TeaP 2012
54. Tagung experimentell arbeitender Psychologen

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

koRpus

ein R-paket zur textanalyse

dipl. psych. m.eik michalke

heinrich heine universität düsseldorf
institut für experimentelle psychologie
abt. für diagnostik & differentielle psychologie

TeaP 2012, universität mannheim
4. april 2012

and now for something

completely different

koRpus

ein R-paket zur textanalyse

wozu brauchen wir textanalyse?

anwendungsgebiete

- text als stimulusmaterial
  - wortlisten
  - altersangemessenheit
  - vergleichbarkeit
- text als AV
  - sprachanalyse bei affektiven störungen
  - sprachentwicklung
  - intelligenz, kreativität
- lesbarkeit von manualen
- ...

R in a nutshell

R ist...

- eine freie (GPL) implementierung von S

  Becker und Chambers (1984); Becker, Chambers und Wilks (1988); Chambers und Hastie (1991); Chambers (1998); R Development Core Team (2012)

- eine objektorientierte programmiersprache
- verfügbar für GNU/linux, MacOS X & Windows
- **erweiterbar über pakete**

R in a nutshell

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

typischer funktionsaufruf in R

```
ergebnis <- funktion(option1="wert", option2=TRUE, ...)
```

http://koRpus.reaktanz.de

entwicklung

- 0.01-1: veröffentlicht am 24.04.2011
  - 49 revisionen
- 0.04-29: veröffentlicht heute abend ;o)
- stable: `http://cran.r-project.org`

was kann koRpus?

beispieltext ~~(Guttenberg, 2009)~~ (Schieren, 2002)

[...] Der EuGH zog das Misstrauen der Mitgliedstaaten vor allem deswegen auf sich, weil er nicht als Hüter der nach dem Prinzip der begrenzten Einzelermächtigung konstruierten Kompetenzordnung der Gemeinschaft erschien, nicht als "neutraler Richter", sondern als das "Integrationsorgan der Europäischen Union". Die Zusicherung des Gerichtshofs, er sei sein eigener Wächter, fand in der Rechtsprechung keine Bestätigung. Selbst J. H. H. Weiler, beileibe kein Kritiker der europäischen Integration (wenngleich auch selten das Florett diplomatischer Differenzierung führend), merkte kritisch an: "Der Gerichtshof nimmt seine Rolle als Schutzmann in Europa nicht wahr. Er sagt nicht nein zur Union, wenn sie ihre Kompetenzen überschreitet." Nicht zuletzt durch diese Kritik in seiner Selbstgewissheit erschüttert, urteilte der EuGH am 5. Oktober 2000 erstmals, dass die Gemeinschaft jenseits ihrer Ermächtigung agiert habe. *[...]*

tokenizing & part-of-speech tagging

tokenizing & POS-tagging

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

- eigener tokenizer
- wrapper für TreeTagger (Schmid, 1994)

  http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

treetag()

```
> taggedText <- treetag("./guttenberg.txt")
> summary(taggedText)

  Sentences: 13
  Words:     217 (16.69 per sentence)
  Letters:   1389 (6.4 per word)

  Word class distribution:

             num       pct
noun          55 25.345622
article       33 15.207373
verb          24 11.059908
pronoun       23 10.599078
adjective     22 10.138249
preposition   19  8.755760
conjunction   14  6.451613
adverb         9  4.147465
particle       9  4.147465
name           7  3.225806
number         2  0.921659
comma         16        NA
fullstop      13        NA
punctuation   12        NA
```

tokenizing & POS-tagging

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

mehr details?

tokenizing & POS-tagging

treetag()

```
> slot(taggedText, "TT.res")
```

	token	tag	lemma	lttr	wclass	desc
1	Der	ART	d	3	article	Artikel
2	EuGH	NE	EuGH	4	name	Eigenname
3	zog	VVFIN	ziehen	3	verb	finites Verb, voll
4	das	ART	d	3	article	Artikel
5	Misstrauen	NN	Mißtrauen	10	noun	Nomen
[...]						
32	nicht	PTKNEG	nicht	5	particle	Negationspartikel
33	als	KOKOM	als	3	conjunction	Vergleichspartikel ohne Satz
34	"	$(	"	1	punctuation	satzinterne Interpunktion
35	neutraler	ADJA	neutral	9	adjective	attributives Adjektiv
36	Richter	NN	Richter	7	noun	Nomen
37	"	$(	"	1	punctuation	satzinterne Interpunktion
38	,	$,	,	1	comma	Komma
39	sondern	KON	sondern	7	conjunction	nebenordnende Konjunktion
40	als	KOKOM	als	3	conjunction	Vergleichspartikel ohne Satz
41	das	ART	d	3	article	Artikel
42	"	$(	"	1	punctuation	satzinterne Interpunktion
[...]						

wortschatzanalyse:

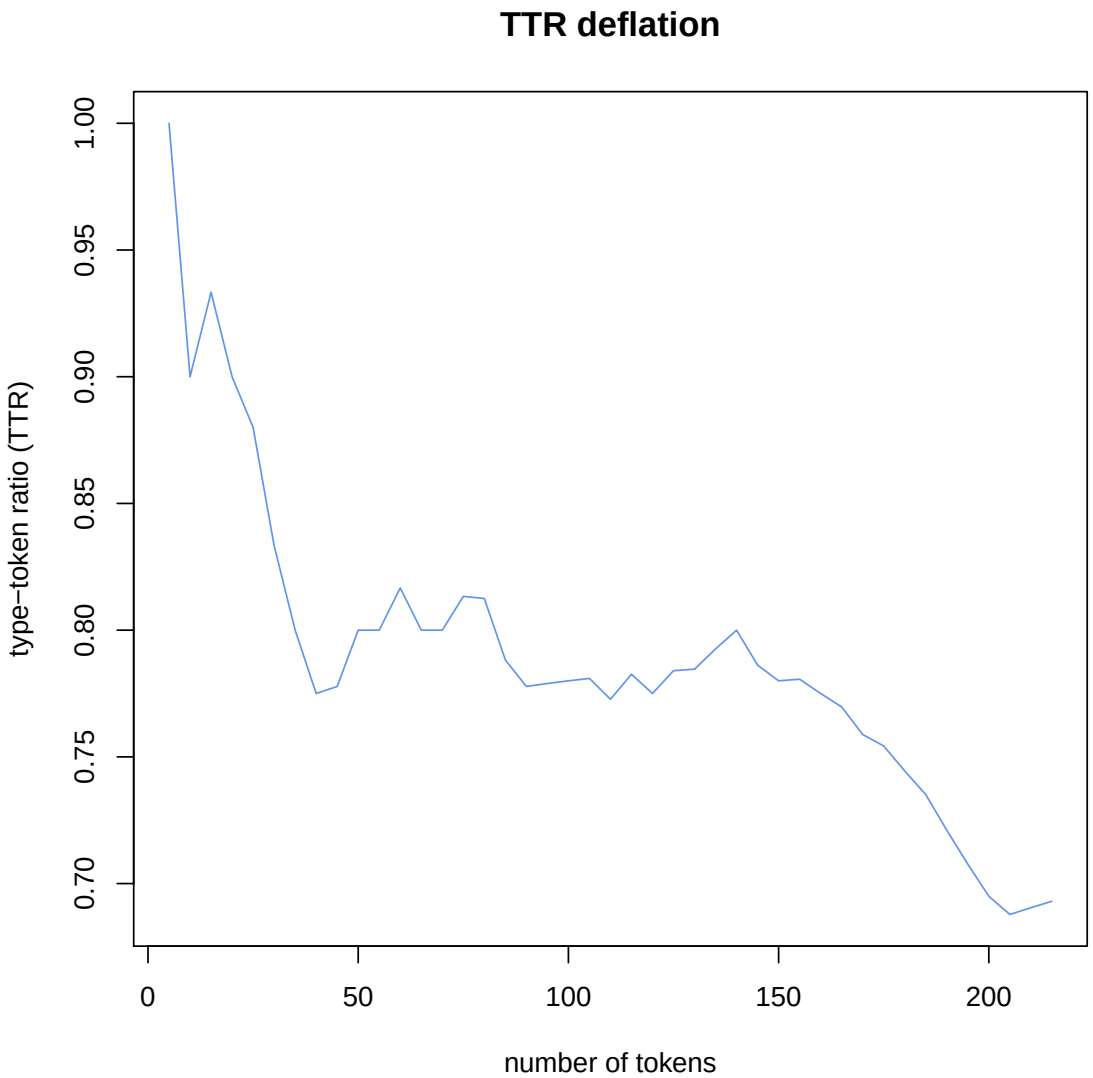
lexical diversity

lexical diversity

[...] Der EuGH zog das Misstrauen der Mitgliedstaaten vor allem deswegen auf sich, weil er nicht als Hüter der nach dem Prinzip der begrenzten Einzelermächtigung konstruierten Kompetenzordnung der Gemeinschaft erschien *[...]*

- types: **25**
  unterschiedliche worte
- tokens: **29**
  alle worte

lexical diversity



lexical diversity: in koRpus implementiert

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

vgl. Tweedie und Baayen (1998); Covington und McFall (2010); McCarthy und Jarvis (2010)

lexical diversity

lex.div()

```
> (taggedTextLd <- lex.div(taggedText))

Total number of tokens: 217
Total number of types:  149

Type-Token Ratio
   TTR: 0.69


Mean Segmental Type-Token Ratio
               MSTTR: 0.79
          SD of TTRs: 0.01
        Segment size: 100
      Tokens dropped: 17

Hint: A segment size of 108 would reduce the drop rate to 1.
      Maybe try ?segment.optimizer()


Moving-Average Type-Token Ratio
               MATTR: 0.83
          SD of TTRs: 0.04
         Window size: 100
[...]
```

lexical diversity

weniger details?

lexical diversity: summary

lex.div()

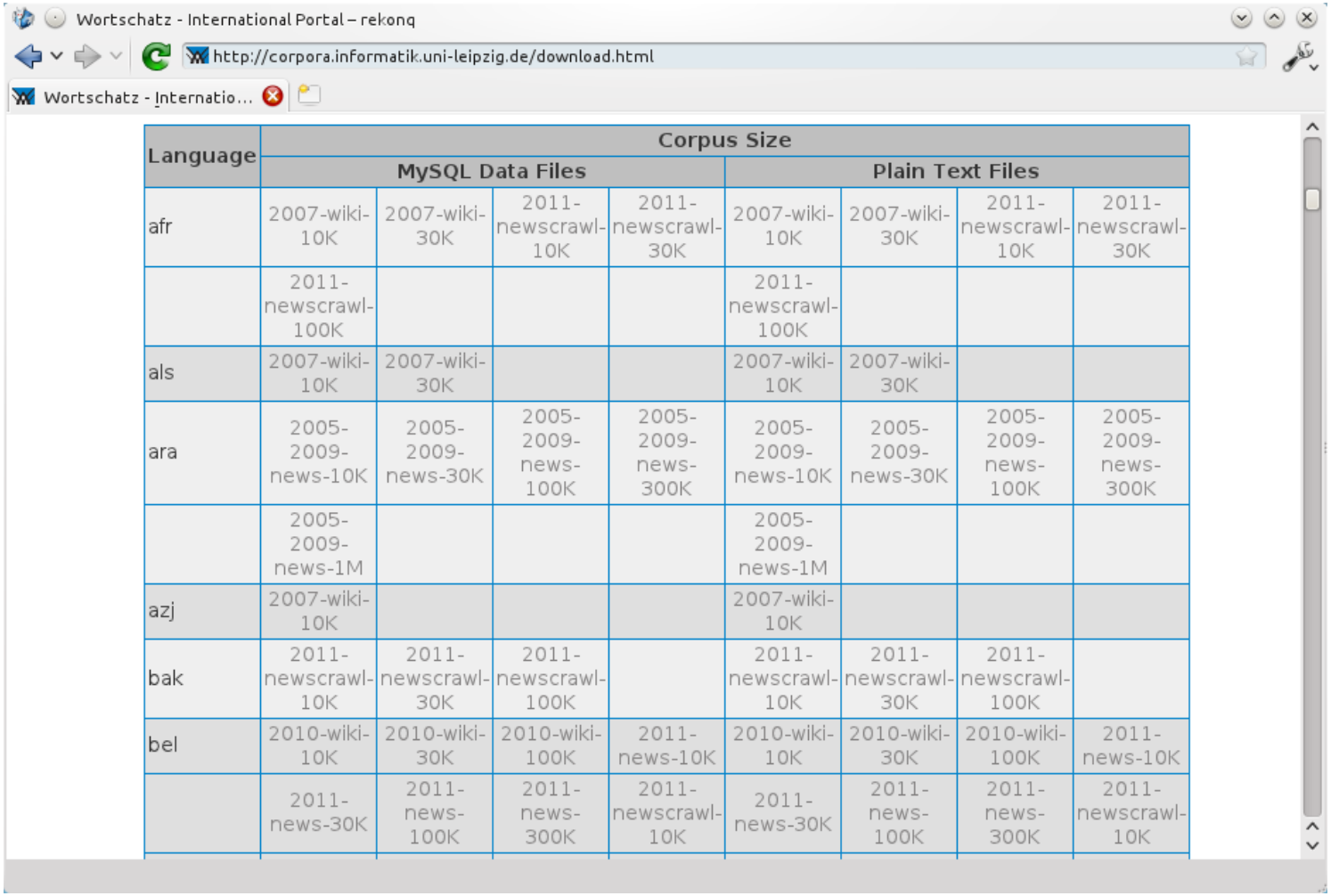
```
> summary(taggedTextLd)

            index  value
1             TTR   0.69
2           MSTTR   0.79
3           MATTR   0.83
4      Herdan's C   0.93
5        Root TTR  10.11
6            CTTR   7.15
7      Uber index  33.43
8          Summer   0.91
9        Yule's K  87.92
10         Maas a   0.17
11      Maas lgV0   5.92
12 HD-D (vocd-D)   36.8
13           MTLD 153.67
```

allgemeiner sprachgebrauch:

frequency analysis

frequency analysis



Wortschatz - International Portal – rekonq

http://corpora.informatik.uni-leipzig.de/download.html

Wortschatz - Internatio...

Language	Corpus Size							
	MySQL Data Files				Plain Text Files			
afr	2007-wiki-10K	2007-wiki-30K	2011-newscrawl-10K	2011-newscrawl-30K	2007-wiki-10K	2007-wiki-30K	2011-newscrawl-10K	2011-newscrawl-30K
	2011-newscrawl-100K				2011-newscrawl-100K			
als	2007-wiki-10K	2007-wiki-30K			2007-wiki-10K	2007-wiki-30K		
ara	2005-2009-news-10K	2005-2009-news-30K	2005-2009-news-100K	2005-2009-news-300K	2005-2009-news-10K	2005-2009-news-30K	2005-2009-news-100K	2005-2009-news-300K
	2005-2009-news-1M				2005-2009-news-1M			
azj	2007-wiki-10K				2007-wiki-10K			
bak	2011-newscrawl-10K	2011-newscrawl-30K	2011-newscrawl-100K		2011-newscrawl-10K	2011-newscrawl-30K	2011-newscrawl-100K	
bel	2010-wiki-10K	2010-wiki-30K	2010-wiki-100K	2011-news-10K	2010-wiki-10K	2010-wiki-30K	2010-wiki-100K	2011-news-10K
	2011-news-30K	2011-news-100K	2011-news-300K	2011-newscrawl-10K	2011-news-30K	2011-news-100K	2011-news-300K	2011-newscrawl-10K

read.corp.LCC() & query()

```
> LCC.de <- read.corp.LCC("deu_newscrawl_2011_1M-text.tar.gz")
Fetching needed files from LCC archive... done.

> query(LCC.de, var="pmio", query=c(700, 750))
```

	num	word	freq	pct	pmio	log10	rank.avg	rank.min	rank.rel.avg	rank.rel.min
184	202	hier	14200	0.0007346965	734	2.865696	682880	682880	99.98309	99.98316
185	203	mich	13889	0.0007186056	718	2.856124	682879	682879	99.98294	99.98302
186	204	soll	13842	0.0007161738	716	2.854913	682878	682878	99.98280	99.98287
187	205	zwei	13820	0.0007150356	715	2.854306	682877	682877	99.98265	99.98272
188	206	ja	13554	0.0007012729	701	2.845718	682876	682876	99.98250	99.98258
189	207	mir	13549	0.0007010142	701	2.845718	682875	682875	99.98236	99.98243

freq.analysis()

```
> taggedTextFreq <- freq.analysis(taggedText, corp.freq=LCC.de)
> slot(taggedTextFreq, "TT.res")
```

	token	tag	lemma	lttr	wclass	[...]	pmio	rank.avg	rank.min
1	Der	ART	d	3	article		3839	99.99582	99.99582
2	EuGH	NE	EuGH	4	name		3	97.72070	97.70378
3	zog	VVFIN	ziehen	3	verb		77	99.85767	99.85761
4	das	ART	d	3	article		8430	99.99840	99.99840
5	Misstrauen	NN	Mißtrauen	10	noun		7	98.63301	98.62889
[...]									
32	nicht	PTKNEG	nicht	5	particle		6993	99.99766	99.99766
33	als	KOKOM	als	3	conjunction		4700	99.99655	99.99655
34	"	$(	"	1	punctuation		16142	99.99889	99.99889
35	neutraler	ADJA	neutral	9	adjective		0	93.07405	92.90016
36	Richter	NN	Richter	7	noun		106	99.89785	99.89785
37	"	$(	"	1	punctuation		16142	99.99889	99.99889
38	,	$,	,	1	comma		54408	99.99963	99.99963
39	sondern	KON	sondern	7	conjunction		600	99.97896	99.97896
40	als	KOKOM	als	3	conjunction		4700	99.99655	99.99655
41	das	ART	d	3	article		8430	99.99840	99.99840
42	"	$(	"	1	punctuation		16142	99.99889	99.99889
[...]									

lesbarkeitsanalyse:

readability()

readability: in koRpus implementiert

- ARI
  NRI, simple
- Dale-Chall
  Powers-Sumner-Kearl, old
- Coleman
- Coleman-Liau
- Danielson-Bryan
- Strain
- Farr-Jenkins-Paterson
  Powers-Sumner-Kearl
- Bormuth
- Degrees of Reading Power
- FORCAST
  precise RGL
- Flesch
  DE, ES, FR, NL, Powers-Sumner-Kearl
- Flesch-Kincaid
- LIX
- RIX
- Harris-Jacobson
- Wheeler-Smith
  DE
- Linsear Write
- neue Wiener Sachtextformeln
- **silbenzählung:**
  LaTeX (Liang, 1983)
- SMOG
  DE, C, simple
- FOG
  Powers-Sumner-Kearl, NRI
- Fucks
- Dickes-Steiwer
- Traenkle-Bailer
- TRI
- Spache
  DE
- Easy Listening Formula

vgl. Klare (1974); DuBay (2004); Bamberger und Vanecek (1984)

welche sprachen werden unterstützt?

sprachunterstützung

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

- out of the box:
  - deutsch
  - englisch
  - spanisch
  - italienisch
  - russisch
- modular & leicht erweiterbar

welche sprache ist das?

language detection

ધૂમકેતુઓ સૂર્યમંડળના અંતમાં આવેલા ઊર્ટ વાદળ માંથી ઊદ્ભવતા હોવાની માન્યતા જાન હેન્ડ્રીક ઊર્ટ નામના વૈજ્ઞાનીકે રજૂ કરી હતી. જ્યારે બરફના થીજેલા આ ગોળાઓ તેમની ભ્રમણકક્ષામાંથી(બાહ્ય ગુરુત્વાકર્શી ખલેલોને કારણે) ચલીત થાય છે ત્યારે તેઓ સૂર્ય તરફ ખેંચાય છે. જ્યારે ધૂમકેતુ સૂર્યમંડળના અંદરના ભાગમાં પ્રવેશ કરે છે ત્યારે સૂર્યના વિકિરણો ને કારણે થીજેલા વાયુઓ પીગળવા માંડે છે. આમ ધૂળ અને વાયુઓ ના મુક્ત થવાથી મોટું વાતાવરણ ધૂમકેતુના કેન્દ્રની આસપાસ રચાય છે જેને ધૂમકેતુનું *કૉમા* કહે છે. સૂર્યના વિકિરણ દબાણ તથા સૂર્ય પવન ની કૉમા પર થતી અસર ને કારણે ધૂમકેતુની લાંબી *પૂંછ* રચાય છે. આ પૂંછ હંમેશા સૂર્યથી વિરૂદ્ધ (ભ્રમણકક્ષાની બહારની) દીશામાં રચાતી હોય છે. ધૂળ તથા વાયુઓ પોત–પોતાની અલગ અલગ પૂંછ રચતા હોય છે. વાયુઓના આયનીકરણ ને કારણે તે પૂંછ સૂર્યના ચુંબકીય ક્ષેત્રની અસર હેઠળ આવે છે જ્યારે ધૂળની પૂંછ સામાન્યરીતે ગુરુત્વાકર્ષણ હેઠળ આવે છે. ધૂમકેતુના ઘન કેન્દ્રને તેનુ *ન્યુક્લીયસ* કહેવાય છે જે સામાન્યરીતે ૫૦ કી.મી.થી નાનું હોય છે. કૉમા તથા તેની પૂંછ ક્યારેક ૧ AU (૧૫૦ મીલીયન કી.મી.) થી પણ વધુ લાંબી હોય છે.

language detection (Benedetto, Caglioti & Loreto, 2002)

- verfahren basiert auf LZ77/gzip
  - vergleicht größenunterschiede nach kompression
- download menschenrechte (`udhr_txt.zip`): `http://unicode.org/udhr`
- erkennt so $>$ 350 sprachen

language detection (Benedetto et al., 2002)

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

guess.lang()

```
> guess.lang("sample_text.txt", udhr.path="udhr_txt.zip")

  Estimated language: Gujarati
          Identifier: gu
             Country: IN (Asia)

377 different languages were checked.
```

text: "komet" von http://gu.wikipedia.org

wissen, wie's geht:

dokumentation

`"SMOG"`: *Simple Measure of Gobbledygook*. By default calculates formula D by McLaughlin (1969):

$$SMOG = 1.043 \times \sqrt{W_{3Sy} \times \frac{30}{St}} + 3.1291$$

If `parameters` is set to `SMOG="C"`, formula C will be calculated:

$$SMOG_C = 0.9986 \times \sqrt{W_{3Sy} \times \frac{30}{St} + 5} + 2.8795$$

If `parameters` is set to `SMOG="simple"`, the simplified formula is used:

$$SMOG_{simple} = \sqrt{W_{3Sy} \times \frac{30}{St}} + 3$$

If `parameters` is set to `SMOG="de"`, the formula adapted to German texts ("Qu", Bamberger & Vanecek, 1984, p. 78) is used:

$$SMOG_{de} = \sqrt{W_{3Sy} \times \frac{30}{St}} - 2$$

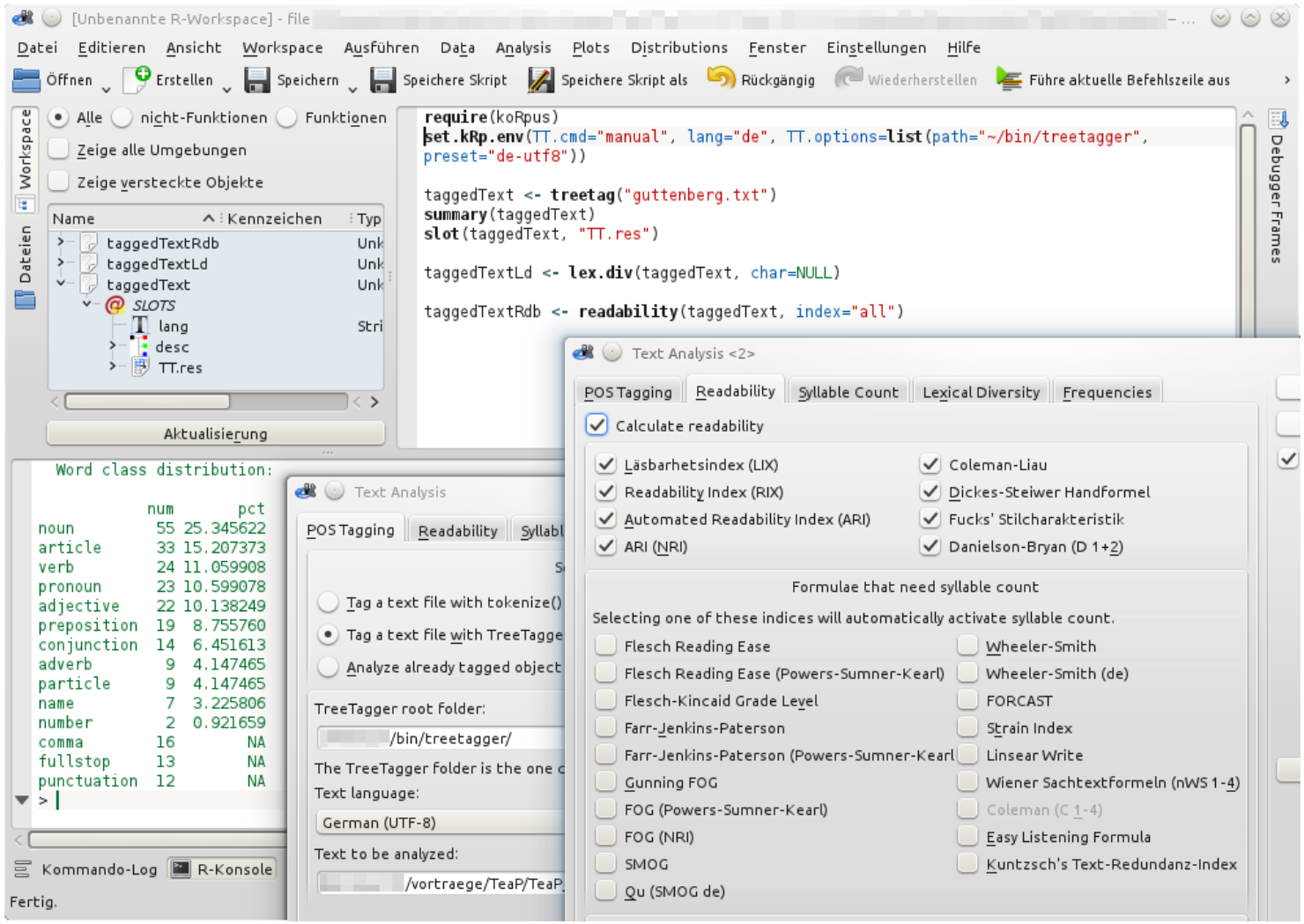
Wrapper function: `SMOG`

grafische benutzeroberfläche:

RKWard-plugin

(Rödiger, Friedrichsmeier, Kapat & Michalke, o. J.)

grafische benutzeroberfläche



ausblick

foto: © basykes, CC-BY 2.0

ausblick

- authorship detection

  Brennan, Afroz und Greenstadt (2011); Brennan und Greenstadt (2009)

- kompatibilität mit `tm`-paket

  Feinerer, Hornik und Meyer (2008)

[ceterum censeo]

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

menschenrechte vs. studiengebühren

internationaler pakt über wirtschaftliche, soziale und kulturelle rechte, artikel 13:

»die vertragsstaaten erkennen an, daß im hinblick auf die volle verwirklichung dieses rechts [...] **der hochschulunterricht** auf jede geeignete weise, **insbesondere durch allmähliche einführung der unentgeltlichkeit**, jedermann gleichermaßen entsprechend seinen fähigkeiten zugänglich gemacht werden muß.«

- völkerrechtlicher vertrag
- am 9. oktober 1968 unterzeichnet
- am 17. dezember 1973 vorbehaltlos ratifiziert
- am 3. januar 1976 in kraft getreten.
- alle bundesländer haben dem beitritt zugestimmt
- durch vertragsgesetz vom 23. november 1973 formelles bundesgesetz
- `http://www.auswaertiges-amt.de/diplo/de/Aussenpolitik/Menschenrechte/Download/IntSozialpakt.pdf`

thx:

Earl Brown *(California State University, Monterey Bay)*
Scott Jarvis *(Ohio University)* & **Phil McCarthy** *(University of Memphis)*
Eleni Miltsakaki *(University of Pennsylvania)*
Alberto Mirisola *(Italian Research Council, Institute for Educational Technology)*
Helmut Schmid *(Universität Stuttgart)*
Laura Hauser *(HHU)*

fragen? ideen? kommentare?

<meik.michalke@hhu.de>

http://koRpus.reaktanz.de

literatur I

Bamberger, R. & Vanecek, E. (1984). *Lesen, Verstehen, Lernen, Schreiben*. Wien: Jugend und Volk.

Becker, R. A. & Chambers, J. M. (1984). *S: An interactive environment for data analysis and graphics*. Belmont, CA: Wadsworth.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books & Software.

Benedetto, D., Caglioti, E. & Loreto, V. (2002). Language trees and zipping. *Physical Review Letters*, *88* (4), 048702.

Brennan, M., Afroz, S. & Greenstadt, R. (2011). *Deceiving authorship detection*. Presentation at the 28th Chaos Communication Congress (28C3), Berlin, Germany.

Brennan, M. & Greenstadt, R. (2009). Practical attacks against authorship recognition techniques. In *Proceedings of the Twenty-First conference on innovative applications of artificial intelligence (IAAI)*. Pasadena, CA.

Chambers, J. M. (1998). *Programming with data: A guide to the S language*. Berlin: Springer.

Chambers, J. M. & Hastie, T. (1991). *Statistical models in S*. London: Chapman and Hall/CRC.

Covington, M. A. & McFall, J. D. (2010). Cutting the gordian knot: The Moving-Average Type-Token ratio (MATTR). *Journal of Quantitative Linguistics*, *17* (2), 94–100.

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.

Feinerer, I., Hornik, K. & Meyer, D. (2008, März). Text mining infrastructure in R. *Journal of Statistical Software*, *25* (5), 1–54. Verfügbar unter `http://www.jstatsoft.org/v25/i05`

Guttenberg, K. F. z. (2009). *Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU* (1. Aufl.). Berlin: Duncker & Humblot.

Klare, G. R. (1974, Januar). Assessing readability. *Reading Research Quarterly*, *10* (1), 62–102.

Liang, F. M. (1983). *Word hy-phen-a-tion by com-put-er*. Unveröffentlichte Dissertation, Stanford University, Dept. Computer Science, Stanford.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42* (2), 381–392.

literatur II

Michalke, M. (2012a). *koRpus: An R package for text analysis (Manual)*. Verfügbar unter `http://reaktanz.de/?c=hacking&s=koRpus`

Michalke, M. (2012b). *Using the koRpus package for text analysis*. Verfügbar unter `http://reaktanz.de/?c=hacking&s=koRpus`

R Development Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Verfügbar unter `http://www.R-project.org`

Rödiger, S., Friedrichsmeier, T., Kapat, P. & Michalke, M. (o. J.). RKWard – A comprehensive graphical user interface and integrated development environment for statistical analysis with R. *Journal of Statistical Software*.

Schieren, S. (2002). Europa zwischen rechtlich-konstitutioneller Konkordanz und politisch-kultureller Vielfalt. *Arbeitspapiere/Mannheimer Zentrum für Europäische Sozialforschung*, *53*.

Schmid, H. (1994). Probabilistic part-of-speec tagging using decision trees. In *International conference on new methods in language processing* (S. 44–49). Manchester, UK.

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.