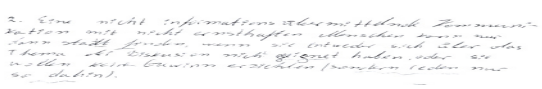
# Ein **f**ehler**a**nnotiertes **L**erner**ko**rpus des Deutschen als Fremdsprache

1. Welche Fehler machen fortgeschrittene Lerner des Deutschen?
2. Wodurch unterscheiden sich Lernertexte von Texten nativer Schreiber?
3. Wie lassen sich Korpora als Hilfsmittel in einer solchen Analyse einsetzen?



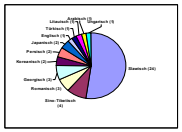
Eine nicht informations übermittelnde Kommunikation mit nicht ernsthaften Menschen kann nur dann stadt finden, wenn sie entweder sich über das Thema der Diskusion nich geeignet haben, oder sie wollen kein Gewinn erziehlen (sondern reden nur so dahin).

| | | | | | | |
|---|---|---|---|---|---|---|
| **Textwort** | oder | sie | wollen | kein | Gewinn | erziehlen |
| **Wortart** | KON | PPER | VMFIN | PIAT | NN | VVINF |
| **Lemma** | oder | sie | wollen | kein | Gewinn | [unknown] |
| **Zielhypothese** | | keinen | Gewinn | erzielen | wollen | |
| **Orthographie** | | | | | | |
| *Identifikation* | | | | x | | x |
| *Klassifikation* | | | | A | | E |
| *Hypothese* | | | | | | Dehnung |
| **Wortstellung** | | | | | | |
| *Identifikation* | | | | x | | |
| *Klassifikation* | | | | MF_RKS | | |
| *Hypothese* | | | | | | |
| **Kongruenz** | | | | | | |
| *Identifikation* | | | | x | | |
| *Klassifikation* | | | | Gen:m_n | | |
| *Hypothese* | | | | GS | | |

## Sammlung

Das **f**ehler**a**nnotierte **L**erner**k**orpus Falko besteht aus den drei Subkorpora
Texte von Schreibern des Deutschen als Fremdsprache (L2)
Texte von nativen Schreibern des Deutschen (L1)
Germanistische Fachtexte (Vorlagen)

**Das Subkorpus L2** enthält in Version 1.0 bislang 44 Texte von Schreibern des Deutschen als Fremdsprache. Diese Texte sind authentische Prüfungstexte, die im Jahre 2004 an der Freien Universität Berlin von ausländischen Germanistikstudierenden verfasst wurden.
Die Prüfung bestand in der Zusammenfassung eines linguistischen bzw. literaturwissenschaftlichen Fachtextes und bildete damit die schriftliche Komponente einer obligatorischen Sprachstandsbestimmung, die zusätzlich zur Zwischenprüfung durchgeführt wird. Die Studierenden haben die DSH-Prüfung erfolgreich absolviert und werden deshalb als fortgeschrittene Lerner (auf dem Niveau C1 - C2 des Europäischen Referenzrahmens) eingestuft. Das Schema 1 zeigt das Spektrum der Herkunftssprachen.




SCHEMA 1: Sprachfamilien/Sprachen in Falko 1.0 (Subkorpus L2)

In einem Fragebogen wurden zudem sprachbiographische Daten erfasst:
- Alter
- Geschlecht
- L1
- Dauer und institutioneller Kontext des L2-, L3-..., Ln-Erwerbs.

Neben den Lernertexten wurde ein **Kontrollkorpus mit nativen Schreibern des Deutschen L1** erstellt. Es umfasst derzeit 30 Texte von Studierenden des Haupstudiums der deutschen Philologie, die anhand identischer Textvorlagen und Aufgabenstellungen verfasst wurden.

Das dritte Subkorpus **Vorlagen** besteht aus den zusammenzufassenden Texten. Die drei angeführten Subkorpora wurden korpuslinguistisch aufbereitet.

| | L2 Deutsch | L1 Deutsch |
|---|---|---|
| Anzahl der Texte | 44 | 30 |
| Anzahl der Token | 17 607 | 9 659 |
| Anzahl der Lemmata pro 100 Wörter (iterativ normalisiert) | 63.41 | 66.89 |

TABELLE 1: Korpusgröße

Die Korpuszusammensetzung ist streng kontrolliert. Die folgenden Parameter in den beiden Subkorpora L1 und L2 sind konstant:
- Textsorte (Zusammenfassung eines linguistischen bzw. literaturwissenschaftlichen Fachtextes)
- Studienphase (Studium der deutschen Philologie in einem fortgeschrittenen Stadium)
- situativer Kontext (Prüfungssituation unter Aufsicht, keine zugelassenen Hilfsmittel, Zeitbegrenzung, handschriftlich verfasste Texte)

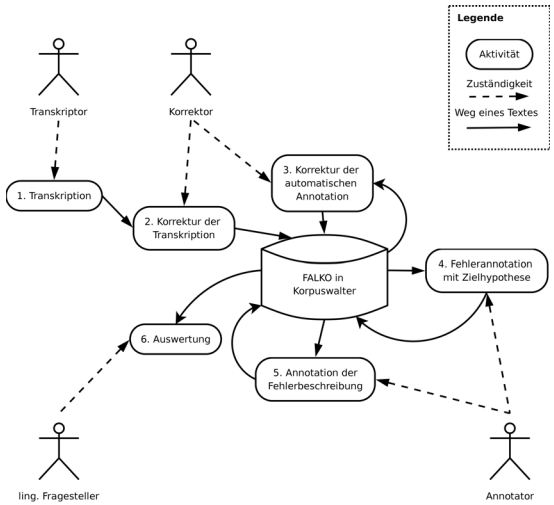
In der Version 1.1 werden derzeit weitere Daten aufbereitet.


**Literatur:**

Belz, J. A. (2004): Learner Corpus Analysis and the Development of Foreign Language Proficiency. In: System: An International Journal of Educational Technology and Applied Linguistics 32.4 , 577-591.
Bolton, K./ Nelson, G./ Hung, J. (2002): A corpus-based study of connectors in student writing. In: International Journal of Corpus Linguistics 7:2 (2002), 165-182.
Breindl, E. (2004): Konnektoren in Übungsgrammatiken. In: Materialien Deutsch als Fremdsprache 66, 426-458.
Ellis, R. (1994): The Study of Second Language Acquisition. Oxford: Oxford University Press.
Granger S./ Hung J./ Petch-Tyson, S. (eds) (2002): Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Language Learning and Language Teaching 6. Amsterdam & Philadelphia: Benjamins.
Granger, S./ Tyson, S. (1996): Connector usage in the English essay writing of native non-native EFL-speakers of English. In: Word Englishes, 15 (1), 17-27.
Granger, S./ Dagneaux, E./ Meunier, F. (eds) (2002): International Corpus of Learner English. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
Pasch, R. / Brauße, U. / Breindl / E. Waßner, H.U. (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). Berlin: Walter de Gruyter.
Pravec, N. A. (2002): Survey of learner corpora. In: ICAME Journal 26, 81-114.
Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing.
Schmidt, T. (2004): Transcribing and annotating spoken language with EXMARaLDA. Proceedings of the LREC-Workshop on XML based richly annotated corpora.
Weinberger, U. (2002): Error Analysis with Computer Learner Corpora: A corpus-based study of errors in the written German of British university students. Master's Thesis. University of Lancaster.


## Aufbereitung

Das Korpus ist auf verschiedenen Ebenen annotiert. Es besitzt eine – im Unterschied zu den meisten bisherigen Lernerkorpora (Pravec 2002)– eine flexible Architektur, welche es ermöglicht, jederzeit Annotationsebenen neu einzufügen und unabhängig von anderen Ebenen zu bearbeiten (*multi-layer stand-off annotation*). Neben den automatisch annotierten Ebenen für Wortart und Lemma verwenden wir spezifische Annotationsebenen für die Auszeichnung von Lernerfehlern.
Tabelle 2 listet die Probleme der Fehlerannotation und ihre Behandlung in flachen Token-Tag-Korpora (z.B. Weinberger 2002) und in Falko auf.




SCHEMA 2: Arbeitsablauf

**Annotationsebenen** In Falko werden Fehler aus unterschiedlichen grammatischen Bereichen (Orthographie, Wortstellung, Kongruenz, Tempus, etc.) unabhängig voneinander annotiert. Jeder Bereich unterteilt sich dabei in 3 Schichten: die Identifikationsschicht dient zur Kennzeichnung des Fehlerbereichs, die Klassifikationsschicht zur Klassifikation des Fehlers und die Hypothesenschicht zur Angabe einer Hypothese über die Ursache eines Fehlers (für die Klassifikations- und Hypotheseschichten wurden jeweils Tagsets entwickelt, siehe URL).

Die **Aufbereitung** eines Textes verläuft in folgenden Schritten:

- Transkription von der handschriftlichen Vorlage
- Überprüfung auf Transkriptionsfehler
- Automatische Annotation von Wortart und Lemma
- Überprüfung auf Annotationsfehler
- Angabe einer expliziten Zielhypothese für fehlerhafte Bereiche (Wort oder Folge von Wörtern)
- Annotation der Fehler auf verschiedenen Fehlerannotationsebenen

Zur automatischen Annotation verwenden wir die Software *TreeTagger* (H. Schmidt 1994). Alle anderen Schritte erfolgen dezentral durch verschiedene Annotatoren. Die Annotation wird mit dem Annotationsprogramm *EXMARaLDA* (T. Schmidt 2004) erstellt. Eine von uns entwickelte webbasierte Software (*Korpuswalter*) ermöglicht es, dass Texte gleichzeitig von verschiedenen Annotatoren bearbeitet werden. *Korpuswalter* führt die automatische Annotation mit Hilfe des *TreeTaggers* durch, sammelt die Annotation der verschiedenen Annotatoren und überprüft sie auf Konsistenz.

| Problem | Beschreibung | ‚flache' Annotation | Falko |
|---|---|---|---|
| **Definition von Fehlern** | Fehler werden oft als Regel- oder Normverstöße beschrieben. Dabei ist es schwierig, zu entscheiden, welche Norm/welches Regelsystem vorliegt. Für einige Gebiete (z.B. Orthographie) liegt ein solches Regelsystem vor, für viele andere Bereiche (z.B. Wortstellung, Ausdruck) gibt es nur Präferenzen, manchmal gibt es verschiedene Regelsysteme. | oft Mischung von verschiedenen Informationstypen wie Fehlerart & Entstehungshypothese, Unterschiede zwischen ‚ungrammatisch' und ‚ungewöhnlich' nur in den Tags erkennbar | verschiedene Typen von Information getrennt, ‚ungrammatische' und ‚ungewöhnliche' Informationen auf unterschiedlichen Ebenen |
| **Zielhypothese** | Um einen Fehler zu erkennen, muss die Lerneräußerung mit einer (angenommenen) Zielhypothese verglichen werden. Bei vielen Lerneräußerungen können unterschiedliche Zielhypothesen angenommen werden. | implizit | explizit, konkurrierende Auszeichnungen möglich |
| **Fehlerkategorien** | Man kann ganz unterschiedlich kategorisieren – nach der betroffenen Wortart, nach dem formalen Fehlertyp (z.B. Auslassung, Vertauschung) oder nach theoretischen Annahmen | Festlegung auf eine Kategorisierung | verschiedene Ebenen/Kategorien möglich, s.u. |
| **Fehlerexponent** | Viele Fehler (z. B. Wortstellungsfehler) kann man nicht eindeutig einem Wort zuweisen; zum Teil ist eine ganze Sequenz falsch. Dazu kommt, dass es mehrere Fehler innerhalb eines Wortes (z.B. ein Orthographiefehler und ein Wortbildungsfehler) oder einer Sequenz geben kann. | Annotationseinheit Token | Annotationseinheit Token oder Sequenz von Tokens, überlappende/konfligierende Annotation möglich |

TABELLE 2: Probleme der Fehlerannotation

## Auswertung

Lernerkorpora können auf zwei Arten ausgewertet werden
- Fehleranalyse
- statistischer Vergleich mit Kontrolltexten (Contrastive Interlanguage Analysis)

Eine **Fehleranalyse** von Falko zeigt, dass auch fortgeschrittene Lerner häufig Fehler machen in Bereichen
- Orthographie: Dehnung, Schärfung, Getrennt- & Zusammenschreibung, Groß- und Kleinschreibung
- Wortbildung: Fugensetzung in Komposita, Verbpräfigierung
- Tempus & Modus: Verwendung des Konjunktivs
- Wortstellung: Satzklammerfehler
- Kongruenz: Subjekt-Finitum-Kongruenz, Fehler in der Nominalflexion
- Ausdruck: Verletzungen der *nativelike selection*

In einer **Contrastive Interlanguage Analysis** wird der Gebrauch der Konnektoren auf der Basis von Pasch et al. 2003 in Falko untersucht. Konnektoren definieren sich durch die folgenden vier Merkmale:
- x ist nicht flektierbar.
- x vergibt keine Kasusmerkmale an seine syntaktische Umgebung.
- Die Bedeutung von x ist eine zweistellige Relation.
- Die Relate der Bedeutung von x sind Sachverhalte.
- Die Relate der Bedeutung von x müssen durch Sätze bezeichnet werden können.

[Pasch, R. et al. 2003: 1]

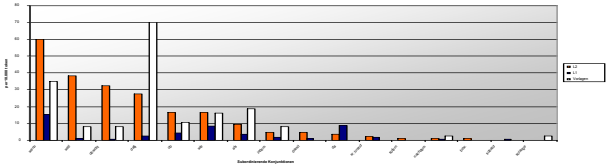
Bislang wurden die Fälle der so genannten Konnektrektion analysiert, bei der die Form eines der beiden Konnekte durch den Konnektor beeinflusst wird (Pasch et al. 2003: 351). Drei Fälle sind dabei zu unterscheiden:

| Subjunktoren | Postponierer | V2-Einbetter |
|---|---|---|
| **Weil** die Karawane weiterzieht, bellen die Hunde. | Die Karawane zieht weiter, **weshalb** die Hunde bellen. | **Angenommen** die Karawane zieht weiter, bellen die Hunde. |

TABELLE 3: Konnektrektion

**Overuse-Hypothese:** Konnektoren werden von fortgeschrittenen Lernern häufiger verwendet als von nativen Schreibern.

Diese Hypothese wird für die drei oben genannten Konnektorenklassen einzeln überprüft. Für die so genannten Postponierer (*weshalb*) und Verbzweitsatzeinbetter (*angenommen*) wurden in den drei Subkorpora keine signifikanten Ergebnisse nachgewiesen. Hierfür wurden die Lemmata einzeln ausgezählt. Für Subjunktoren konnte folgende Heuristik angewendet werden: Die zuvor automatisch getaggten subordierenden Konjunktionen werden mit einer CQP-Abfrage zusammengestellt und ausgezählt.



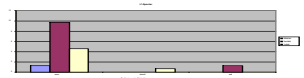
| | L2 | L1 | Vorlage |
|---|---|---|---|
| Anzahl standardisiert auf 10.000 Token | 221,1 | 52,9 | 180,7 |

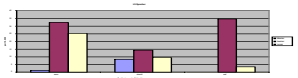
TABELLE 4: Subordinierende Konjunktionen

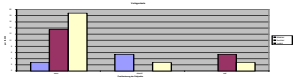
Die Ergebnisse - dargestellt in der Tabelle 4 - stützen die Overuse-Hypothese.
Die Korpusanalyse gibt zudem ein weitaus differenzierteres Bild: Die L2-Sprecher präferieren *wenn, weil* und *obwohl*. L1-Sprecher hingegen gebrauchen verstärkt die Subjunktion *da*. Die Fälle *dass* und *ob* müssen noch genauer untersucht werden, da sie nur vereinzelt als Konnektoren gebraucht werden.

**Nachfeld-Hypothese:** Subjunkte werden von Lernern präferiert im Nachfeld positioniert.

Zur Überprüfung der Hypothese wurden die drei häufigsten Subjunktoren *wenn, weil* und *obwohl* bezüglich ihrer Positionierung analysiert. Dabei wurde unterschieden, ob die Subjunkte im Vorfeld, im Mittelfeld oder im Nachfeld des Trägerkonnekts lokalisiert wurden.







Zusammenfassend lässt sich feststellen:
- *Weil*-Subjunkte erscheinen nicht im Mittelfeld
- L2-Schreiber präferieren die Nachfeldbesetzung (Stützung der Hypothese)
- L2-Schreiber nutzen die Positionierungsmöglichkeiten stärker aus
- Der Einfluss der Textvorlagen muss jedoch noch systematisch untersucht werden.


Maik Walter, Anke Lüdeling, Karin Schmidt, Emil Kroymann, Peter Adolphs | http://www.linguistik.hu-berlin.de/korpuslinguistik/projekte/falko/