GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Sommersemester 2016
Forschungsorientiertes Lehren und Lernen (FoLL)

GÖTTINGEN CAMPUS PLUS

Du bist, wie du schreibst?

CAPS: Computergestützte Analyse von persönlichem Stil

Das Selbst als terminologisches Konstrukt

Theoretischer Ausgangspunkt der Untersuchungen ist das sprachanalytische Tool LIWC von James Pennebaker. Die Idee dabei ist, dass ein direkter Zusammenhang zwischen den Worten, die ein Autor wählt, und seiner/ihrer Persönlichkeit besteht. Ausgehend von der theoretischen Grundlage Pennebakers nehmen wir an, dass das Selbst ein terminologisches Konstrukt ist, das als Oberbegriff die psychologischen Konstrukte Aufmerksamkeit, Emotionalität, soziale Beziehungen, Denkstile und individuelle Unterschiede (Persönlichkeit) synthetisiert. Das führt zu der Annahme, dass natürliche Sprache die Qualität und Quantität dieser Konstrukte abbilden kann (Abbildungsbeziehung zwischen Sprache und Schemata des Denkens/Fühlens/Sozialer Welt). Insofern erlaubt unsere Definition das Generieren von Forschungshypothesen bezüglich des Zusammenhangs der obenstehenden strukturierenden, ausführenden und emotionalen Funktionen. Das Selbst in diesem Sinne ist ein Sammelbegriff für die unterschiedlichen Funktionen.

SELBST
+ **Aufmerksamkeit**
+ **Emotionalität**
+ **soziale Beziehungen**
+ **Denkstile**
+ **individuelle Unterschiede**

DENKEN
FÜHLEN
SOZIALE WELT
→ QUALITÄT / QUANTITÄT → **SPRACHE**

TEXT
SATZ
WORT
→ COMPUTERANALYSE →

POS-TAGS	PARSING	DICTIONARY
PPER	SUBJ	NEGEMO
NN	OBJ	COGMECH
VVFIN	ADV	INCL/EXCL

Workshops

Nach der Festlegung einer Arbeitsdefinition für unseren Selbst-Begriff gingen wir unser Vorhaben mit einer Vielzahl von Workshops an, zu denen wir verschiedene Experten einluden.

- **Prof. Dr. Stefan Evert** (Friedrich-Alexander-Universität Erlangen-Nürnberg): Statistik-Workshop zur Auswertung quantitativer Analysen.
- **Prof. Dr. Caroline Sporleder** (Georg-August-Universität Göttingen): Multiword Expressions (MWEs) und Idiome in den Digitalen Humanwissenschaften und korpusbasiertes maschinelles Lernen.

Statistik-Workshop zur Auswertung quantitativer Analysen mit Prof. Dr. Stefan Evert im Lichtenberg-Kolleg.

- **Dr. des. Matt Munson** (Universität Leipzig): Programmiersprache Python, insbesondere Natural Language Tool Kit (NLTK).
- **Dr. Markus Wolf** (Ruprecht-Karls-Universität Heidelberg): Workshop zum Tool LIWC.
- **Dr. Zsofia Demjen** (Open University England): Tagebuch-Forschung Selbst, zusätzliche Aspekte der Selbstreferenz, Mixed Method Approach.
- **Markus Paluch, B. A.** (Georg-August-Universität Göttingen): Einführung in statistische Methoden.

CAPS – Das gesamte Team.

The Zen of Python

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Flat is better than nested.
- *[...]*

https://www.python.org/dev/peps/pep-0020/

LIWC

- steht für „Linguistic Inquiry and Word Count“
- basiert auf sozialpsychologischer Forschung (Pennebaker 1993)
- digitales Tool zur Erfassung von Wortdistribution
- ordnet Wörter soziolinguistischen Kategorien zu (insg. 68)

Korpuserstellung

Unser Forschungsinteresse gilt selbstbezüglichen Ausdrucksweisen und ihren Zusammenhängen zur Persönlichkeit. Wir wählten deshalb ein Korpus, das die Genres „Tagebücher“ und „Briefe“ umfasst. Es stand dabei die Annahme im Vordergrund, dass hier der Schreibstil der Autoren und Autorinnen – mit gewissen Einschränkungen – authentisch ist und nicht durch eine figurenorientierte Sprechweise vermittelt wird.
Zudem sind unsere Untersuchungsmethoden so gewählt, dass eine literarisch konstruierte und eine Alltagssprache gleichermaßen als Grundlage geeignet sind.
Unser Korpus umfasst ca. 4.6 Millionen Wörter, verteilt auf 38 Autoren und 13 Autorinnen. Um möglichst wenige Sprachwandelprozesse einzufangen, liegt der Hauptveröffentlichungszeitraum zwischen 1850 – 1950.
Nicht alle Verfasser unserer Korpustexte sind Schriftsteller. Für die Erstellung des Korpus haben wir verschiedene Quellen genutzt, die jeweils ihre eigenen Hürden hatten. Die **Akquise bei Verlagen** hat vor allem rechtliche Fragestellungen hervorgebracht, **Ebooks** kosten zum Teil Geld, Inhalte von **Webseiten** zu extrahieren ist technisch anspruchsvoll und mit **OCR** bearbeitete Texte sind nicht immer fehlerfrei.
Für die computergestützte Analyse mussten die Texte von allen autorenfremden Inhalten befreit werden (editorische und verlegerische Notizen, Fußnoten, gliedernde Elemente). Bestimmte technische Grundlagen wie Dateiformate, Textcodierung und die Benennung der einzelnen Files mussten vereinheitlicht werden.
Der technische Prozess hat einen enormen zeitlichen – in diesem Maße nicht erwarteten – Aufwand bedeutet.

LIWC + python™

Analysetools

Für die Analyse verwendeten wir hauptsächlich zwei Programme: ein diktionärbasiertes Tool zum Wörterzählen und Einordnen in Kategorien namens LIWC sowie die Programmiersprache Python. LIWC ordnet Wörter eines Textes bestimmten Kategorien zu, welche zuvor von Psychologen als aussagekräftig für die jeweiligen Kategorien eingestuft wurden.
Python als Programmiersprache bietet hier deutlich umfassendere und personalisierbarere Möglichkeiten als LIWC. Hierfür teilten wir uns in zwei Arbeitsgruppen auf, eine LIWC- und eine Python-Gruppe. Natürlich kooperierten wir auch innerhalb dieser Gruppen untereinander, so z. B. nicht nur beim Einrichten und Instandhalten der Technik, sondern auch bei der Erstellung individueller packages des Korpus für die weiterführende Analyse und der Visualisierung der Ergebnisse.

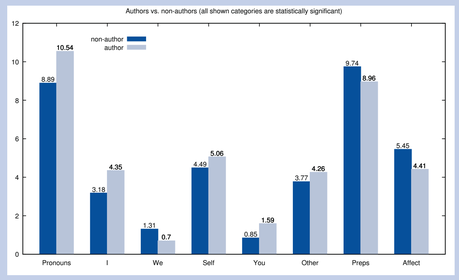
Kommunikationskanäle

- **Wiki** von Stud.IP als Kommunikationsplattform und Festhalten des Fortschrittes
- **Wiki** der Git-Plattform zum Datenaustausch (Korpus sowie Analysefortschritt und -ergebnisse)
- **Gitlab**-Plattform der GDWG (gitlab.gwdg.de) zum kollaborativen Arbeiten
- Blog caps.**wordpress**.com und **Twitter** als öffentliche Kommunikationskanäle

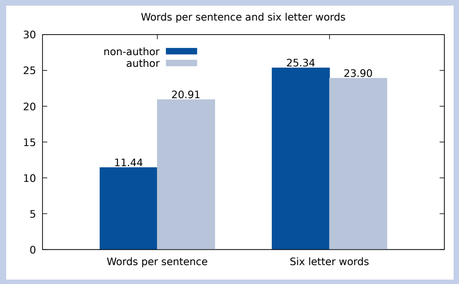
KORPUS

- **Kernzeitraum: 1850 bis 1950**
- **51 Autoren gesamt**
- **38 männlich, 13 weiblich**
- **Genre Briefe:**
  - **~ 160.000 Sätze**
  - **~ 2.6 Millionen Wörter**
- **Genre Tagebücher:**
  - **~ 162.000 Sätze**
  - **~ 2.0 Millionen Wörter**

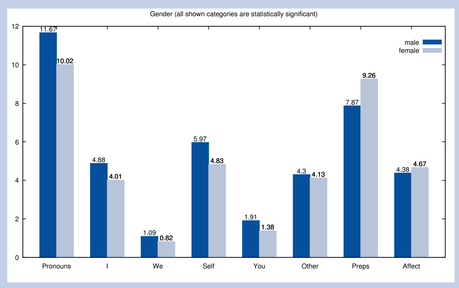
Resultate



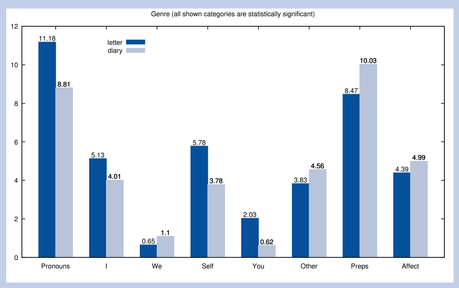
Kategorienverteilung für Schriftsteller/innen und Nicht-Schriftsteller/innen.



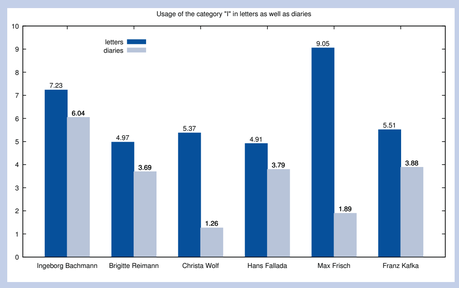
„Six letter words“: Wörter mit sechs oder mehr Buchstaben.



Kategorienverteilung nach Geschlecht.



Kategorienverteilung nach Genre.



Verwendung der Kategorie „I“ (siehe rechts) verteilt über beide Genres nach Autoren.

„Word count strategies are based on the assumption that the words people use convey psychological information over and above their literal meaning and independent of their semantic context.“

(Pennebaker et al.: Psychological Aspects of Natural Language Use. 2003, 550)

Wortdistributionen nach LIWC

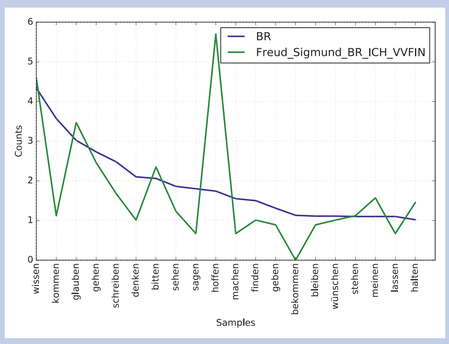
Die Balkendiagramme zeigen für die Untersuchungen nach Gender, Genre sowie Schriftstellertum die unterschiedlichen Häufigkeiten im schriftlichen Gebrauch:

- aller Pronomen
- der Pronomen ich, mir, mich usw. **(„I“)**
- der Pronomen wir, uns, unsere usw. **(„we“)**
- aller selbstreferentiellen Pronomen „I“ + „we“ **(„Self“)**
- des Pronomens du, dich, dir usw. **(„You“)**
- anderer, noch nicht genannter Pronomen (ihr, sie, Namen usw.) **(„Other“)**
- von Präpositionen **(„Preps“)**
- von Emotionswörtern (positiv + negativ) **(„Affect“)**

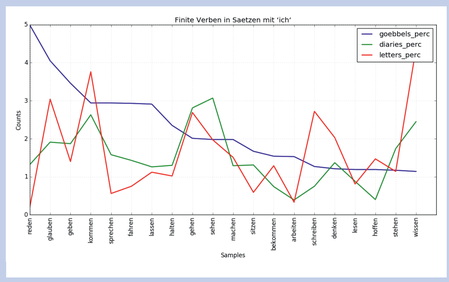
Alle Werte sind in Prozent angegeben.
Für den Vergleich Schriftsteller/ nicht-Schriftsteller wurden außerdem die Kategorien „Anzahl der Wörter pro Satz“ sowie die „Anzahl an Wörtern mit sechs oder mehr Buchstaben“ gegenübergestellt. Die Werte entsprechen den Durchschnittswerten der absoluten Zahlen.

Verbverwendung in Sätzen mit „Ich“ als Subjekt

Diese Tabellen stellen exemplarisch für einige Autoren und Autorinnen die prozentuale Verteilung der häufigsten finiten Verben in Sätzen mit „Ich“ als Subjekt dar (Verben sind der Übersichtlichkeit halber infinit angegeben).
Auffällig ist, dass manche Autoren wie z. B. Franz Kafka sogar in beiden Genres ähnliche Kurven wie die Referenzkorpora zeigen und andere stark abweichen wie z. B. Brigitte Reimann, die bestimmte Verben in Ich-Sätzen nicht benutzt. Hier bieten sich weitere Untersuchungen an.

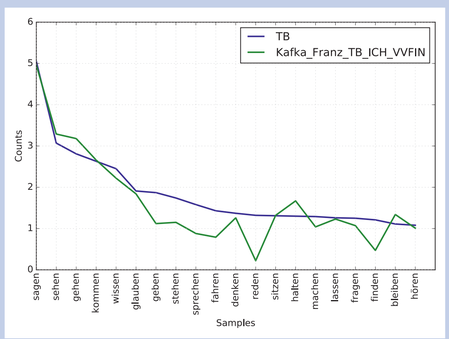


Autor: Sigmund Freud; Genre: Brief.

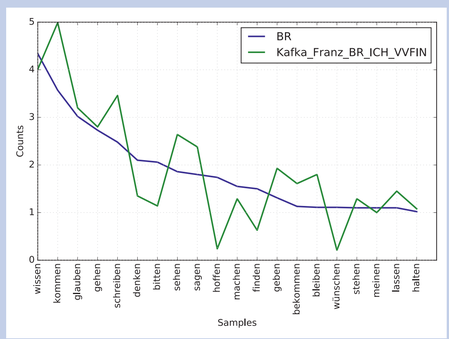


Autor: Joseph Goebbels; Genre: Tagebuch.

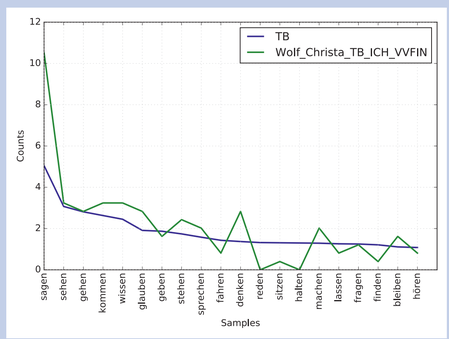
Hier zeigt sich, dass die Verteilung der häufigsten Verben für beide Referenzkorpora relativ ähnlich verläuft. Ausnahme: „Wissen“ und „schreiben“ als Ausreißer bei den Briefen.
Auffällig im Verbgebrauch von Joseph Goebbels sind die hohen Werte für „reden“, „glauben“ und „fahren“.



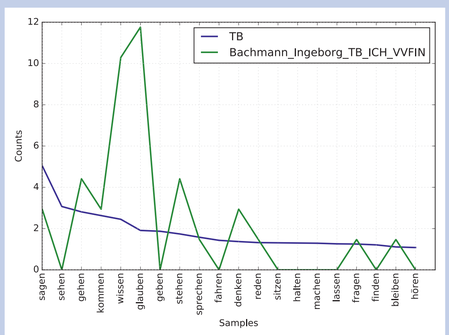
Autor: Franz Kafka; Genre: Tagebuch.



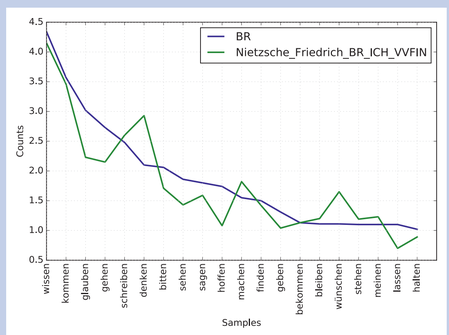
Autor: Franz Kafka; Genre: Brief.



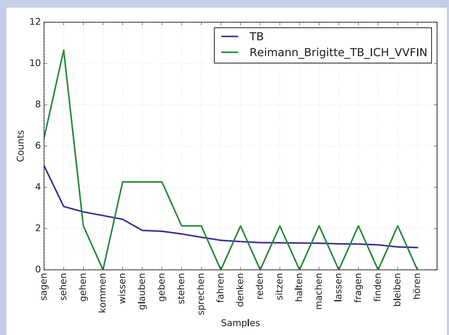
Autorin: Christa Wolf; Genre: Tagebuch.



Autorin: Ingeborg Bachmann; Genre: Tagebuch.



Autor: Friedrich Nietzsche; Genre: Brief.



Autorin: Brigitte Reimann; Genre: Tagebuch.

Ansprechpartnerin für FoLL:
Susanne Wimmelmann, Hochschuldidaktik
Susanne.wimmelmann@zvw.uni-goettingen.de
www.uni-goettingen.de/forschendeslernen

Gemeinsames Bund-Länder-Programm für bessere Studien-bedingungen und mehr Qualität in der Lehre

Dieses Vorhaben wird aus Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01PL11061 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Betreuende: Berenike Herrmann, Gerhard Lauer.
Studierende: Imme Bageritz, Lisa Hannemann, Phillipp Heimann, Julia Kleinschmidt, Cosima Mattner, Georg Roth-Kleyer, Ronald Weller.
Wir danken Markus Paluch für seine signifikante Unterstützung!

Niedersächsisches Ministerium für Wissenschaft und Kultur