TeaP
60th Conference of
Experimental Psychologists

»Entschuldigen Sie, dass ich Ihnen einen komplizierten Artikel schreibe, für einen lesbaren habe ich keine Zeit«

Textanalyse mit den R-Paketen `koRpus` & `tm.plugin.koRpus`

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

http://koRpus.reaktanz.de

meik.michalke@hhu.de
Heinrich-Heine-Universität Düsseldorf
Diagnostik & Differentielle Psychologie

Lesbarkeit

Die Lesbarkeit von Texten ist im Informationszeitalter von zunehmender Bedeutung. Es genügt nicht, Wissen zu verschriftlichen; die resultierenden Texte müssen auch verständlich sein.
Zahlreiche Formeln wurden vorgeschlagen, um die Lesbarkeit eines Textes aus leicht zu erfassenden Merkmalen zu errechnen (Bamberger & Vanecek, 1984; DuBay, 2004; Klare, 1974). Dazu gehören die durchschnittliche Wort- und Satzlänge oder die Häufigkeit von Wörtern mit bestimmten Eigenschaften, z. B.:

$$\text{LIX (Läsbarhetsindex)} = \frac{\text{Worte}}{\text{Sätze}} + \frac{\text{Worte}_{\geqq 7\,\text{Zeichen}} \times 100}{\text{Worte}}$$

`koRpus` ist ein `R`-Paket für **Natural Language Processing (NLP)**, das diese Formeln einfach nutzbar macht.

Beispiel: Corpusanalyse zur Finanzkrise

Fragestellung: *»Berichten deutsche Medien über deutsche und griechische Positionen zur Finanzkrise Griechenlands vergleichbar verständlich?«*

```
library(koRpus.lang.de) # koRpus < 0.11: library(koRpus)
library(tm.plugin.koRpus)
# Lokaler Pfad zu TreeTagger und Textsprache werden definiert
set.kRp.env(
  TT.cmd="manual",
  TT.options=list(path="~/bin/treetagger", preset="de"),
  lang="de"
)
# Themen (Positionen der Finanzminister) haben eigene Verzeichnisse
themen <- c(
  Schaeuble="~/text/schaeuble", Varoufakis="~/text/varoufakis"
)
# Darin die Texte aus den Quellen jeweils in Unterverzeichnissen
medien <- c(
  Bild="bild", N24="n24",
  SpOn="spiegel_online", SZ="sueddeutsche"
)
# Mit den o.g. Informationen können nun Tokenizing und POS-Tagging
# für den ganzen Textcorpus vorgenommen werden
texts <- topicCorpus(paths=themen, sources=medien)
# Schließlich Berechnung der Lesbarkeit für die aufbereiteten Texte
texts <- readability(texts)
```

Die Lesbarkeit der Texte lässt sich nun vergleichen. Höhere LIX-Werte stehen für eine schwierigere Sprache:

```
# Laden der Pakete für ANOVA und Interaktionsplot
library(ez)
library(phia)
# ANOVA über die LIX-Werte
anova.results.LIX <- ezANOVA(
  data=corpusSummary(texts),
  dv=.(LIX),
  wid=.(doc_id),
  between=.(topic, source),
  observed=.(topic),
  type=3,
  detailed=TRUE,
  return_aov=TRUE
)
# Interaktionsplot des Modells
plot(
  interactionMeans(
    anova.results.LIX[["aov"]]
  )
)
```

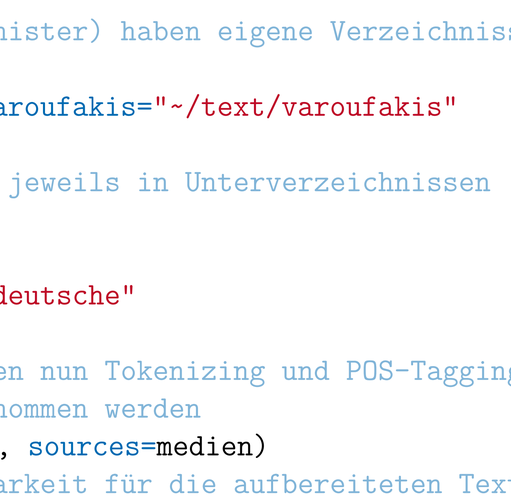

Abb. 1: Haupteffekte und Interaktionen mit Standardfehlern

Kernfeatures von `koRpus`

Part-of-Speech-Tagging

`koRpus` bietet mit `treetag()` einen Wrapper für `TreeTagger` (Schmid, 1994). So lassen sich Satzzeichen und Wortarten von Texten bestimmen und diese Information in `R` nutzen.

Silbenzählen

Der sprachunabhängige Silbentrennungsalgorithmus von LaTeX (Liang, 1983) wurde in `R` implementiert. Er ist inzwischen als `sylly` separat veröffentlicht.

Lesbarkeitsformeln (Bamberger & Vanecek, 1984; DuBay, 2004; Klare, 1974)

- ARI
  NRI, simple
- Bormuth
- Coleman
- Coleman-Liau
- Dale-Chall
  Powers-Sumner-Kearl, old
- Danielson-Bryan
- Strain
- Farr-Jenkins-Paterson
  Powers-Sumner-Kearl
- Degrees of Reading Power
- FORCAST
  precise RGL
- Flesch
  DE, ES, FR, NL, Szigriszt, Powers-Sumner-Kearl
- Flesch-Kincaid
- LIX
- RIX
- Harris-Jacobson
- Wheeler-Smith
  DE
- Linsear Write
- neue Wiener Sachtextformeln
- SMOG
  DE, C, simple
- FOG
  Powers-Sumner-Kearl, NRI
- Fucks
- Dickes-Steiwer
- Tränkle-Bailer
- TRI
- Tuldava
- Spache
  DE, old
- Easy Listening Formula

Lexikalische Diversität (Tweedie & Baayen, 1998)

- Type-Token Ratio (TTR)
- Herdan's C
- Summer's S
- Mean Segmental TTR (MSTTR)
- Guiraud's Root TTR
- Carroll's Corrected TTR
- Dugast's Uber Index (U)
- Moving Average TTR (MATTR)
- Yule's K
- Maas
- HD-D
- MTLD
- MTLD-MA

Unterstützte Sprachen

Sprachsupport ist modular über Zusatzpakete (`koRpus.lang.**`) wählbar:

- de: Deutsch
- en: Englisch
- fr: Französisch
- it: Italienisch
- nl: Niederländisch
- pt: Portugiesisch
- ru: Russisch
- es: Spanisch

Interdisziplinäre Anwendungsbereiche

`koRpus` wurde bisher u. a. in den Forschungsgebieten Kognition (Bannert, Sonnenberg, Mengelkamp & Pieger, 2015), Kommunikation (Shulman & Sweitzer, 2018), Medizin (Brulet, Llorca & Letrilliart, 2015), Politik (Correa & Camargo, 2017), Informatik (Blohm, Riedl, Füller & Leimeister, 2016), Psychoakkustik (Lindborg, 2015), Wirtschaft (Taborda, 2015), Reaktorsicherheit (Kovesdi & Joe, 2017), Bildung (Zimmermann, 2016) & Linguistik (Klaussner & Vogel, 2015) eingesetzt.

Literatur

Bamberger, R. & Vanecek, E. (1984). *Lesen, verstehen, lernen, schreiben*. Jugend u. Volk.

Bannert, M., Sonnenberg, C., Mengelkamp, C. & Pieger, E. (2015). Short- and long-term effects of students' self-directed metacognitive prompts on navigation behavior and learning performance. *Computers in Human Behavior*, *52*, 293–306. doi: 10.1016/j.chb.2015.05.038

Blohm, I., Riedl, C., Füller, J. & Leimeister, J. M. (2016). Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing. *Information Systems Research*, *27* (1), 27–48. doi: 10.1287/isre.2015.0605

Brulet, A., Llorca, G. & Letrilliart, L. (2015). Medical wikis dedicated to clinical practice: a systematic review. *Journal of medical Internet research*, *17* (2). doi: 10.2196/jmir.3574

Correa, J. C. & Camargo, J. E. (2017). Ideological Consumerism in Colombian Elections, 2015: Links Between Political Ideology, Twitter Activity, and Electoral Results. *Cyberpsychology, Behavior, and Social Networking*, *20* (1), 37–43. doi: 10.1089/cyber.2016.0402

DuBay, W. H. (2004). *The principles of readability*. Impact Information.

Klare, G. R. (1974). Assessing Readability. *Reading Research Quarterly*, *10* (1), 62–102. Zugriff auf `http://www.jstor.org/stable/747086` doi: 10.2307/747086

Klaussner, C. & Vogel, C. (2015). Stylochronometry: Timeline Prediction in Stylometric Analysis. In *Research and Development in Intelligent Systems XXXII* (S. 91–106). Springer.

Kovesdi, C. & Joe, J. (2017, Juni). A novel tool for improving the data collection process during control room modernization human-system interface testing and evaluation activities. In *Proceedings of the 10th International Conference on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies (NPIC & HMIT 2017)* (S. 1261–1271). San Francisco, California: ANS.

Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-put-er*. Department of Computer Science, Stanford University.

Lindborg, P. (2015). Psychoacoustic, physical, and perceptual features of restaurants: A field survey in Singapore. *Applied Acoustics*, *92*, 47–60. doi: 10.1016/j.apacoust.2015.01.002

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Shulman, H. C. & Sweitzer, M. D. (2018). Advancing Framing Theory: Designing an Equivalency Frame to Improve Political Information Processing. *Human Communication Research*. doi: 10.1093/hcr/hqx006

Taborda, R. (2015). Procedural transparency in Latin American central banks under inflation targeting schemes. A text analysis of the minutes of the Boards of Directors. *Ensayos sobre Política Económica*, *33* (76), 76–92. doi: 10.1016/j.espe.2015.01.002

Tweedie, F. J. & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32* (5), 323–352.

Zimmermann, S. (2016). Entwicklung einer computerbasierten Schwierigkeitsprädiktion von Leseverstehensaufgaben. *NEPS Working Paper*, *64*. doi: 10.13140/RG.2.1.2462.6962

Für die Aufarbeitung des Beispielcorpus Dank an | Rabea Viermann