Empirisch ermittelte Muster in Rechtschreibfehlern für die Automatisierung qualitativer Rechtschreibdiagnostik

Jan Thomas Röhrig | jan.roehrig@uni-due.de | Universität Duisburg-Essen

Ziel

- Ziel ist es, einen sprachdidaktischen Beitrag zur Entwicklung eines Programms zu leisten, das Rechtschreibfehler in freien Texten qualitativ analysieren kann.
- Hauptmotivation dafür ist der hohe Arbeitsaufwand, den qualitative Fehleranalysen erfordern[1].

Annahmen

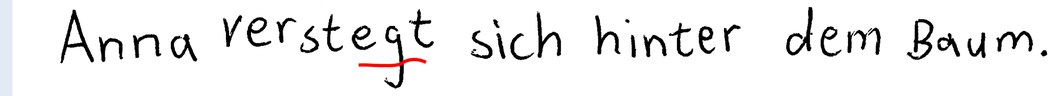
- Rechtschreibfehler in frei geschriebenen Texten von Kindern treten nicht willkürlich, sondern systematisch auf.
- Diese Systematizität lässt sich in den Rechtschreibfehlern in Form von prototypischen Mustern nachweisen, die für die automatische Erkennung und Analyse von Rechtschreibfehlern genutzt werden können.

Methode

- Grundlage für die Untersuchung ist das LitKey-Korpus der Ruhr-Universität-Bochum, das etwa 1900 Kindertexte aus Grundschulen in NRW enthält[2].
- Alle Rechtschreibfehler des Korpus werden entlang der Fehlerkategorien der AFRA[3] analysiert und kategorisiert.
- Die Annotationen werden auf Muster in den Fehlern hin verglichen.

- **Musterhaft sind Fehler, die**
  - **(A)** in mehreren Texten unterschiedlicher Kinder auftreten
  - **(B)** unter den gleichen Fehlergruppen der AFRA annotiert werden und
  - **(C)** signifikant häufiger auftreten als andere Fehler in der gleichen AFRA-Fehlerkategorie
- Ein Beispiel für ein solches Fehlermuster zeigt Abb.1. Das Fehlermuster <g> statt <ck> nach Vokalgraphem und vor Flexionsendung tritt signifikant häufiger als andere Fehler der entsprechenden Kategorien im LitKey-Korpus auf.

Abb. 1



Fehlerbeschreibung: *<g> statt <ck> nach kurzem Vokalgraphem und vor konsonantischer Flexionsendung

(A) kommt auch in anderen Texten im Korpus vor,

(B) wird immer in den gleichen Fehlerkategorien annotiert (konkret: KV_D+ und MS),

(C) und tritt signifikant häufiger als andere Fehler in KV_D+ und MS auf (siehe Abb. 2)

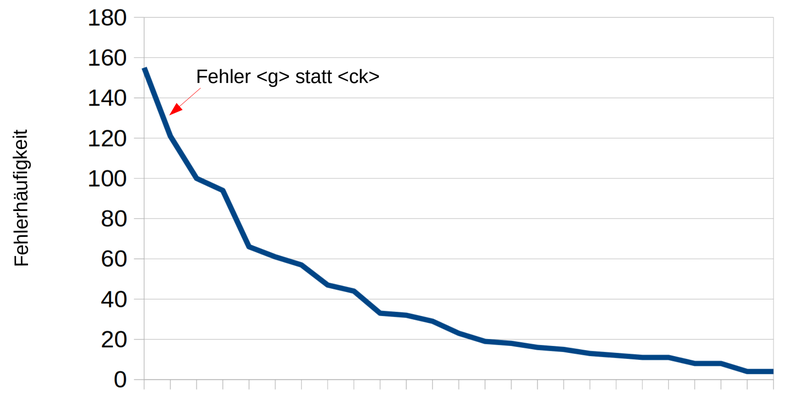


Abb. 2: Fehlerverteilung Kategorien KV_D+ und MS

Ablauf der automatischen Fehleranalyse

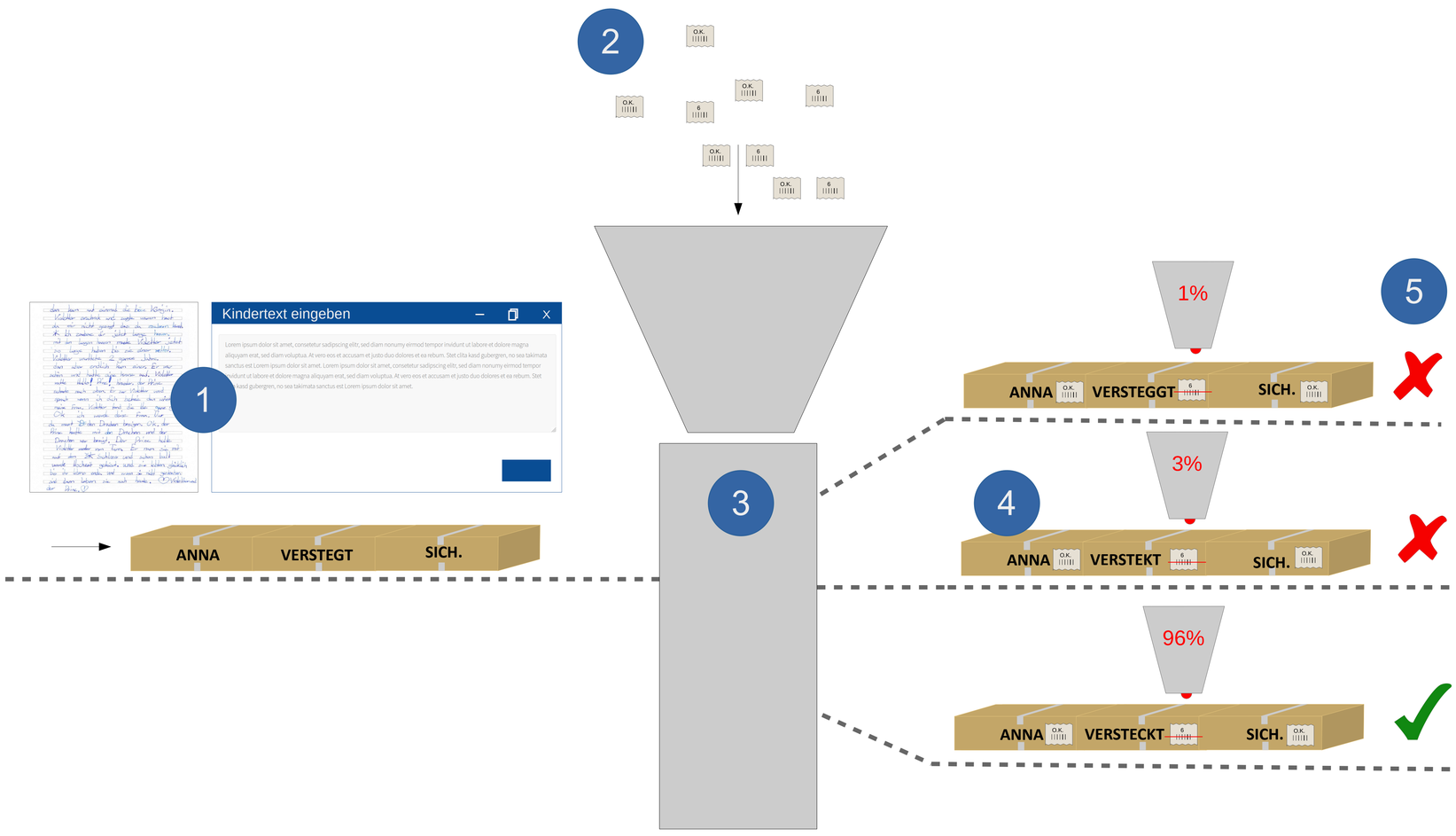


Abb. 3

Ergebnisse

- Die erwartete Systematizität in den Rechtschreibfehlern zeigt sich im Korpus.
- Die ermittelten Fehlermuster lassen sich in drei Gruppen unterscheiden:
  - (1) **morphemspezifisch**, d.h. Fehler treten wort- oder morphemspezifisch auf,
  - (2) **umgebungsspezifisch**, d.h. Fehler treten in einer bestimmten graphematischen Umgebung auf, und
  - (3) **unspezifisch**, d.h. Fehler treten unabhängig von einer bestimmten graphematisch- oder morphologischen Struktur auf.
- Die Fehlermuster machen es möglich, von einer fehlerhaften Schreibung durch „Suchen und Ersetzen“ automatisiert auf die orthografisch korrekte Version zu schließen (Abb. 3).

Ausblick

- Im Vordergrund steht die (weitere) technische Umsetzung des Verfahrens und die anschließende umfassende Evaluation, die dadurch erst möglich wird.
- Eine Erweiterung des Verfahrens um syntaktische Auswertungen ist wünschenswert, damit mehr Fehler in der automatischen Analyse erfasst werden (z.B. Groß- und Kleinschreibung).
- Die empirisch ermittelten Fehlermuster können auch für eine Weiter- und Neuentwicklung von qualitativen Fehleranalyseverfahren unabhängig von einer Automatisierung genutzt werden.

1 Der handgeschriebene Kindertext muss zuerst für die automatisierte Auswertung digitalisiert werden. Dafür wird er inkl. Rechtschreibfehler in die Programmoberfläche eingegeben.

2 Grundlage für die Fehlererkennung sind die empirisch ermittelten Fehlermuster. Jedes Fehlermuster besteht aus
- einer Fehlerkategorie für die Analyse,
- der Häufigkeit des Fehlermusters,
- den Graphemen der Originalschreibung inkl. graphematischer Umgebung und
- den Graphemen der intendierten korrekten Schreibung.

3 Der eingegebene Kindertext wird nach den definierten Fehlermustern durchsucht und Fehlerstellen werden durch die korrekten Schreibungen aus den definierten Fehlermustern ersetzt (Suchen und Ersetzen). Jede Ersetzung wird mit dem jeweiligen Fehlermuster markiert.

4 Durch das Suchen und Ersetzen der Fehlermuster ergeben sich mehrere mögliche Korrekturen. Jede Korrektur ist durch die Markierung mit denjenigen Fehlermustern verknüpft, die zur Korrektur geführt haben. Aus den verschiedenen Möglichkeiten wird die wahrscheinlichste Korrektur ausgewählt. Die Wahrscheinlichkeit ergibt sich aus der Häufigkeit des Fehlermusters und kann um Wahrscheinlichkeitswerte aus N-Gramm-Sprachmodellen ergänzt werden.

5 Die Rechtschreibfehler werden mit der jeweils wahrscheinlichsten Korrektur korrigiert. Alternativ kann der Nutzer selbst die passende Korrektur auswählen. Die Markierung der Fehlermuster in der korrigierten Version erlaubt Rückschlüsse auf die Fehlerkategorie und macht damit die automatische qualitative Fehleranalyse neben der Korrektur möglich.

Vorschau:

Kontakt:

Jan Thomas Röhrig

Universität Duisburg-Essen
Institut für Germanistik
Universitätsstr. 12
D-45141 Essen

Campus Essen
R11 T05 D01

Tel.: 0201 183-2869
Fax: 0201 183-2262
jan.roehrig(at)uni-due.de

Beiträge zur Automatisierung qualitativer Fehleranalyse (Auswahl):

Fay, Johanna; Berkling, Kay; Stüker, Sebastian (2012): Automatische Analyse von Rechtschreibfähigkeit auf Basis von Speech-Processing-Technologien. In: Didaktik Deutsch 19, S. 14–37.
Laarmann-Quante, Ronja (2017): Towards a Tool for Automatic Spelling Error Analysis and Feedback Generation for Freely Written German Texts Produced by Primary School Children. In: International Speech Communication Association (ISCA) (Hg.): 7th ISCA Workshop on Speech and Language Technology in Education, 25-26 August 2017. ISCA: ISCA, S. 36–41.
Thelen, Tobias (2010): Automatische Analyse orthographischer Leistungen von Schreibanfängern. Dissertation. Universität Osnabrück, Osnabrück.

Literatur:

[1] **Fay, Johanna; Berkling, Kay; Stüker, Sebastian (2012)**: Automatische Analyse von Rechtschreibfähigkeit auf Basis von Speech-Processing-Technologien. In: Didaktik Deutsch 19, S. 14–37.
[2] **Laarmann-Quante, Ronja; Ortmann, Katrin; Ehlert, Anna; Masloch, Simon; Scholz, Doreen; Belke, Eva; Dipper, Stefanie (2019):** The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. In: Behavior Research Methods.
[3] **Herné, Karl-Ludwig; Naumann, Carl Ludwig (2016)**: Aachener Förderdiagnostische Rechtschreibfehler-Analyse. 5., überarbeitete Auflage, Version 5. Aachen: Alfa Zentaurus (AFRA).