

SAMPLE APPLICATIONS

TERMINOLOGY ANALYSIS

- Wanted: words that appear much more frequently in domain-specific texts (or only there) than in other texts
- Various approaches:
 - Fixed comparison parameters, e.g. frequency classes
 - tf.idf
 - Significance tests, e.g. via Log Likelihood

1. FREQUENCY CLASSES

Absolute word frequencies are hard or impossible to compare. Several approaches:

- Relative word frequency
- Frequency classes: words with similar frequency are put in same frequency class, e.g. via rounded logarithm dual of the frequency of the most common word divided by the frequency of word w ($\rightarrow FCL(w) = \text{round}(\log_2(\text{freq}_{\max} / \text{freq}_w))$)
 - Example based on some German word frequencies:
der (14,935,099), die (14,271,215), Auto (108,857), Rennstall (2,000)
 - $FCL(\text{der}) = \log_2(\text{freq}(\text{der}) / \text{freq}(\text{der})) = 0$
 - $FCL(\text{die}) = \log_2(\text{freq}(\text{der}) / \text{freq}(\text{die})) = 0.0656 = 0$
 - $FCL(\text{Auto}) = \log_2(\text{freq}(\text{der}) / \text{freq}(\text{Auto})) = 7.1 = 7$
 - $FCL(\text{Rennstall}) = \log_2(\text{freq}(\text{der}) / \text{freq}(\text{Rennstall})) = 12.9 = 13$

FREQUENCY CLASSES – SAMPLE DISTRIBUTION

FCL	Types	FCL	Types	FCL	Types	FCL	Types
0	3	4	22	8	525	12	5714
1	2	5	67	9	1166	13	9600
2	15	6	119	10	1986	14	15837
3	24	7	280	11	3450	15	25804

Frequency Dictionary German - Häufigkeitswörterbuch Deutsch. Uwe Quasthoff, Sabine Fiedler and Erla Hallsteindóttir (eds.).
Leipziger Universitätsverlag, 2011

TERMINOLOGY EXTRACTION – EXAMPLE SAP

- Most frequent 100 words in a German Web corpus:

der, die, und, in, den, von, zu, das, mit, sich, des, auf, für, ist, im, dem, nicht, ein, Die, eine, als, auch, es, an, werden, aus, er, hat, daß, sie, nach, wird, bei, einer, Der, um, am, sind, noch, wie, einem, über, einen, Das, so, Sie, zum, war, haben, nur, oder, aber, vor, zur, bis, mehr, durch, man, sein, wurde, sei, In, Prozent, hatte, kann, gegen, vom, können, schon, wenn, habe, seine, Mark, ihre, dann, unter, wir, soll, ich, eines, Es, Jahr, zwei, Jahren, diese, dieser, wieder, keine, Uhr, seiner, worden, Und, will, zwischen, Im, immer, Millionen, Ein, was, sagte

TERMINOLOGY EXTRACTION – EXAMPLE SAP

- Most frequent 100 words with SAP texts:

die, Sie, der, und, in, werden, den, für, das, im, können, wird, zu, eine, auf, des, %N%, Die, ist, mit, ein, von, dem, the, oder, nicht, an, einer, aus, sind, In, einen, zur, als, über, System, kann, bei, einem, Wenn, Das, auch, nur, diesem, sich, eines, müssen, Daten, Der, daß, zum, to, haben, diese, alle, B, durch, z, R, wenn, nach, es, Feld, dann, of, wählen, Funktion, bzw, um, dieser, Wählen, Im, a, wie, is, Informationen, Diese, Bei, for, muß, and, vom, so, Für, Mit, unter, sein, keine, ob, soll, definieren, Es, verwendet, automatisch, Tabelle, Geben, wurde, finden, you, beim

TERMINOLOGY EXTRACTION – EXAMPLE SAP

- Difference (min FCL: 8, factor: 16) SAP/German Web corpus:
etc (314), TCP (164), INDX (28), dsn (25), Nachfolgeposition (24), SHIFT (24),
TRANSLATE (24), entreprise (24), Abrechnungskostenart (23), Alternativmengeneinheit
(23), Anordnungsbeziehung (23), Anwendungssicht (23), Bandstation (23), Banf-Position
(23), Berichtsspalte (23), Berichtszeile (23), CO-PC (23), DBSTATC (23), DSplit (23),
Datumsart (23), ELSE (23), ENDDO (23), Entries (23), Freigabecodes (23), Hauptkondition
(23), Leiterplanstelle (23), Merkmalswertekombination (23), Nachfolgematerial (23),
Nettoberechnung (23), ...

ABAP, Advanced Business Application Programming

2. TF.IDF

- As already seen...

3. SIGNIFICANCE TESTS

- As already seen...

EXAMPLE

TEXT ANALYSIS ERNST JÜNGER

ERNST JÜNGER?

- German author (1895 – 1998)
- Most famous book: „In Stahlgewittern“ (1920)
- Controversial relationship with NSDAP
 - „Wegbereiter des Nationalsozialismus“?
 - „most controversial German writer of the 20th century“?



Bundesarchiv, B 145 Bild-F073370-0006 / Wegmann, Ludwig CC-BY-SA 3.0

RESEARCH QUESTION

- Are there differences in the use of vocabulary over time (diachronic)?
- How does the vocabulary used differ from contemporary texts?
- Approach:
 - Acquisition of digital versions of Jünger's texts
 - Building a reference corpus
 - Calculating (sub-)corpus similarity
 - Analysis of the differences

REFERENCE CORPUS – DWDS KERNKORPUS

- Types of text
 - Fiction (26%)
 - Newspaper (27%)
 - Scientific texts (25%)
 - Practical literature („Gebrauchsliteratur“) (22%)
- Extent
 - Tokens: 121M
 - Types: 1.94M
 - Documents: 79,116

<https://www.dwds.de/d/korpora/kern>


(PRE)PROCESSING

- Processing steps: tokenization, pos-tagging, frequency analysis
- Analysis:
 - Similarity matrix based on word list similarity
 - Clustering / Dendrogram
 - Frequency over time / timelines
- Visual analysis / Distant Reading

Dirk Goldhahn, Thomas Eckart, Thomas Gloning, Kevin Dreßler und Gerhard Heyer: Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case. In: CLARIN Annual Conference 2015 in Wrocław, Poland, 2015.

ANALYSIS I

Corpus Comparison
Similarity Matrix
Word Frequency Analysis
en de



Configuration

Length of Wordlist (10000)

Request title

Corpora

Jünger 1927 ✕
Jünger 1928 ✕
Jünger 1929 ✕
Newspaper 1927 ✕
Newspaper 1928 ✕
Newspaper 1929 ✕

Similarity measure

Cosine - based on rank

Cosine - based on frequency

Cosine - based on logarithm of frequency

Job Selection

Jünger '19-33

Jünger 1927,
Jünger 1928,
Jünger 1929,
11 more

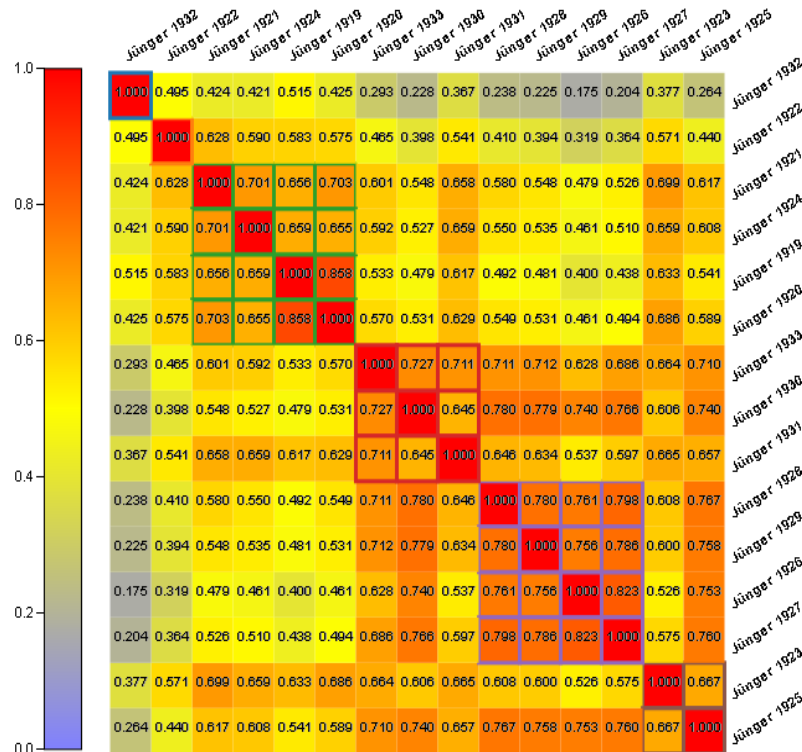
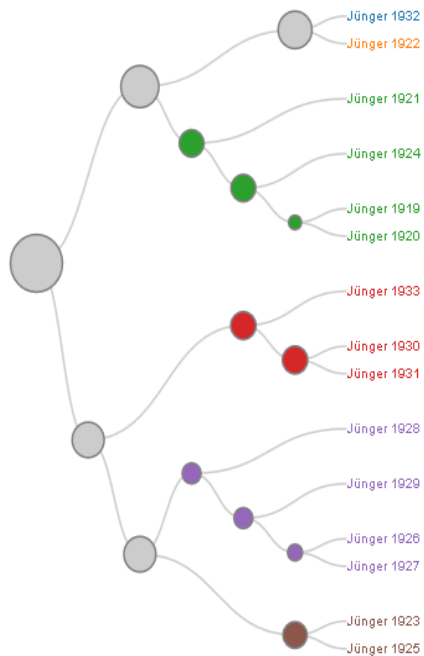
Newspaper '19-33

Newspaper 1927,
Newspaper 1928,
Newspaper 1929,
11 more

Overview '27-29

Jünger 1927,
Jünger 1928,
Jünger 1929,
9 more

ANALYSIS II - CLUSTERING



More frequent in Jünger 1920

Word

Feuer

standen

zurück

zusammen

kurzen

Schatten

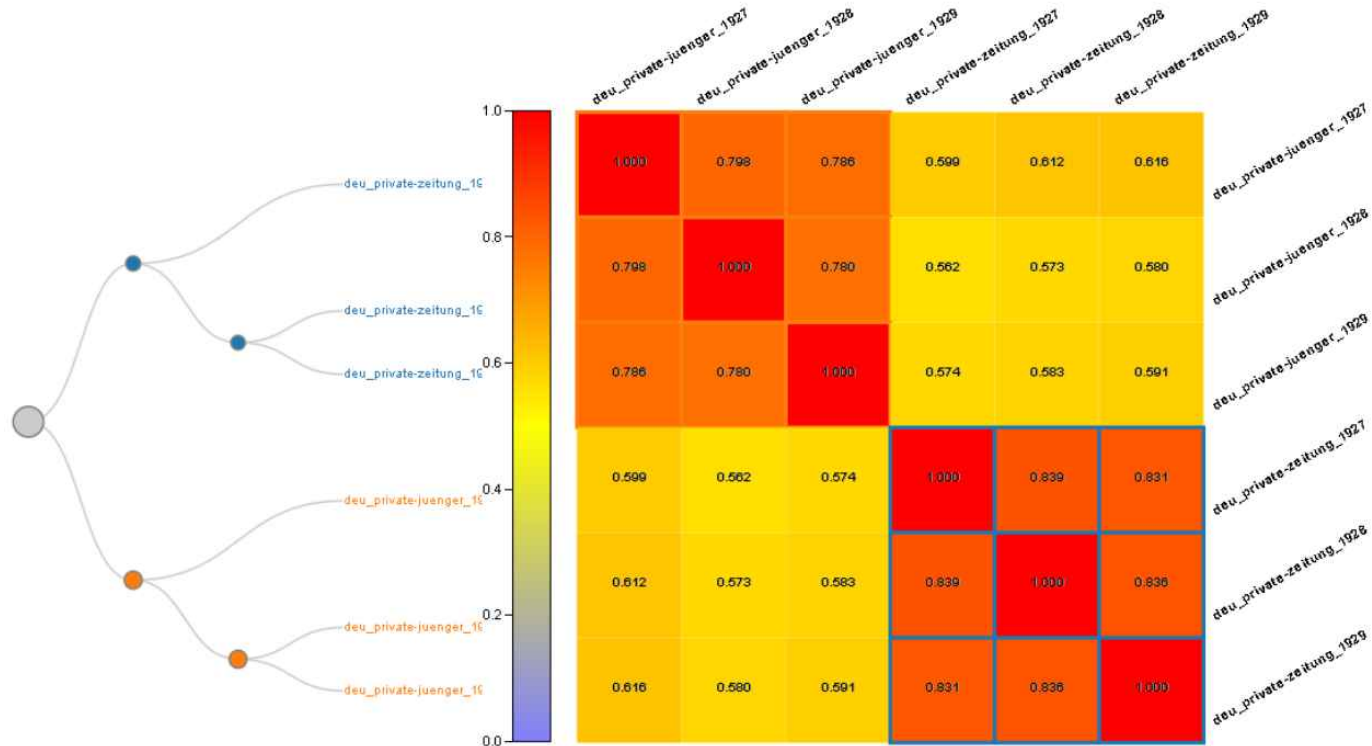
Kein

springt

größeres

überragende

ANALYSIS III - CLUSTERING



ANALYSIS IV – JÜNGER 1929 VS. JÜNGER 1925

More frequent in Jünger 1929

Word	POS	Frequency in Jünger 1929	Frequency in Jünger 1925
<input type="text" value="Search word..."/>	<input type="text" value="NOUN"/>		
Buch	NOUN	29	1
Nationalismus	NOUN	39	2
Ordnung	NOUN	12	1
Gestalten	NOUN	12	1
Jahren	NOUN	11	1
Verhältnis	NOUN	11	1
Willens	NOUN	11	1
Glauben	NOUN	10	1
Verantwortung	NOUN	19	2
Unterschied	NOUN	9	1

First < 1 of 47 > Last

More frequent in Jünger 1925

Word	POS	Frequency in Jünger 1929	Frequency in Jünger 1925
<input type="text" value="Search word..."/>	<input type="text" value="NOUN"/>		
Frontsoldaten	NOUN	4	21
Material	NOUN	1	6
Maschine	NOUN	1	6
Fragen	NOUN	1	6
Pflicht	NOUN	1	7
Ziele	NOUN	1	7
Massen	NOUN	1	7
Frontsoldat	NOUN	3	25

First < 16 of 16 > Last

ANALYSIS V – JÜNGER 1929 VS. ZEITUNG 1929

Nomen – nur bei Jünger

Willens
Elementare
Verwesung
Mißverhältnis
Schauwecker
Ideologie
Kriegserlebnis
Zone
Dämon
Frontsoldat

Nomen – häufiger bei Jünger

Nationalismus
Liberalismus
Gestalten
Erstaunen
Erlebnis
Bestände
Bindungen
Schärfe
Chaos
Unruhe

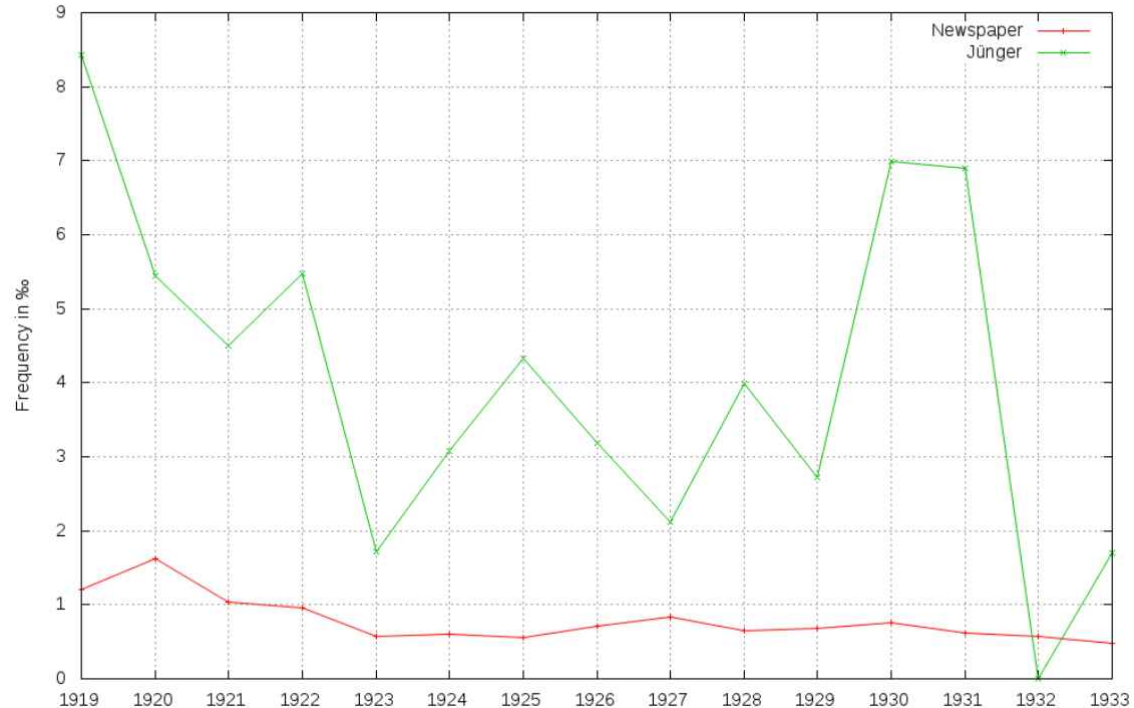
ANALYSIS VI – „REVOLUTION“ (DIACHRON)

Relative Frequency of 'Revolution*' in German Newspaper and Ernst Jünger Texts from 1919 - 1933



ANALYSIS VII – „KRIEG“ (DIACHRON)

Relative Frequency of 'Krieg' in German Newspaper and Ernst Jünger Texts from 1919 - 1933



COMPARISON OF RANKINGS – RANK CORRELATION

- Common problem: comparison of sorted lists
 - e.g. features and their frequency in two documents / corpora (words, n-grams, etc.)
- Question: How strong do these lists correlate based on frequency ranks?

→ Rank correlation coefficient

- e.g.
 - Spearman's Rho
 - Kendall's Tau

Feature	Rank _{d1}	Rank _{d2}	Rank _{d3}
f ₁	1	1	23
f ₂	2	5	4
f ₃	3	2	33
f ₄	4	7	13
f ₅	5	9	15
...

KENDALL'S TAU

- Idea:
 - Use of the rank differences between two lists X and Y of length n, sorted by X ($x_i < x_j$, for $i < j$)
 - Compare pairs (x_i, x_j) and (y_i, y_j) , for $i = 1 \dots n$ and $j = i+1 \dots n$
 - $n * (n-1) / 2$ comparisons
 - If order for the respective pair identical in both lists ($y_i < y_j$)
 - „concordant“
- $\tau = |\text{concordant pairs}| - |\text{discordant pairs}| / |\text{comparisons}|$
- Value: [-1,1]

KENDALL'S TAU

– Example:

- $(f_1f_2) = \text{concordant } (1 < 2 \text{ vs. } 1 < 4)$
- $(f_1f_3) = \text{concordant } (1 < 3 \text{ vs. } 1 < 2)$
- $(f_1f_4) = \text{concordant } (1 < 4 \text{ vs. } 1 < 3)$
- $(f_2f_3) = \text{discordant } (2 < 3 \text{ vs. } 4 > 2)$
- $(f_2f_4) = \text{discordant } (2 < 4 \text{ vs. } 4 > 3)$
- $(f_3f_4) = \text{concordant } (3 < 4 \text{ vs. } 2 < 3)$

– $\tau = (4 - 2) / 6 = 1/3$

Feature	Rank _{d1}	Rank _{d2}
f ₁	1	1
f ₂	2	4
f ₃	3	2
f ₄	4	3

KENDALL'S TAU

– Extrema:

Feature	Rang _{d1}	Rang _{d2}
f ₁	1	1
f ₂	2	2
f ₃	3	3
f ₄	4	4

Concordant pairs: {f₁f₂, f₁f₃, f₁f₄, f₂f₃, f₂f₄, f₃f₄}

$$\tau = (6 - 0) / 6 = 1$$

Feature	Rang _{d1}	Rang _{d2}
f ₁	1	4
f ₂	2	3
f ₃	3	2
f ₄	4	1

Concordant pairs: \emptyset

$$\tau = (0 - 6) / 6 = -1$$

Feature	Rang _{d1}	Rang _{d2}
f ₁	1	1
f ₂	2	4
f ₃	3	3
f ₄	4	2

Concordant pairs: {f₁f₂, f₁f₃, f₁f₄}

$$\tau = (3 - 3) / 6 = 0$$

KENDALL'S TAU – APPLICATION LINGUISTIC TYPOLOGY I

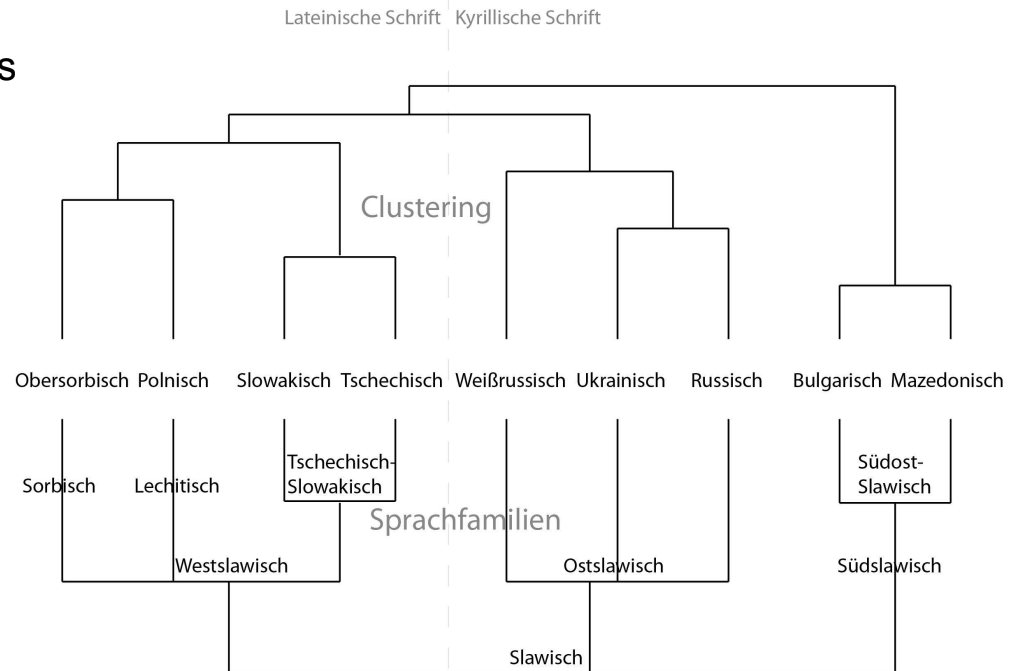
- Goal of linguistic typology: classification of languages based on structural properties
- Idea: use of corpus-based language statistics
 - (e.g.) Comparison based on stop words and character 3-grams
- Slightly adjusted measure for non-identical element sets

Rang	Wort im schwedischen Korpus	Wort im dänischen Korpus
1	och	i
2	i	at
3	eft	og
4	på	er
5	som	til
6	är	en
7	en	af
8	det	på
9	för	det
10	med	for
11	av	der
12	til	med
13	har	de
14	inte	har
15	den	den
16	de	ikke
17	om	som
18	ett	et
19	Det	om
20	jag	Det

Dirk Goldhahn: Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken. 2013

KENDALL'S TAU – APPLICATION LINGUISTIC TYPOLOGY II

- Result: Cluster vs. Language families



Dirk Goldhahn: Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken. 2013

EXAMPLE

TEXT ANALYSIS RICHMOND TIMES

RICHMOND TIMES DISPATCH?

- Newspaper from Richmond (Virginia, USA)
- Digital text corpus by Perseus project
 - Daily corpora between 1.11.1860 – 30.12.1865
- American civil war: 12.04.1861 – 23.06.1865
- Comparison:
 - Aggregation to monthly corpora
 - Differences in frequency classes (-5 – 5) based on relative word frequencies

<https://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:RichTimes>

COMPARISON DECEMBER 1860 – MARCH 1865

Differential analysis

Select Database: **Richmond1** Compare: **Work** Reference: **RichmondTimes 1860 Decem** Analysis: **RichmondTimes 1865 March** **Process**

Reference	references	-5	-4	-3	-2	-1	0	1	2	3	4	analysis
C	Union											her
how	Va.											rights
Dr.	N											She
street	Main											December
hand	Negroes											25
things	oh											angels
fathers	blessed											Oh
England	ready											quite
Federal	likely											young
County	slave											Thomas
submitted	remain											save
find	house											Cook
sweet	Personal											knew
room	Hall											told
pretty	stream											Franklin
Office												

COMPARISON DECEMBER 1860 – MARCH 1865

Differential analysis

Select Database: **Richmond1** Compare: **Work** Reference: **RichmondTimes 1860 Decem** Analysis: **RichmondTimes 1865 March** **Process**

Reference	references	-5	-4	-3	-2	-1	0	1	2	3	4	analysis
army	General											Secretary
hundred	speech											tobacco
forward	War											Senators
reported	papers											Chief
remarks	Times											platform
scene	means											Russian
resolved	consideration											Lincoln
Justice	minutes											deserters
commands	capture											payable
river	set											Emperor
meeting	seems											adjourned
instant	resolutions											presented
increase	soldier											secret
captured	Charlottesville											crowd
Bonds												

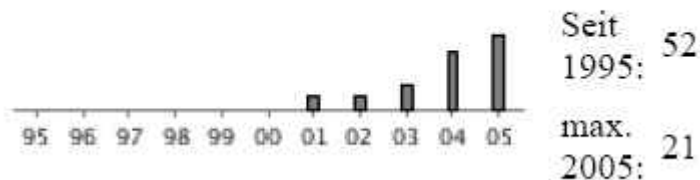
EXAMPLE DIACHRONIC COMPARISON

TYPES OF NEOLOGISMS

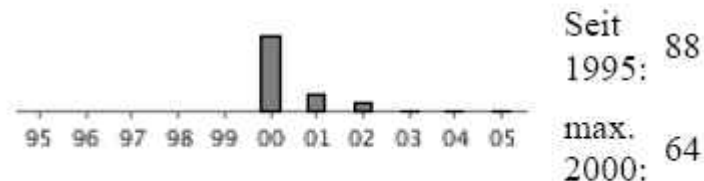
1. „Real“ neologism, completely new creation (*Energiepreisbremse*)
 - a) Growing trend or
 - b) After a strong start, a downward trend
2. Previously low frequency word,
 - a) whose frequency increases or
 - b) which suddenly (e.g. due to an event) becomes noticeable in general language (*Acrylamid, Feinstaub*)
3. Event-driven words that occur at intervals with great frequency (*Weihnachten*)
4. Word with new meaning
5. Identifier for individual objects or groups
6. Words with a short lifespan

NEOLOGISMS

DSL-Kunden

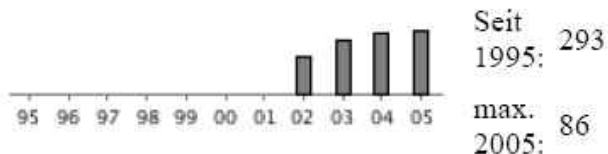


UMTS-Auktion



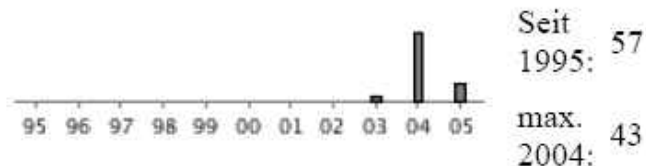
Defizitverfahren *Politik*

Von der EU eingeleitetes Verfahren gegen Staaten mit zu hohem Haushaltsdefizit



Alkopops *Ernährung, Gesellschaft*

Alkoholische Mischgetränke



Quelle: Uwe Quasthoff (Hrsg.) *Deutsches Neologismenwörterbuch*. De Gruyter. 2007.

EXAMPLE DIACHRONIC COMPARISON (ARABIC)

ARABIC NEWSTEXTS (2007 – 2012) – WORD RANKS

<i>Englisch</i>	<i>Term</i>	<i>2007</i>	<i>2008</i>	<i>2009</i>	<i>2010</i>	<i>2011</i>	<i>2012</i>
<i>Democracy</i>	الديمقراطية	631	721	453	655	347	500
<i>Israel</i>	إسرائيل	168	118	88	99	114	195
<i>Obama</i>	أوباما	10485	173	93	195	187	630
<i>Elections</i>	الانتخابات	170	141	97	153	138	158
<i>Rights</i>	الحقوق	683	2063	1180	1590	2892	1507
<i>Iran</i>	إيران	141	190	104	147	215	291
<i>Freedom</i>	الحرية	1635	1372	1175	656	699	636
<i>Gaddafi</i>	القذافي	1959	1894	2134	3804	79	589
<i>Brotherhood</i>	الاخوان	6556	5763	23147	15725	5122	2895

IN GENERAL: COMPARISON FOCUSES

	synchron	diachron
Same text type	Terminology extraction, authorship and source identification	Monitoring and trend analysis of technological developments, public opinion, literary categories
Different text type	Determination of genre and media-specific differences	Influence analysis between genres and media

Chris Biemann, Gerhard Heyer & Uwe Quasthoff: Wissensrohstoff Text: Eine Einführung in das Text Mining, Springer Vieweg, 2022.

