių valentingumu (pvz., *ieškoti / ieškojimas ko, priešintis / priešinimas kam, susižavėti / susižavėjimas kuo*), o kitų – skiriasi (pvz., *gauti ką – gavimas ko, domėtis kuo – dėmesys kam, neapkesti ko – neapykanta kam*). Kodėl veiksmažodžių ir iš jų padarytų daiktavardžių sintaksinis valentingumas kartais skiriasi, nėra ištirta. Taip pat trūksta tyrimų, kurie padėtų nustatyti, kas daro įtaką tokiam skirtingam sintaksiniams valentingumui. Be to, svarbu nustatyti ne tik būtinajį bei išprastą valentingumą, bet ir sintaksinio valentingumo variantus, parodyti visus įmanomus valentinius palydovus. Tuo pačiu nustatyti, ar leksinis ir semantinis tokiu veiksmažodžių ir daiktavardžių valentingumas panašus, ar skiriasi. Sisteminių daiktavardžių valentingumo tyrimai galėtų padėti nustatyti tik daiktavardžių valentingumui būdingas sintaksines, leksines bei semantines ypatybes. Taip pat išaiškinti kitokio vartojimo polinkius bei priežastis. Šie tyrimai praverstų ruošiant praktines lietuvių kalbos gramatikas bei rašant lietuvių kalbos žodynus, nustatant kalbos normas. Jie būtų naudingi lingvistų anotuotiemis duomenynams kurti, taip pat automatiniams vertimui, papildytų dirbtinio intelekto žinias.

Šio straipsnio tikslas – remiantis valentingumo teorijos principais ir verbocentrinio sakinio modelio samprata, aprašyti daiktavardžio *pyktis* sintaksinį, semantinį ir leksinį valentingumą. Taip pat siekiama nustatyti, kas daro įtaką daiktavardžio *pyktis* sintaksiniams valentingumui. Kad gautume visą semantinio ir ypač sintaksinio valentingumo realizavimo vaizdą, turime remtis pakankama medžiaga, iliustruojančia atskirų žodžių vartoseną, kurią galime gauti tekstyñ pagalba (plg. Sližienė 1978, 117). Todėl tyrimo pagrindu tapo empirinė medžiaga, surinkta iš Kauno Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro sudaryto Dabartinės lietuvių kalbos tekstyno (DLKT)³. Daiktavardžio *pyktis* pavyzdžiai buvo peržiūrėti visi. Kitų žodžių – pasirinktinai: po 100 vienetų, jeigu sakiniai informatyvūs, po daugiau vienetų, jeigu tiriamas žodis retai vartojamas su papildymais.

³ DLKT pavyzdžiai tyrimui rinkti 2025 m. sausio mėn.

Tikslus surinktų pavyzdžių skaičius nurodomas aptariant kiekvieno nagrinėjamo žodžio valentingumą. Kadangi DLKT, anot naujų standartų, nėra labai didelis duomenų rinkinys, negalima atesti net vienos tame užfiksuotos konstrukcijos, verta jas irgi fiksuoti, ypač siekiant parodyti visus įmanomus daiktavardžio valentinius palydovus.

Keliami tokie uždaviniai: a) apžvelgti daiktavardžio *pyktis* ir veiksmažodžio *pykti(s)* reikšmes bei valentingumą lietuvių kalbos žodynose (žr. 1 dalį); b) aprašyti daiktavardžio *pyktis* vartoseną DLKT junglumo požiūriu (žr. 2 dalį); c) sintaksiniu aspektu aptarti žodžių, įeinančių į minimalų semantinį daiktavardžio *pyktis* lauką, valentingumą (žr. 3 dalį). Straipsnio pabaigoje pateikiamas apibendrinimas ir išvados (žr. 4 dalį). Darbe taikyti interpretacijos, aprašomasis analitinis, gramatinės, semantinės, leksinės analizės metodai.

2 Daiktavardžio *pyktis* ir veiksmažodžio *pykti* reikšmės ir valentingumo aspektai lietuvių kalbos žodynose

Visuotinėje lietuvių enciklopedijoje (VLE) pyktis apibūdinamas kaip „impulsyvi emocienė reakcija, kylanti dėl nepasitenkinimo, nesėkmės arba konflikto. [...] Pykti gali sukelti įvairūs išorės veiksniai: įžeidimas, draudimas, vertimas atlikti tai, kas prieštarauja žmogaus norams, nepritarimas, apgaulė, kritika, išdavystė ir kita. [...] Nelabai smarkus, sąmonės kontroliuojamas pyktis padidina pasitikėjimą savimi, suteikia drąsos ir jėgų ginti teisėtus savo ir kitų žmonių interesus“ (<https://www.vle.lt/straipsnis/pyktis/4>). *Lietuvių kalbos žodyne* (LKŽe) leksema *pyktis* apibūdinama kaip ‘įsirdimas, užsirūstimas’; *Dabartinės lietuvių kalbos žodyne* (DLKŽe) pyktis apibréžiamas kaip ‘priešiškas nusistatymas dėl patirtos skriaudos, piktumas, susierzinimas, rūstis’; *Bendrinės lietuvių kalbos žodyne* (BLKŽe) žodis *pyktis* įtrauktas į žodyno antraštyną, tačiau informacija dar renkama. *Sinonimų žodyne* (SŽe) pyktis apibūdinamas kaip ‘priešiškas jausmas dėl skriaudos, įžeidimo ir pan.’. Teikiami tokie iliustracinių pavyzdžiai:

- (1) *Pykti* turėjo J.Jabl. (LKŽe)
- (2) *Numirė iš pykčio* Š. (LKŽe)
- (3) *Pyktis* ima, žiūrint į tokį darbą (SŽe)
- (4) *Viską nori pykčiu* padaryti (DLKŽe)
- (5) *Visą pykti* ant manęs išliejo (DLKŽe)
- (6) *Aš tuo pykčiu* ir rasiu geltoną grūšnią Ds. (LKŽe)
- (7) *Jis ant man turi pykti* Vb. (LKŽe)
- (8) *Sunku dabar ir bepasakyti, nuo kada tarp jų tie pykčiai* Sb. (LKŽe)

⁴ Interneto prieiga: <https://www.vle.lt/straipsnis/pyktis/>. Žiūrėta: 2025-12-12.

(9) *Mergos pykčio iš dantų nepažinsi rš.* (LKŽe)

(10) *Giminių pykčiai lig vartų (giminės ilgai nesipyksta) Jnš.* (LKŽe)

Kaip rodo žodynų pavyzdžiai, daiktavardis *pyktis* dažnai vartojamas nevalentiškai. Kartais jis atveria pozicijas kilmininko papildymui (žr. 9–10), konstrukcijoms *ant* + kilmininkas (žr. 7) ir *tarp* + kilmininkas (žr. 8). Nors DLKŽe ir SŽe *pykčio* apibrėžimai vardina (kaip ir VLE) to jausmo atsiradimo priežastis, tačiau nei viename iš žodynų nepateikta nė vieno pavyzdžio, kuriame būtų nurodyta *pykčio* priežastis.

Kaip žinoma, daiktavardis *pyktis* yra veiksmo pavadinimas (veiksmažodžio abstraktas), priesagos *-tis* vedinys, padarytas iš veiksmažodžio *pykti* ir reiškiantis būseną (Ambrazas, red. 1996, 94–95). LKŽe teikiamos trys veiksmažodžio *pykti* reikšmės: 1. jausti *pyktį*: *Už ką pyksti ant jo?* Dkš. 2. būti su kuo piktuoju, vienam ant kito rūstauti: *Jei nenori su kaimynu pykt, neskolink pinigų Vdn.* 3. refl. bartis: *Pykos kunesgas, kad neposnykau[ja] Vdn.* DLKŽe veiksmažodis *pykti* apibūdinamas kaip ‘jausti *pyktį*, susierzinimą, norėti atkeršyti už nuoskaudą’: *Ko (už ką) tu pyksti ant manęs? Jau atlyžo, nepyksta.* BLKŽe kol kas veiksmažodžio *pykti* apibrėžimo nerasime, nes informacija dar renkama.

Anot Nijolės Sližienės *Lietuvių kalbos veiksmažodžių junglumo žodyno* (LKJVV), veiksmažodis *pykti* yra trivalentis: reikalauja vardininku žymimo percipiente⁵ (Pcp), konstrukcija *ant* + Gen žymimo kontentyvo⁶ (Con) ir konstrukcija *dėl* + Gen arba *už* + Acc žymimo kauzatyvo⁷ (Caus) arba motyvo⁸ (M) (žr. 11–13), t. y. jo sintaksinis valentingumas yra Nom (antGen) (*dėlGen/užAcc*), o semantinis – [Pcp Con Caus/M] (Sližienė 1998, 174). Jo leksinis valentingumas yra Nom Hum, antGen +Anim, *dėlGen +Anim/-Anim/Abstr*, *užAcc +Anim/-Anim/Abstr* (ten pat). Nors prie reikalaujamų aktantų neminima, kad veiksmažodis *pykti* gali prisijungti šalutinį sakinių, tačiau pateikiama pora tai iliustruojančių pavyzdžių (žr. 14–15). Sangrąžinio veiksmažodžio *pyktis* valentingumas kiek kitoks – jis gali būti vartojamas pagal du modelius: a) kaip trivalentis: reikalauja vardininku žymimo percipiente, konstrukcija *su* + Instr žymimo komitatyvo⁹ (M) ir konstrukcija *dėl* + Gen žymimo kauzatyvo (žr. 16–17), t. y. jo sintaksinis valentingumas yra Nom suInstr (*dėlGen*), o semantinis – [Pcp Com Caus]; b) kaip dvivalentis: reikalauja daugiskaitos vardininku

⁵ Percipientas – psichinės būsenos patyrėjo funkcija, būdinga gyvoms būtybėms, suvokiančioms tikrovės dalykus, turinčioms tam tikrą nusiteikimą jų atžvilgiu, psichologiskai į juos reaguojančiomis (Sližienė 1994, 20–21).

⁶ Kontentyvas – veiksmo ar būsenos turinio funkcija, būdinga įvairiems daiktams bei reiškiniams (Sližienė 1994, 22).

⁷ Kauzatyvas – veiksmo ar būsenos priežasties funkcija (Sližienė 1994, 21).

⁸ Motyvas – motyvavimo funkcija, būdinga daiktams, reiškiniams, kurie suvokiami kaip veiksmo ar būsenos pamatas (Sližienė 1994, 22–23).

⁹ Komitatyvas – antrėilio agento, patiento, percipiente ar beneficiente funkcija, pasirodanti greta atitinkamos pagrindinės funkcijos ir sakinyje dažniausiai atitinkanti prielinksnię konstrukciją suInstr (Sližienė 1994, 22).

žymimo percipiento ir konstrukcija *dėl* + Gen žymimo kauzatyvo (žr. 18–19), t. y. jo sintaksinis valentingumas yra Nom_{Plur} (dėlGen), o semantinis – [Pcp Caus] (Sližienė 1998, 175). Veiksmažodžio *pyktis* leksinis valentingumas yra Nom Hum, dėlGen +Anim/-Anim/Abstr, suInstr Hum (ten pat).

- (11) *Pykstu ant tavęs dėl arklio, už sudaužytą mašiną* (LKVIJŽ 1998, 174)
- (12) *Sūnus irgi (už)pyko ant motinos* (LKVIJŽ 1998, 174)
- (13) *Bet pykti jis niekuomet nepyko Krv* (LKVIJŽ 1998, 174)
- (14) *Supyko, kad neatėjau* (LKVIJŽ 1998, 174)
- (15) *Supykau ant katės, kam paliejo pieną* (LKVIJŽ 1998, 174)
- (16) *Jis amžinai pykstasi dėl ežių su kaimynais* (LKVIJŽ 1998, 175)
- (17) *Su ponu pykstamės, nebeleis į kirtimus* Pt (LKVIJŽ 1998, 175)
- (18) *Dažnai jie pykdavosi dėl vaikų, dėl jų auklėjimo, dėl pinigų* (LKVIJŽ 1998, 175)
- (19) *Jie ir vėl susipyko* (LKVIJŽ 1998, 175)

DLKT peržiūra parodė, kad šalutinis sakinių turėtų būti įtrauktas į veiksmažodžio *pykti* reikalaujamų aktantų sąrašą, nes tokią pavyzdžių gausu (žr. 20–22). Taip pat veiksmažodis *pykti* atveria poziciją prieveiksmio papildymui, kurio negalima laisvai praleisti, nes pasikeičia sakiniu teikiama informacija (žr. 22–24).

- (20) *Pykstu, kad jie tokie laimingi.*
- (21) – *Pykstu, nes gaištu savo laiką.*
- (22) *Aš beveik pykstu, kad šitaip elgiesi.*
- (23) *Supykstu labai greitai, pykstu žiauriai, ir ... greitai atlėgstu.*
- (24) *Pavyzdžiui, kai jūs sakote: „Aš taip pykstu, jog galėčiau užsukti tau ausj“, tai dar nereiškia, kad tą ir padarytumėte.*

Taip pat DLKT peržiūra parodė, kad veiksmažodis *pyktis* irgi atveria poziciją prieveiksmio papildymui (žr. 25–26) bei šalutiniams sakiniui (žr. 27–28).

- (25) *Paaugliai pykdavosi rečiau nei suaugusieji.*
- (26) *Johnny ir Winona’os gyvenimas nebuvo ramus: jie vaidydamosi, pykdavosi, tai-kydamosi ir vėl pykdavosi...*
- (27) *Su juo dažnai pykstuosi dėl to, kad jis mamai prišneka apie mane visokių nesąmonių [...]*
- (28) *Pykdavosi tik tada, kai E. Ivanauskas prasižengdavo, bet tai buvo gana seniai.*

Apibendrinant tai, kas buvo pasakyta, tikėtina, kad daiktavardžio *pyktis* valentingumo lauke turėtų būti mažiausiai keturi valentiniai aktantai (jausmo patyrėjas, jausmo sukėlėjas, jausmo sukilimo priežastis arba motyvas ir antraeilis jausmo patyrėjas), kuriuos lemtų veiksmažodžių *pykti* ir *pyktis* valentingumai. Taip pat daiktavardžio *pyktis* valentingumą galėtų veikti paties daiktavardžio semantika ir gramatika (*priešiškas nusistatymas / jausmas*) bei jo sinonimų (*jširdimas, užsirūstimas, piktumas, susierzinimas, rūstis*) semantika ir gramatika¹⁰.

3 Daiktavardžio *pyktis* valentingumas

DLKT teikia 6690 daiktavardžio *pyktis* įvairių formų pavartojo pavyzdžių. Pavyzdžiai rinkti pateikus sutrumpintas ieškomo žodžio šaknis (*pykt* ir *pykč*), kad konkordančias apimtų visas linksnių formas. Peržiūrėti visi sakiniai, ir atrinkti 292, kur *pyktis* reikalauja papildymo. Palyginus nedidelis atrinktų sakinių skaičius rodo, kad dažniausiai šis daiktavardis vartojoamas be valentinių palydovų, pvz.: *Pyktis* gadina jo kraują; *Geriau būtų pyktis, neapykanta, tik ne toks svetumumas; Ji tiesiog pasiuto iš pykčio.* I tyrimą neįtraukti sakiniai, kuriuose *pyktis* vartojoamas su fakultatyviu(ais) pažyminiu(ais), pvz.: *Turi Rita ir kitą trūkumą: trūksta jai sveiko pykčio, kuris varytų į priekį; Bet paskui, kai atvėso visas pyktis, Veronika atsiminė batukus, kuriuos jai nupirko Julius; Ramzis sunkiai suvaldė užvirusį pyktį: – Jūs tuo abejojate?* Dabar apžvelkime, su kokiais valentinių aktantais vartojoamas žodis *pyktis*.

3.1 Kilmininko aktantas

Iš pradžių pasakytina, kad daiktavardis *pyktis* dažnai reikalauja vardažodžio, žymimo kilmininko linksniu ir nurodančio, kas jaučia tą pyktį (105 atvejai). „Subjekto (siauriau – veikėjo) kilmininkas – rodo asmenį, daiktą ar reiškinį, kuris pats daro kam nors kokį poveikį, nukreipia į ką nors veiksmą“ (Šukys 1998, 116). Anot Jono Šukio, „[k]ilmininkai, vartojoami su veiksmažodiniais daiktavardžiais, nesiskiria nuo bet kurio kito linksnio pažymimųjų variantų [...], o skiriasi tik didele tokų linksnių gausa ir svarbiausia tuo, kad čia būtinai yra pakitęs linksnis. [...] iš subjekto vardininko gauti kilmininkai paprastai vadinami veikiančiojo dalyko kilmininkas. Čia dar labai ryški ir subjekto reikšmė“ (ten pat, 120). DLKT pavyzdžiai rodo, kad toks vardažodžio kilmininkas gali būti reiškiamas asmeniniu įvardžiu (žr. 29), tikriniu daiktavardžiu (žr. 30), bendriniu daiktavardžiu, reiškiančiu asmenį (žr. 31). Kaip rodo rasti pavyzdžiai, *pyktis* dažniausiai būdingas žmonėms (žr. 32–37), tačiau gali būti priskirtas ir daiktui (žr. 32) bei dievui (žr. 33–34). Vadinas, daiktavardis *pyktis* gali prijungti skirtinges reikšmės žodžius. Jie sakinyje atlieka percipiente funkciją.

(29) *Su kiekvienu žodžiu jos pyktis auga.*

(30) *Nenorėjau, kad Miro pykčiai visa tai sugriautų.*

¹⁰ Apie tai, kad tam tikro žodžio valentingumui gali turėti įtakos kitų į jo minimalų leksinių semantinių laukų įeinančių žodžių valentingumai, buvo kalbėta straipsnio pradžioje.

- (31) *Vaiko pyktį visada reikia priimti.*
- (32) *Spaudžiu pilvo apacią ir dėkoju jam, čiurškiančiam per kateterio angelę: nešk mano kraujo pyktį žemei, nešk.*
- (33) *Perkūno pyktis* jau buvo atlėgės.
- (34) *Tokiais momentais pasirodo baisus Dievo pyktis, o kartu su juo visa kūrinija.*

Tarpusavio santykiai tarp dviejų veikėjų galėtų būti išreikšti ir konstrukcija su prielinksniu *tarp*: „Prielinksnio *tarp* konstrukcijos kartais reiškia objektą, tiksliau – objektus, tarp kurių yra koks neatitikimas, vyksta konfliktai ar bendravimai.“ (Šukys 1998, 515) Tokiais atvejais tarpusavio situacijos dalyviai reiškiami kaip vienas aktantas, t. y. atskaitos tašku pasirenkami visi situacijoje dalyvaujantys asmenys, o ne kuris vienas iš jų (Haspelmath 2007, 2091)¹¹, ir nurodoma, kad visi situacijos dalyviai yra vienodai svarbūs. DLKT rastas tik vienas toks pavyzdys: [...] *ryškėja abiejų partijų trintis dėl kandidatų į ministrus ir kitus postus, vis įsižiebia pykčiai periferijoje tarp liberalų ir socialliberalų.* Tokį vartojimą rodo ir LKŽE pavyzdys (žr. 8). Kad toks vartojimas galimas, rodytų ir šis sakiny: *Pykčiai tarp sesių ir brolių atrodo neatsiejamas dalykas, tačiau psichologė pataria, kaip galima to išvengti*¹².

3.2 Naudininko ir jo variantų aktantai

Taip pat kartu su daiktavardžiu *pyktis* dažnai nurodomas šios būsenos turinys (86 atvejai). Jis gali būti nurodytas skirtingais būdais: dažniausiai vardažodžio naudininku (50 atvejų) arba konstrukcija su *ant* (30 atvejų), retai – konstrukcija su *prieš* (5 atvejai) ir polinksniškojo *atžvilgiu* konstrukcija (1 atvejis). Primintina, kad veiksmažodžiui *pykti* iš ką tik pateiktų papildymu būdinga tik konstrukcija su *ant*. Iš pradžių aptarkime sakinius su vardažodžio naudininku. Kaip žinoma, „[o]bjekto naudininku pasakomas asmuo ar daiktas, kurio interesais – naudai ar nenaudai – atliekamas veiksmas, kuriam kas nors skiriama ar kurio atžvilgiu reiškiasi kokia nors ypatybė“ (Ambrasas, red. 1996, 513). Anot V. Grabausko, nors naudininkas yra netiesioginis objektas, „[t]ačiau jo sintaksinė funkcija yra pagrindinė, svarbiausia. Netiesioginis objektas paprastai rodo tiesioginio objekto gavėją-vartotoją (kai kuriais atvejais nusako, kam adresuotas veiksmas)“ (Grabauskas 1971, 55). DLKT pavyzdžiai rodo, kad daiktavardžio *pyktis* reikalaujamas vardažodžio naudininkas reiškiamas asmeniniu įvardžiu (žr. 35, 39), tikriniu daiktavardžiu (žr. 36), bendriniu daiktavardžiu, įvardinančiu gyvą asmenį (žr. 37, 39), daiktą (žr. 38) ar abstraktą (žr. 39). Vadinas, daiktavardis *pyktis* gali prijungti skirtingos reikšmės

¹¹ Citiuojama pagal Vaičiulytė-Seménienė 2016, 1.

¹² Interneto prieiga: <https://www.facebook.com/tavovaikas/posts/pyk%C4%8Dai-tarp-sesi%C5%B3-ir-broli%C5%B3-atrodo-neatsiejamas-dalykas-ta%C4%8Diau-psicholog%C4%97-patar/3468843446566834/>. Žiūrėta: 2025-12-12.

žodžius. Kartais pykčio jausmas gali būti abipusis, tokiais atvejais situacija reiškiama reciproke¹³ anafora *vienas kitam* (žr. 40); nurodoma, kad abu situacijos dalyviai yra vienodai svarbūs ir jaučia tą patį. Sakinyje vardažodžio naudininkas atlieka kontentyvo funkciją.

- (35) *Mano pyktis Jam didėjo, ir aš pasibaisėjau pagalvojės, kad tuoj tuoj Jis apie tai sužinos ir Pats ant manęs supyks.*
- (36) *Jefreitorius Nesterovas neslepija pykčio Grošiui, bet jis negali jam nieko padaryti, nes kisdamas išsinarino petį [...]*
- (37) *Kartu nešesi širdyje kerštą ir pyktį lietuviam.*
- (38) – *Iškart skambink, kai tik jis atvažiuos, – neslepija pykčio visam pasauliui Olga ir išjungia telefoną.*
- (39) *Daugiau nei trečdalis tėvų (38 proc.) jautė didelį pyktį sau, Dievui / likimui arba sutuoktinui [...]*
- (40) *Nesuvokiamas „išrinktųjų“ pyktis vienų kitiems.*

Kontentyvo funkciją sakinyje atlieka ir prielinksnio *ant* konstrukcijos. Kaip rašo J. Šukys, šios „konstrukcijos gali reikšti objektą, į kurį nukreiptas veiksmas“ (Šukys 1998, 360). DLKT pavyzdžiai rodo, kad tokiose konstrukcijose vartojami asmeniniai įvardžiai (žr. 41), tikriniai daiktavardžiai (žr. 41), bendariniai daiktavardžiai, įvardinantys gyvą asmenį (žr. 41–42), daiktą (žr. 43) ar abstraktą (žr. 44). Vadinas, prielinksnio *ant* konstrukcijose irgi vartojami skirtinges reikšmės žodžiai.

- (41) *Rytų vėjas kaskart įpūsdavo jį, pyktį ant amtmono, kiekviena proga prie jo kimbancio, pyktį ant savęs, įsileidusio į beprasmį ginčą, pyktį ant Reginos, nutylėjusios jam tą atsitikimą.*
- (42) *Kartais dėl sovietų vadovų pykčio ant nepaklusniųjų baltijiečių pasitaikydavo net komiškų epizodų.*
- (43) *Dilgius brido vis daugiau iš vandens, jį ginė pyktis ant to mažo šmėžuojančio daikto.*
- (44) *Anas jau, regis, nebe kad loska, o su pykčiu ant Dievo: ištrauk dūsią, ir gana.*

Kodėl vienais atvejais pasirenkamas naudininkas, o kitais – konstrukcija su *ant*, nėra aišku. Ir vienu, ir kitu atveju vartojami skirtinges reikšmės žodžiai. Kilo prielaida, kad naujininko pasirinkimą galėtų įtakoti sakinio veiksmažodis: jei veiksmažodžio valentiniu palydovu gali būti naudininkas, tikėtina, kad su daiktavardžiu *pyktis* irgi bus vartojamas

¹³ Plačiau apie leksinius reciprokus žr. Vaičiulytė-Seménienė 2016.

naudininkas (pvz., su veiksmažodžiais (*ne*)*jausti*, *nešti*, *reikšti*; žr. 37, 39). Dar pastebėta tai, kad, jei sakinyje daiktavardis *pyktis* vartojamas kilmininko linksniu, tuomet jo papildymas paprastai nurodomas naudininko linksniu (21 atvejis, beveik pusė visų rastų pavyzdžių su naudininku; žr. 36, 38). Daiktavardžio *pyktis* kilmininkas ir konstrukcija su *ant* buvo pavartoti tik 2 kartus, abiem atvejais daiktavardžio *pyktis* kilmininkas buvo reikalaujamas prielinksnio *dėl* (žr. 42). Matyt, tokiais atvejais naudininkas pasirenkamas vengiant dviejų kilmininkų. Taip pat naudininkas gali būti pasirenkamas pagal analogiją su kitais daiktavardžiais (apie tai plačiau 3 skyriuje).

Kaip minėta, buvo rasti tik 5 pavyzdžiai, kuriuose daiktavardis *pyktis* reikalauja papildymo, reiškiamo prielinksnio *prieš* konstrukcija. Kaip nurodo, J. Šukys, šios „konstrukcijos gali reikšti objektą. Jomis pasakoma, į ką nukreiptas priešiškas veiksmas, kam rodomas palankumas, nusijemėminimas arba kieno atžvilgiu reiškiasi kieno būsena“ (Šukys 1998, 482). DLKT pavyzdžiai rodo, kad tokiose konstrukcijose vartojami asmeniniai įvardžiai (žr. 45), tikriniai daiktavardžiai (žr. 46), bendriniai daiktavardžiai, įvardinantys gyvą asmenį (žr. 47) ar abstraktą (žr. 48). Vadinas, prielinksnio *prieš* konstrukcijose vartojami skirtinges reikšmės žodžiai, tačiau tekstyne nerasta pavyzdžių, rodančių, kad šiose konstrukcijose būtų vartojami daiktą reiškiantys daiktavardžiai. Prielinksnio *prieš* konstrukcija sakinyje irgi atlieka kontentyvo funkciją.

- (45) *Kai draugo akys nebedvelks šalčiu ir pykčiu prieš jus, tačiau jniršis vis dar valdys jo veiksmus, suprasite [...].*
- (46) *[...]jis įsitikinęs, jog Ašlis neturėjo jokio pykčio prieš velionį poną Lansingą.*
- (47) *[...] nebūna nė vienos dienos, kad galėtum taip sau pagulėti lovoje, kaupdamas dykų jėgą ir pyktį prieš kitus žmones.*
- (48) *Bejėgiškas pyktis prieš Lietuvos nepriklausomybę kalba jo lūpomis.*

Kaip rodo DLKT, pykčio objektą įmanoma išreikšti ir polinksniškojo *atžvilgiu* konstrukcija¹⁴. Nors rastas tik vienas pavyzdys, jo negalima atmeti, apie jį verta užsiminti šiame straipsnyje. Ši konstrukcija sakinyje irgi atlieka kontentyvo funkciją: *Tokie išgyvenimai sukelia stiprius savęs niekinimo, kitų vyru įtarinėjimo, pykčio moterų atžvilgiu ir savigarbos praradimo jausmus.*

¹⁴ Primintina, kad tokios konstrukcijos laikomos vėlyvu suintelektintos kalbos reiškiniu. Šios konstrukcijos laikomos taisyklingomis, kai jos yra skirtos nurodyti ko nors pagrindą pagal pašalinį, veiksmą arba būseną sąlygojantį objektą, o netinka tiesioginiam santykui su objektu reišksti (Akelaitis 2005). Apie *atžvilgiu* konstrukcijas valentingumo aspektu žr. Vaičiulytė-Semėnienė 2022.

3.3 Konstrukcijų *dėl* ir *už* aktantai

Taip pat kartu su daiktavardžiu *pyktis* dažnai nurodoma šios būsenos priežastis arba motyvas (69 atvejai). Ji gali būti nurodyta dviem būdais: dažnesnės konstrukcijos su *dėl* (57 atvejai), retesnės – konstrukcijos su *už* (12 atvejų). Primintina, kad tokios konstrukcijos būdingos ir veiksmažodžiui *pykti*. Anot J. Šukio, „[*už*] su galininku dažniausiai nusako įvairius atpildo, atsilyginimo objektus, taigi čia būdinga objekto reikšmė. [...] Kai objektas, už kurį atsilyginama, sukelia neigiamas atlyginimo (paprastai – bausmės) pasekmes, jis įgauna ir šias pasekmes sukeliančios priežasties atspalvį. [...] Retkarčiais priežasties atspalvis atsiranda ir nesant neigiamų pasekmių“ (Šukys 1998, 524–525). *Dabartinės lietuvių kalbos gramatikoje* nurodoma, kad viena iš konstrukcijų *už* su galininku reikšmiu – jausmo, vertinimo pagrindas, motyvas (Ambrazas, red. 1996, 454). DLKT pavyzdžiai rodo, kad tokiose konstrukcijose vartojami parodomieji įvardžiai (žr. 49), bendariniai daiktavardžiai, įvardinantys gyvą asmenį (žr. 50), daiktą (žr. 51) ar abstraktą (žr. 52). Vadinasi, prielinksnio *už* konstrukcijose vartojami skirtinges reikšmės žodžiai. Sakinyje šios konstrukcijos atlieka motyvo funkciją.

- (49) *Sūnus, neakivaizdžiai jaučiantis tėvų kontrolę ir su ja visokeriopai kovojantis, duktė, nešiojanti savyje pyktį už tai, kad vaikystėje nebuvo suprasta, nesugeba tėvams prisipažinti esą mylinti ar mylintis.*
- (50) [...] aš iššoviau ir, matyt, pataikiau banditui į pilvą, nes jis susilenkė, bet vis dėlto suspėjo iššauti ir susvyrauęs kareivis nugriuvo (...) **Pyktis už kareivį** dusino mane.
- (51) *Esmę mėginome išsaugoti – šventas pyktis „už reikalą“, noras „patvarkyti“.*
- (52) [...] yra varžovas, ant kurio aikštės šeimininkai galės išlieti susikaupusį pyktį už apmaudžias **nesėkmes** Eurolygoje.

Kaip minėta, būsenos priežastis gali būti išreikšta konstrukcija su *dėl* (Ambrazas, red. 1996, 441; Šukys 1998, 380). Elena Valiulytė manė, kad prielinksnių *dėl* ir *už* pasirinkimą su veiksmažodžiu *pykti* „gali lemti reiškiamos situacijos skirtumai. Jei reiškiamų emocijų priežastis yra subjekto siekiamas dalykas, vartojamas prielinksnis *dėl*, jei turimas – prielinksnis *už*“ (Valiulytė 1978, 29). Plg. su Šukio nuomone dėl daiktavardžių junginių su prielinksniais *už* ir *dėl*: „[k]ai pasakomas ne tiek objektas ar priežastis, kiek aktualus kalbamuoju momentu veiksmo tikslas, siekiamas dalykas, kurio dar neturima, dažniau vartojamos ne prielinksnio *už*, bet prielinksnio *dėl* konstrukcijos“ (Šukys 1998, 525). Tačiau N. Sližienė manė, kad šios konstrukcijos su veiksmažodžiu *pykti* vartojamos sinonimiškai (Sližienė 1998, 175). DLKT rasti sakiniai rodo, kad konstrukcijos su *dėl* nurodo jau turimą priežastį. DLKT pavyzdžiai rodo, kad tokiose konstrukcijose vartojami parodomieji įvardžiai (žr. 53), bendariniai daiktavardžiai, įvardinantys gyvą

asmenį (žr. 54), daiktą (žr. 55) ar abstraktą (žr. 56). Vadinasi, prielinksnio *dėl* konstrukcijose irgi vartojami skirtingos reikšmės žodžiai.

- (53) *Gal pagaliau normaliems žmonėms užeis pyktis dėl to, kas vyksta Lietuvoj.*
- (54) *Pyktis dėl saviškių yra ne mylinčios, bet silpnos sielos požymis.*
- (55) *Demokratų partija yra pagrindinis protestuotojų pykčio dėl prarastų santaupų objektas.*
- (56) *Šeimos nariai išlieja ant jo susikaupusių pyktijų dėl visų savo ir šeimos bėdų, negandų.*

Sakinyje šios konstrukcijos atlieka kauzatyvo funkciją.

3.4 Konstrukcijos su aktantas

Kartu su daiktavardžiu *pyktis* gali būti pavartota ir konstrukcija su prielinksniu *su* (6 atvejai; tokia konstrukcija būdinga tik sangrąžiniam veiksmažodžiui *pyktis*). Pagrindinė tokijų konstrukcijų reikšmė – draugės, buvimo kartu, turėjimo (Ambrazas, red. 1996, 452; Šukys 1998, 493). Jomis žymimas tarpusavio būsenos antraelis veikėjas, t. y. nurodomas mažesnio svarbumo veikėjas. DLKT pavyzdžiai rodo, kad tokiose konstrukcijose vartojami tik bendriniai daiktavardžiai, įvardinantys gyvą asmenį (vartojama daiktavardžio *pyktis* daugiskaita, rodomi tarpusavio santykiai, žr. 57–58) arba abstraktą (vartojama daiktavardžio *pyktis* vienaskaita, rodomas draugės santykis, žr. 59–60). Tai rodo, kad konstrukcijose su prielinksniu *su* vartojami tik dviejų reikšmių žodžiai, ir tie žodžiai žymi skirtingus šios konstrukcijos santykius su daiktavardžiu *pyktis*. Sakinyje konstrukcijos su prielinksniu *su* atlieka komitatyvo funkciją.

- (57) *Dėl mąstymo, kad aš geresnis nei kiti, galimi pykčiai netgi su kaimynais.*
- (58) *Kai jam atsibodo amžinių pykčiai su seseria Hetavei ir nesibaigiantys amerikiečių apendicitai, metė darbą [...].*
- (59) *Pyktis su gailesčiu sumišo manyje.*
- (60) *Elizabet Kuber-Ross savo knygoje „Apie mirtį ir mirimą“ aprašo penkias stadijas, kurias žmogus pergyvena tapęs invalidu, sužinojęs, koks žiaurus likimas jolaukia: po žinios ištinka šokas ir „žado netekimas“, ji pakeičia pyktis su neapykanta ir kančia, po to apima depresija [...].*

3.5 Šalutinis sakiny

Daiktavardis *pyktis*, kaip ir veiksmažodžiai *pykti* ir *pyktis*, gali reikalauti ir šalutinio pázymimojo (žr. 61), aiškinamojo (žr. 62) arba nusakomojo (žr. 63–64) sakinio.

- (61) *Tavoji gentis sumažina pyktį, kurį jie nuolat nešiojasi su savimi.*
- (62) *Tačiau suėmė pyktis, nes dabar nebegalėjo šio judesio sulaikyti neapsijuokęs.*
- (63) *Jo akyse žaižaravo tokia neapykanta ir pyktis, kaip ir mano.*
- (64) *Likusių žemės ūkio bendrovėje kaimynų pavydas, pyktis, kam ūkininkai „atima“ (iš tikrujų atsiūma) atidalintą turtą, buvo fonas.*

Apibendrinant galima sakyti, kad daiktavardis *pyktis* atveria penkias priklausomųjų dėmenų pozicijas, nors gali būti pavartotas pats vienas. Tai, kad daiktavardžio papildymai sakinje praleidžiami daug dažniau nei veiksmažodžiu, pastebėta nė vieno kalbininko (pvz., Panevová 2014, 6–7; Kolářová 2014, 27; Holvoet, Judžentis 2003a, 34; Pakerys 2006, 134, 146; Vaičiulytė-Seménienė 2016, 23; Sližienė 1978, 111; Laužikas 1986, 72; Vaskeleitė 2003, 153). To priežastys gali būti apibendrinta arba anaforinė (numanomoji) vartosena, noras išvengti komunikacinio pertekliaus, konteksto ekonomija. Taigi daiktavardžio *pyktis* sintaksinio valentingumo struktūra yra Gen(*tarp* + Gen/Gen *ir* Gen)-N-Dat(*ant* + Gen/*arieš* + Acc/Gen + *atžvilgiu*)/*dėl* + Gen(*už* + Acc)/*su* + Instr/S, semantinio valentingumo struktūra – [Pcp Con Caus Com/M], o leksinio – Gen Hum/+Anim/-Anim/Abstr, Dat ir *ant* + Gen Hum/+Anim/-Anim/Abstr, *arieš* + Acc Hum/+Anim/-Anim, *dėl* + Gen ir *už* + Acc Hum/+Anim/-Anim/Abstr, *su* + Instr Hum /Abstr. Vadinas, daiktavardžio *pyktis* sintaksinio valentingumo struktūra kiek skiriasi nuo veiksmažodžio *pykti(s)* sintaksinės struktūros: veiksmažodžio reikalaujamas vardininkas nominalizacijos atveju virsta kilminku (pirmoji transformacijos taisyklė, žr. Karlík, Nübler 1998, 3; Kolářová 2014, 25); prielinksninės konstrukcijos *ant* + Gen, *dėl* + Gen, *už* + Acc, *su* + Instr ir šalutinis sakinys išlieka nepakitę (antroji transformacijos taisyklė, žr. Karlík, Nübler 1998, 4; Kolářová 2014, 26); daiktavardžio *pyktis* sintaksinio valentingumo struktūra dar atveria pozicijas naudininko papildymui ir konstrukcijoms *arieš* + Acc, Gen + *atžvilgiu* ir *tarp* + Gen/Gen *ir* Gen, kas patvirtina teiginį, kad galimų nominalizacijos papildymų skaičius gali būti didesnis už atitinkamo veiksmažodžio papildymų skaičių (plg. Kolářová 2014, 25). Semantinio daiktavardžio *pyktis* valentingumo struktūra lieka tokia pat, kaip ir veiksmažodžio *pykti(s)* struktūra. Leksinio daiktavardžio *pyktis* valentingumo struktūra pakitusi visai nežymiai.

4 Žodžių, įeinančių į daiktavardžio *pyktis* semantinį lauką, sintaksinis valentingumas

Kaip buvo minėta, tam tikro žodžio valentingumui gali turėti įtakos paties daiktavardžio semantika ir gramatika bei kitų į jo minimalų leksinį semantinį lauką įeinančių žodžių valentingumai. Primintina, kad daiktavardžio *pyktis* atveju tokie žodžiai yra *priešiškas nusistatymas* / *jausmas*, *jširdimas*, *užsirūstimas*, *piktumas*, *susierzinimas*, *rūstis*. Kadangi daiktavardis *pyktis* akivaizdžiai skiriasi nuo veiksmažodžio *pykti(s)* sintaksiniu valentingumu, toliau bus kalbama tik apie kitų žodžių sintaksinį valentingumą.

4.1 Būdvardžio *priešiškas* ir daiktavardžio *nusistatymas* sintaksinis jungumas

Remiantis L. Vaičiulytės-Seménienės (2001) būdvardžių valentingumo aprašu, būdvardis *priešiškas* priklausytų būdvardžių grupei, kurių sintaksinio valentingumo struktūra yra N-Adj-Dat. Peržiūrėjus 100 DLKT sakinių, nustatyta, kad būdvardis *priešiškas* gali būti vartojamas su vardininko ir naudininko valentiniuose palydovais (29 sakiniuose, žr. 65–66). Taip pat pastebėta, kad vietoj naudininko gali būti pavartota polinksniškojo *atžvilgiu* konstrukcija (žr. 67–68). Peržiūrėjus visus 2247 DLKT sakinių su būdvardžiu *priešiškas*, nustatyta, kad ši konstrukcija pavartota 21 kartą. Todėl galima teigti, kad būdvardžio *priešiškas* sintaksinis valentingumas yra N-Adj-Dat/Gen + *atžvilgiu*¹⁵.

- (65) *Per visą antikinės Graikijos istoriją Mesenijos gyventojai buvo priešiški Spartai.*
- (66) *Ar vokiečiai nori, kad lietuvių tauta taptų jiems priešiška, kaip belgai ir olandai?*
- (67) *Visata gyvybės atžvilgiu yra priešiška.*
- (68) *Todėl ir šiandien moteris nedrįsta konkuruoti su priešišku jos atžvilgiu* vyrų pasauliu, negalėdama įveikti jai iškieptyt „bejėgiškumo ir pasyvumo“ savybių.

Peržiūrėjus 100 DLKT sakinių su daiktavardžiu *nusistatymas*, nustatyta, kad jis gali reikalauti papildymų, išreikštų konstrukcija su prielinksniu *prieš* (žr. 69), konstrukcija su *dėl* (žr. 70), bendratimi (žr. 71), polinksniškojo *atžvilgiu* konstrukcija (žr. 72) ir kilmininku (žr. 71–72). Vadinas, daiktavardžio *nusistatymas* sintaksinis valentingumas yra Gen-N-*prieš* + Acc/*dėl* + Gen/Inf/Gen + *atžvilgiu*.

- (69) *Vienpusiškas nusistatymas prieš svetimybes.*
- (70) *Visi jam buvo pareiškę neigiamą nusistatymą dėl pasilikimo buvusiouose postuose.*
- (71) *Ji nepritarė ponios Ašli nusistatymui nesirodyti miesto gatvėse.*
- (72) *Taip mes išreikšim savo nusistatymą tikrovės atžvilgiu.*

Buvo bandyta nustatyti ir žodžių junginio *priešiškas* *nusistatymas* valentingumą. Rasti 27 sakiniai: 7 buvo reikalaujama kilmininko (žr. 73, 76), 9 – polinksniškojo *atžvilgiu* konstrukcijos (žr. 74), 4 – konstrukcijos su *prieš* (žr. 75) ir 2 – naudininko (žr. 76).

- (73) *Bet šis dėl kurių karininkų priešiško nusistatymo tu pareigų atsisakė.*
- (74) – *Tas priešiškas nusistatymas jūsų atžvilgiu, atrodo, nesibaigė vien šiuo atveju.*

¹⁵ Vieno tyrimo metu buvo nustatyta, kad daiktavardis *priešišumas* atveria dvi priklausomų dėmenų pozicijas: kilmininkui ir naudininkui arba polinksniškojo *atžvilgiu* konstrukcijai, t. y. jo sintaksinio valentingumo struktūra yra Gen-N-Dat/Gen + *atžvilgiu* (Čižik-Prokaševa 2024, 163–167).

- (75) *Kaip matome, tam tikri apribojimai, numatyti mūsų Konstitucijoje, nėra priešiško nusistatymo prieš demokratiją pasekmė.*
- (76) *Henriko Plauenio užtikrinimas, kad jis norėj vėl išpirkti visas užstatytas gėrybes, buvo skeptiškai įvertintas ir nesušvelnino balijų valdovų priešiško nusistatymo didžiojo magistro planui.*

Tai rodo, kad būdvardžio *priešiškas* ir daiktavardžio *nusistatymas* sintaksinis valentingumas gali turėti įtakos naudininko ir konstrukcijų su *prieš* ir polinksniškuoju *atžvilgiu* vartojimui su daiktavardžiu *pyktis*.

4.2 Daiktavardžio *jausmas* sintaksinis junglumas

DLKT buvo tikrintas ir daiktavardžio *jausmas* junglumas. Paaiškėjo, kad šis daiktavardis irgi gali reikalauti kilmininko (žr. 77), naudininko (žr. 78), polinksniškojo *atžvilgiu* konstrukcijos (žr. 79) ir šalutinio sakinio (žr. 80). Vadinasi jo sintaksinis valentingumas yra Gen-N-Dat/Gen + *atžvilgiu/S*.

- (77) *Gal mano jausmai nenormalūs?*
- (78) *Viskų jai papasakos. Nors pavėluota nuo pat vestuvių, gal jos širdyje jausmai jam dar rusena.*
- (79) *Tad mano jausmai jų atžvilgiu buvo gana prieštarangi.*
- (80) *Vykite šalin mintis ir jausmus, kurie gali pakenkti kitiem.*

Buvo bandyta nustatyti ir žodžių junginio *priešiškas jausmas* valentingumą. DLKT rasta tik 11 sakinį ir tik 3 iš jų turėjo papildymą: du – naudininką (žr. 81), ir vienas – šalutinį sakinį (žr. 82).

- (81) *O aš nejaučiu tokios psichologinės būtinybės ir tokios moralinės pareigos, nors jokių priešiškų jausmų tiems mano buvusiems draugams ir pažystamiems nejaučiu.*
- (82) *Pakartotinos tėvo vedybos sukelia dukrai neigiamų ir priešiškų jausmų, kurie puikiai aprašyti pasakose apie Sigutę, Pelenę ir kt.*

Visa tai irgi rodo, kad daiktavardžio *jausmas* sintaksinis valentingumas gali turėti įtakos naudininko ir konstrukcijos su polinksniškuoju *atžvilgiu* vartojimui su daiktavardžiu *pyktis*.

4.3 Daiktavardžių *susierzinimas* ir *rūstis* sintaksinis jungumas

Dabar aptarkime daiktavardžių *susierzinimas* ir *rūstis* sintaksinį valentingumą. DLKT teikia tik 210 sakinį su daiktavardžiu *rūstis*. Paaiškėjo, kad šis daiktavardis retai reikalauja papildymų (10 atvejų). Jis gali reikalauti kilmininko (žr. 83), konstrukcijos su *už* (žr. 84) ir šalutinio sakinio¹⁶ (žr. 85), t. y. jo sintaksinis valentingumas yra Gen-N-*už* + Acc/S.

- (83) *O dar tėvų, o dar auklėtojų rūstis...*
- (84) [...] *jie ims jos nekesti už tai, kad primetė jų valdovui išdaviko likimą, – galiausiai baimintis dievų rūsties už tai, kad paklusso savo vyrui ir nepaisė švento įstatymo.*
- (85) *Šitas rūstis, gimęs iš tiek daugelio blogybių, yra didelė ir nepagydoma nuodėmė.*

Daiktavardžio *susierzinimas* pavartojimo atvejų DLKT yra 472, visi jie buvo peržiūrėti. Nustatyta, kad jis gali reikalauti kilmininko papildymo (žr. 86), konstrukcijos su *dėl* (žr. 87) ir šalutinio sakinio (žr. 88).

- (86) *Džandaro susierzinimas kaipmat praėjo.*
- (87) *Pasimatymų metu visiškai galimas dalykas, jog ji slėps savo nuomonę, kad jis neskonkingai rengiasi, o jis slėps savo susierzinimą dėl jos nuolatinio vėlavimo.*
- (88) *Jos šypsenai profesionaliai slėpė susierzinimą, nes jau antras keleivis tiek užsi- svajojo, kad pamiršo užgesinti cigarečę.*

Taigi daiktavardžių *susierzinimas* ir *rūstis* sintaksinis valentingumas negali veikti daiktavardžio *pyktis* polinkio reikalauti naudininko papildymo, taip pat konstrukcijų su *atžvilgiu* ir su *arieš*.

4.4 Daiktavardžio *piktumas* sintaksinis jungumas

Dar reikėtų aptarti daiktavardžio *piktumas* sintaksinį valentingumą. Kadangi šio daiktavardžio pavyzdžiu DLKT nedaug (451) ir jis gan retai vartojamas su papildymu, buvo peržiūrėti visi sakiniai. Nustatyta, kad *piktumas* gali reikalauti papildymų, išreikštų kilmininku (žr. 89–90), naudininku (žr. 89–90), konstrukcija su *ant* (žr. 91), konstrukcija su

¹⁶ Šalutiniais sakiniais laikytini ir sakiniai su dalyvinėmis formomis, pvz.: *Šitas rūstis, gimęs iš tiek daugelio blogybių, yra didelė ir nepagydoma nuodėmė* (plg. *rūstis, kuri gimė iš tiek daug blogybių*). Lietuvių kalbos gramatikos darbuose tokie sakiniai vadinami nesantykiniais būdvardiniai sakiniai, nes jie neturi santykinio žodžio (žr. Holvoet, Judžentis 2003b, 171–172). Anksčiausiuose gramatikos darbuose tokie sakiniai laikomi vientisiniais su derinamaisiais (išplėstiniiais arba sudėtiniais) pažyminiai, einančiai po pažymimo žodžio (Balkevičius 1963, 166–167; Ulvydas, red. 1976, 414–415; Ambrazas, red. 1996, 556; Labutis 1998, 323–324).

dėl (žr. 92), konstrukcija su *prieš* (žr. 93), konstrukcija su *tarp* (žr. 94) ir šalutiniu sakiniu (žr. 95). Vadinas daiktavardžio *piktumas* sintaksinis valentingumas yra Gen-N-Dat/ *ant* + Gen/*dėl* + Gen/*prieš* + Acc/*tarp* + Gen/S. Tai rodo, kad daiktavardžio *piktumas* sintaksinis valentingumas gali turėti įtakos naudininko ir prielinksnio *tarp* konstrukcijos vartojimui su daiktavardžiu *pyktis*.

- (89) *Ir dar žinok: piktumas mano tau seniai praėjo.*
- (90) *Tai ugdo šuns piktumą žvériui.*
- (91) *Bet kartu ir toks piktumas ant mamos, kad sergėk, Dieve!*
- (92) *Iš pradžių buvęs piktumas dėl klaikaus savaitgalio praėjo.*
- (93) *Viešpatie Okkupirmai, kartočiau: kodėl pasipylė toks piktumas prieš lietuvišką stabmeldybę, kai iš esmės tas pačias problemas savo romane iškėlė ir „stabmedys“ Alantas?*
- (94) *Tėtė atsikėlė, išėjo į kiemą ir iš karto suprato, kad jokio tikro piktumo tarp abiejų „stovyklų“ nėra.*
- (95) *O kai atsibudau, įmė piktumas, kad mane visaip tąso ir bado švirkštais.*

Daiktavardis *užsirūstimas* DLKT pavartotas vieną kartą ir be valentinių palydovų; daiktavardis *įširdimas* – tris kartus ir irgi be valentinių palydovų. Todėl, remiantis DLKT, neįmanoma nustatyti, ar jie turi įtakos daiktavardžio *pyktis* valentingumui.

5 Apibendrinimas ir išvados

Straipsnyje išnagrinėti 292 DLKT sakiniai su skirtingomis daiktavardžio *pyktis* formomis. Kaip rodo LKŽe ir DLKŽe pavyzdžiai, daiktavardis *pyktis* kartais atveria pozicijas kilmininko papildymui ir konstrukcijoms su *ant* + Gen bei *tarp* + Gen. Tačiau tyrimo metu nustatyta, kad daiktavardžio *pyktis* sintaksinis valentingumas įvairesnis, jis iš dailes sutampa su veiksmažodžio *pykti(s)* sintaksiniu valentingumu. Sutampa, nes, kaip ir veiksmažodis, gali prisijungti konstrukcijas *ant* + Gen, *dėl* + Gen, *už* + Acc, *su* + Instr ir šalutinių sakinjų. Skiriasi, nes vyksta vardininko linksnio perkodavimas: veiksmažodžiu *pykti(s)* valdomas vardininkas virsta daiktavardžiu *pyktis* valdomu kilmininku. Taip pat daiktavardis *pyktis*, skirtingai nuo veiksmažodžio, gali prisijungti naudininką ir konstrukcijas *prieš* + Acc, Gen + *atžvilgiu* bei *tarp* + Gen/Gen *ir* Gen. Nustatyta, kad šios daiktavardžio *pyktis* savybės susijusios su jo reikšmėje minimų raktinių žodžių (būdvardžio *priešiškas*, daiktavardžių *nusistatymas* ir *jausmas*) ir jo sinonimo (daiktavardžio *piktumas*) valentingumu. *Pyktis*, kaip ir *priešiškas*, *jausmas*, *nusistatymas* ir *piktumas*, reikalauja naudininko, gali reikalauti ir polinksniškojo *atžvilgiu* konstrukcijos. *Pyktis*, kaip ir *nusistatymas*, gali prisijungti konstrukcijas su prielinksniais *prieš* ir *tarp*. Todėl

naudininkas, konstrukcijos *prieš* + Acc, Gen + *atžvilgiu* ir *tarp* + Gen/Gen *ir* Gen traktuotini kaip daiktavardžio *pyktis* sintaksinio valentingumo variantai.

Kai sakinyje minimas / akcentuojamas pykti jaučiantis asmuo, daiktavardis *pyktis* dera su kilmininko ar prielinksnio *tarp* konstrukcijos percipientu. Kai kartu su daiktavardžiu *pyktis* nurodomas šios būsenos turinys, jis dera su kontentyvu (reiskiama naudininku, konstrukcija su *ant*, su *prieš*, kartais su *atžvilgiu*). Kai aktualizuojama emocinės būsenos priežastis, *pyktis* dera su kauzatyvu (prielinksnio *dėl* konstrukcija), kai aktualizuojama būsenos skatinamoji priežastis, pagrindas – su motyvu (prielinksnio *už* konstrukcija). Kai aktualizuojamas ir tarpusavio pykčio antraelis veikėjas, daiktavardis *pyktis* dera su komitatytu (prielinksnio *su* konstrukcija). Vadinasi, daiktavardžio *pyktis* semantinio valentingumo struktūra išlieka tokia pat, kaip ir veiksmažodžio *pykti(s)* struktūra, tik atsiranda daugiau būdų išreikšti kontentyvą ir percipientą.

Daiktavardžio *pyktis* leksinio valentingumo struktūra kiek pakitusi: kilmininko vardažodis gali žymėti ne tik žmogų, bet ir daiktą bei abstraktą; prielinksnių konstrukcija *ant* + Gen gali žymėti ne tik žmogų, bet ir negyvą daiktą bei abstraktą; prielinksnių konstrukcija *su* + Instr gali žymėti ne tik žmogų, bet ir abstraktą. Naudininko vardažodis gali žymėti žmogų, daiktą ar abstraktą, prielinksnio *prieš* konstrukcija – žmogų ar abstraktą. Vadinasi, leksinis daiktavardžio *pyktis* valentingumas yra įvairesnis nei leksinis veiksmažodžio *pykti(s)* valentingumas.

Kaip rodo tyrimas, daiktavardis *pyktis* ne tik paveldėjo veiksmažodžių *pykti* ir *pyktis* valentingumo savybes, bet ir įgijo naujų, nebūdingų šiems veiksmažodžiams.

Santrumpų sąrašas

Abstr – abstraktas, Acc – galininkas, Adj – būdvardis, Adv – prieveiksmis, +Anim – gyvas daiktas, -Anim – negyvas daiktas, Caus – kauzatyvas, Com – komitatytas, Con – kontentyvas, Dat – naudininkas, Gen – genitive, Hum – žmogus, Inf – bendratis, Instr – įnagininkas, M – motyvas, N – daiktavardis, Nom – nominative, Pcp – percipientas, Plur – daugiskaita, S – sakinys.

Šaltiniai

BLKŽe *Bendrinės lietuvių kalbos žodynas*. 2013–2023. Danutė Liutkevičienė, vyr. red., e. variantas. Vilnius: Lietuvių kalbos institutas, <https://ekalba.lt/bendrinės-lietuvių-kalbos-zodynės/>.

DLKT *Dabartinės lietuvių kalbos tekštynas*. Kaunas: Vytauto Didžiojo universitetas. <http://tekstynas.vdu.lt/tekstynas/>.

- DLKŽe *Dabartinės lietuvių kalbos žodynas*. 2021. 8-as patais. ir papild. leidimas. Stasys Keinys, vyr. red., e. variantas. Vilnius: Lietuvių kalbos institutas. <https://ekalba.lt/dabartines-lietuviu-kalbos-zodynus/>.
- LKŽe *Lietuvių kalbos žodynas 1–20 (1941–2002)*. 2005 (atnaujinta versija 2017). Gertrūda Naktinienė, vyr. red., e. variantas. Vilnius: Lietuvių kalbos institutas. <https://ekalba.lt/lietuviu-kalbos-zodynus/>.
- LKVJŽ Sližienė, Nijolė. 1998. *Lietuvių kalbos veiksmažodžių junglumo žodynas 2 (1)*. Vilnius: Mokslo ir enciklopedijų leidykla.
- SŽ Lyberis, Antanas. 2002. *Sinonimų žodynas*. 2-asis patais. leidimas. Vilnius: Lietuvių kalbos institutas. Elektroninis variantas 2015. <https://ekalba.lt/sinonimu-zodynus/>.
- VLE *Visuotinė lietuvių enciklopedija*. <https://www.vle.lt/>.

Literatūra

- Akelaitis, Gintautas. 2005. Konstrukcijos su *atžvilgiu, požiūriu* administracinėje kalboje. *Kalbos kultūra* 78, 67–78.
- Alexiadou, Artemis, Monika Rathert, eds. 2010. *The Syntax of Nominalizations Across Languages and Frameworks*. Berlin: De Gruyter Mouton. (*Interface Explorations* 23)
- Ambrasas, Vytautas, red. 1996. *Dabartinės lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidykla.
- Balkevičius, Jonas. 1963. *Dabartinės lietuvių kalbos sintaksė*. Vilnius: Valstybinė politinės ir mokslinės literatūros leidykla.
- Čižik-Prokaševa, Voslava. 2024. Daiktavardžio *neapykanta* ir su juo semantiškai susijusių žodžių valentingumas dabartinėje lietuvių (rašto) kalboje. *Taikomoji kalbotyra* 21, 158–177.
- Durme, Karen van, ed. 1997. *The Valency of Nouns*. Odense: Odense University Press. (*Odense Working Papers in language and communication* 15)
- Geniušienė, Emma. 1971. Lietuvių kalbos veiksmažodžių sintaksinė klasifikacija. *Kalbotyra* 23 (1), 7–16.
- Grabauskas, V. 1971. Dabartinės lietuvių kalbos naudininko reikšmės ir funkcijos. *Kalbotyra* 22 (1), 49–59.
- Grimshaw, Jane. 1990. *Argument Structure*. Cambridge MA: The MIT Press.
- Haspelmath, Martin. 2007. Further remarks on reciprocal constructions. *Reciprocal Construction* 4. Vladimir P. Nedjalkov, ed. Amsterdam & Philadelphia: John Benjamins Publishing Company. 2087–2116.
- Helbig, Gerhard. 1992. *Probleme der Valenz- und Kasustheorie*. Tübingen: M. Niemeyer.
- Holvoet, Axel, Artūras Judžentis. 2003a. Sintaksinių ryšių tipai. *Sintaksinių ryšių tyrimai*. Axel Holvoet, Artūras Judžentis, red. Vilnius: Lietuvių kalbos institutas. 11–35. (*Lietuvių kalbos gramatikos darbai* 1)

- Holvoet, Axel, Artūras Judžentis. 2003b. Sudėtinio prijungiamojo sakinio aprašymo pagrindai. *Sintaksinių ryšių tyrimai*. Axel Holvoet, Artūras Judžentis, red. Vilnius: Lietuvių kalbos institutas. 115–172. (*Lietuvių kalbos gramatikos darbai* 1)
- Karlík, Petr, Norbert Nübler. 1998. Poznámky k nominalizaci v češtině. *Slovo a slovesnost* 59, 105–112.
- Kolářová, Veronika 2014. Special valency behavior of Czech deverbal nouns. *Noun valency*. Olga Spevak, ed. Amsterdam: John Benjamins Publishing Company. 19–59.
- Labutis, Vitas. 1998. *Lietuvių kalbos sintaksė*. Vilnius: Vilniaus universiteto leidykla.
- Laužikas, Laimutis. 1986. Subjektinis determinantas neasmenuojamujų veiksmažodžių predikatinėse konstrukcijose. *Kalbotyra* 37 (1), 67–76.
- Mackenzie, J. Lachlan. 1997. Nouns are avalent – and nominalizations too. *The Valency of Nouns*. Karen van Durme, ed. Odense: Odense University Press. 89–118. (*Odense Working Papers in Language and Communication* 15)
- Pakerys, Jurgis. 2006. Veiksmo pavadinimo konstrukcija lietuvių kalbos gramatikoje. *Daiktavardinio junginio tyrimai*. Axel Holvoet, Rolandas Mikulskas, red. Vilnius: Lietuvių kalbos institutas. 121–149. (*Lietuvių kalbos gramatikos darbai* 4)
- Panevová, Jarmila. 2014. Contribution of valency to the analysis of language. *Noun valency*. Olga Spevak, ed. Amsterdam: John Benjamins Publishing Company. 1–17.
- Rathert, Monika, Artemis Alexiadou, eds. 2010. *The Semantics of Nominalizations Across Languages and Frameworks*. Berlin: De Gruyter Mouton. (*Interface Explorations* 22)
- Sližienė, Nijolė. 1978. Veiksmažodžių valentingumo klausimu. *Gramatinės kategorijos ir jų raida*. 107–123. (*Lietuvių kalbotyros klausimai* 18)
- Sližienė, Nijolė. 1994. *Lietuvių kalbos veiksmažodžių junglumo žodynas* 1. Vilnius: Mokslo ir enciklopedijų leidykla.
- Sližienė, Nijolė. 1998. *Lietuvių kalbos veiksmažodžių junglumo žodynas* 2 (1). Vilnius: Mokslo ir enciklopedijų leidykla.
- Sližienė, Nijolė. 2004. *Lietuvių kalbos veiksmažodžių junglumo žodynas* 2 (2). Vilnius: Lietuvių kalbos instituto leidykla.
- Spevak, Olga, ed. 2014. *Noun valency*. Amsterdam: John Benjamins Publishing Company.
- Šukys, Jonas. 1998. *Lietuvių kalbos linksniai ir prielinksniai: vartosena ir normos*. Kaučius: Šviesa.
- Tekorienė, Dalija. 1983. Lietuvių kalbos būdvardžio semantinis valentingumas. *Kalbotyra* 34 (1), 132–139.
- Tekorienė, Dalija. 1985. Lietuvių kalbos būdvardžių būtinojo semantinio valentingumo raiška. *Kalbotyra* 36 (1), 72–83.
- Tesniére, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
- Ulydyas, Kazys, red. 1976. *Lietuvių kalbos gramatika* 3 (*Sintaksė*). Vilnius: Mokslas.

- Vaičiulytė-Seménienė, Loreta. 2001. *Lietuvių kalbos būdvardžių valentingumas*. Dakto-ro disertacija. Vilnius.
- Vaičiulytė-Seménienė, Loreta. 2016. Teigiamų alelinių būdvardžių ir jų vedinių aktantu raiška. *Bendrinė kalba* 89, 1–26. <https://journals.lki.lt/bendrinekalba/article/view/170/175>.
- Vaičiulytė-Seménienė, Loreta. 2020. Daiktavardžiai *atitiktis, atitikimas* administracinėje kalboje. *Bendrinė kalba* 93, 1–35. <https://journals.lki.lt/bendrinekalba/article/view/2059>.
- Vaičiulytė-Seménienė, Loreta. 2022. *Atžvilgiu* konstrukcija valentingumo aspektu. *Acta Linguistica Lithuanica* 87, 245–269. <https://journals.lki.lt/actalinguisticalithuanica/article/view/2194/2298>.
- Vaičiulytė-Seménienė, Loreta. 2023a. *Prieraišus, prieraišumas* valentingumas dabartinėje lietuvių (rašto) kalboje. *Acta Linguistica Lithuanica* 88, 181–205. <https://journals.lki.lt/actalinguisticalithuanica/article/view/2245/2338>.
- Vaičiulytė-Seménienė, Loreta. 2023b. *Atvirauti, atviravimas* valentingumas dabartinėje lietuvių kalboje. *Acta Linguistica Lithuanica* 89, 75–92. <https://journals.lki.lt/actalinguisticalithuanica/article/view/2279/2350>.
- Valiulytė, Elena. 1978. Veiksmažodžio *pykti* bei jo sinonimų junginiai su prielinksniais *dėl* ir *už*. *Kalbos kultūra* 35, 26–30.
- Vaskelaitė, Ramunė. 2003. Naudininkas veiksmažodinių daiktavardžių junginiuose. *Kalbos kultūra* 76, 43–46.

Melanie Andresen. *Computerlinguistische Methoden für die Digital Humanities: Eine Einführung für Geisteswissenschaftler:innen*. Tübingen: Narr Francke Attempto (= narr STUDIENBÜCHER), 2024, 240 S., ISBN 978-3-8233-8579-0, EAN 978382338579.

Rezensiert von / Reviewed by **Christa Dürscheid**

Deutsches Seminar
Universität Zürich
Schönberggasse 9
8001 Zürich, Schweiz
E-Mail: duerscheid@ds.uzh.ch
ORCID iD: 0000-0001-9141-7562
<https://ror.org/02crff812>

Vorbemerkungen

„Die Digital Humanities ersetzen – selbstverständlich – die Geisteswissenschaften nicht, sondern erweitern ihr Methodenarsenal, und zwar dann, wenn Forschungsfragen durch empirische Arbeit mit großen Datensammlungen bearbeitet werden können“ (Krämer 2025, 11). Diese Aussage steht zu Beginn eines Essays, das den markanten Titel „Der Stachel des Digitalen: Geisteswissenschaften und Digital Humanities“ trägt. Darin setzt sich die Philosophieprofessorin Sybille Krämer kritisch mit dem Selbstbild der Geisteswissenschaften auseinander und zeigt auf, in welchem Verhältnis diese zu den Methoden der Digital Humanities stehen. Doch um welche Methoden handelt es sich dabei? Und welche Vorteile können sich daraus ergeben, wenn traditionelle Verfahren der Textanalyse, d. h. Textinterpretation und Hermeneutik, eine Erweiterung ihres „Methodenarsenals“ erfahren? In dem Studienbuch von Melanie Andresen, das hier zur Rezension

Submitted: 10/11/2025. **Accepted:** 17/11/2025

Copyright © 2025 Christa Dürscheid. Published by Vilnius University Press

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

steht, werden diese Fragen nicht nur auf theoretischer Ebene erörtert; es wird eine praxisorientierte Einführung präsentiert, die im Unterschied zu ähnlich gelagerten Überblicksdarstellungen (z. B. Alpaydin 2022; Mitkov 2022) den Schwerpunkt explizit auf die Anwendung computerlinguistischer Methoden in den Geisteswissenschaften legt.

Allerdings ist der Untertitel des Buches – „Eine Einführung für Geisteswissenschaftler:innen“ – etwas zu weit gefasst. Denn diese Formulierung lässt vermuten, dass sich die Verfasserin auf Anwendungsfelder aus verschiedenen geisteswissenschaftlichen Disziplinen bezieht (so z. B. auch auf die Geschichtswissenschaft, die Kunstgeschichte, die Theologie oder die Musikwissenschaft), tatsächlich aber hat sie vor allem die Sprach- und Literaturwissenschaft im Blick. Dass die Verfasserin einen solchen Schwerpunkt setzt, verwundert nicht. Sie hat an der Universität Hamburg ein Masterstudium in germanistischer Linguistik absolviert, und auch die Dissertation, die 2022 unter dem Titel *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen* erschien, hat einen germanistisch-linguistischen Hintergrund. Nach der Promotion war sie als Post-Doc am Institut für Maschinelle Sprachverarbeitung an der Universität Stuttgart, seit 2024 arbeitet sie bei der Firma *DeepL* an der Verbesserung maschineller Übersetzung (so die Autor:inneninformation auf der Webseite des Narr-Verlags). Aus dem Vorwort zum Buch entnimmt man zudem, dass Melanie Andresen während ihrer Tätigkeit an der Universität Stuttgart Lehrveranstaltungen unter dem Titel „Computerlinguistische Methoden für die Digital Humanities“ durchgeführt hat. Diese dienten ihr als Grundlage für das vorliegende Studienbuch.

Der Werdegang von Melanie Andresen wurde an dieser Stelle auch deshalb kurz skizziert, um deutlich zu machen, dass die Verfasserin vertiefte Kenntnisse in der germanistischen Linguistik und in der Computerlinguistik hat und über Unterrichtserfahrung in den Digital Humanities verfügt. All das kommt ihrem Einführungswerk, wie noch zu zeigen sein wird, zugute. Doch im Titel, im Klappentext und in der Einleitung hätte deutlich werden müssen, dass es im Buch vor allem um die Anwendung computerlinguistischer Methoden in den Philologien (und hier insbesondere in der Germanistik) geht. Zwar mag man einwenden, dass die von Frau Andresen vorgestellten Methoden auch auf andere geisteswissenschaftliche Disziplinen übertragbar sind. Das trifft aber nur dann zu, wenn in diesen Disziplinen Forschungsfragen gestellt werden, die sich auf der Basis von Textanalysen beantworten lassen. Darüber hinaus wäre es interessant gewesen, im Buch etwas über solche Methoden zu erfahren, die zur Bildanalyse oder zur Analyse von Audiodateien (z. B. in der Musikwissenschaft) eingesetzt werden können. Darauf geht die Verfasserin nicht ein. Auch mit der Analyse gesprochener Sprache (z. B. in der Interaktion mit Sprachassistenten) befasst sie sich nicht. Eine Begründung dafür wird nicht gegeben, auf S. 15 heißt es dazu nur: „In diesem Buch wird es nur um die schriftliche Form von Sprache gehen.“

Aufbau des Buches

Das Kernstück des Buches bilden zwei Teile, die insgesamt mehr als 170 Seiten umfassen (Teil 1: „Linguistische Ausgangspunkte“; Teil 2: „Methoden“). Alle Kapitel im ersten Teil enden mit „Beispielstudien“ (z. B. zu Wahlprogrammen, Interviews und literarischen Werken) und „Übungen“, wobei zu den Übungen auf der Webseite des Narr-Verlags auch Musterlösungen angeboten werden (siehe dazu aber weiter unten). Ergänzt werden diese beiden Themenblöcke um einen dritten, wesentlich kürzeren Teil (elf Seiten), der unter der etwas vagen Überschrift „Gesellschaft“ steht und nur ein Kapitel beinhaltet. Den drei Teilen vorangestellt sind ein Vorwort und eine Einleitung, die – obwohl insgesamt recht kurz (neun Seiten) – bereits in wichtige Grundbegriffe einführt. Ein Verzeichnis der Skripts und der vielen interessanten digitalen Ressourcen, auf die im Laufe des Textes verwiesen wird (Korpora, Tools zur Korpusanalyse, Programmierumgebungen für Python etc.), folgt am Ende des Buches (alle Quellen wurden zuletzt im Dezember 2023 geprüft). Auch finden sich hier ein umfangreiches, sorgfältig aufbereitetes Literaturverzeichnis und ein Sachregister, das vom Stichwort „Abkürzungen“ (warum im Plural?) bis zu „Zufallsstichprobe“ reicht. So informativ dieser Service-Teil (der insgesamt über 30 Seiten umfasst) auch ist, es zeigt sich spätestens jetzt, welche Nachteile damit einhergehen, wenn man für die Lektüre des Studienbuches nur die Printausgabe vorliegen hat. Will man die entsprechenden Webseiten im Internet aufrufen, müssen die Links manuell eingetippt werden. In der E-Book-Ausgabe des Buches stellt sich dieses Problem nicht. Als Open Access-Publikation steht das E-Book aber leider nicht zur Verfügung. Im Interesse einer kostenfreien Bereitstellung von Forschungsergebnissen hätte man sich das anders gewünscht.

Zu den einzelnen Kapiteln

In der Einleitung trägt die Verfasserin viele Informationen zusammen, die hilfreich für die Lektüre der Folgekapitel sind (Kap. 1). Dazu gehört nicht nur, wie in Einleitungen zu wissenschaftlichen Monographien üblich, dass sie Inhalt und Aufbau des Buches knapp skizziert, sie behandelt schon hier interessante thematische Aspekte. So geht sie auf die (nicht triviale) Frage ein, welche Gemeinsamkeiten und Unterschiede zwischen Computerlinguistik und Korpuslinguistik bestehen und was man unter Termini wie *Metadaten*, *Annotationen* und *Tagsets* versteht. Auch stellt die Verfasserin in Kap. 1 – allerdings nur sehr kurz – das Korpus vor, auf das sie sich im Folgenden immer wieder bezieht. Es ist das Foodblog-Korpus, das 150 Texte aus 15 verschiedenen Blogs umfasst (vgl. Andressen & Zinsmeister 2018). Die Einleitung ist insgesamt sehr informativ, sie ist – wie der gesamte Text – flüssig geschrieben und sie beginnt wie alle folgenden Hauptkapitel auch: Zunächst legt Melanie Andressen in einigen wenigen Zeilen dar, welche Themen-schwerpunkte sie im Folgenden setzen wird, dann erst beginnt sie mit ihren inhaltlichen

Ausführungen. Das ist eine lesefreundliche Vorgehensweise, zudem ist dieser einleitende Passus jeweils recht eingängig formuliert – und dies konsequent im *Wir*-Duktus. Das liest sich dann so: „Wir klären außerdem, was genau die Computerlinguistik ist“ (S. 11). Oder an späterer Stelle: „Wir schauen uns einerseits lexikonbasierte Ansätze an, [...], andererseits betrachten wir Ansätze mit maschinellem Lernen“ (S. 81).

Teil 1: Linguistische Ausgangspunkte

Wie die Überschrift zu Teil 1 bereits vermuten lässt, steht in diesem ersten Teil des Buches die Linguistik (und nicht die Computerlinguistik) im Zentrum. Als „Ausgangspunkte“ dienen verschiedene linguistische Beschreibungsebenen, die von der Lexik über die Syntax bis zur Semantik und Pragmatik reichen (vgl. Kap. 2 bis 8). Die ersten Kapitel lesen sich fast wie eine Bachelor-Einführung in die Sprachwissenschaft – ergänzt um Informationen zu Grundbegriffen der Korpusanalyse (z. B. *Tokenisierung*, *Lemma-tisierung*, *Type-Toke-Ratio*). Inhaltlich sind die Ausführungen in den Kap. 2 bis 4 sehr gut nachvollziehbar. Stellenweise sind sie sogar so elementar, dass sich zunächst nicht erahnen lässt, wie anspruchsvoll der Text in der Folge noch werden wird. So beginnt das Kapitel zur Lexik mit einem Überblick über verschiedene Definitionen zum Wort-Konzept (Kap. 2.1), das Wortarten-Kapitel listet einleitend eine Reihe von Kriterien auf, nach denen sich die Wortarten im Deutschen unterscheiden lassen (Kap. 3.1). Im Syntax-Kapitel informiert die Verfasserin über die Grundlagen der Konstituentengrammatik und Dependenzgrammatik, dann leitet sie über zu der Frage, welche Möglichkeiten sich auf Basis dieser beiden syntaktischen Theorien für die automatisierte Textanalyse bieten. Sie stellt verschiedene Annotationsschemata vor (z. B. das Schema der *Hamburg Dependency Treebank*, HDT), beschreibt die Speicherung und Weiterverarbeitung von Dependenzannotationen (z. B. im CoNLL-Format) und gibt einen Überblick zur computerbasierten Syntaxanalyse (Kap. 4.3). In diesem Zusammenhang geht sie auch auf das Dependenzparsing und die Graphentheorie ein und nennt Tools, mit denen syntaktische Strukturen analysiert werden können (z. B. CoreNLP). Zum Schluss präsentiert Melanie Andresen – wie in allen Kapiteln von Teil 1 – „Beispielstudien“ (hier z. B. zur Analyse von literarischen Texten) und „Übungen“. Letztere starten mit einfachen Fragen zur Satzgliedbestimmung, dann folgen Aufgaben zur automatischen Annotation und zur Arbeit mit Musterskripten (Kap. 4.5).

Die folgenden drei Kapitel zur Semantik tragen die Überschriften „Wortfelder“ (Kap. 5), „Sentimentanalyse“ (Kap. 6) und „Distributionelle Semantik“ (Kap. 7). Zunächst erinnert die Lektüre wieder an eine Bachelor-Einführung in die Sprachwissenschaft. Zu Beginn von Kap. 5 kündigt die Verfasserin an, dass man sich nun „mit der linguistischen Teildisziplin der Semantik [befassen werde], die die Bedeutung von Wörtern und ihre semantischen Relationen behandelt“ (S. 71). Im Anschluss daran geht sie ausführlich auf

Ähnlichkeitsbeziehungen zwischen Wörtern und auf die Wortfeldtheorie ein und weist auf nützliche Tools zur automatisierten Wortschatzanalyse hin (z. B. *GermaNet*). Anspruchsvoller wird die Lektüre in den Kapiteln 6 und 7, in denen Melanie Andresen das Erstellen von Sentimentwörterbüchern und die Möglichkeiten zur computergestützten Analyse von Emotionen thematisiert. Auch stellt sie hier die Grundlagen der distributionellen Semantik vor und erklärt, wie die semantischen Relationen zwischen Wörtern auf der Basis von Vektorwerten und Kosinusähnlichkeiten berechnet werden können. Damit schafft sie die Grundlagen, die nötig sind, um in den Kapiteln 7.3.1 bis 7.3.5 ihren Ausführungen zum *Word Embedding* folgen zu können. Veranschaulicht wird dieses Verfahren wieder mit Beispielen aus dem Foodblog-Korpus. Es wird gezeigt, wie die Wörter *kochen* und *Minuten* als Vektorwerte dargestellt werden können und wie sich automatisiert ermitteln lässt, welche Wörter im Kontext anderer Wörter vorkommen. Zum Schluss des Kapitels stellt die Verfasserin einige Arbeiten vor, in denen in den Geisteswissenschaften mit *Word Embeddings* gearbeitet wurde, um „Wordfelder“ [sic] zu erstellen, um Synonyme zu einem Zielwort zu ermitteln oder die Veränderung von Wortbedeutungen zu untersuchen (vgl. S. 106 f.): Sie verweist auf eine Untersuchung zur automatischen Erkennung von Shakespeare-Referenzen in modernen Texten, auf eine diachrone Studie zum Wortschatz in der englischen Wissenschaftssprache und auf eine Arbeit zur Verwendung des Wortes *national* in verschiedenen Korpora. Die „Übungen“ in Kap. 7.5 umfassen vier Aufgaben, von denen sich zwei abermals auf das Foodblog-Korpus beziehen (hier auf die Berechnung der Kosinusähnlichkeit zwischen Wortpaaren wie *backen* und *Butter*, *Mehl* und *Salz*).

Das Kapitel „Pragmatik“ (so die Überschrift) steht am Ende des ersten Themenblocks. Die Verfasserin erläutert zunächst die Termini *Entität* und *Referenz* und geht dann auf das computergestützte Verfahren der Eigennamenerkennung (*Named Entity Recognition*) und das Vorgehen bei der Annotation von Koreferenzrelationen ein. Wie praktisch diese Methoden sind, um Texte inhaltlich zu erschließen, macht sie wieder in den „Beispielstudien“ (Kap. 8.4) deutlich. Dieser Abschnitt fällt mit drei Seiten sogar etwas länger aus als die Beispielstudien in den vorangehenden Kapiteln. Die Verfasserin stellt hier eine Untersuchung zur Figurencharakterisierung in einem Roman von Juli Zeh vor, die sie selbst zusammen mit Michael Vauth durchgeführt hat, sie geht aber auch auf andere interessante Studien ein (z. B. Braun & Ketschik 2019).

Teil 2: Methoden

Der zweite Teil des Studienbuchs fällt im Vergleich zu Teil 1 wesentlich kürzer aus, er umfasst nur vier Kapitel und ca. 70 Seiten. Wer sich gezielt über aktuelle korpus- und computerlinguistische Methoden informieren möchte, sei direkt auf diesen Teil verwiesen. In Kap. 9 erklärt Melanie Andresen auf didaktisch geschickte Weise, was reguläre

Ausdrücke sind; auch die Informationen zur Berechnung von Frequenzen, zur deskriptiven Statistik und zur Interenzstatistik sind gut nachvollziehbar. Sehr nützlich ist weiter, dass sie verschiedene Arten von Visualisierungen vorstellt (Säulendiagramme, Tortendiagramme, Boxplots u. a.) und deutlich macht, wie wichtig solche Darstellungen für die Textanalyse und die Ergebnispräsentation sind. Auch hier stehen Übungen am Ende des Kapitels (etwa zur Bestimmung des Skalenniveaus und zur Analyse von Beispielskripten), doch nur zu drei der vier Aufgaben finden sich im Internet auch Musterlösungen. Auf eine Zusammenstellung von „Beispielstudien“ wird verzichtet.

In Kap. 10 erläutert die Verfasserin die Unterschiede zwischen manueller und automatischer Annotation, die Funktion von Annotationsrichtlinien und die verschiedenen Verfahren zur Qualitätskontrolle (z. B. das Inter-Annotator-Agreement). Thematisch schließt das Kapitel an Teil 1 der Einführung an, zum Verständnis ist es aber nicht nötig, die vorangehenden Ausführungen im Detail gelesen zu haben. Zwar fehlt auch hier ein separates Kapitel mit Beispielstudien, im Verlauf des Kapitels werden gelegentlich aber Bezüge zu spezifischen Forschungsfragen in den Geisteswissenschaften hergestellt (so etwa zu der Frage, wie der Mediendiskurs über ein bestimmtes Ereignis ausgewertet werden kann).

Die folgenden Kapitel befassen sich mit dem Maschinellen Lernen (Kap. 11) und seiner Weiterentwicklung, dem Deep Learning (Kap. 12), also mit solchen Methoden, die im Gegensatz zu regelbasierten Systemen mit KI-basierten Techniken arbeiten und auch in den Digital Humanities immer wichtiger werden. Insofern ist es besonders lobenswert, dass die Verfasserin dem maschinellen Lernverfahren und der Frage, wie künstliche neuronale Netze funktionieren, gleich zwei Kapitel widmet. Dabei ist sie durchweg bemüht, eine verständliche Einführung in diese komplexe Thematik zu geben. Das gelingt ihr über weite Strecken, doch wird diese Aufgabe immer anspruchsvoller – und das nicht zuletzt deshalb, weil ein solides mathematisches Grundlagenwissen erforderlich ist, um den Ausführungen folgen zu können. Die Inhalte der beiden Kapitel können hier nicht im Detail vorgestellt werden, es seien nur einige Aspekte genannt, die zur Sprache kommen und von der Verfasserin Schritt für Schritt erläutert werden. Dazu gehören (ausgewählten Überschriften in Kap. 11 und 12 folgend): die Unterscheidung zwischen überwachtem und unüberwachtem Lernen (Kap. 11.2), der Musterablauf einer Klassifikation (Kap. 11.3), das Training eines Deep-Learning-Modells (Kap. 12.3), Recurrent Neural Networks (Kap. 12.5) und Transformer (Kap. 12.6). Welche Forschungsfragen damit bearbeitet werden können, rückt hier allerdings in den Hintergrund, der Fokus liegt darauf, die Methoden so weit verständlich zu machen, dass man selbst künstliche neuronale Netze programmieren und trainieren oder auf einem Server damit arbeiten kann. Doch immerhin wird in den „Übungen“ in Kap. 12.8 danach gefragt, welche Anwendungsfälle es für Deep Learning in den Geisteswissenschaften geben könnte. Auf der Webseite des

Narr-Verlags findet man dazu aber keine Hinweise. Das gilt auch für die anderen beiden Übungen, die hier angegeben sind (z. B. zum Modell *German BERT*). Auf der Webseite steht dazu nur, dass keine Musterlösungen vorhanden seien.

Teil 3: Gesellschaft

Dieser Teil besteht nur aus einem Kapitel; es trägt die Überschrift „Computerlinguistik und Ethik“ (Kap. 13). Die Verfasserin nimmt hier vor allem solche Fragen in den Blick, die ethische Probleme beim Einsatz von KI-basierten Tools betreffen (etwa die Reproduktion von Diskriminierungen). Sie geht aber – anders als es die Überschrift erwarten lässt – auch auf ökologische Aspekte ein (so z. B. auf den enormen Energieverbrauch, der aus dem Training und der Nutzung von Sprachmodellen resultiert). In Kap. 13.5 problematisiert sie abschließend den Umstand, dass gesellschaftliche Entscheidungsprozesse davon abhängen können, welche Daten überhaupt für die Textanalyse zur Verfügung stehen. Wenn in historischen Textkorpora bestimmte Personengruppen über- oder unterrepräsentiert sind (z. B. Frauen), könne dies, so legt sie dar, dazu führen, dass in den Digital Humanities bestimmte Texte weniger Berücksichtigung finden (vgl. S. 208). So seien im literarischen Kanon Männer stark überrepräsentiert, Texte von Frauen würden kaum Beachtung finden. Mit solch kritischen Überlegungen leitet die Verfasserin zum Ende ihres Studienbuchs über und plädiert im letzten Satz „für eine differenzierte Sicht auf Daten und Analysen“ (S. 208).

Fazit

Melanie Andresen ist es mit ihrem Studienbuch hervorragend gelungen, in ein „Methodenarsenal“ (siehe das obige Zitat) einzuführen, das in den Geisteswissenschaften immer mehr an Bedeutung gewinnt. Sie führt ihre Leserinnen und Leser souverän durch alle 13 Kapitel und achtet dabei immer darauf, dass der Text, der zunehmend an Komplexität gewinnt, verständlich bleibt. Auf formaler Ebene gibt es nichts zu beanstanden; das gesamte Buch wurde sorgfältig gestaltet, der Text ist sehr gut ausformuliert, mit vielen anschaulichen Abbildungen und interessanten Zusatzmaterialien. Die wenigen Kritikpunkte, die zu nennen sind, beziehen sich vor allem auf das, was fehlt: Nicht zu allen Übungsaufgaben gibt es Musterlösungen; auch hätte man sich ein Abkürzungsverzeichnis gewünscht, um die vielen Kürzel (TTR, STTR, STTS, POS, HMM, HDT, CoNLL, NER etc.) nachschlagen zu können. Auf inhaltlicher Ebene wäre es bereichernd gewesen, wenn mehr Bezüge zu geisteswissenschaftlichen Forschungsfragen hergestellt worden wären und auch auf die Analyse nicht-geschriebener Daten eingegangen worden wäre.

Diese Anmerkungen sollen aber nicht die enorme Leistung schmälern, die hinter dem Studienbuch steht, sondern lediglich als Hinweise darauf verstanden werden, was in ei-

ner zweiten Auflage noch ergänzt werden könnte. Und eine zweite Auflage ist dem Buch zu wünschen. Bis hinein in die vielen weiterführenden Fußnoten stellt Melanie Andresen ihr profundes Fachwissen unter Beweis und arbeitet auch schwierige Inhalte didaktisch geschickt auf. Im Resultat liegt ein Werk vor, das mehr ist als ein Studienbuch; es ist ein Grundlagentext für die Digital Humanities. Die Lektüre wird allen empfohlen, die sich über aktuelle computerlinguistische Methoden informieren möchten.

Literatur

- Alpyadin, Ethem. 2022. *Maschinelles Lernen*. 3., aktualisierte und erweiterte Auflage. Berlin: De Gruyter.
- Andresen, Melanie. 2022. *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen. Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen*. Tübingen: Narr Francke Attempto.
- Andresen, Melanie & Michael Vauth. 2020. Figurenrelationen und Figurencharakterisierung. Interdisziplinarität zwischen Literaturwissenschaft und Computerlinguistik am Beispiel der Text- und Genreanalyse. *Kultur und Technik. Interdisziplinäre Perspektiven*. Dominik Orth & Margarete Jarchow, Hrsg. Kiel/Hamburg: Wachholtz. 43–62.
- Andresen, Melanie & Heike Zinsmeister. 2018. *Foodblog-Korpus*. Zenodo. <https://doi.org/10.5281/zenodo.1410445>
- Braun, Manuel & Nora Ketschik. 2019. Soziale Netzwerkanalysen zum mittelhochdeutschen Artusroman oder: Vorgreiflicher Versuch, Märchenhaftigkeit des Erzählens zu messen. *Das Mittelalter* 24 (1), 54–70.
- Krämer, Sybille. 2025. *Der Stachel des Digitalen. Geisteswissenschaften und Digital Humanities*. Frankfurt: Suhrkamp.
- Mitkov, Ruslan (Hrsg.). 2022. *The Oxford Handbook of Computational Linguistics*. 2. Auflage. Oxford: Oxford University Press.

Atmena autoriams

Žurnalas ***Kalbotyra*** skirtas įvairių kalbų aspektų (taip pat ir tarpkalbinių) tyrimams. Jame spausdinami mokslo straipsniai, knygų recenzijos, pranešimai apie konferencijas. Gali būti spausdinama ir konferencijų medžiaga.

Pateikiami straipsniai ir recenzijos neturi būti publikuoti anksčiau ar atiduoti publikuoti kituose leidiniuose. Kiekvieną iš jų recenzuoja bent du anoniminiai recenzentai.

Autoriai, norėdami įteikti straipsnį, turi registruotis žurnalo leidybos sistemoje ([nuoroda](#)) ir ten pat įkelti straipsnio rankraščius dviem formatais: MS Word (*.docx) ir PDF (angl. *Portable Document Format*, *.pdf). Įrašius dokumentą PDF formatu rekomenduojama patikrinti, ar teksto ir iliustracijų formatavimas išliko nepakitęs. Kad būtų užtikrintas anoniminis recenzavimas, autorius(-iai) privalo pateikti du variantus: vieną straipsnio tekštą kaip reikalaujama atmenoje, antrą variantą – be nuorodų ar užuominų į autorystę. Straipsnio failas turi būti pateiktas taip, kad Jame neliktu duomenų, galinčių padėti identifikuoti autorij (būtina pašalinti informaciją iš dokumento skilties *properties*).

Publikacijos apimtis paprastai neviršija 8 000 žodžių; kai kuriais atvejais gali būti siūlomi ir ilgesni straipsniai.

Pateikiami rankraščiai turi būti parengti pagal toliau išdėstytaus reikalavimus viena iš šių kalbų: anglų, lietuvių, prancūzų ar vokiečių. Jei straipsnio kalba autorui nėra gimtoji, toks tekstas gali būti teikiamas tik suredaguotas gimtakalbio specialisto.

Autorius(-iai) prisiima atsakomybę už tai, kad galutinis pateikiamo publikuoti rankraščio tekstas visiškai atitiktų toliau išdėstytaus žurnalo reikalavimus.

Autorius(-iai) privalo garantuoti, kad jų autoriniame darbe nėra pažeistos trečiųjų asmenų autorinės teisės ir kad tiesiogiai ar netiesiogiai panaudodami kitų autorij mintis jie pateikia nuorodas į šaltinius. Žurnale ***Kalbotyra*** publikuoti straipsniai nekomerciniais tikslais, nurodant autorij ir pirminj šaltinj, gali būti naudojami pagal Kūrybinių bendrijų (*Creative Commons*) licenciją CC Attribution 4.0.

1 Struktūra ir forma

Pateikiami straipsniai turi atitikti bendruosius straipsniams keliamus reikalavimus. Juose turi būti suformuluotas tyrimo klausimas/problema ir tikslas, apžvelgti ankstesni tiriavimos srities darbai, apibūdinti duomenys ir metodai, pateikti rezultatai ir argumentuotos išvados bei nurodyti duomenų šaltiniai ir naudota literatūra. Darbai, neatitinkantys šių reikalavimų, bus grąžinami autoriams taisyti.

Straipsnių tekstas teikiamas A4 formato lape, kurio paraštės yra tokios: 1,5 cm dešinėje, 2,5 cm viršuje, kairėje ir apačioje. Puslapiai numeruojami nuo pirmo iki paskutinio vir-

šutiniame dešiniajame puslapio kampe. Tekstas rašomas tarp eilučių paliekant 1,5 intervalą, 12 pt Times New Roman šriftu (jei nereikalaujama kitaip), lygiuojamas tik kairėje.

Straipsnio pradžioje pateikiami:

(1) straipsnio pavadinimas, 14 pt, paryškintas,

(2) pilnas(-i) autoriaus(-ių) vardas(-ai) ir pavardė(-s) paryškintu šriftu; visa informacija rašoma 12 pt šriftu publikacijos kalba; nurodomas pilnas institucijos pavadinimas (įskaitant katedrą, centrą, institutą ar kt.), institucijos adresas; taip pat nurodomas autoriaus elektroninio pašto adresas (be pabraukimo), autoriaus ORCID numeris (orcid.org), institucijos ROR identifikatorius (ror.org).

Pavyzdžiui:

Evidenciniai būdvardžiai lietuvių akademiniame diskurse

Anna Ruskan

Anglų filologijos katedra

Anglistikos, romanistikos ir klasikinių studijų institutas

Filologijos fakultetas

Vilniaus universitetas

Universiteto g. 5

LT-01131 Vilnius, Lietuva

El. paštas: anna.ruskan@flf.vu.lt

ORCID iD: <https://orcid.org/0000-0001-5547-0075>

<https://ror.org/03naddee84>

Jei straipsnį rengė keli autoriai, visi straipsnio autoriai turėtų pateikti šiuos duomenis. Vienas autorius turės būti nurodytas kaip kontaktinis asmuo, pateikimo formoje (žurnalo sistemoje) pateikiant autorius el. pašto adresą. Atkreipkite dėmesį, kad po to, kai straipsnis bus priimtas publikuoti, tekste nebegalima pakeisti darbovietės.

Taip pat, kai autorių daugiau nei vienas, prašome straipsnio pabaigoje pateikti paaiškinimą, ką kas konkretiai atliko vadovaujantis CreDiT kriterijais ([nuoroda](#)). Pavyzdžiui (dažniausiai pateikiama anglų kalba):

Author contributions

John Jonsered: conceptualization, methodology, formal analysis, investigation, writing – original draft, writing – review & editing, visualization. **Lucy Schneider:** conceptualization, methodology, formal analysis, investigation, writing – original draft, writing – review & editing.

(3) Santrauka

Ji turi būti anglų kalba (250 žodžių). Santraukoje turi būti pristatomas tyrimo objektas, problema, metodas ir pagrindiniai rezultatai bei apibendrinimai. Galima pateikti santrau-

ką ir lietuvių ar kita žurnalo kalba. Prieš santrauką kita nei straipsnio kalba nurodomas straipsnio pavadinimas paryškintai 12 pt ir žodis Abstract / Santrauka ta kalba.

(4) **Raktažodžiai:** 5–7 žodžiai, atskirti kableliais, pateikiami po kiekviena santrauka atitinkama kalba.

2 Tekstas

Tekstas skirstomas į skyrius ir poskyrius, nurodant numerį (1, 1.1, 1.1.1) ir pavadinimą. Numeris ir pavadinimas rašomi **paryškintai**, bet ne didžiosiomis raidėmis. Visame tekste ištisai pastraipos rašomos be įtraukų pirmosiose eilutėse. Prieš pastraipas paliekamas 12 pt tarpas.

Paveikslai ir lentelės (tekstas rašomas 12 pt) numeruojami atskirai, jų numeriai ir pavadinimai pateikiami apačioje (po pavadinimo taškas nededamas) be paryškinimo. Illustracijos turi būti pritaikytos juodai baltais spaudai, jų rezoliucija turi būti ne mažesnė nei 300 dpi.

Kursyvu tekste rašomi svetimos kalbos žodžiai. Tai, kas norima pabrėžti, rašoma **paryškintai**. Pabraukimai nenaudojami. Laužtiniuose skliaustuose [taip] rašomi autoriaus(-ių) papildymai.

Citatos. Trumpos citatos tekste rašomos išskiriant jas dvigubomis kabutėmis „štai taip“ (*citata originalo kalba pateikiama skliausteliuose*). Citatos citatose arba perfrazuoti svetimos kalbos žodžiai ir pateikti vertimai rašomi viengubose kabutėse ‘štai taip’. Ilgesnės nei trijų eilučių citatos (apie 40 žodžių) rašomos kursyvu atskiroje įtrauktoje pastraipoje (įtrauka – 5 pt).

Įvardijimo eilės tvarka žymima mažosiomis raidėmis su vienu skliausteliu, kiekviena sąvoka pateikiama atskiroje įtrauktoje (5 pt) eilutėje, pavyzdžiui:

- a) pirma frazė / sakiny
- b) antra frazė / sakiny ...

Pavyzdžiai (žodžiai, frazės, sakiniai ir t. t.) teikiami *kurssyvu* ir numeruojami ištisai visame tekste; skaičiai rašomi skliaustuose be įtraukų: (1), (2) ir t. t. Būtina nurodyti pavyzdžių šaltinius, galima naudoti santrumpas.

(1) *Šeimininkas akivaizdžiai suglumo.* (LKT)

Vertimas ir **glosos** būtini visoms citatoms/pavyzdžiams, kurie pateikiami ne publikacijos kalba. Vertimas (rašomas ne kurssyvu) paprastai žymimas viengubomis kabutėmis, pavyzdžiui: *evidently* ‘matyt’. Žodžiai sulygiuojami vertikaliai, naudojant tabuliavimo (angl. *tab*), o ne tarpo (angl. *space bar*) klavišą. Gramatinė informacija (NOM.SG.F) pateikiama sumažintomis didžiosiomis raidėmis (angl. *small caps*), pavyzdžiui:

- (2) *Jai reikia eiti namo.*
she.DAT.SG need.3PRS go.INF home.ADV
'She has to go home.'

Daugiau apie glosas galima rasti interneto svetainėje adresu: <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

Išnašos puslapio apačioje galimos norint trumpai paaiškinti (10 pt, iki 3 eilučių ilgio). Jos numeruojamos ištisai.

Padėka pateikiama straipsnio gale prieš literatūros sąrašą.

Santrumpų sąrašas pateikiamas prieš **Duomenų šaltinius** arba **Literatūros sąrašą**.

Po visu straipsniu nurodoma jo įteikimo redaktorių kolegijai data.

Straipsnio pateikimo metu autorius, kuris yra ir kontaktinis asmuo, gali pateikti ir trumpą biografinę informaciją (ne daugiau kaip 200 žodžių), o kai straipsnis priimamas publikuoti, tokią trumpą biografiją gali pateikti ir visi bendraautoriai. Ši informacija, autoriams pageidaujant, bus matoma publikuoto straipsnio nukreipimo puslapyje – prie autoriaus pavardės atsiranda įrašas Bio, kurį atsidarius matoma pateikta informacija.

3 Literatūros nuorodos tekste

Visos nuorodos tekste pateikiamas reikiamoje vietoje skliaustuose (autorius pavardė ar publikacijos pavadinimas ir metai, kablelis, puslapis(-iai), jei reikia), pavyzdžiui, (Howarth 1998, 27–28). Skirtingų autorų nuorodos skiriamos kabliataškiais, pavyzdžiui, (Aijmer 1996; van der Auwera, Schalley & Nuyts 2005). Nelotyniška abécéle (kirilica ir kt.) parašytos autorų pavardės ir publikacijų pavadinimai turi būti transliteruojami.

4 Duomenų šaltiniai ir naudotos literatūros sąrašas

Straipsnio pabaigoje abécélės tvarka turi būti pateikiami darbe analizuojamų duomenų šaltiniai ir tik darbe cituojamų naudotos literatūros šaltinių sąrašas. Kiekvienas šaltinis pateikiamas atskira pastraipa be padidinto tarpo tarp atskirų įrašų; antroji pastraipos eilutė įtraukama 10 pt. Straipsnių pavadinimuose didžiaja raide rašomas tik pirmasis pavadinimo žodis ir tikriniai žodžiai. Jei straipsnis rašomas ne lietuvių kalba, literatūros sąraše teikiamus knygų ir straipsnių pavadinimus lietuvių, latvių, rusų, lenkų kalba būtina išversti į straipsnio kalbą ir pateikti juos laužiniuose skliaustuose. Literatūros sąraše nelietuviškos pavardės, prasidedančios raidėmis Q, W, X, Y, teikiamos pagal anglų kalbos abécélę. Prašome rašyti pagal pateiktus pavyzdžius:

Duomenų šaltiniai

- BNC *The British National Corpus*. Davies, M. 2004–. BYU–BNC. Interneto prieiga: <http://corpus.byu.edu/bnc>.
- CorALit *Lietuvių mokslo kalbos tekstyner*. Interneto prieiga: <http://www.coralit.lt/>.

Literatūra

- Ambrasas, Vytautas. 1990. *Sravnitel'nyj sintaksis pričastij baltijskikh jazykov*. [Comparative syntax of participles in Baltic languages]. Vilnius: Mokslas.
- Barbieri, Federica. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12 (1), 58–88.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bolinger, Dwight. 1965. The atomization of meaning. *Language* 41, 555–573.
- Gansel, Christina, Frank Jürgens. 2007. *Textlinguistik und Textgrammatik. Eine Einführung*. Göttingen: Vandenhoeck & Ruprecht.
- Haß, Ulrike, Hg. 2005. *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin & New York: Walter de Gruyter.
- Holvoet, Axel, Loreta Seménienė, red. 2004. *Gramatinių kategorijų tyrimai*. [Studies in grammatical categories]. Vilnius: Lietuvių kalbos institutas.
- Huddleston, Rodney & Geoffrey K. Pullum, eds. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Ide, Nancy & James Pustejovsky, eds. 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-024-0881-2>
- Jaszczolt, Katarzyna. 2009. Default Semantics. *The Oxford Handbook of Linguistic Analysis*. Bernd Heine & Heiko Narrog, eds. Oxford: Oxford University Press. 193–221.
- Kleiber, Georges. 1990. *La sémantique du prototype: catégories et sens lexical*. Paris: Presses Universitaires de France.
- Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Vol. 2. Descriptive application*. Stanford: Stanford University Press.
- Nau, Nicole, Peter Arkadiev. 2015. Towards a standard of glossing Baltic languages: The Salos Glossing Rules. *Baltic Linguistics* 6, 195–241.
- Šinkūnienė, Jolanta. 2011. *Autoriaus pozicijos švelninimas rašytiniame moksliniame diskurse: gretinamasis tyrimas*. [Hedging in written academic discourse: A cross-linguistic and cross-disciplinary study], ms. Humanitarinių mokslų daktaro disertacija. Vilnius: Vilniaus universitetas.
- Trbojević-Milosević, Ivana. 2010. On innocence and experience: Modal hedges in health care products instructions in English and Serbian. Presentation in *The 4th international conference 'Modality in English 4'*, Madrid.
- Wiesmann, Eva. 2022. Rechtsbezogene Wissensvermittlung. Zum Einsatz von GeLeCo in der Übersetzungsausbildung. *Korpora Deutsch als Fremdsprache* 2 (1), 42–60. Abrufbar unter: <https://cris.unibo.it/retrieve/e038653d-c674-42ad-a5f6-50b030f45fc2/kordaf-56-wiesmann.pdf>. Zugriff: 7. August 2025.

5 Prieinamumas

Nuo 2025 m. įsigaliojo Europos pritaikomumo aktas, kuriuo išreiškiamas reikalavimas, kad el. publikacijos būtų pritaikytos visiems, taip pat ir neregiamams, ir silpnaregiamams. Mes paruošime el. rinkmenas tinkamai, tačiau reikia, kad savo straipsnyje pateiktas lenteles ir iliustracijas aprašytumėte taip, kad jų turinį suprastų ir neregintis. Iliustracijų aprašymo pavyzdžius rasite čia ([nuoroda](#)). Savo aprašymą pateikite po lentele ar iliustracija [laužtiniuose skliaustuose]. Jo galutiniame variante nesimatys, tačiau šio teksto reikia gamybos metu, kad jį matytų kalbos redaktoriai ir maketuotojai.

Style sheet

The journal of linguistics ***Kalbotyra*** focuses on research into various aspects of language studies as well as the ones addressing cross-linguistic issues. It publishes articles, reviews of books and reports of conferences. Proceedings of conferences are also invited.

Papers submitted for publication should not have been published or submitted for publication elsewhere. They are reviewed by at least two anonymous referees following the double-blind refereeing procedure.

In order to submit an article, authors must register in the journal's publishing system ([link](#)) and upload the article manuscripts there in two formats: MS Word (*.docx) and PDF (*Portable Document Format*, *.pdf). After saving the document in PDF format, it is recommended to check whether the formatting of the text and illustrations has remained unchanged. To ensure anonymous review, the author(s) must submit two versions: one with the text of the article as required in the reference, the second version without references or hints to authorship. The article file must be submitted in such a way that it does not contain data that could help identify the author (it is necessary to remove information from the document's *properties* section).

Papers should not normally exceed 8,000 words in length; only in exceptional circumstances can significantly longer papers be considered.

Papers should be prepared according to the requirements set out below in one of the following languages: English, French, German or Lithuanian. If the language of the paper is not a native language of the author(s), the paper should be proof-read by a native-language specialist to check its correctness.

It is the authors' responsibility to ensure that the final version of their paper fully conforms to this style sheet.

The author(s) warrant that their paper is original and no property rights (including copyright or other intellectual property rights) of any third parties have been violated. ***Kalbotyra*** follows the policy of screening for plagiarism. Articles published in ***Kalbotyra*** are distributed under a Creative Commons Attribution-NonCommercial 4.0 Licence.

1 Structure and form

Papers submitted for publication should correspond to the general requirements of research papers and cover the following points: the research question/problem, review of previous research on the subject, data and methods, research findings/results (evaluated and validated), evidence (documented), conclusions and references. Papers that do not conform to the requirements will be returned to the authors for revision before further processing.

Papers should be prepared on A4 paper size with a 1.5 cm margin on the right and 2.5 cm margins on the top, left and bottom; the pages should be numbered beginning with the title page at the top right corner of the page. The authors should use 1.5 spacing between the lines throughout the paper. The font is 12 pt Times New Roman. The text should be justified left.

The paper should contain:

- (1) the title of the paper, 14 pt, bold
- (2) full name(s) and surname(s) of the author(s) in bold; all information is written in 12 pt font in the language of publication; the full name of the institution (including department, centre, institute, etc.), the address of the institution are indicated; the author's email address (without underlining), the author's ORCID number (orcid.org), the institution's ROR identifier (ror.org) are also indicated.

Evidential adjectives in Lithuanian academic discourse

Anna Ruskan

Department of English Philology
Institute of English, Romance and Classical Studies
Faculty of Philology
Vilnius University
Universiteto g. 5
LT-01131 Vilnius, Lithuania
E-mail: anna.ruskan@flf.vu.lt
ORCID iD: <https://orcid.org/0000-0001-5547-0075>
<https://ror.org/03nadee84>

If the article has multiple authors, all authors of the article should provide this information. One author should be listed as the contact person, providing the author's email address in the submission form (in the journal system). Please note that once the article has been accepted for publication, the place of employment cannot be changed in the text.

Also, when there is more than one author, please provide an explanation at the end of the article of what each author has done in accordance with the CreDiT criteria ([link](#)). For example (usually provided in English):

Author contributions

John Jonsered: conceptualization, methodology, formal analysis, investigation, writing – original draft, writing – review & editing, visualization. **Lucy Schneider:** conceptualization, methodology, formal analysis, investigation, writing – original draft, writing – review & editing.

(3) Abstract

All articles must have an abstract in English (250 words). An abstract should clearly describe the purpose of the research, data and methodology, the main results and the principal conclusions. The second abstract in Lithuanian or any other language of the journal is optional. Abstracts in languages other than the language of the publication should bear the title (in bold, 12 pt) and the words Abstract / Santrauka.

(4) **Keywords:** a list of 5–7 key words separated by commas is provided below every abstract in the language of the abstract. For example, articles written in English should have keywords in English.

2 The text

The text should be divided into sections and subsections, each of them decimaly numbered beginning with 1 (e.g.: 1, 1.1, 1.1.1, etc.) and titled. The number and title should be in **bold** type. The block organisation of paragraphs (not indented) should be used throughout the whole text with spaces of 12 pt before each new paragraph.

Figures and tables (12 pt) should be numbered and titled separately under the figure/table. The illustrations will be printed black and white, their resolution should not be less than 300 dpi.

Use *italics* for foreign words; use **bold** face for emphasis. Use square brackets [like this] for personal additions.

Quotations. Short quoted sections in the running text should be enclosed in double quotation marks “like this” (*the original citation in italics is given in round brackets*). Use single quotes for special forms, for quotations within quotations, and for glosses and paraphrases of (foreign) words. Quotations longer than three lines (ca. 40 words) should be given in a separate indented paragraph (5 pt) in italics.

Listings for the purpose of classification should be written in a new indented (5 pt) line, e.g.:

- a) the first model
- b) the second model ...

Examples (words, phrases, sentences, etc.) are not indented, they are given in italics and numbered consecutively throughout the article; the numbers (regular) are enclosed in round brackets, e.g.: (1), (2), e.g.:

(1) *Šeimininkas akivaizdžiai suglumo.* (LKT)

‘The host **evidently** became confused.’

References for cited examples should be indicated, translation correspondences of all language data in a language other than the language of the paper should be given in single commas, e.g.: *eiti* ‘to go’.

Translation and word-by-word glosses are provided for all quotations/examples from languages other than the language of the article. Translation is given in single quotation marks. Words are aligned vertically using tab key rather than space bar key. Use small caps to indicate grammatical information (NOM.SG.F). The glosses and the translation should be left-aligned with the example text as in the example below:

- (2) *Jai reikia eiti namo.*
she.DAT.SG need.3PRS go.INF home.ADV
‘She has to go home.’

For more details about glossing refer to: <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

Footnotes set in 10 pt can be used only for very brief explanatory remarks. They should be numbered consecutively throughout the text.

Acknowledgements

They follow the main text of the paper.

List of Abbreviations should precede **Data sources** or **References**.

Below the body of the article, the date of its submission for publication should be indicated.

When submitting an article, the author, who is also the contact person, may also provide a short biographical information (no more than 200 words), and when the article is accepted for publication, all co-authors may also provide such a short biography. This information, if requested by the authors, will be visible on the landing page of the published article – a Bio entry will appear next to the author’s name, which, when opened, will display the information provided.

3 References in the text

All references should be given at the appropriate point in the text in brackets (author’s name or title of publication, year of publication, comma, page(s) referred to, if relevant), like this: (Howarth 1998, 27–28). Different sources of reference should be separated by semi-colons (Aijmer 1996; van der Auwera, Schalley & Nuysts 2005). If letters of Slavic or some other non-Latin alphabet have been used, the names and titles should be transliterated.

4 Reference list

All data sources and works cited in the text, and only those, should be listed alphabetically at the end of the paper in separate sections under the headings **Data Sources** and **References**. Each reference entry is given in a separate paragraph; the second line of the paragraph is indented by 10 pt. All lexical words are capitalized only in the *Names of Periodicals*; only the first word is capitalized in the *Titles of books* (proper names, etc. are exceptions). Papers written in languages other than Lithuanian should provide translations of Lithuanian, Latvian, Russian, Polish book and article titles in brackets. Please follow the pattern given below:

Data sources

BNC *The British National Corpus*. Davies, M. 2004–. BYU–BNC. Available at: <http://corpus.byu.edu/bnc>.

CorALit *Lietuvių mokslo kalbos tekstyны*. [Corpus Academicum Lithanicum]. Available at: <http://www.coralit.lt/>.

References

- Ambraszas, Vytautas. 1990. *Sravnitel'nyj sintaksis pričastij baltijskikh jazykov*. [Comparative syntax of participles in Baltic languages]. Vilnius: Mokslas.
- Barbieri, Federica. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12 (1), 58–88.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Bolinger, Dwight. 1965. The atomization of meaning. *Language* 41, 555–573.
- Gansel, Christina, Frank Jürgens. 2007. *Textlinguistik und Textgrammatik. Eine Einführung*. Göttingen: Vandenhoeck & Ruprecht.
- Haß, Ulrike, Hg. 2005. *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin & New York: Walter de Gruyter.
- Holvoet, Axel & Loreta Seménienė, red. 2004. *Gramatinių kategorijų tyrimai*. [Studies in grammatical categories]. Vilnius: Lietuvių kalbos institutas.
- Huddleston, Rodney & Geoffrey K. Pullum, eds. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Ide, Nancy & James Pustejovsky, eds. 2017. *Handbook of Linguistic Annotation*. Dordrecht: Springer. <https://doi.org/10.1007/978-94-024-0881-2>
- Jaszczolt, Katarzyna. 2009. Default Semantics. *The Oxford Handbook of Linguistic Analysis*. Bernd Heine & Heiko Narrog, eds. Oxford: Oxford University Press. 193–221.
- Kleiber, Georges. 1990. *La sémantique du prototype: catégories et sens lexical*. Paris: Presses Universitaires de France.

- Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Vol. 2. Descriptive application*. Stanford: Stanford University Press.
- Nau, Nicole, Peter Arkadiev. 2015. Towards a standard of glossing Baltic languages: The Salos Glossing Rules. *Baltic Linguistics* 6, 195–241.
- Šinkūnienė, Jolanta. 2011. *Autoriaus pozicijos švelninimas rašytiame moksliniame diskurse: gretinamasis tyrimas*. [Hedging in written academic discourse: A cross-linguistic and cross-disciplinary study], ms. Humanitarinių mokslų daktaro disertacija. [PhD dissertation]. Vilnius: Vilniaus universitetas.
- Trbojević-Milosević, Ivana. 2010. On innocence and experience: Modal hedges in health care products instructions in English and Serbian. Presentation in *The 4th international conference 'Modality in English 4'*, Madrid.
- Wiesmann, Eva. 2022. Rechtsbezogene Wissensvermittlung. Zum Einsatz von GeLeCo in der Übersetzungsausbildung. *Korpora Deutsch als Fremdsprache* 2 (1), 42–60. Abrufbar unter: <https://cris.unibo.it/retrieve/e038653d-c674-42ad-a5f6-50b030f45fc2/kordaf-56-wiesmann.pdf>. Zugriff: 7. August 2025.

5 Accessibility

The European Accessibility Act, which requires e-publications to be accessible to all, including blind and partially sighted individuals, came into force in 2025. We will ensure that the e-files are properly prepared; however, you need to describe the tables and illustrations in your article in a way that a blind person can understand. For examples of illustration descriptions, refer to the Guide to Image Descriptions ([link](#)). Place your description under the table or illustration in square brackets [abcd] to provide a clear explanation. These descriptions will not be visible in the final version but are essential for language editors and layout artists during production.

Viršelio dailininkas *Gediminas Markauskas*
Maketuotoja *Vida Vaidakavičienė*

Už straipsnių turinį
atsako autoriai

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt
www.knygynas.vu.lt, www.zurnalai.vu.lt

Spausdino UAB „Baltijos kopija“
Kareivių g. 13b, LT-09109 Vilnius

Tiražas 50 egz. 13 aut. l.

Cover design *Gediminas Markauskas*
Layout *Vida Vaidakavičienė*

Authors are responsible for
the content of their work

Vilnius University Press
9 Saulėtekio Av., III Building, LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt/en/
www.knygynas.vu.lt, www.zurnalai.vu.lt

Printed by UAB „Baltijos kopija“
13b Kareivių St., LT-09109 Vilnius

Print run 50 copies. 13 author's sheet

