KorDaF Korpora Deutsch als Fremdsprache

Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 2 (Dezember 2023)
ISSN 2749-4411

# NaLeKo
# Ein komplex annotiertes Lernerkorpus mit schriftlichen Erzähltexten des Deutschen als Erst- und Zweitsprache

Hagen Hirschmann, Humboldt-Universität zu Berlin
Anja Binanzer, Leibniz Universität Hannover
Miriam Langlotz, Universität Kassel

**Abstract**

NaLeKo ist ein online zugängliches und auswertbares Lernerkorpus des schriftsprachlichen Deutschen als Erst- und Zweitsprache, entstanden aus dem Forschungsbestreben, den Erwerb und die Entwicklung narrativer Kompetenzen korpusbasiert zu untersuchen. Das Korpus besteht aus Erzählungen von Schüler*innen der Klassenstufen zwei bis neun. Es enthält allgemeine linguistische Annotationen, eine Beschreibung von lernerbedingten Abweichungen durch Zielhypothesen und Fehlertags und wurde außerdem differenziert hinsichtlich der in den Texten verwendeten Junktoren annotiert. Somit liegt mit der ersten Korpusveröffentlichung NaLeKo V1.0 ein tief annotiertes Mehrebenenkorpus mit aktuell 288 Narrationen vor.

**Keywords**: Lernerkorpus; Erstspracherwerbsforschung; Zweitspracherwerbsforschung; Mehrsprachigkeit; Sprachentwicklung; Mehrebenenkorpus; Korpusarchitektur

**Abstract**

NaLeKo is an online accessible and analyzable learner corpus of written texts of German as a first and second language. The underlying research aim is to analyze the written acquisition and development of narrative competence using a corpus-based approach. The corpus consists of narrative texts written by students in grades two to nine and contains general linguistic annotations, descriptions of learner-based deviations, expressed via target hypotheses and error tags. So far, the corpus has also been annotated for connectives used in the texts. As a result, the first corpus publication NaLeKo V1.0 is a deeply annotated multidimensional corpus with currently 288 narratives.

**Keywords:** learner corpus; first language acquisition; second language acquisition; multilingualism; language development; multi-layer corpus; corpus architecture

## 1. Einleitung

Das Erst- und Zweitspracherwerbskorpus NaLeKo[1] besteht aus Narrationstexten von Schüler*innen der Klassenstufen 2, 3, 4, 5, 7 und 9 mit Deutsch als L1 und L2. Den Ausgangspunkt für das Korpusprojekt bildeten bereits digitalisierte, ursprünglich handschriftlich erhobene Texte. Das Korpus basiert auf Daten, die bereits in (weiter unten erwähnten) Forschungsprojekten an Schulen in Deutschland erhoben wurden. Durch NaLeKo sollen diese Forschungsdaten längerfristig gesichert und der Forschungsgemeinschaft zur Verfügung gestellt werden. Hierdurch soll im Sinne nachhaltigen Forschungsdatenmanagements der offene Umgang mit Forschungsdaten und ebenso die Transparenz, Reproduzierbarkeit und Nachnutzung der Forschungsdaten gefördert werden.

In einem ersten Schritt der Datenaufbereitung war es das Ziel, die Fähigkeit zur Satzverknüpfung (Junktion) in diesen Daten systematisch auswerten zu können, so dass u. a. Sprachentwicklungsverläufe nachgezeichnet und kontrastive L2-L1-Vergleiche angestellt werden können (vgl. auch Binanzer / Langlotz 2019). Die entsprechenden linguistischen Annotationen im Korpus erfolgten im

[1] **Na**rratives **Le**rner**ko**rpus, https://hu-berlin.de/naleko/ (01.12.2023).

Wesentlichen anhand der von Raible (1992), Ágel (2010) und Langlotz (2014) erarbeiteten Analysekonzepte zur Junktion im Deutschen. In diesem Sinn ermöglicht es die hier vorgestellte erste Veröffentlichung NaLeKo V1.0, den Erwerb der lexikalischen Mittel zur Textverknüpfung unter Berücksichtigung der ausgedrückten Bedeutung sowie syntaktischen Verwendungsmustern abzubilden. Erste Analyseergebnisse mit einem Fokus auf temporale Junktoren liegen bereits vor (vgl. Binanzer/ Hirschmann / Langlotz 2022).

Zusätzlich zu den spezifischen Junktionsanalysen ist das Korpus mit allgemeinen korpusüblichen Annotationen versehen und entspricht lernerkorpusmethodologischen Standards. Zu diesen gehören Annotationen, welche spracherwerbsbedingte Abweichungen abbilden (s. Abschnitt 3). Das Korpus ist nach Erwerbsstufen stratifizierbar, um Entwicklungstendenzen zu den im Korpus verfügbaren Annotationen erfassen zu können. Die nachfolgend beschriebenen Details zur Datenaufbereitung und Korpusarchitektur richten sich nach diesen Grundsätzen.

## 2. Grundlegende Merkmale der Lernertexte (Erhebungsspezifika, Metadaten)

Die dem Korpus zugrunde liegenden Lernertexte wurden in den Jahren 2012-2022 von Anja Binanzer (vgl. Binanzer 2017) und Miriam Langlotz (vgl. Langlotz 2014 sowie Langlotz / Späth 2022) erhoben. Hierzu wurden in den Klassenstufen 2, 3, 4, 5, 7 und 9 an insgesamt sechs Grundschulen und einem Gymnasium in den Bundesländern Nordrhein-Westfalen, Niedersachsen und Hessen unterschiedliche, altersangepasste Schreibanlässe genutzt, um narrative Texte zu elizitieren:

- In den Klassenstufen 2-4 wurde der Bildimpuls „Hund und Katze“ genutzt – die beiden Tiere schmiegen sich harmonisch aneinander. Der Schreibauftrag bestand darin, eine Geschichte zu verfassen, die die auf dem Bild dargestellte Freundschaft begründet (vgl. Binanzer 2017: 238).
- In denselben Klassenstufen 2-4 wurden Texte zum Bildimpuls „Eine schlaflose Nacht“ genutzt – das Bild zeigt eine Frau und einen Mann (offensichtlich Mutter und Vater), die besorgt auf ein schreiendes Baby schauen, das sie in ihrer Mitte halten. Die Frau hält außerdem einen Teddybären, der Mann einen Ball. Hier besteht der Schreibauftrag darin, eine Geschichte zu verfassen, die die aktuelle Situation erklärt und eine Lösung derselben aufzeigt (vgl. Binanzer 2017: 238).
- In den Klassenstufen 5-9 bestand der Schreibauftrag (ohne Bildimpuls) darin, eine schulbezogene Geschichte zu verfassen (vgl. Langlotz 2014: 70).

Die Texte wurden in vorgegebenen Zeitrahmen (max. 45min.) handschriftlich im Klassenraum verfasst. Elterliche Einverständniserklärungen zur forschungsbezogenen Nutzung der Daten wurden eingeholt.

Im Nachgang dieser Erhebungen wurden zur Erstellung von NaLeKo v1.0 288 Schülertexte transkribiert (hierbei digitalisiert) und entsprechend dem nachfolgend dargestellten Aufbereitungsverfahren korpuslinguistisch weiterverarbeitet. Zur Typisierung der Proband*innen lässt sich sagen, dass 162 Texte von 135[2] Schüler*innen mit Deutsch als L1 vorliegen. Hierunter fallen 14 Texte von Schüler*innen mit einer weiteren L1, die im Korpus als bilingual aufwachsend eingestuft werden. Des Weiteren finden sich im Korpus 126 Texte von 72[3] Schüler*innen mit Deutsch als Zweitsprache (DaZ). Die Erstsprachen sind im Wesentlichen Russisch (64 Texte), Türkisch (41 Texte) und Italienisch (14 Texte). Zusätzlich sind die L1 Armenisch (vier Texte), Polnisch (vier Texte), Bosnisch, Bulgarisch, Kasachisch, Kroatisch, Kurdisch, Portugiesisch, Rumänisch, Serbisch und Twi (jeweils zwei Texte) vertreten.

[2] Von 27 Proband*innen mit Deutsch L1 liegen jeweils zwei Texte vor.
[3] Von 54 Proband*innen mit Deutsch L2 liegen jeweils zwei Texte vor.

Neben diesen Spracherwerbshintergründen und der Klassenstufe sind in Korpusmetadaten das Alter (7-17 Jahre), der Aufgabentyp, das Geschlecht, die Schule sowie der Schultyp, im Erwerb befindliche Fremdsprachen und die in der Familie dominierende Umgangssprache abgebildet. Für 118 Dokumente sind die Ergebnisse von Sprachstandstests (C-Test) verzeichnet.

Abbildung 1 zeigt einen exemplarischen Textausschnitt von einer zehnjährigen Schülerin der vierten Klasse mit L1 Türkisch und L2 Deutsch.

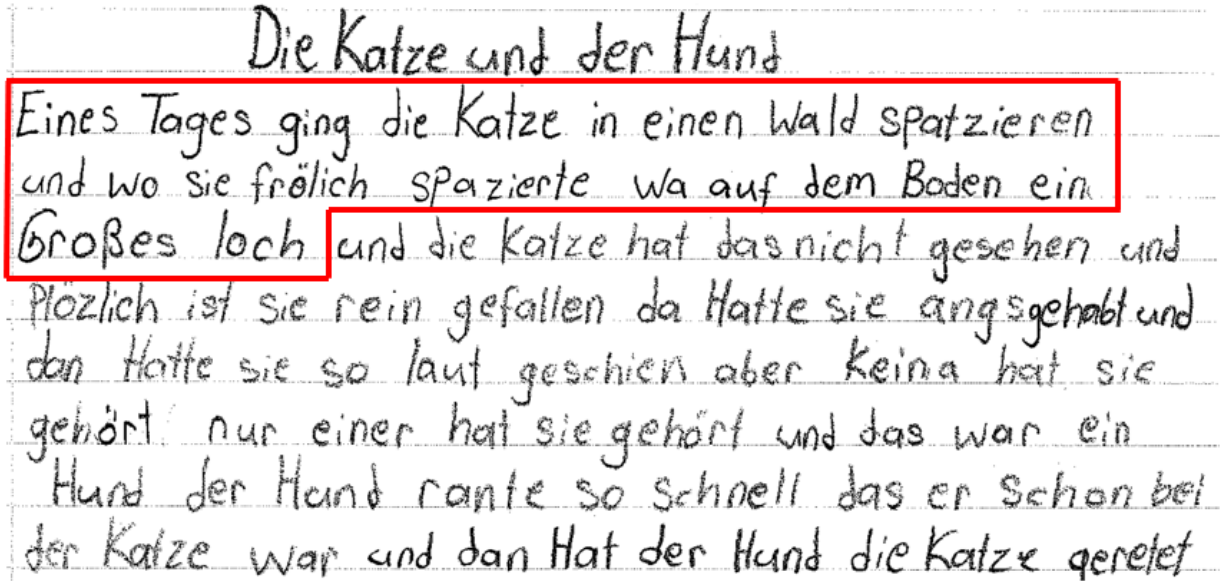
Die Katze und der Hund
Eines Tages ging die Katze in einen Wald spatzieren
und wo sie frölich spazierte wa auf dem Boden ein
Großes loch und die Katze hat das nicht gesehen und
Plötzlich ist sie rein gefallen da Hatte sie angsgehabt und
dan Hatte sie so laut geschien aber keina hat sie
gehört nur einer hat sie gehört und das war ein
Hund der Hund rante so schnell das er schon bei
der Katze war und dan Hat der Hund die Katze geretet

Abbildung 1
Exemplarischer Textausschnitt

Der rot umrandete Textteil wird für die nachfolgenden Erläuterungen zur Datenaufbereitung in NaLeKo verwendet.

## 3. Das Kerngerüst: Korpusaufbereitung mit der Dulko-Transformationspipeline in EXMARaLDA

Im Rahmen des Projekts Dulko (Deutsch-ungarisches Lernerkorpus), das maßgeblich an der Universität Szeged durchgeführt wurde und die Erstellung eines fehlerannotierten Korpus mit ungarischen Lernenden des Deutschen als Fremdsprache zum Ziel hat, wurde eine Datenaufbereitungsprozedur für Lernerkorpora entwickelt, die vollständig im Partitur-Editor EXMARaLDA (http://www.exmaralda.org/, vgl. Schmidt / Wörner 2014) durchführbar ist. Die Ressource wird auf der Webseite https://sr.ht/~nolda/exmaralda-dulko/ beschrieben; der damit beschrittene Aufbereitungsweg für das Dulko-Korpus ist beschrieben in Nolda (2023) und Hirschmann / Nolda (2019). Die einzelnen Verarbeitungsschritte orientieren sich an der Korpusarchitektur des Lernerkorpus Falko (https://hu-berlin.de/falko/, vgl. Hirschmann et al. 2022).

Nutzer*innen können sich aus den verfügbaren Verarbeitungsschritten eine eigene Aufbereitungsprozedur zusammenstellen. Zusammengefasst besteht die für NaLeKo verwendete Prozedur aus den folgenden Verarbeitungsschritten:

- Tokenisierung der transkribierten Lernertextdaten;
- Tagging der tokenisierten Daten: Lemmatisierung und Wortartenzuweisung mittels TreeTagger[4] (vgl. Schmid 1994);
- Hinzufügung von Satzspannen mittels Erkennung von Satzbeendungszeichen.

Aus diesen Verarbeitungsschritten ergibt sich der in Abbildung 2 dargestellte Datenaufbau.

| [word] | Eines | Tages | ging | die | Katze | in | einen | Wald | spatzieren | und | wo | sie | frölich | spazierte | wa | auf | dem | Boden | ein | Großes | loch |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [S] | s1 | | | | | | | | | | | | | | | | | | | | |
| [pos] | ART | NN | VVFIN | ART | NN | APPR | ART | NN | VVINF | KON | PWAV | PPER | ADJD | VVFIN | NE | APPR | ART | NN | ART | NN | VVIMP |
| [lemma] | eine | Tag | gehen | die | Katze | in | eine | Wald | spatzieren | und | wo | sie | frölich | spazieren | Wa | auf | die | Boden | eine | Große | lochen |

Abbildung 2
Screenshot von der Annotation des Lernertexts (Hund1_T1_SOEK) im EXMARaLDA-Partitureditor mit Satzspannen (Zeile [S]), Wortarten (Zeile [pos]) und Lemmata (unterste Zeile [lemma])

Die Lemmatisierung des TreeTaggers (Ebene [lemma]) erfolgt mittels eines Vollformlexikons. Bei unbekannten Wortformen dupliziert das System die Oberflächenform. Die Wortartenkürzel, die ebenso mittels TreeTagger auf der Ebene [pos] vergeben werden, entstammen dem STTS-Tagset (vgl. Schiller et al. 1999, www.cis.lmu.de/~schmid/tools/TreeTagger/data/STTS-Tagset.pdf). In der in Abbildung 2 gezeigten Analyse finden sich bestimmte Analyseprobleme, wenn eine normabweichende Form auftritt. Die auffälligsten Probleme im Beispielsatz sind diesbezüglich die Formen *wa* (anstelle eines Verbs als Eigenname analysiert) und *loch* (anstelle eines Nomens als Imperativform des Verbs *lochen* analysiert). Die Annotation der Satzspannen (Ebene [S]) erfolgt nach der (von den Lernenden) gesetzten Interpunktion (Satzbeendungszeichen bewirken eine Segmentgrenze). Solche Probleme lassen wir zunächst unberührt.

Die anschließende Datenverarbeitung betrifft die folgenden Schritte:

- Hinzufügung einer Zielhypothesenspur, die anschließend manuell bearbeitet wird.
- Hinzufügung von Spuren, die im Fall lernerbedingter Abweichungen die Art der Abweichung beschreiben (verschiedene Fehlertags werden mittels eines Annotationspanels manuell hinzugefügt).

Hieraus ergibt sich der in Abbildung 3 aufgezeigte Aufbau.

[4] Der TreeTagger wird durch die Transformation in EXMARaLDA angesteuert und muss zu diesem Zweck vorher installiert worden sein.

| [word] | Eines | Tages | ging | die | Katze | in | einen | Wald | spatzieren | und | wo | sie | frölich | spazierte | | wa | auf | dem | Boden | ein | Großes | loch | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [S] | s1 | | | | | | | | | | | | | | | | | | | | | | |
| [pos] | ART | NN | VVFIN | ART | NN | APPR | ART | NN | VVINF | KON | PWAV | PPER | ADJD | VVFIN | | NE | APPR | ART | NN | ART | NN | VVIMP | |
| [lemma] | eine | Tag | gehen | die | Katze | in | eine | Wald | spatzieren | und | wo | sie | frölich | spazieren | | Wa | auf | die | Boden | eine | Große | lochen | |
| [ZH] | **Eines** | **Tages** | **ging** | **die** | **Katze** | **in** | **einem** | **Wald** | **spazieren** | **und** | **als** | **sie** | **fröhlich** | **spazierte** | **,** | **war** | **in** | **dem** | **Boden** | **ein** | **großes** | **Loch** | **.** |
| [ZHDiff] | | | | | | | CHA | | CHA | | CHA | | CHA | | CHA | CHA | CHA | | | | CHA | CHA | INS |
| [ZHS] | s1 | | | | | | | | | | | | | | | | | | | | | | |
| [ZHpos] | ART | NN | VVFIN | ART | NN | APPR | ART | NN | VVINF | KON | KOUS | PPER | ADJD | VVFIN | $, | VAFIN | APPR | ART | NN | ART | ADJA | NN | $. |
| [ZHlemma] | eine | Tag | gehen | die | Katze | in | eine | Wald | spazieren | und | als | sie | fröhlich | spazieren | , | sein | in | die | Boden | eine | groß | Loch | . |
| [FehlerOrth] | | | | | | | | | WS | | | | WS | | ZS | WS | | | | | GKS | GKS | ZS |
| [FehlerMorph] | | | | | | | | | | | | | | | | | | | | | | | |
| [FehlerSyn] | | | | | | | ValAP | | | | | | | | | | | | | | | | |
| [FehlerLex] | | | | | | | | | | | Lex | | | | | | | | | | | | |
| [FehlerSem] | | | | | | | | | | | | | | | | | SemRel | | | | | | |

Abbildung 3
Screenshot von der Annotation des Lernertexts (Hund1_T1_SOEK) mit den zusätzlichen Spuren [ZH], [ZHDiff], erneutem Tagging von Satzspannen, Wortarten und Lemmata ([ZHS], [ZHpos] und [ZHlemma]) sowie fünf Abweichungsebenen[5]

Die Zielhypothese (Ebene [ZH]) stellt eine Annahme darüber dar, welche Form in einem Text standardisierter Schriftlichkeit gewählt worden wäre. Hierbei handelt es sich um einen notwendigen Verarbeitungsschritt, der die Grundlage für die nachfolgenden Beschreibungen der Abweichungen im Lernertext bietet (vgl. Lüdeling 2008). Wie weit die Zielhypothese in den Lernertext eingreift und welche Abweichungen nach welchen Maßgaben behandelt werden, ist grundsätzlich arbiträr und abhängig von den jeweiligen Analysezielen. Um später auch bestimmte stilistische, lexikalische und semantische Auffälligkeiten strukturell finden zu können, wurden in NaLeKo nicht nur orthographische und grammatikalitätsbeeinflussende Abweichungen ‚korrigiert', sondern eben auch subtilere Abweichungen wie das nebensatzanschließende *wo* im Beispielausschnitt. Die Zielhypothese hat ausdrücklich nicht den Anspruch, über die Korrektheit von Äußerungsteilen zu urteilen, sondern bestimmte Typen von Abweichungen abbilden zu können. Die tokengenaue Annotation der Zielhypothese erlaubt auf der Analyseebene [ZHDiff] die automatische Hinzufügung der Information darüber, ob ein Token geändert wurde (CHA), eingefügt (INS), entfernt (DEL) oder aufgespalten (SPLIT) wurde oder ob zwei Token der Ebene [word] zusammengeführt wurden (MERGE). Die Klassifikation der Abweichungen mitsamt den Ebenenbezeichnungen („Fehler...") ist aus dem Dulko-Projekt übernommen und wird mithilfe eines in EXMARaLDA implementierten Annotationspanels umgesetzt. Bei jedem der benannten Einträge auf der Ebene [ZHDiff] werden auf einer oder mehreren der [Fehler...]-Spuren spezifische Interpretationen der Abweichung eingetragen. Dies erlaubt es später, mit einem Suchbefehl alle Vorkommen bestimmter Typen von Abweichungen auffinden zu können.

Die automatische Verarbeitung der Zielhypothesenebene [ZH] (analog zur vorigen Verarbeitung der Lernertextebene [word]) ist nun aufgrund der Normalisierung, die die Erstellung der Zielhypothese mit sich bringt, deutlich akkurater. Verbleibende Probleme werden, sofern sie erkannt werden, manuell korrigiert, so dass die Zuverlässigkeit der Analysen auf den Ebenen [ZHpos] und [ZHlemma] als sehr hoch anzusehen ist (diesbezügliche Evaluationen stehen noch aus).

[5] Erläuterungen zu den Kürzeln: [ZH] steht für „Zielhypothese"; [ZHDiff] steht für die Differenz zwischen der Ebene [ZH] und der Ebene [word] (der Lernertextebene); [FehlerOrth] steht für orthographische Abweichungen; [FehlerMorph] steht für wortbildungs- und flexionsmorphologische Abweichungen; [FehlerSyn] steht für syntaktisch interpretierbare Abweichungen; [FehlerLex] steht für lexikalische Abweichungen; [FehlerSem] steht für semantisch interpretierbare Abweichungen.

## 4. Analyse von Junktion

Junktoren werden auf vier Beschreibungsebenen analysiert. Für den exemplarischen Satz ergibt sich das in Abbildung 4 dargestellte Analysebild.

| [word] | Eines | Tages | ging | die | Katze | in | einen | Wald | spatzieren | und | wo | sie | frölich | spazierte | | wa | auf | dem | Boden | ein | Großes | loch | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [pos] | ART | NN | VVFIN | ART | NN | APPR | ART | NN | VVINF | KON | PWAV | PPER | ADJD | VVFIN | | NE | APPR | ART | NN | ART | NN | VVIMP | |
| [lemma] | eine | Tag | gehen | die | Katze | in | eine | Wald | spatzieren | und | wo | sie | frölich | spazieren | | Wa | auf | die | Boden | eine | Große | lochen | |
| [ZH] | Eines | Tages | ging | die | Katze | in | einem | Wald | spazieren | und | als | sie | fröhlich | spazierte | , | war | in | dem | Boden | ein | großes | Loch | . |
| [ZHDiff] | | | | | | | CHA | | CHA | | CHA | | CHA | | INS | CHA | CHA | | | | CHA | CHA | INS |
| [ZHpos] | ART | NN | VVFIN | ART | NN | APPR | ART | NN | VVINF | KON | KOUS | PPER | ADJD | VVFIN | $, | VAFIN | APPR | ART | NN | ART | ADJA | NN | $. |
| [ZHlemma] | eine | Tag | gehen | die | Katze | in | eine | Wald | spazieren | und | als | sie | fröhlich | spazieren | , | sein | in | die | Boden | eine | groß | Loch | . |
| [SyntaxJunktion] | | | | | | | | | | KON | SUB | | | | | | | | | | | | |
| [SemantikJunktion] | | | | | | | | | | KOP | TEMPGZ | | | | | | | | | | | | |
| [StellungJunktion] | | | | | | | | | | | VF | | | | | | | | | | | | |
| [NSFunktion] | | | | | | | | | | | ADV | | | | | | | | | | | | |

Abbildung 4
Screenshot von der Annotation des Lernertexts (Hund1_T1_SOEK) mit den zusätzlichen Spuren [SyntaxJunktion], [SemantikJunktion], [StellungJunktion] und [NSFunktion], auf denen die Junktoren im Korpus annotiert werden. Einige der zuvor benannten Annotationsebenen wurden hier zugunsten der besseren Übersicht weggelassen.

Im Beispielsatz existieren für die Junktionsanalyse zwei relevante Wörter (*und* und *wo* bzw. *als*). Auf der Annotationsebene [SyntaxJunktion] wird die syntaktische Funktion des jeweiligen Junktors spezifiziert. Im Wesentlichen werden subordinierende (SUB) und koordinierende (KON), aber auch Adverbjunktoren (AP), bestimmte Partikeln (PTK) und wenige andere Konstruktionen unterschieden. Diese Typen werden auf der Ebene [SemantikJunktion] auf ihre Bedeutung hin unterschieden. Im konkreten Fall werden die Verwendung des Konjunktors *und* mit kopulativ (KOP) und die Verwendung des nebensatzeinleitenden *wo* bzw. der normalisierten Form *als* mit temporal-gleichzeitig (TEMPGZ) analysiert. Eingebettete bzw. einbettende Junktoren werden auf der Ebene [StellungJunktion] dahingehend untergliedert, ob sie im Vorfeld (VF), Mittelfeld (MF) oder Nachfeld (NF) stehen. Im Falle von nebensatzeinbettenden Junktoren werden die Funktionen Komplementierung (KOMP), Adverbial (ADV) und Attribut (ATTR) differenziert. (Wie Abbildung 4 zeigt, werden die letzten beiden Spezifizierungen bei nebenordnenden Junktoren und anderen Fällen, in denen die Klassifikationskriterien nicht anwendbar sind, weggelassen.). Zusammengefasst werden im Beispielsatz *und* als kopulativer Konjunktor und *wo*, das als Variante zur Subjunktion *als* interpretiert wird, als adverbialer Subjunktor mit temporal-gleichzeitiger Bedeutung, der im Vorfeld des Matrixsatzes steht, ausgewiesen.

## 5. Zugänglichkeit und Nutzungsperspektiven

Das Korpus ist in einer Instanz des ANNIS-Suchinterfaces (https://corpus-tools.org/annis/, vgl. Krause / Zeldes 2016) für Lernerkorpora (https://korpling.org/annis/lc/) veröffentlicht und ist dort analysierbar. Die Konversion der Korpusdaten vom EXMARaLDA- in das ANNIS-Format erfolgte mittels des Konversionsframeworks ‚Pepper' (vgl. Zipser / Romary 2010, https://corpus-tools.org/pepper/). Im Folgenden werden einige Such- und Auswertungsszenarien skizziert, welche die in Abschnitt 3 und 4 dargestellten Analysen aufgreifen. Die exemplarischen Suchanfragen können im Eingabefenster (oben links im oben genannten ANNIS-Interface) eingegeben werden. Das Korpus NaLeKo_V1.0 muss hierbei ausgewählt (blau markiert) sein. Ein Kurzlink zum Korpus im Suchinterface ist https://hu.berlin/annis-naleko/.

- Suche nach einer bestimmten Lernerform – hier beispielhaft die Form *haten* im gesamten Korpus: word="haten".[6] Auf diese Weise können sämtliche Lernerformen direkt abgefragt werden. Reguläre Ausdrücke werden folgendermaßen in Schrägstriche gesetzt:
- Suche nach allen Lernerformen mit dem Bestandteil "spil" bzw. „Spil": word=/.*(S|s)pil.*/.
- Suche nach bestimmten normalisierten Formen, die der Lernerform entsprechen (hier alle Vorkommen von normalisiert „hatten", die auch so von den Lernenden realisiert wurden: ZH="hatten" _=_ word="hatten".
- Suche nach normalisierten Formen, die sich von den Lernerformen unterscheiden: ZH="hatten" _=_ ZHDiff="CHA" (hier wird normalisiert „hatten" gesucht, das mit dem Tag „CHA" auf der Ebene ZHDiff abdeckt, was besagt, dass die Lernerform abgeändert wurde); ZH="hatten" _=_ word & #1 != #2 (hier wird nach normalisiert „hatten" gesucht, das mit einer Lernerform (word) abdeckt und ungleich dieser Form ist).
- Suche nach Nomina (außer Eigennamen), die in ihrer Verschriftlichung von einer Abweichung im Bereich der Groß-/Kleinschreibung betroffen sind: ZHpos="NN" _=_ FehlerOrth="GKS".
- Suche nach kausalen Subjunktoren: SyntaxJunktion="SUB" _=_ SemantikJunktion="KAUS".
- Metadaten (hier als Beispiel die Einschränkung auf Schüler*innen der Klassenstufe vier oder auf Schüler*innen mit L2 Deutsch): Der gegebenen Suche z. B. den Ausdruck @* Klasse="4" (für Klassenstufe 4) oder @* L2="D" (für L2 Deutsch) hinzufügen.
- Frequenztabellen online erstellen – hier werden orthographische Abweichungen bei Nomina aufgelistet: Geben Sie eine Suche ein, die die auszuwertende Variable (hier Wortformen auf der Lernerformebene „word") enthält – für das gegebene Beispiel ist das ZHpos="NN" _=_ word _=_ FehlerOrth. Um eine Variable (hier „word") auszuwerten, wählt man die Funktion „More" > „Frequency Analysis", löscht dann die für die Auswertung irrelevanten Variablen (hier „ZHpos" und „FehlerOrth") durch Anklicken und die Funktion „Delete selected row(s)" und wählt dann die Funktion „Perform frequency analysis".
- Ergebnisse exportieren – hier sollen alle Lernerformen (Ebene „word"), die den Bestandteil „Spiel" bwz. „spiel" enthalten und orthographisch abweichend sind, exportiert werden: Zur durchlaufenen Suchanfrage ZH=/.*(S|s)piel.*/ _=_ FehlerOrth wählt man die Funktion „More" > „Export", wählt (den für den genannten Zweck passenden) „GridExporter", spezifiziert den Kontext mit „0" und gibt bei „Annotation Keys" „word" ein, um Lernerformen zu exportieren. Die Funktion „Perform Export" liefert die herunterladbare Datenmenge.

Diese Anwendungsbeispiele decken nur einen Bruchteil des mit den Daten Möglichen ab. Prinzipiell sind alle strukturellen Beziehungen in den komplex annotierten Daten in der Anfragesprache des ANNIS-Suchinterfaces ausdrückbar. Für eine generellere Einführung in die Korpussuche (vergleichend für verschiedene Suchsysteme, u. a. ANNIS) vgl. auch Hirschmann (2019: 108-153).

Zukünftig soll NaLeKo sowohl in die Tiefe wachsen und mehr Annotationen erhalten als auch breiter werden, indem dem Korpus unter Einhaltung der bisher geltenden Erhebungsstandards mehr Textdaten von Schüler*innen hinzugefügt und nach der hier dargestellten Prozedur aufbereitet werden. Derzeit werden Analysen von Tempuskonstruktionen erarbeitet. Bereits erhobene Daten von DaZ-Schüler*innen werden zukünftig eingepflegt.

## Literatur und Ressourcen

Ágel, Vilmos (2010): Explizite Junktion. Theorie und Operationalisierung. In: Ziegler, Arne / Braun, Christian (Hrsg.): *Historische Textgrammatik und historische Syntax des Deutschen: Traditionen, Innovationen, Perspektiven: Vol. 2: Frühneuhochdeutsch, Neuhochdeutsch.* Berlin / New York: de Gruyter, 897-936.

[6] Bitte im Interface die jeweilige grau unterlegte Suchanfrage eingeben.

Binanzer, Anja (2017): *Genus – Kongruenz und Klassifikation: Evidenzen aus dem Zweitspracherwerb des Deutschen.* Berlin / New York: de Gruyter.

Binanzer, Anja / Hirschmann, Hagen / Langlotz, Miriam (2022): Narrative Funktionen temporaler Junktoren – Lernertexte und Kinder- und Jugendliteratur im korpuslinguistischen Vergleich. In: Mesch, Birgit / Uhl, Benjamin (Hrsg.): *Tempus und Temporalität. Empirische Zugänge zum Erwerb von Zeitlichkeit.* Münster: Waxmann, 175-198.

Binanzer, Anja / Langlotz, Miriam (2019): Junktion und Narration – Schreibentwicklungsprozesse ein- und mehrsprachiger Kinder. In: Binanzer, Anja / Langlotz, Miriam / Wecker, Verena (Hrsg.): *Grammatik in Erzählungen – Grammatik für Erzählungen. Erwerbs-, Entwicklungs- und Förderperspektiven*. Baltmannsweiler: Schneider, 125-150.

Hirschmann, Hagen (2019): *Korpuslinguistik. Eine Einführung*. Stuttgart: Metzler.

Hirschmann, Hagen / Lüdeling, Anke / Shadrova, Anna / Bobeck, Dominique / Klotz, Martin / Akbari, Roodabeh / Schneider, Sarah / Wan, Shujun (2022): FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. In: *KorDaF (Korpora Deutsch als Fremdsprache)* 2: 2, 139-148.

Hirschmann, Hagen / Nolda, Andreas (2019): Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus. In: Eichinger, Ludwig / Plewnia, Albrecht (Hrsg.): *Neues vom heutigen Deutsch: Empirisch – methodisch – theoretisch. Institut für Deutsche Sprache: Jahrbuch 2018.* Berlin / Boston: de Gruyter, 339-342.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities* 31: 1, 118-139.

Langlotz, Miriam (2014): *Junktion und Schreibentwicklung: eine empirische Untersuchung narrativer und argumentativer Schülertexte.* Berlin u.a.: de Gruyter.

Langlotz, Miriam / Späth, Sina (2022): *Interpunktion im bilingualen Schriftspracherwerb Deutsch/Italienisch – eine Untersuchung freier Texte aus dem 4. Jahrgang.* In: Noack, Christina / Nimz, Katharina / Schmidt, Karsten (Hrsg.): *Mehrsprachigkeit und Orthographie. Empirische Studien an der Schnittstelle von Linguistik und Sprachdidaktik.* Baltmannsweiler: Schneider Hohengehren, 143-162.

Lüdeling, Anke (2008): Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. In: Walter, Maik / Grommes, Patrick (Hrsg.): *Fortgeschrittene Lernervarietäten.* Tübingen: Niemeyer, 119-140.

Nolda, Andreas (2023): Fehlerannotation und Fehleranalyse am Beispiel des deutsch-ungarischen Lernerkorpus Dulko. In: *Jahrbuch für internationale Germanistik* 10, 747-755.

Raible, Wolfgang (1992): *Junktion: eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration.* Heidelberg: Winter.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technischer Bericht. Universität Stuttgart. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (01.12.2023).

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*. Manchester: United Kingdom, 44-49.

Schmidt, Thomas / Wörner, Kai (2014): EXMARaLDA. In: Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (Hrsg.): *The Oxford Handbook on Corpus Phonology*. Oxford: Oxford University Press, 402-419.

Zipser, Florian / Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/ (01.12.2023).

**Biographische Notiz**: Hagen Hirschmann ist Mitarbeiter am Institut für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin, Fachbereich für Korpuslinguistik und Morphologie.

**Kontaktanschrift:**
Hagen Hirschmann
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Unter den Linden 6
10099 Berlin
hagen.hirschmann@hu-berlin.de

**Biographische Notiz**: Anja Binanzer ist Professorin für Deutsch als Zweitsprache an der Leibniz Universität Hannover.

**Kontaktanschrift:**
Anja Binanzer
Leibniz Universität Hannover
Deutsches Seminar
Königsworther Platz 1
30167 Hannover
anja.binanzer@germanistik.uni-hannover.de

**Biographische Notiz**: Miriam Langlotz ist Professorin für Didaktik der deutschen Sprache und Literatur mit Schwerpunkt Grundschule an der Universität Kassel.

**Kontaktanschrift:**
Miriam Langlotz
Universität Kassel
Institut für Germanistik
Kurt-Wolters-Straße 5
34125 Kassel
m.langlotz@uni-kassel.de