



Kleinvieh macht auch Mist:

Was eine Lernerkorpussammlung wie PORTA leisten kann

Andrea Abel (unibz – Eurac Research) Aivars Glaznieks (Eurac Research)



Fahrplan

- Die Lernerkorpussammlung PORTA
- Anwendungsbeispiele
 - 1. Porta in der Angewandten Linguistik (AL)
 - 2. Porta in der Lernerkorpusforschung (LCR)
- Fazit & Diskussion













LEONIDE

Longitudinal IEarner cOrpus iN Italiano, Deutsch and English A longitudinal trilingual corpus of young learners of Italian, German and English



Die Lernerkorpussammlung PORTA

- Steckbrief:
 - Fokus: Lernerkorpora
 - Quelle: Eurac-Institut f
 ür Angewandte Sprachforschung & Partner
 - Zugriff auf Hintergrundinformationen zum Entstehungskontext
 - Zugriff auf Dokumentation (Transkriptions- und Annotationshandbücher, Aufgabenstellungen etc.)
 - Zugriff auf die Korpora direkt über ANNIS oder zum Download (über ERCC)



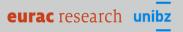
https://www.porta.eurac.edu/





https://clarin.eurac.edu/

https://commul.eurac.edu/annis/learnercorpora



Die Lernerkorpussammlung PORTA



коко:

Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenz bei Lernenden im deutschen Sprachraum



scuolan sprachtes
rachkom tudenti soci
aferenzrahme ompetenz
genit ompetenz
genit ompetenz
anto Schüler Schule cor
iti sociolinguistidingue ompetenze enter competenze enter com

KOLIPSI Korpus-Familie

Größe: 950.000 Tokens

Texte: 1.503 Argumentationen von

1.503 SuS deut. Oberschulen

(DEU, AUT, STIR) **Sprache**: Deutsch

Jahr: 2011

Größe: 340.000 Tokens

Texte: 2.246 Texte

unterschiedlicher Textsorten von

2.246 Erwachsenen

Sprache: Deutsch, Italienisch,

Tschechisch
Jahr: 2012

Größe: 1.171.000 Tokens

Texte: 5.871 Argumentationen und

Narrationen von 2.993 SuS deut.

und ital. Oberschulen

Sprache: Deutsch, Italienisch

Jahr: 2007/08, 2014/15

Die Lernerkorpussammlung PORTA



LEONIDE

Longitudinal IEarner cOrpus iN Italiano, Deutsch and English A longitudinal trilingual corpus of young learners of Italian, German and English



Größe: 64.000 Tokens

Texte: 332 Texte unterschiedlicher

Textsorten von 141 SuS und

50 Erwachsenen

Sprache: Italienisch

Jahr: 2007, 2012

Größe: 237.000 Tokens

Texte: 2.510 Argumentationen und Narrationen von 163 SuS

deut. u. ital. Mittelschulen

Sprache: Deutsch, Italienisch,

Englisch

Jahr: 2015-2018

Größe: 425.000 Tokens

Texte: 635 Argumentationen von

635 SuS ital. Oberschulen

Sprache: Italienisch

Jahr: 2022

Die Lernerkorpussammlung PORTA - WIP

COMPLeTT

Competenzes Plurilinguistiches Ladin Tudesch Talian

Größe: 87.000 Tokens (x3)

Texte: 611 Texte unterschiedlicher

Textsorten von 205 SuS

Sprache: Deutsch (+Italienisch,

+Ladinisch)

Jahr: 2024 (+2026, +2028)



KOLIPSI Korpus-Familie

Größe: 1.171.000 Tokens (+???)

Texte: 5.871 Argumentationen und Narrationen von 2.993 SuS deut.

und ital. Oberschulen

Sprache: Deutsch, Italienisch

Jahr: 2007/08, 2014/15 (+2025)

Gemeinsamkeiten der Korpora in PORTA

- **Größe der Korpora**: klein (< 100.000 Tokens) bis mittelgroß (> 100.000, < 1.000.000 Tokens) (?Kolipsi Korpus-Familie)
- **Sprachproduktionsbedingung**: institutionalisierter Bildungskontext, überwiegend aus der Schule (*Merlin)
- Ursprung der Texte: überwiegend aus Südtirol (*Merlin)
- Modalität: geschriebene Texte

Unterschiede zwischen den Korpora von PORTA

Datensammlung:

- Querschnitt vs. Längsschnitt
- Aufgabenstellung
- Sprachen im Korpus:
 - Deutsch vs. Italienisch (Englisch, Tschechisch)
 - Lernumgebung: L1 vs. L2 vs. L3
- Anzahl der Sprachen im Korpus:
 - eine Sprache vs. mehre Sprachen

Unterschiede zwischen den Korpora von PORTA

- Textproben/Person: eine vs. mehr als eine bzgl.
 - Textsorte (argumentativ + narrativ)
 - Sprachen (echtes Mehrsprachigkeitskorpus vs. Lernerkorpus)
 - Erhebungszeitpunkte (echt longitudinal)
- Personenmerkmale:
 - Erstsprachen,
 - Sprachbiografie
 - Altersgruppe ...

Umgang mit Heterogenität in PORTA

Ursprung der Heterogenität

- Reaktion auf soziale Veränderungen und bildungspolitische Realitäten der mehrsprachigen Gesellschaft Südtirols
- Veränderung von Forschungsinteressen

Herstellung von Homogenität bzgl. Design & Entwicklung

- standardisierte Datenerfassung (vgl. zu Metadaten Paquot et al. 2024)
- Vereinheitlichung von Workflows (vgl. Glaznieks et al. 2014)
- Vereinheitlichung von Formaten (Frey, J.-C., König, A., & Fišer 2020).

• Ziele:

- Bildungspolitische und forschungspezifische Relevanz der Korpora
- Vergleichbarkeit der Daten aus unterschiedlichen Korpora

Anwendungsbeispiele

- Dimensionen:
 - 1. PORTA in der Angewandten Linguistik (AL)
 - 2. PORTA in der Lernerkorpusforschung (LCR)

- Verortung in der AL (Fokus: Arbeit mit Lernerkorpora)
 - Bezugnahme u. a. auf die "Würzburger Thesen zur Angewandten Sprachwissenschaft" (Schulz/Schwebler 2025)
 - Kernelemente der AL (als Sprachwissenschaft PLUS) u. a.:
 - **Lösungsorientierung**: "[...] Ihr Ausgangspunkt ist problemorientiert, sie agiert aber lösungsorientiert" (10), vgl.:
 - "[...] the theoretical and empirical investigation of real-world problems in which language is a central issue" (Brumfit 1995: 27)
 - [...] trying to resolve language-based problems that people encounter in the real world, whether they be learners, teachers, supervisors, academics, lawyers, service providers, those who need social services, test takers, policy developers, dictionary makers, translators, or a whole range of businessclients" (Grabe 2002: 9)

eurac research unibz

- Handlungs- und Problemfelder im Rahmen der AL im Bildungsbereich z. B.
 - Handlungsfeld: Kommunikation in der L2;
 Problemfeld: aktive Teilhabe am Alltagsleben in der L2
 - Handlungsfeld: Leben im Kontext sprachlicher und kultureller Vielfalt;
 Problemfeld: Management zunehmender sprachlicher Vielfalt im Bildungsbereich
- Tätigkeitsfelder der AL:
 - "(1) Beraten,
 - (2) Lehren,
 - (3) Professionalisieren,
 - (4) Diagnostizieren und Therapieren sowie
 - (5) Aufklären und
 - (6) Informieren und Dokumentieren [...]" (9)

- Gesellschaftliche Ebenen im Bildungsbereich:
 - Mikroebene \rightarrow Kleingruppe \rightarrow z. B. Schulklasse
 - Mesoebene → Organisation → z. B. Schule, Verlag,
 Forschungseinrichtung
 - Makroebene → Gesellschaft → z. B. Bildungsbehörden, Bildungspolitik

(vgl. Henecka 2015: 34ff.)

- Gesellschaftliche Ebenen im Bildungsbereich:
 - Mikroebene \rightarrow Kleingruppe \rightarrow z. B. Schulklasse
 - Mesoebene → Organisation → z. B. Schule, Verlag,
 Forschungseinrichtung
 - Makroebene → Gesellschaft → z. B. Bildungsbehörden, Bildungspolitik

(vgl. Henecka 2015: 34ff.)

Anwendungsbeispiel 1: COMPLeTT

COMPLeTT

Competënzes Plurilinguistiches Ladin Tudësch Talian

Eckdaten Korpusfamilie COMPLeTT (Work in Progress):

	COMPLETT (DE)
Größe:	~86800 Token
Texte:	615
Texttypen:	Anleitung, Bericht, informelle WhatsApp-Nachricht
Sprachen:	deutsch (ladinisch, italienisch, deutscher Dialekt)
Erhebungsjahr:	2024

	COMPLETT (IT)
Größe:	
Texte:	
Texttypen:	Anleitung, Bericht, informelle WhatsApp-Nachricht
Sprachen:	italienisch (ladinisch, deutsch, deutscher Dialekt)
Erhebungsjahr:	2026

	COMPLETT (LAD)
Größe:	
Texte:	
Texttypen:	Anleitung, Bericht, informelle WhatsApp-Nachricht
Sprachen:	ladinisch (deutsch, italienisch, deutscher Dialekt)
Erhebungsjahr:	2028

Anwendungsbeispiel 1: COMPLeTT

COMPLeTT

Competenzes Plurilinguistiches
Ladin Tudesch Talian

- Projektrahmen: "COMPLeTT: Competënzes Plurilinguistiches Ladin Tudësch Talian"
- Ziel: Entwicklung einer praktikablen und nachhaltigen Methode zur Erhebung, Bewertung und Dokumentation sowohl einzelsprachlicher als auch mehrsprachiger Kompetenzen in der Schriftsprache der Schüler:innen der ladinischen Täler in Südtirol (im Auftrag der Landesevaluationsstelle für das ladinische Bildungssystem)

Anwendungsbeispiel 1: COMPLeTT

COMPLeTT

Competenzes Plurilinguistiches
Ladin Tudesch Talian

- Beispiel Makroebene Bildungsbehörde (Schulamt)
 - Handlungsfeld: Leben im Kontext sprachlicher und kultureller Vielfalt
 - Problemfeld: Management sprachlicher Vielfalt im Bildungsbereich (Normen)
 - Tätigkeitsfelder: Dokumentieren, Professionalisieren, Aufklären

Anwendung:

- digitale Langzeitarchivierung
- Erkenntnisse f
 ür die Schulentwicklung (Evaluationsstelle)
- Bewusstseinsbildung bzgl. Mehrsprachigkeitskompetenzen

Anwendungsbeispiel 2: KOLIPSI



KOLIPSI Korpus-Familie

• Eckdaten Korpusfamilie KOLIPSI (Fokus DE & IT L2)

	KOLIPSI 1 DE L2		
Größe:	~87.000		
Texte:	523		
Texttypen:	Argumentation, Narration		
Sprachen:	deutsch		
Erhebungsjahr:	2008		

	KOLIPSI 2 DE L2			
Größe:	~106.000			
Texte:	700			
Texttypen:	Argumentation, Narration			
Sprachen:	deutsch			
Erhebungsjahr:	2015			

	KOLIPSI 3 IT L2
Größe:	
Texte:	
Texttypen:	Argumentation, Narration
Sprachen:	italienisch

	KOLIPSI 1 IT L2
Größe:	~387.000
Texte:	1.990
Texttypen:	Argumentation, Narration
Sprachen:	italienisch
Erhebungsjahr:	2008

	KOLIPSI 2 IT L2			
Größe:	~400.000			
Texte:	2.063			
Texttypen:	Argumentation, Narration			
Sprachen:	italienisch			
Erhebungsjahr:	2015			

KOLIPSI 3 DE L2
Argumentation, Narration
deutsch
2025

Erhebungsjahr: 2025

Anwendungsbeispiel 2: KOLIPSI



KOLIPSI Korpus-Familie

- Projektrahmen: "KOLIPSI: Die Südtiroler SchülerInnen und die Zweitsprache: eine linguistische und sozialpsychologische Untersuchung":
 - Auflage 1 (2007/08)
 - Auflage 2 (2014/15)
 - Auflage 3 (2025/26)
- Ziele: Erhebung und Analyse der Zweitsprachkompetenzen (deutsch und italienisch als L2) der Südtiroler Schüler:innen der Sekundarstufe II mit Bezug auf den GeRS und unter Berücksichtigung soziolinguistischer und sozialpsychologischer Aspekte (Momentaufnahmen & diachroner Vergleich) (in Kooperation mit den Landesschulämtern)

Anwendungsbeispiel 2: KOLIPSI



KOLIPSI Korpus-Familie

- Beispiel Makroebene Bildungsbehörde (Schulamt)
 - Handlungsfeld: Kommunikation in der L2
 - Problemfeld: aktive Teilhabe am Alltagsleben in der L2
 - Tätigkeitsfelder: Dokumentieren, Aufklären, Beraten

Anwendung:

- Monitoring der Zweitsprachkompetenzen
- Berichte mit Maßnahmenempfehlungen in Verbindung mit konkreten Sprachdaten (→ Mesoebene)

eurac research unibz

Ein Zwischenfazit:

- Heterogenität der Korpusressourcen als Reaktion auf bildungspolitische Bedürfnisse (Sprachen ...) und Einschränkungen (Belastung Schulen ...)
- Homogenität der Korpusressourcen wo nötig bzw. möglich in Bezug auf Aufgabenstellung, Alter der Lernenden, Setting der Datenerhebung, Metadaten

- Kaskadierbarkeit von Ergebnissen anhand von Lernerdaten: Makro- > Meso- > Mikroebene
- erleichterter Austausch zwischen bildungspolitischen Akteuren (Sprachdaten)

Ziele der Lernerkorpusforschung (u.a. Meunier 2020, Götz & Granger 2024):

- Analyse von Sprachgebrauch
- Analyse von Fehlern
- Analyse von Entwicklungsmustern
- Analyse von Sprachkompetenzen
- Analyse von Variablen, die Lernprozesse und Spracherwerb beeinflussen

A: Analyse einzelner Korpora:

- Beantwortung einer Forschungsfrage, für die das Korpus zusammengestellt wurde (Bsp. KoKo, Abel & Glaznieks 2019)
- Beantwortung von Forschungsfragen, für die sich das Korpus zusätzlich eignet (Bsp. LEONIDE, Reitbrecht 2025, Schlüter 2025)



KOLIPSI Korpus-Familie

B: Analyse mehrerer (Sub-)Korpora der Korpussammlung

B.1 Vergrößerung der Datenmenge: Bsp. Kolipsi Korpus-Familie

- Studie zur Intensivierung von Adjektiven bei Lernenden von Italienisch und Deutsch als L2 (Projekt IWL2-it)
- Teil 1: Italienische als L2 (Spina et al. 2025)
- Abhängige Variablen: Intensivierungstyp (morphologisch vs. syntaktisch)
- Prädiktorvariablen: L1, CEFR level, Sprachliche Umgebung

Korpusname	Jahr	Größe	Texte	Schreiber*innen		Intensivierte Adjektive (Texte)
Kolipsi-1_L2_IT	2007/08	ca. 387,000	1,990	1,000	16-18 years	1,889 (757)



KOLIPSI Korpus-Familie

- B. Analyse mehrerer (Sub-)Korpora der Korpussammlung
- B.1 Vergrößerung der Datenmenge: Bsp. Kolipsi Korpus-Familie
- Studie zur Intensivierung von Adjektiven bei Lernenden von Italienisch und Deutsch als L2 (Projekt IWL2-it)
- Teil 2 Deutsch als L2 (Spina & Glaznieks eingereicht)
- Dieselben unabhängligen und abhängigen Variablen

Korpusname	Jahr	Größe	Texte	Schreiber*innen	Alter	Intensivierte Adjektive (Texte)
Kolipis-1_L2_DE	2007/08	ca. 87,000	523	267	16-18 years	548
Kolipsi-2_L2_DE	2014/15	ca. 106,000	700	357	16-18 years	450
Kolipsi_DE_total		ca.193,000	1,223	624		998

eurac research unibz



KOLIPSI Korpus-Familie

B. Analyse mehrerer (Sub-)Korpora der Korpussammlung

B.1 Vergrößerung der Datenmenge: Bsp. Kolipsi Korpus-Familie

Möglich durch Homogenität der Korpusressourcen:

- Aufgabenstellung (moderat modernisiert ...)
- Setting der Datenerhebung
- Alter der Lernenden
- Metadaten

Fazit: Homogenität ermöglicht das "Auffüllen" der Daten, damit eine kritische Menge des linguisitschen Phänomens erreicht werden kann.

- B. Analyse mehrerer (Sub-)Korpora der Korpussammlung (zu einer Fragestellung)
- B.2 Erweiterung der Untersuchungsgruppen, Bsp: Glaznieks et al. (2023)
- Studie zum Gebrauch von *Weil* (und *Wenn*-)Sätzen bei unterschiedlichen Lernenden des Deutschen (L1, L2-immersiv, L2-nicht-immersiv)
- Abhängige Variablen: Wortstellung in Weil- (und Wenn-)Sätzen
- Prädiktor-Variablen: Alter, Lernergruppe, Semantik der Weil-Sätze

LEONIDE

Longitudinal IEarner cOrpus iN Italiano, Deutsch and English A longitudinal trilingual corpus of young learners of Italian, German and English



KOLIPSI Korpus-Familie

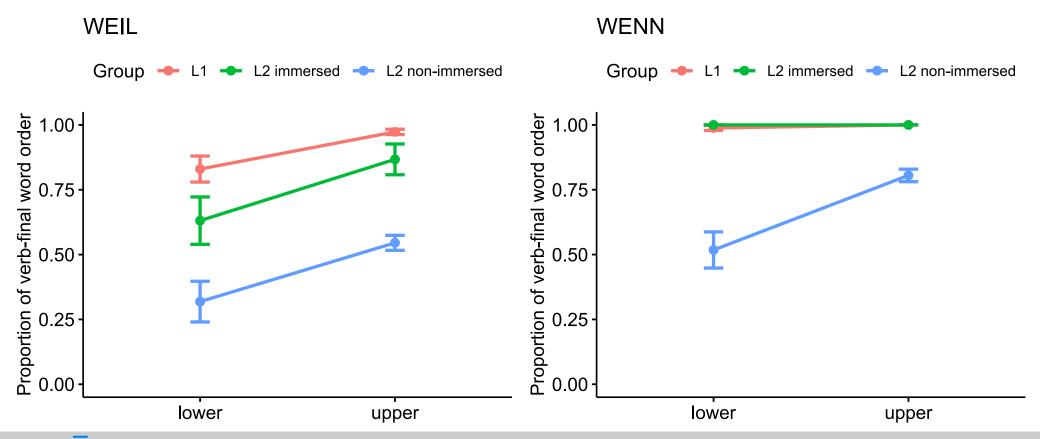


- B. Analyse mehrerer (Sub-)Korpora der Korpussammlung (zu einer Fragestellung)
- B.2 Erweiterung der Untersuchungsgruppen, Bsp: Glaznieks et al. (2023)

	# occurences		# pupils		# texts	
Secondary school stage	lower upper		lower	upper	lower	Upper
NS	295	353	58	189	143	189
NNS immersed	143	52	24	29	70	32
NNS not-immersed	134 389		42 198		98 268	
Total	1366		540		800	

B. Analyse mehrerer (Sub-)Korpora der Korpussammlung (zu einer Fragestellung)

B.2 Erweiterung der Untersuchungsgruppen, Bsp: Glaznieks et al. (2023)



- B. Analyse mehrerer (Sub-)Korpora der Korpussammlung (zu einer Fragestellung)
- B.2 Erweiterung der Untersuchungsgruppen, Bsp: Glaznieks et al. (2023)

Möglich durch Homogenität der Korpusressourcen:

- Aufgabenstellung (argumentative Texte)
- Lernergruppen (immersive vs. nicht-immersive Lernende)
- Anwendung vergleichbarer Erhebungssettings in unterschiedlichen Projekten
- Metadaten

Fazit: Homogenität ermöglicht, "zusammenfügen" der Daten, zu einem Datenset für Gruppenvergleiche

Fazit & Diskussion

- Lernerkorpussammlung wie PORTA
 - Stärken, z. B.:
 - Mehrfachverwendung von Daten (Bildungspolitik, Spracherwerbsforschung ...)
 - Erhöhung der Datenmenge
 - Erhöhung der Datenvielfalt (Variablenspektrum)
 - Profilierung soziolinguistischer Kontexte (Heterogenität)
 - Vergleichbarkeit der Produktionsbedingungen
 - umfassende Dokumentation & standardisierte Datenerfassung (Metadaten ...)
 - Grenzen, z. B.:
 - Aussagekraft (Kontextspezifik)
 - Vergleichbarkeit der Daten (Heterogenität)
 - Modalität (geschriebene Daten)

eurac research unibz

Vielen Dank für die Aufmerksamkeit!

Andrea Abel
anabel@unibz.it
andrea.abel@eurac.edu

Aivars Glaznieks
aivars.glaznieks@eurac.edu



Literaturauswahl

- Abel, A., Glaznieks, A., Nicolas, L. & Stemle, E. W. (2014): KoKo: an L1 Learner Corpus for German. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, 26–31 May, 2014, 2414-2421. [http://www.lrec-conf.org/proceedings/lrec2014/pdf/934 Paper.pdf]
- Abel, Andrea & Aivars Glaznieks (2017): KoKo: Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenz bei Lernenden im deutschen Sprachraum ein Ergebnisbericht. [http://www.korpus-suedtirol.it/KoKo/Documents/Ergebnisse Dokumentation gesamt FINAL.pdf]
- Frey, J.-C., König, A., & Fišer, D. (2020). Creating a learner corpus infrastructure: Experiences from making language learner data available. Proceedings of the ICTeSSH Conference 2020. [https://doi.org/10.1051/itmconf/20203303006]
- Gilquin, G. (2021). Learner Corpora. In M. Paquot & S. Th. Gries (Eds.), A Practical Handbook of Corpus Linguistics (pp. 283–303). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1 13
- Glaznieks, A., Nicolas, L., Stemle, E. W., Abel, A. & Lyding, V. (2014): Establishing a Standardised Procedure for Building Learner Corpora. In: *Apples Journal of Applied Language Studies* 8 (3), 5-20, Special Issue on *Learner Language, Learner Corpora: From corpus compilation to data analysis*, Jarmo Harri Jantunen, Sisko Brunni & Marianne Spoelmann (eds). [http://apples.jyu.fi/article/view/97869]
- Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L. & Nicolas, L. (2022): LEONIDE: A longitudinal trilingual corpus of young learners of Italian, German and English. In: *International Journal of Learner Corpus Linguistics* 8.1, 97-120. [https://doi.org/10.1075/ijlcr.21004.gla]
- Glaznieks, A., Frey, J.-C., & Abel, A. (2023). Weil-Sätze bei Lernenden des Deutschen. Vergleich zwischen immersiv und nicht immersiv Deutschlernenden in Südtirol. In M. Beißwenger, E. Gredel, L. Lemnitzer, & R. Schneider (Eds.), *Korpusgestützte Sprachanalyse: Grundlagen, Anwendungen und Analysen* (pp. 401–423). Francke Narr Attempto.
- Glaznieks, A., Frey, J.-C., Abel, A., Nicolas, L. & Vettori, C. (2023): The Kolipsi Corpus Family: Resources for learner corpus research in Italian and German. In: *Italian Journal of Computational Linguistics* 9(2). [https://doi.org/10.4000/ijcol.1210]
- Götz, S. & Granger, S. (2024): Learner corpus research for pedagogical purposes. An overview and some research perspectives. *Intenational Journal of Learner Corpus Research* 10(1), 1-38.

Literaturauswahl

- Hanecka, H. P. (2015): Grundkurs Soziologie. 10. Auflage. Tübingen.
- Meunier, F. (2020): Introduction to Learner Corpus Research. In: Tracy-Ventura, N. & Paquot, M. (Eds.): *The Routledge Handbook of Second Language Acquisition and Corpora*, S. 23-36.
- Paquot, M., König, A., Stemle, E. W. & Frey, J.-C. (2024): The Core Metadata Schema for Learner Corpora (LC-meta). Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research* 10 (2). [https://doi.org/10.1075/ijlcr.24010.paq]
- Reitbrecht, S. (2025): Anführungszeichen in der Lerner*innensprache zwischen zielsprachlicher Norm und kommunikativer Funktion. Vortrag in der Sektion "Korpora in der angewandten Linguistik" beim IVG-Kongress 2025 an der Universität Graz. Graz, 21.-23.7.2025.
- Schlutter, L. (2025): Genus- und Kasusfehler bei DaZ-Lernenden mit italienischer L1: Eine korpuslinguistische Analyse einfacher Nominalphrasen. Posterpräsentation bei der GAL-Jahrestagung an der Universität Hildesheim. Hildesheim, 17.-19.9.2025.
- Schulz M & Schwebler M. (2025): Würzburger Thesen zur Angewandten Sprachwissenschaft. *Zeitschrift für Angewandte Linguistik* 82, 2025, 130-171. [https://doi.org/10.1515/zfal-2025-2005]
- Spina, S., Glaznieks, A. & Abel, A. (2025): Intensification in Written L2 Italian: Insights from the multilingual region of South Tyrol. *International Journal of Learner Corpus Research* 11(2). [https://doi.org/10.1075/ijlcr.23041.spi]
- Spina, S. & Glaznieks, A. (forthcoming): Morphological and syntactic adjective intensification in L2 Italian and German in a multilingual context. *Folia Linguistica*.