Zum Einfluss statistischer Worteigenschaften auf die Wortschreibung:
Eine korpusbasierte Untersuchung zum Orthographieerwerb in der Grundschule

Ronja Laarmann-Quante, Stefanie Dipper & Eva Belke
{laarmann-quante|dipper|belke}@linguistics.rub.de

Motivation

In vielen Schulen spielt die Anlauttabelle nach wie vor eine zentrale Rolle bei der Orthographievermittlung. Sie ist verbunden mit der Annahme, dass sich basierend auf einfachen Laut-Buchstaben-Zuordnungen orthographisches Wissen implizit, also beiläufig und ohne Intention, einstellt. Ob implizites Lernen jedoch stattfindet, wenn Kinder über einen längeren Zeitraum viele orthographisch inkorrekte Spontanschreibungen produzieren und (in älteren Altersgruppen) schon über erstes explizites Wissen verfügen, ist unseres Wissens bisher noch nicht erforscht.

Fragestellung & Hypothese

Wie beeinflussen verschiedene Worteigenschaften die Wahrscheinlichkeit eines Rechtschreibfehlers in Abhängigkeit von der Gesamtfehlerrate des Kindes?

Hypothese: Gute Rechtschreibleistungen erfordern ein solides Fundament an implizit erworbenem Wissen. Kinder, die insgesamt wenige Rechtschreibfehler produzieren, verfügen über solches Wissen: ihre Fehler zeigen einen stärkeren Zusammenhang mit den statistischen und orthograpischen Eigenschaften der Wörter auf als es bei Kindern mit insgesamt hoher Rechtschreibfehlerrate der Fall ist. Dies testen wir im Litkey-Projekt ("Literacy as the Key to Social Participation") empirisch auf dem umfassend annotierten Litkey-Korpus.

Litkey-Korpus

© Schroff (2000)

- **1922 Texte** produziert von **251 Kindern** in 15 Klassen (7 Schulen) in NRW
- longitudinales Design: **bis zu 10 Texte pro Kind** (∅ 7,7 ± 2,1) aus der 2. - 4. Klasse oder 3. - 4. Klasse, gesammelt zwischen 2010 und 2012 (Frieg, 2014)
- Beschreibung einer **Bildergeschichte**
- 189.394 Wörter (= Token mit mind. 1 alphabetischem Zeichen, 6.202 Wort-Typen)
- **Textlänge: ∅ 98,5 ± 43,6 Wörter**
- manuelle Transkription
- manuell erstellte **Zielhypothese**, die ausschließlich orthographische Fehler korrigiert
- **37.904 Rechtschreibfehler** insgesamt
- Fehlerrate pro Text: ∅ 21,7% ± 11,7

Original	Dodo	hate	zwar	angst	von	den	Gereusch	,	aber
Zielhypothese	Dodo	hatte	zwar	Angst	von	den	Geräusch	,	aber

(Semi-)automatische Annotationen

- **Eigenschaften des Zielworts**
  - Phoneme, Silben, Morpheme (*G2P;* Reichel & Kisler, 2014), POS (Stanford Tagger; Toutanova et al., 2003)
  - *Key Orthographic Features* (KOFs), z.B. Konsonantenverdopplung, Dehnungs-h, Auslautverhärtung, vokalisiertes <r>
  - Lexikalische Eigenschaften (aus *childLex*; Schroeder et al., 2015): z.B. Wortfrequenz, summierte Bigrammfrequenz, Nachbarschaftsmaße
- **Informationen über den Fehler**, z.B.
  - Fehlerkategorie aus dem Litkey-Schema mit 80 feinen Kategorien (basierend auf orthographischen Prinzipien nach Eisenberg (2006))
  - Rolle von Morphemkonstanz bei der korrekten Schreibung
  - Bleibt die Aussprache des Wortes trotz des Fehlers erhalten?
- **Metadaten des Kindes**
  - Alter, Geschlecht, Klassenstufe
  - Sprachlicher Hintergrund (z.B. Mehrsprachigkeit, geboren in Deutschland)

Herausforderungen bei der statistischen Analyse von Korpusdaten

- fehlende Datenpunkte
- unterschiedliche Textlängen
- unterschiedliche verwendete Wörter

Pilotstudien

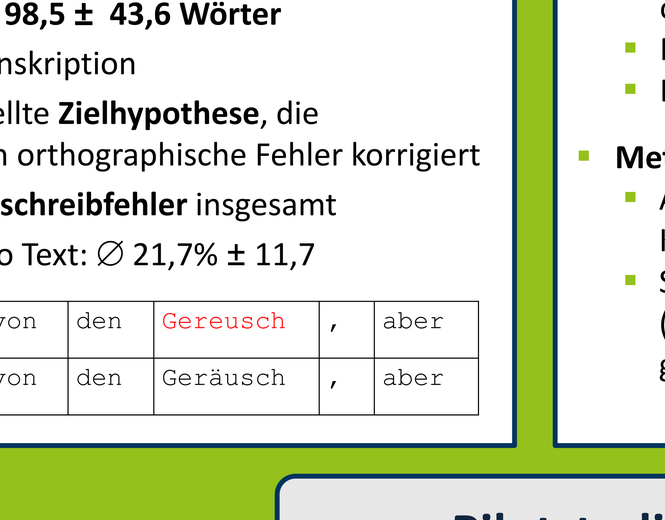
basierend auf rund 75.000 Token aus dem Litkey Korpus

Studie 1: Einfluss einer statistischen Worteigenschaft (mittlere summierte Bigrammfrequenz) auf die Wahrscheinlichkeit, dass ein Wort falsch geschrieben wird

Wahrscheinlichkeit für einen Fehler basierend auf der mittleren summierten Bigrammfrequenz des Zielwortes und der mittleren Fehlerrate des Kindes mit dem Zielwort als zufälligem Einflussfaktor

R: Generalized Linear Mixed Effects Model (binomial)
`ERROR ~ sqrt_bf_C * MEAN_CHILD_ER + (1+ MEAN_CHILD_ER|TARGET)`

Ergebnis:
Die mittlere summierte Bigrammfrequenz (BF) eines Wortes hat im Mittel keinen signifikanten Einfluss auf die Wahrscheinlichkeit, dass es falsch geschrieben wird. Doch je höher die mittlere Fehlerrate des Kindes, desto größer wird dieser Einfluss: Schwächere Schreiber schreiben eher Wörter mit hoher BF richtig als Wörter mit niedriger BF; für stärkere Schreiber ist dieser Zusammenhang weniger stark ausgeprägt
→ entspricht nicht der Hypothese




```
Fixed effects:
                          Estimate  StdErr  z-value
Intercept                -2.53507 0.05453 -46.49***
sqrt_bf_C                -0.04586 0.05532  -0.83
MEAN_CHILD_ER            10.27255 0.23445  43.82***
sqrt_bf_C:MEAN_CHILD_ER -0.55743 0.21491  -2.59**
```

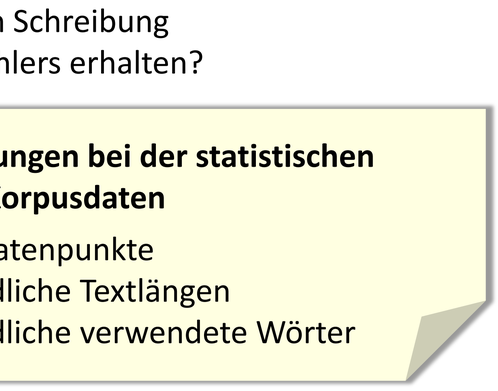
Studie 2: Einfluss des Vorhandenseins eines bestimmten orthographischen Phänomens (Konsonantenverdopplung) auf die Wahrscheinlichkeit, dass ein Wort falsch geschrieben wird

Wahrscheinlichkeit für einen Fehler basierend auf dem Vorhandensein einer Doppelkonsonantenschreibung im Zielwort (z.B. <bellen>) und der mittleren Fehlerrate des Kindes mit dem Zielwort als zufälligem Einflussfaktor

R: Generalized Linear Mixed Effects Model (binomial)
`ERROR ~ DCS * MEAN_CHILD_ER + (1+ MEAN_CHILD_ER|TARGET)`

Sample: 7806 Zielwörter (< 9 Buchstaben) mit Doppelkonsonanten und gleiche Anzahl zufällig gezogener Wörter ohne Doppelkonsonanten

Ergebnis: Wie erwartet steigt die Fehlerwahrscheinlichkeit mit der mittleren Fehlerrate des Kindes und ist höher für Wörter mit Doppelkonsonantenschreibung als für Wörter ohne Doppelkonsonantenschreibung. Der Effekt der Gesamtfehlerrate ist für Wörter mit Doppelkonsonantenschreibung signifikant größer als für Wörter ohne.




```
Fixed effects:
                  Estimate  StdErr   z-value
Intercept        -1.50442 0.07278 -20.670***
DCS               1.99539 0.11519  17.281***
MEAN_CHILD_ER     0.85772 0.04677  16.971***
DCS:MEAN_CHILD_ER 0.37374 0.06779   6.669***
```

Zwischenfazit: Die Pilotstudien legen nahe, dass isolierte Prädiktoren relevante Ergebnisse erzeugen; die abschließende Studie muss aber sowohl die Wirkung der statistischen als auch der orthographischen Eigenschaften eines Wortes und ihrer Interaktionen berücksichtigen.

Referenzen: Eisenberg, P. (2006). *Das Wort* (3rd ed.). *Grundriss der deutschen Grammatik: Band 1*. Stuttgart: J.B. Metzler. • Frieg, H. (2014). Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion (Dissertation). Ruhr-Universität Bochum. Abgerufen von http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/FriegHendrike/diss.pdf. • Reichel, U. D., & Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. In R. Hoffmann (Ed.), *Elektronische Sprachverarbeitung: Studientexte zur Sprachkommunikation 71* (pp. 42–49). TUDpress. • Schroff, C. (2000). Lea, Lars und Dodo: Bilderbox. SCHUBI Lernmedien. • Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A., & Kliegl, R. (2015). childLex: A lexical database of German read by children. *Behavior Research Methods*, *47*(4), 1085–1094. • Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In M. Hearst & M. Ostendorf (Eds.), *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics.

Mehr Informationen:
https://www.linguistics.rub.de/litkey/Scientific/Corpusanalysis