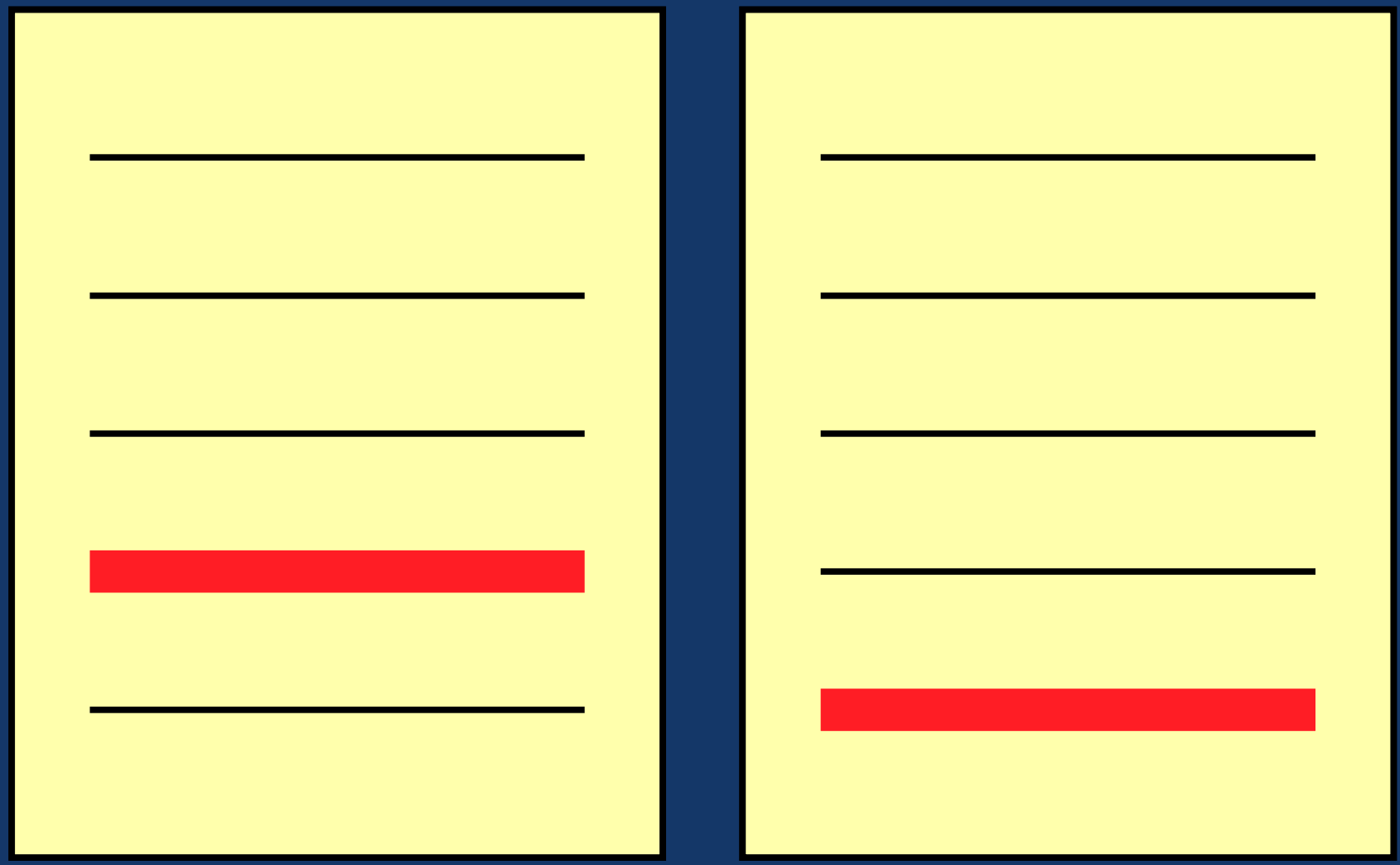
Workshop „Thematic Corpora“

Noah Bubenhofer, Zürich

„Language and Modern Technologies“, Tbilisi, Georgia, September 14, 2015

Noah Bubenhofer

Why Corpus Linguistics in Thematic Corpora?



Example Mountaineering in 1888

"Wir hielten uns möglichst hoch auf der nördlichen, nach Süden fallenden Seite von Vaplona, gewannen das kleine Tobel, das gegen den Punkt 2547 m ansteigt, wateten durch dasselbe hinaus und erreichten den genannten Punkt um 8 Uhr 45 Min. Hier sahen wir etwa 200 m unter uns den Schottensee mit Schnee und Eis bedeckt. Schottenseefurke wäre vielleicht, der Wildseefurke (2515 m) entsprechend, der passendste Name für diese Scharte zwischen den Punkten 2647 m und 2650 m. Wir blieben nicht lange hier, sondern stiegen bald wieder links auf, theils über Schnee, theils über Verrucanotrümmer, und betraten den Gipfelpunkt 2650 m um 9 Uhr 15 Min. Es ging ein schwacher Windzug, und das Thermometer zeigte auf -20 C."

Jahrbuch 1888-1889: Eine Sectionsfahrt auf den Piz Sol (J. J. Schiesser)

Example Mountaineering in 1932

"[N]ochmal wird angegriffen. Und endlich gelingt es mir, für die linke Hand ganz oben einen guten Griff zu schaffen. Die Rechte bohrt sich mit dem Eisbeil in der ersehnten Rampe ein Loch, der rechte Fuss steigt nochmal nach, und ich sehe über den Rand, indessen der Körper schwer nach aussen hängt. Sich ganz auf die rechte Hand verlassend, fährt die linke blitzschnell weit über den Rand in den Firn, ein Ruck, und ich liege verschnaufend auf dem Bauch, während die Füsse in der Luft baumeln."

Die Alpen 1932: Piz Bernina-Nordostflanke (Karl Schneider)

The Pact with the Devil

what we really need is a little pact with the devil: we know how to read texts, now let's learn how *not* to read them. Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems.

Moretti (2000: 57)

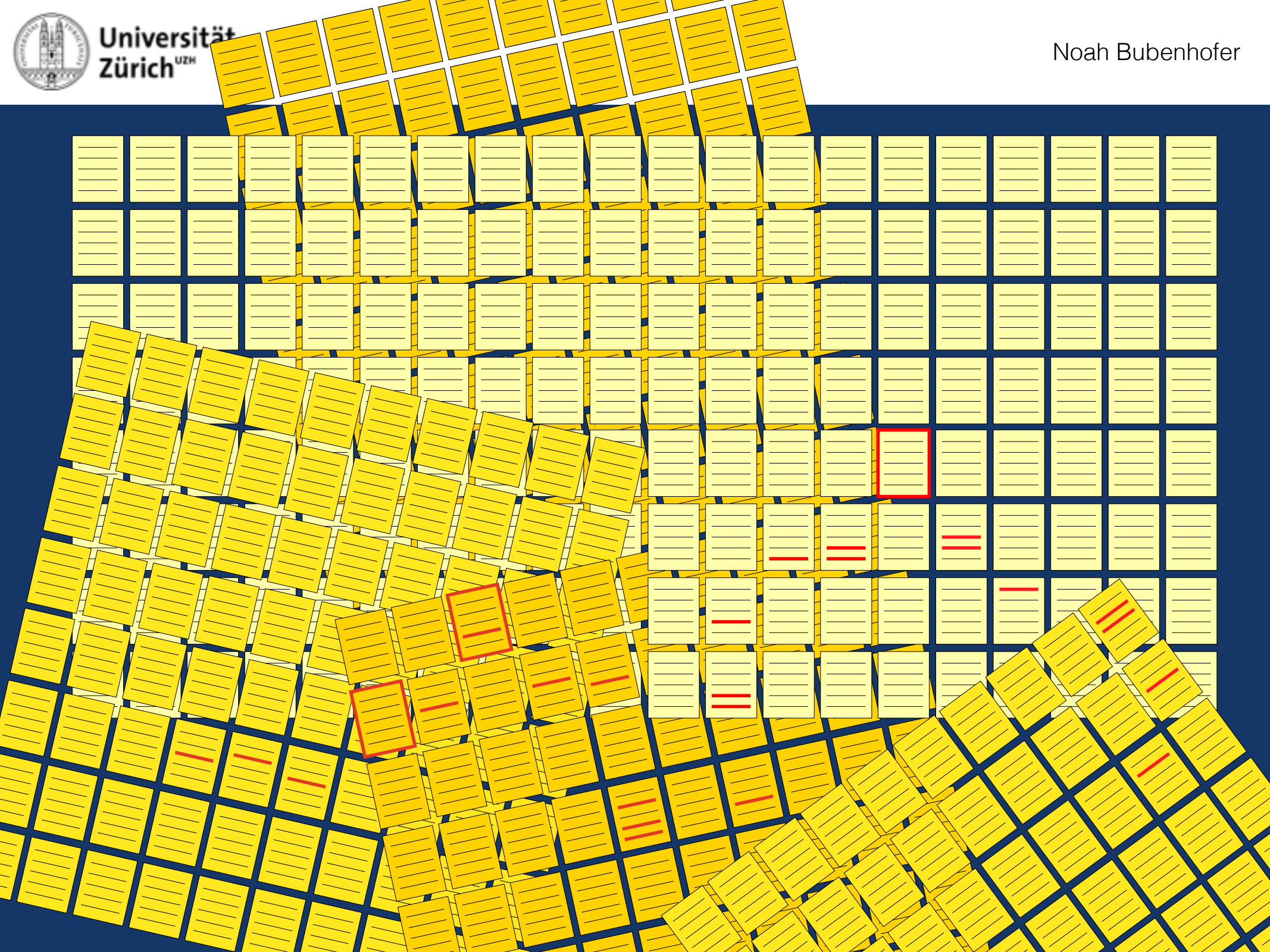
New Questions

Perhaps the single most important issue in effecting transformation is scale. […] digitised texts that can be searched, analysed, and correlated by machine algorithms number in the hundreds of thousands (now, with Google books, a million and more), limited only by ever-increasing processor speed and memory storage. Consequently, machine queries allow questions that would simply be impossible by hand calculation.

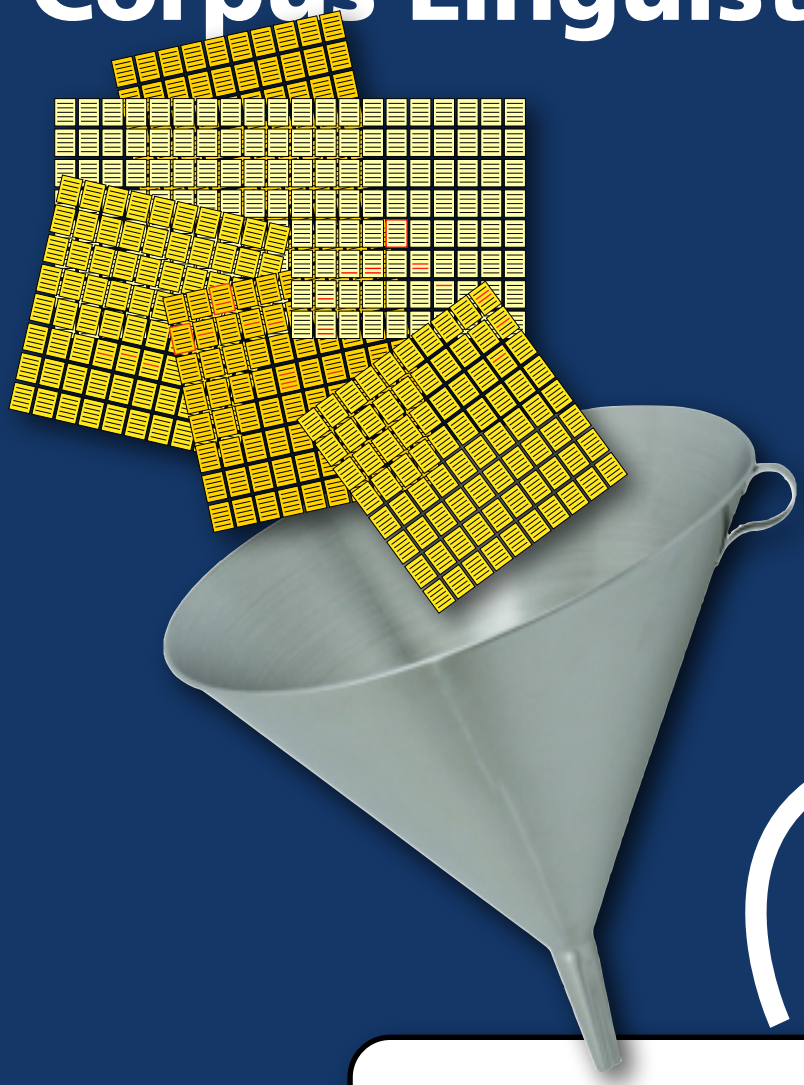
Hayles (2012: 45)

Universität Zürich[UZH]

Noah Bubenhofer



Corpus Linguistics



Text Collection

NLP

Corpus

Analysis

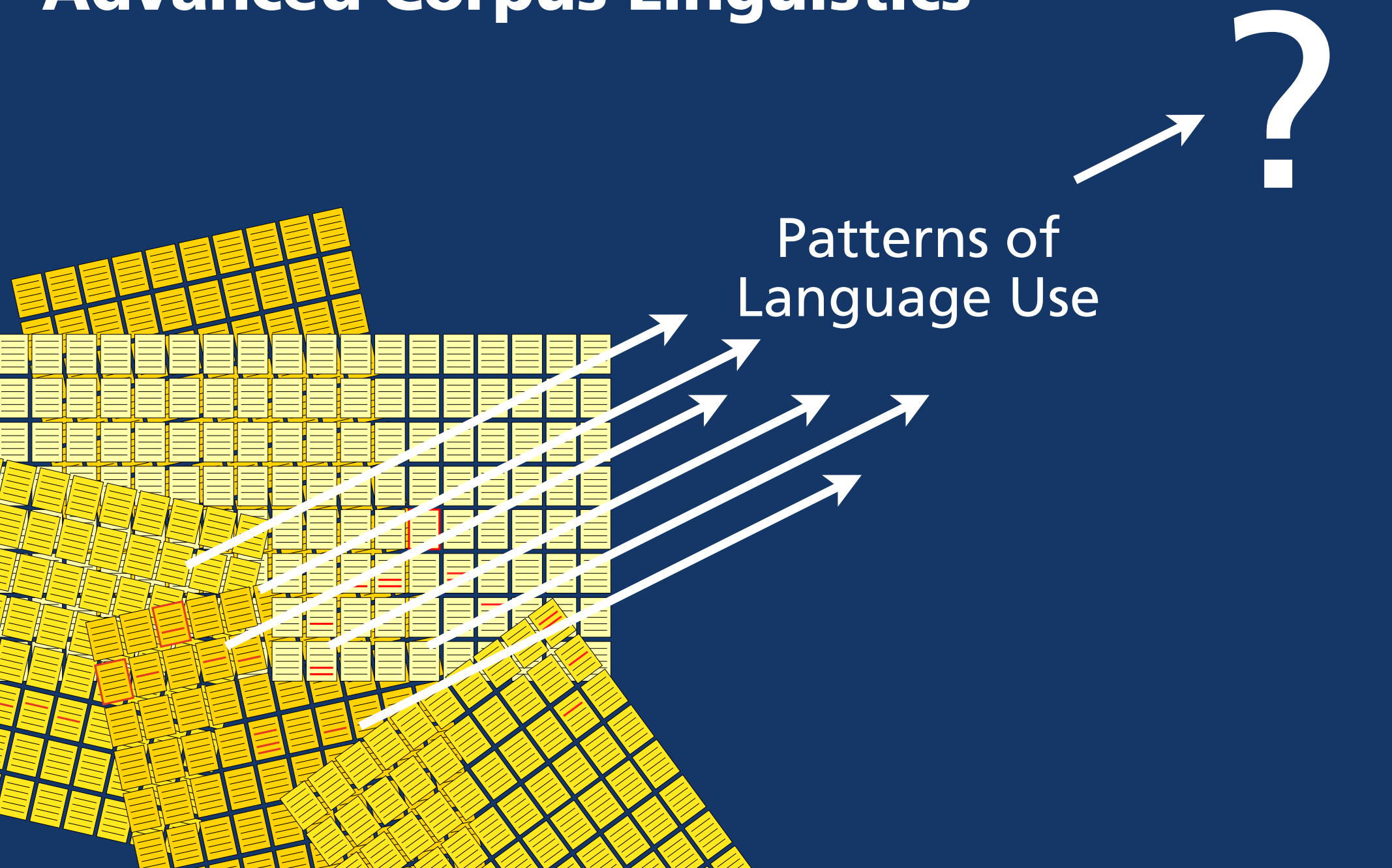
Naïve Corpus Linguistics

Let's find an example where someone uses the expression „war against terrorism"!

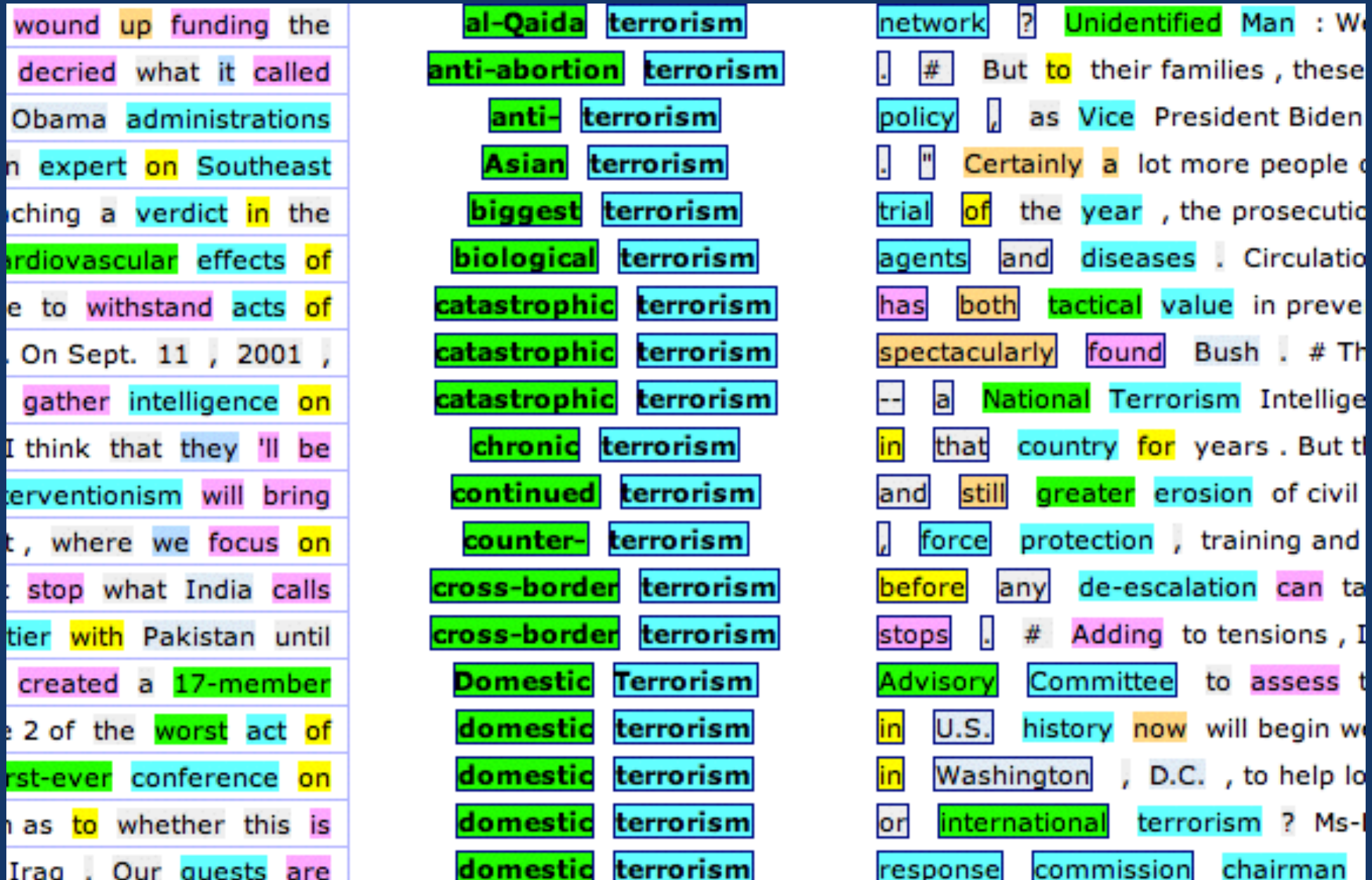
gy that is needed for the	war against terrorism	already exists , unlike
t the central front on our	war against terrorism	and all the dynamics
ribe as satisfactory . The	war against terrorism	and bilateral trade pro
g an important role in the	war against terrorism	and has agreed to tak
to rationalize a perpetual	war against terrorism	and keeps places like
d to be set with us in the	war against terrorism	and support us in the
ng Bush 's first term , the	war against terrorism	and the extensive milit
for conducting America 's	war against terrorism	are being planned toda
overnment should use our	war against terrorism	as an excuse to perse

Noah Bubenhofer

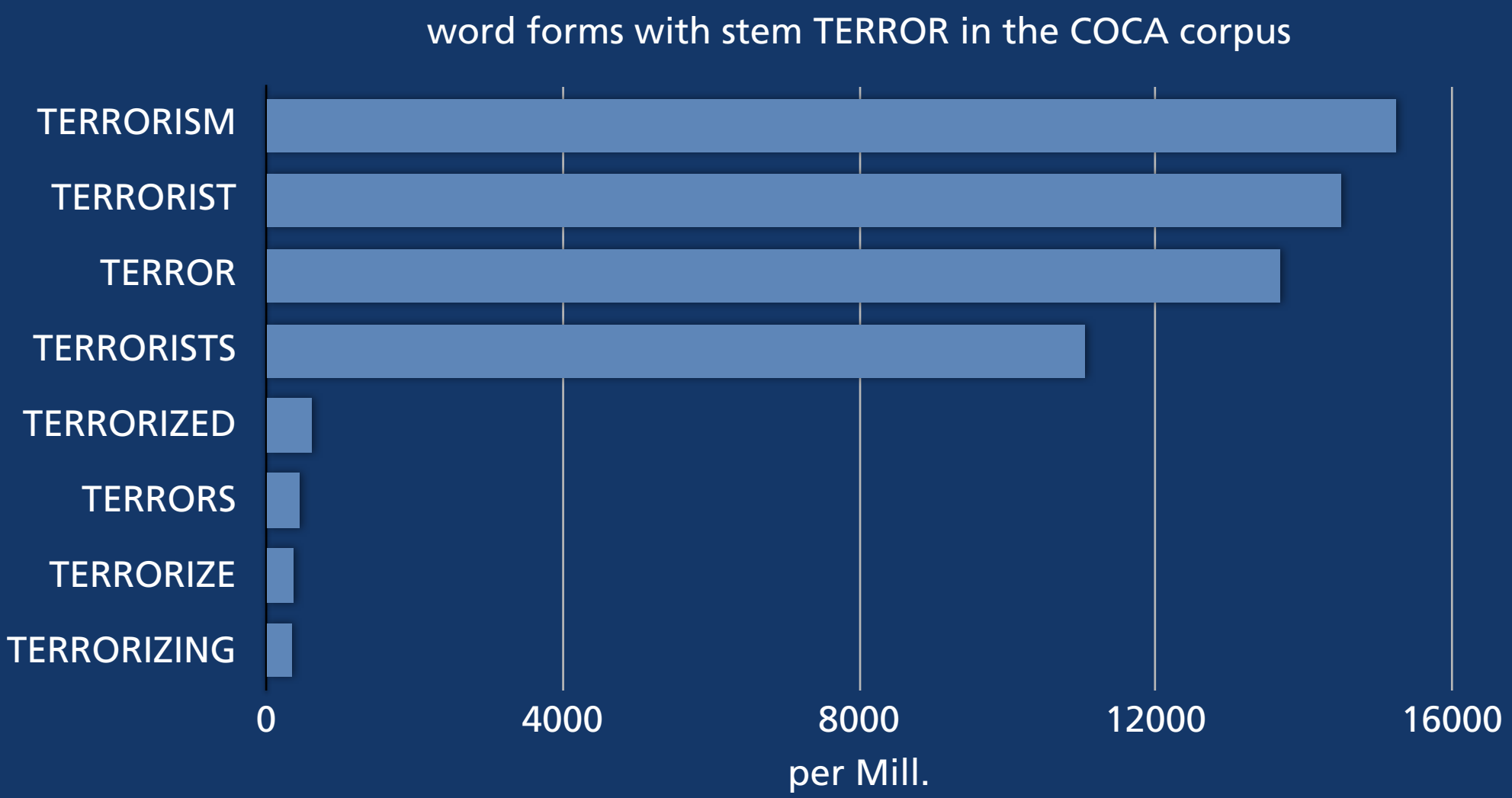
Advanced Corpus Linguistics



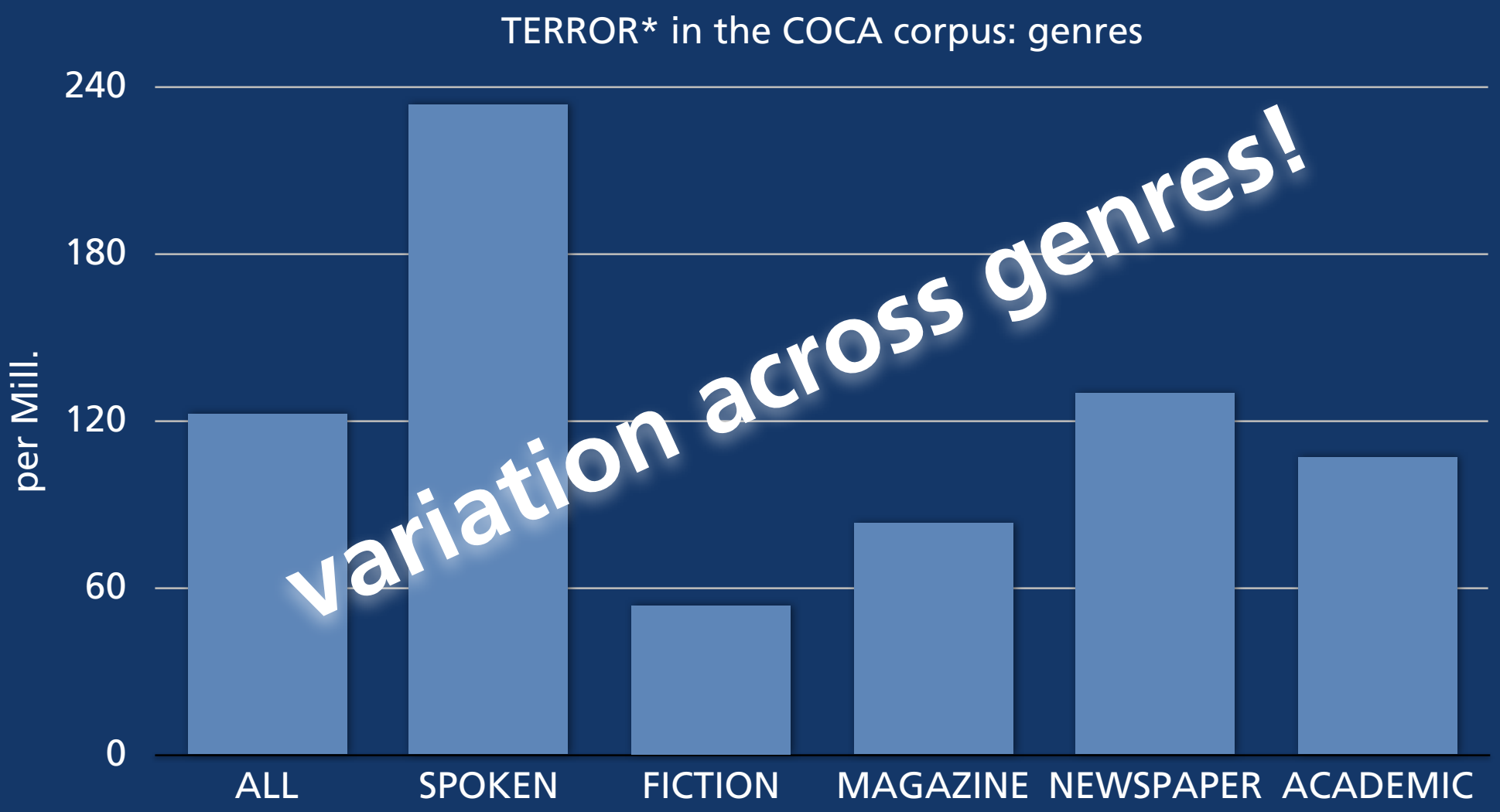
What are Patterns of Language Use?



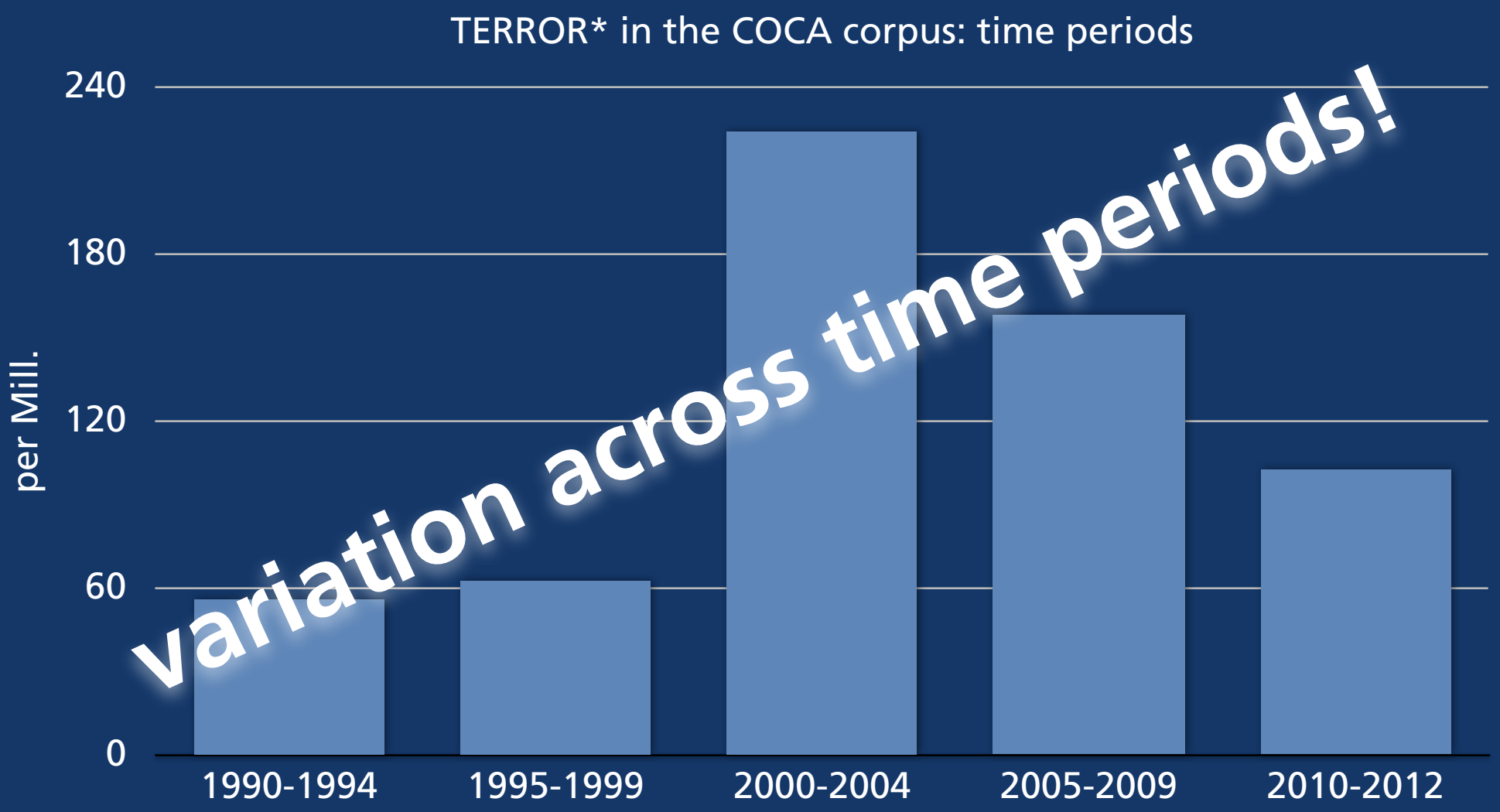
What are Patterns of Language Use?



What are Patterns of Language Use?



What are Patterns of Language Use?



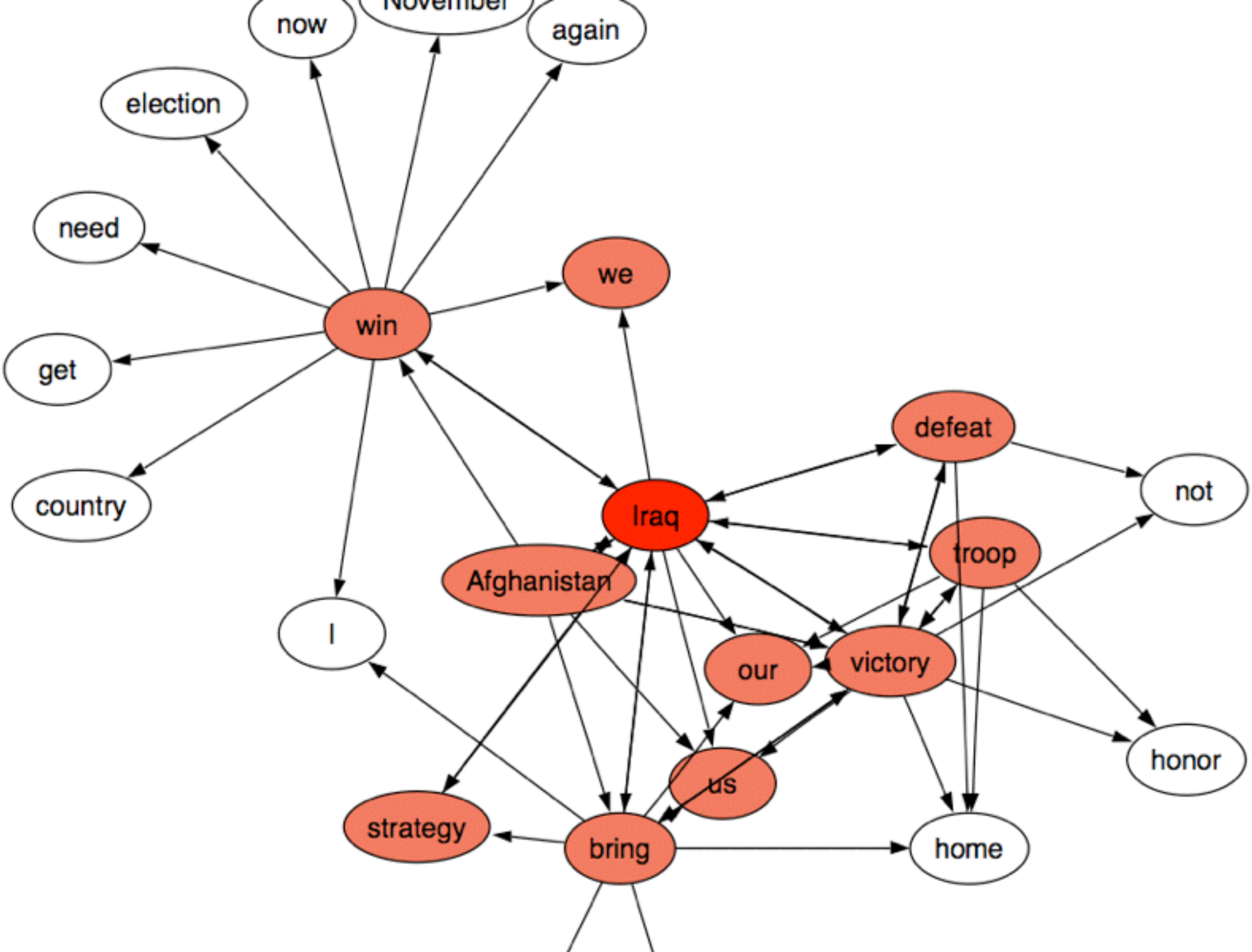
What are Patterns of Language Use?

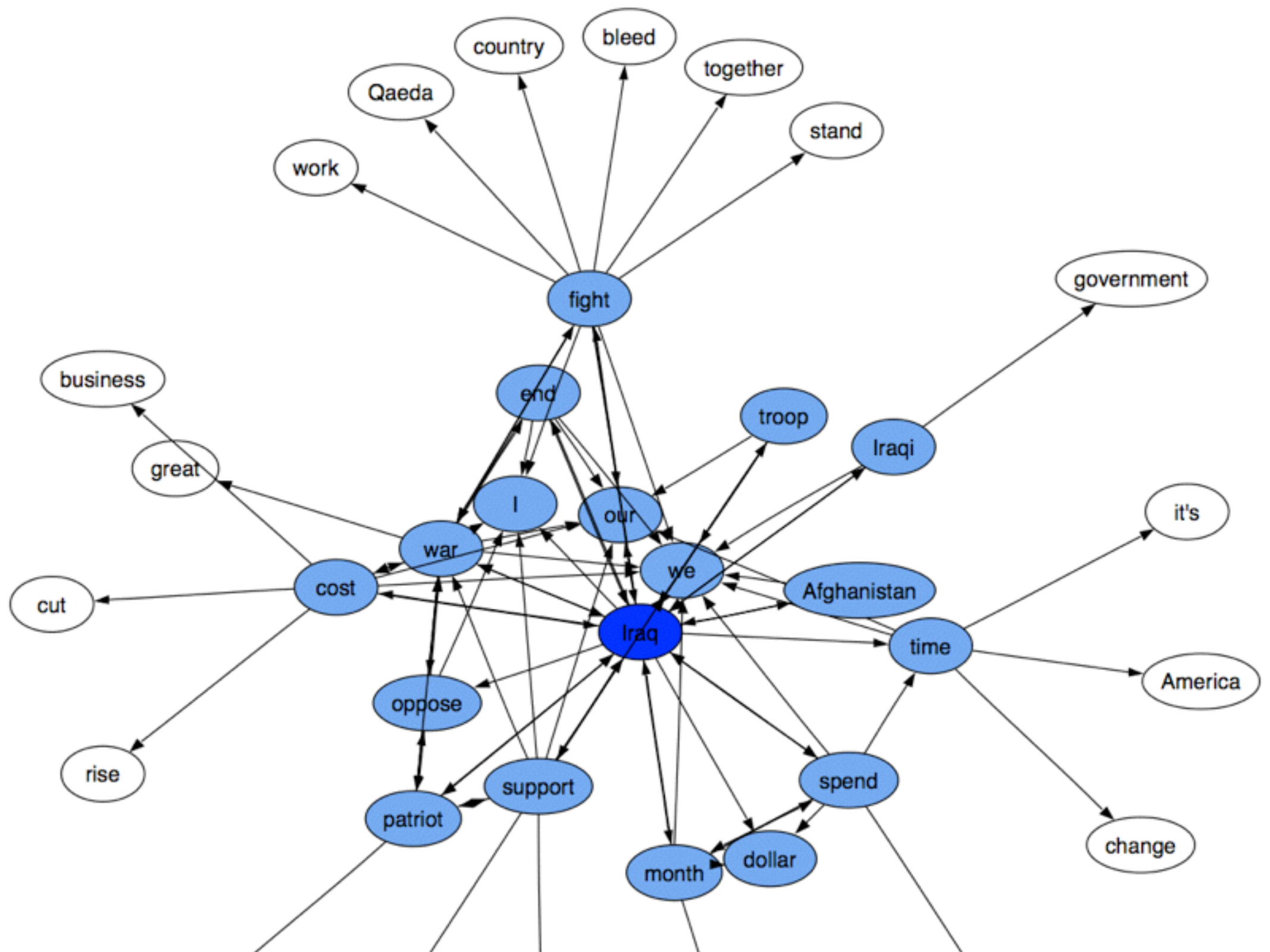
[HELP...]

WORD 2 (W2): **WAR** (13.03)

E		WORD	W2	W1	W2/W1	SCORE
4	1	II	14479	1	14,479.0	1,110.9
1	2	GULF	5981	1	5,981.0	458.9
3	3	VIETNAM	3986	1	3,986.0	305.8
3	4	POST-COLD	1130	0	2,260.0	173.4
7	5	CIVIL	11252	5	2,250.4	172.7
5	6	POST-WORLD	687	0	1,374.0	105.4
4	7	KOREAN	1342	1	1,342.0	103.0
3	8	PERSIAN	1296	1	1,296.0	99.4
3	9	1991	572	0	1,144.0	87.8
1	10	HERO	549	0	1,098.0	84.2
1	11	JAPAN	500	0	1,000.0	76.7
0	12	ERA	906	1	906.0	69.5
9	13	IRAN-IRAQ	450	0	900.0	69.1
3	14	DECLARATION	430	0	860.0	66.0
3	15	PRISONERS	810	1	810.0	62.1
3	16	KOREA	307	0	614.0	47.1
7	17	TUG	298	0	596.0	45.7
7	18	PRIOR	282	0	564.0	43.3
7	19	1812	281	0	562.0	43.1
7	20	ZONES	265	0	530.0	40.7

Obama vs. McCain 2008


now
November
again
election
need
win
we
get
country
defeat
not
Iraq
troop
Afghanistan
I
our
victory
honor
us
strategy
bring
home


country
bleed
Qaeda
together
stand
work
fight
government
business
end
troop
Iraqi
great
I
our
it's
war
we
cost
Afghanistan
cut
Iraq
time
America
oppose
rise
support
spend
patriot
change
dollar
month

Obama vs. McCain 2008

	Pronoun	Significance Level	Relative Frequency Factor
McCain	my	< 0.0001	1.66
	their	< 0.0001	1.19
	he	0.00016	1.26
	I	0.00776	1.08

	Pronoun	Significance Level	Relative Frequency Factor
Obama	we	< 0.0001	1.4
	you	< 0.0001	1.5
	us	< 0.0001	1.33
	yourself	0.00028	6.17
	they	0.03107	1.1

Leadership vs. Intellectual

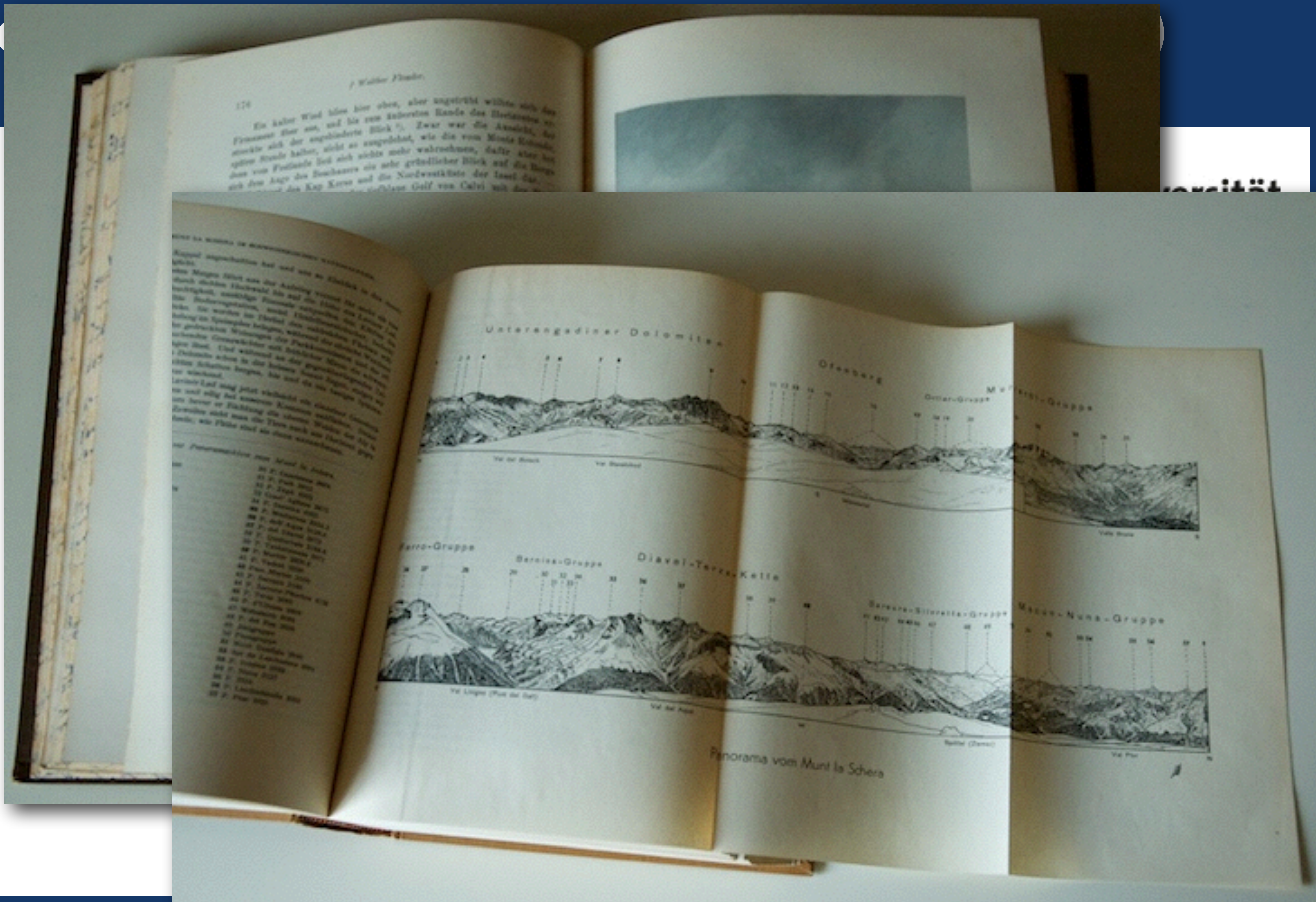
McCain

LLR	Phrase
12.21	Joe the plumber
10.68	We need to have
10.68	the floor of the (Senate)
9.16	So the point is
9.16	whether it be the
9.16	the that Senator Obama
9.16	Senator Obama wants to
9.16	the wealth around
7.63	because they have to
7.63	the point is that/I
7.63	in the United States
7.63	Well let me (just) say
7.63	I will [...] I will
6.11	the DC school system
6.11	let me just
6.11	the size of government
6.11	And not the fact
6.11	the fact that Senator Obama
6.11	to spread the wealth
6.11	will millions of jobs in
6.11	people in the United
6.11	And by the way
6.11	you wanted to spread

Obama

LLR	Phrase
17.58	going to be/have to
11.3	to make sure that
11.3	in order to give
11.3	I think that we
10.04	John let me [...] let
8.79	Senator McCain and I
8.79	and to the American people
8.79	And I think that
8.79	think going to be
7.53	when it comes to
7.53	I think important to
7.53	we are going to do is
6.28	important for us to
6.28	over the last eight years
6.28	to be important to
6.28	I just described what
6.28	I think [...] I think
6.28	the [...] of the president
6.28	the fact of [...] is
5.02	economy for the next
5.02	on policy on spending
5.02	and happen very often
5.02	to invest in the

variation across speakers!


Unterengadiner Dolomiten
Ofenberg
Bernina-Gruppe
Panorama vom Munt la Schera

Text+Berg Corpus

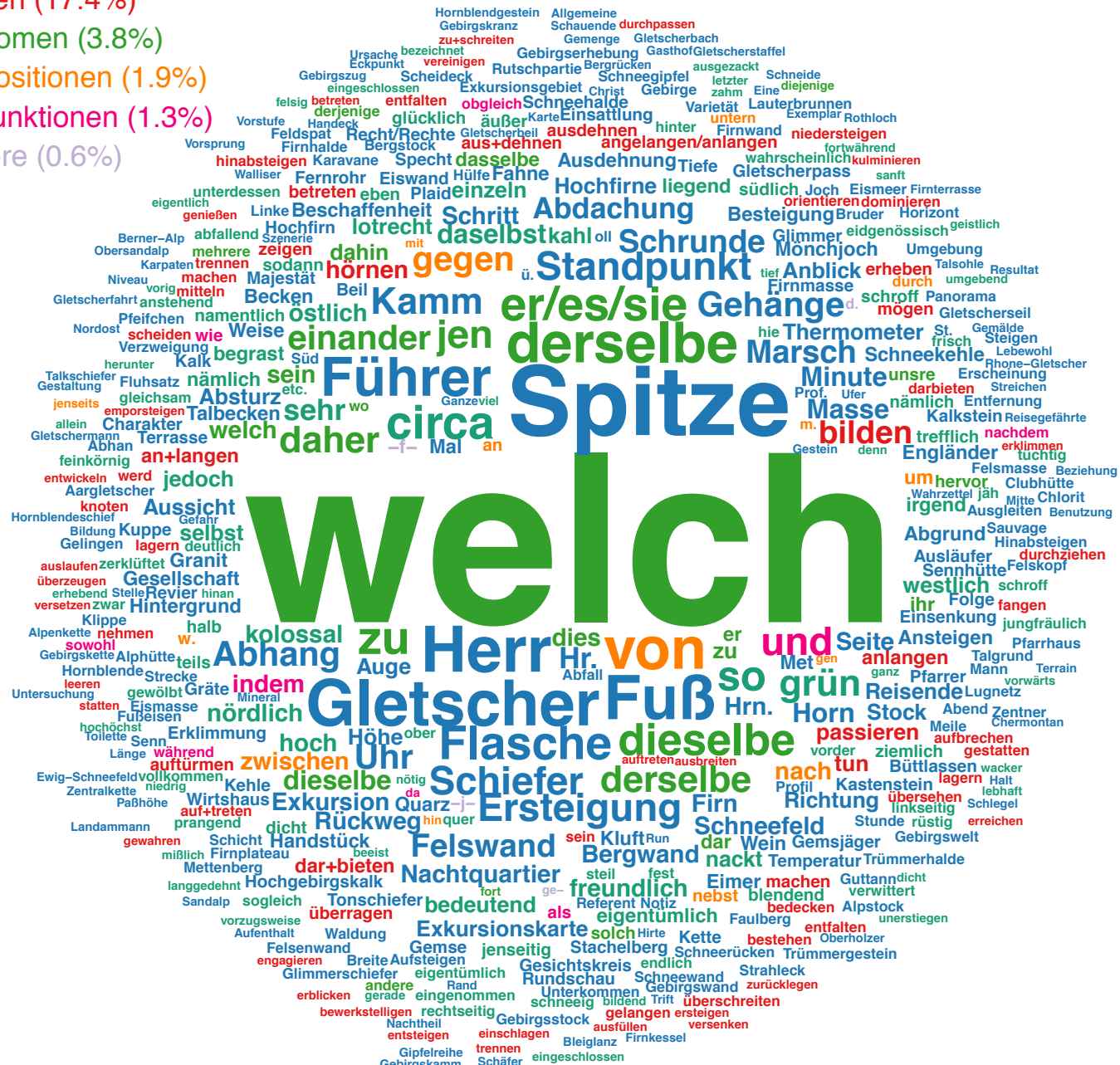
- 302 volumes of year books of:
  - Swiss Alpine Club, 1864 to today (ongoing)
  - British Alpine Club, 1969 to 2008
- 52.2 million words
- German, French, Italian, Romansh; English; partly translated (parallel corpus)
- In collaboration with Volk et al. UZH CL

Time Specific Vocabulary

Vocabulary which appears significantly more often in one period than in another.

Time Period	Tokens	Time Period	Tokens
1864 to 1879	1 035 187	1960 to 1979	1 015 849
1880 to 1899	1 541 063	1980 to 1999	393 811
1900 to 1919	1 146 313	2000 to 2011	207 260
1920 to 1939	1 120 492		
1940 to 1959	1 102 298	Total	7 562 273

Nomen (51.2%)
Adjektive/Adverbien (23.8%)
Verben (17.4%)
Pronomen (3.8%)
Präpositionen (1.9%)
Konjunktionen (1.3%)
Andere (0.6%)



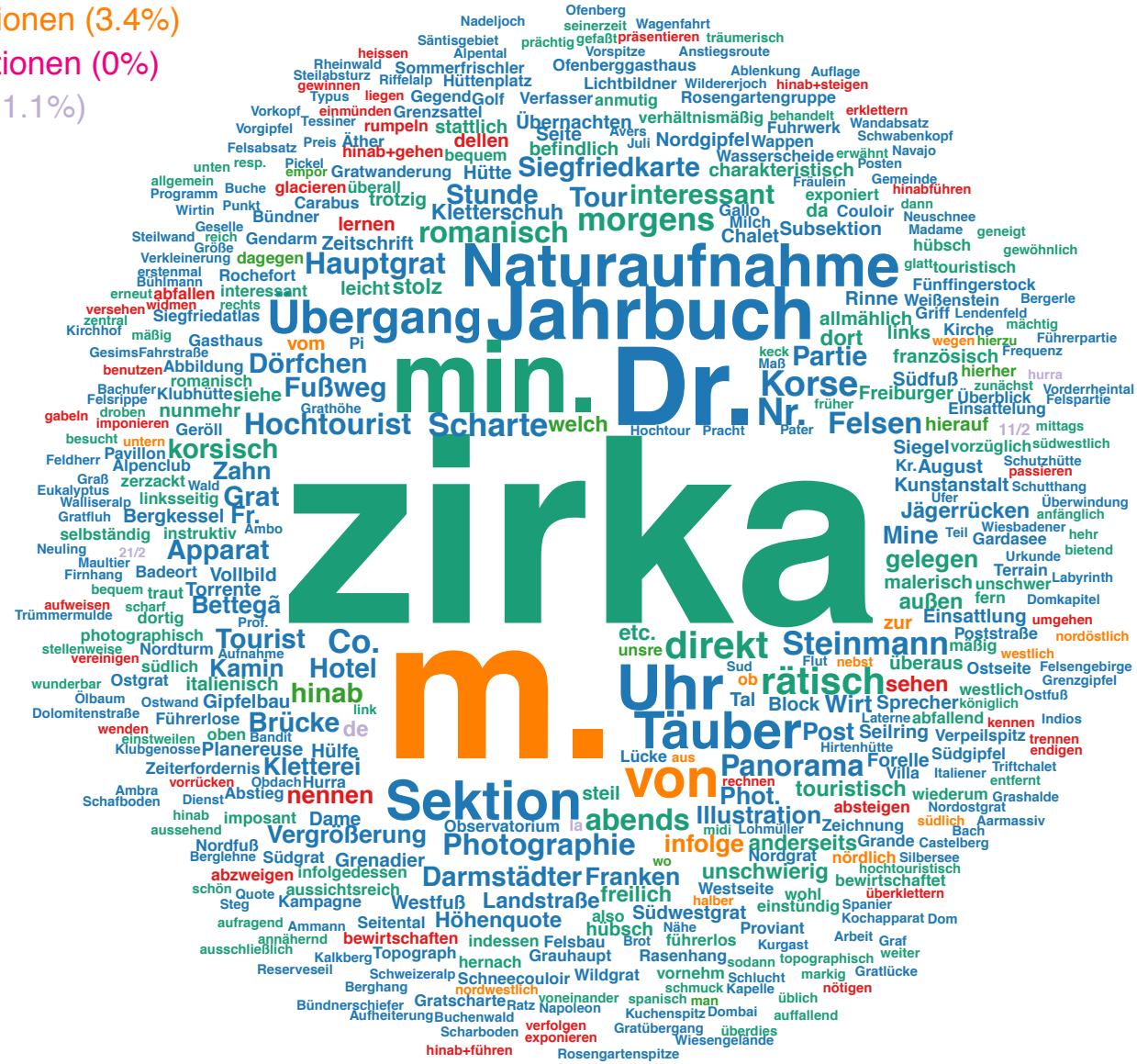
Typisches Vokabular

1860-1879

Nomen (46%)
Adjektive/Adverbien (28.5%)
Verben (15.6%)
Pronomen (5.2%)
Präpositionen (2.4%)
Konjunktionen (1.6%)
Andere (0.8%)

Typisches Vokabular
1880-1899

Nomen (58.4%)
Adjektive/Adverbien (26.8%)
Verben (8.3%)
Pronomen (2.1%)
Präpositionen (3.4%)
Konjunktionen (0%)
Andere (1.1%)



Typisches Vokabular

1900-1919

Nomen (56.6%)
Adjektive/Adverbien (19.4%)
Verben (18%)
Pronomen (3.4%)
Präpositionen (1.4%)
Konjunktionen (1%)
Andere (0.2%)

Typisches Vokabular

1920-1939

Nomen (57.4%)
Adjektive/Adverbien (23.2%)
Verben (12.6%)
Pronomen (4.4%)
Präpositionen (1.3%)
Konjunktionen (0.5%)
Andere (0.7%)

Typisches Vokabular

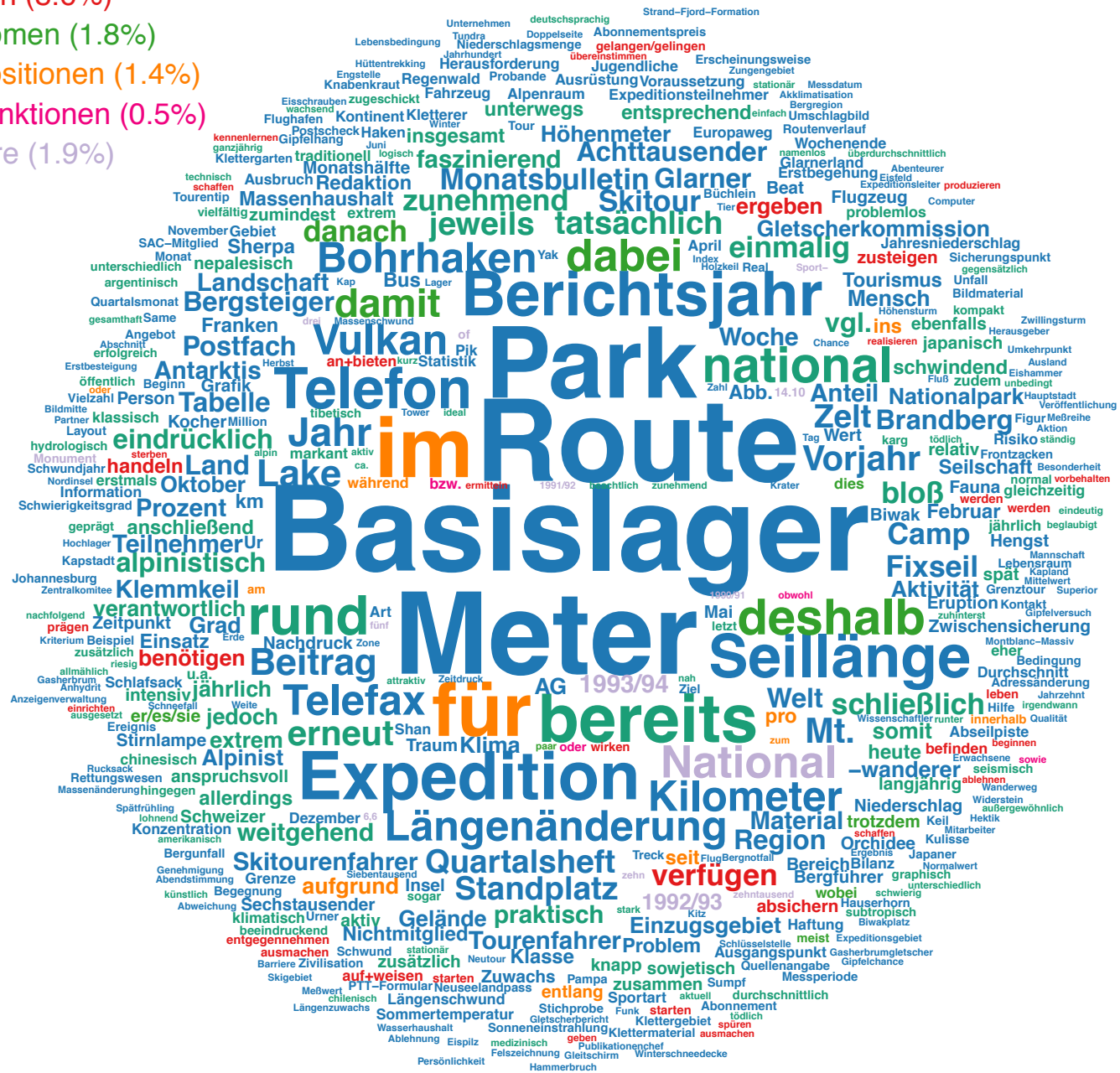
1940-1959

Nomen (56%)
Adjektive/Adverbien (20.9%)
Verben (14.3%)
Pronomen (4.3%)
Präpositionen (1.4%)
Konjunktionen (1.1%)
Andere (1.9%)

Typisches Vokabular

1960-1979

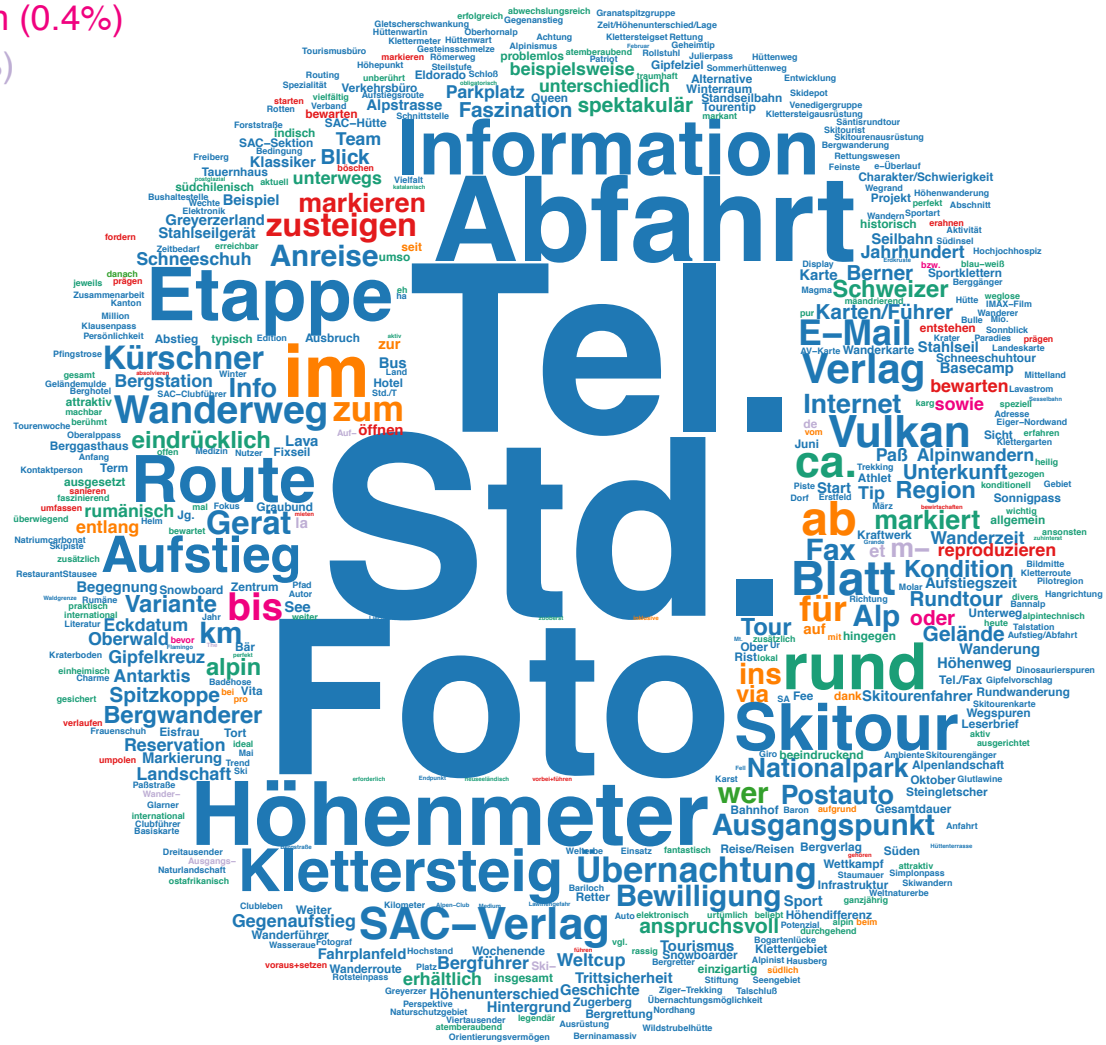
Nomen (62.2%)
Adjektive/Adverbien (23.4%)
Verben (8.6%)
Pronomen (1.8%)
Präpositionen (1.4%)
Konjunktionen (0.5%)
Andere (1.9%)



Typisches Vokabular

1980-1999

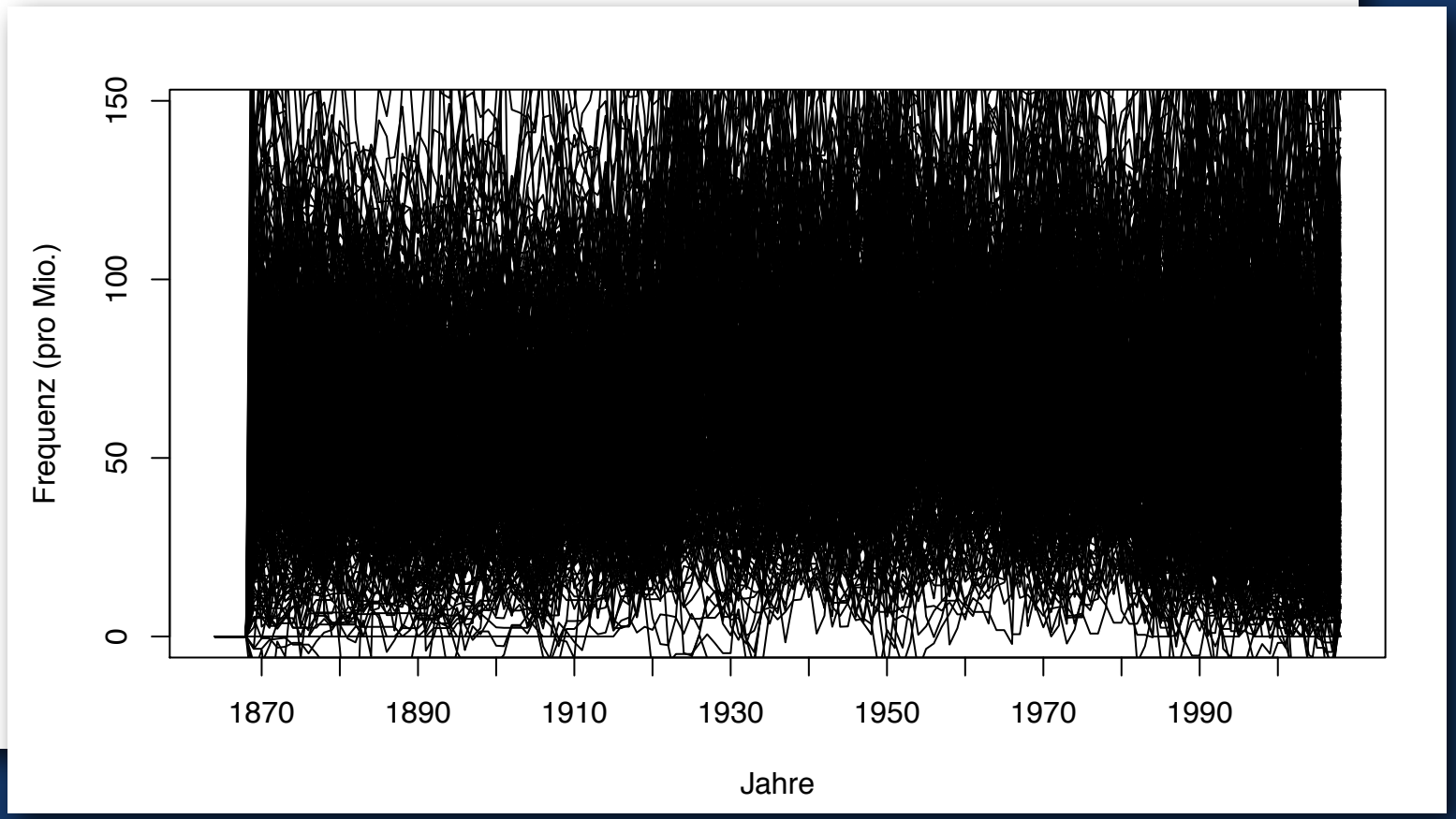
Nomen (71.2%)
Adjektive/Adverbien (18.6%)
Verben (5.2%)
Pronomen (0.9%)
Präpositionen (2.3%)
Konjunktionen (0.4%)
Andere (1.5%)

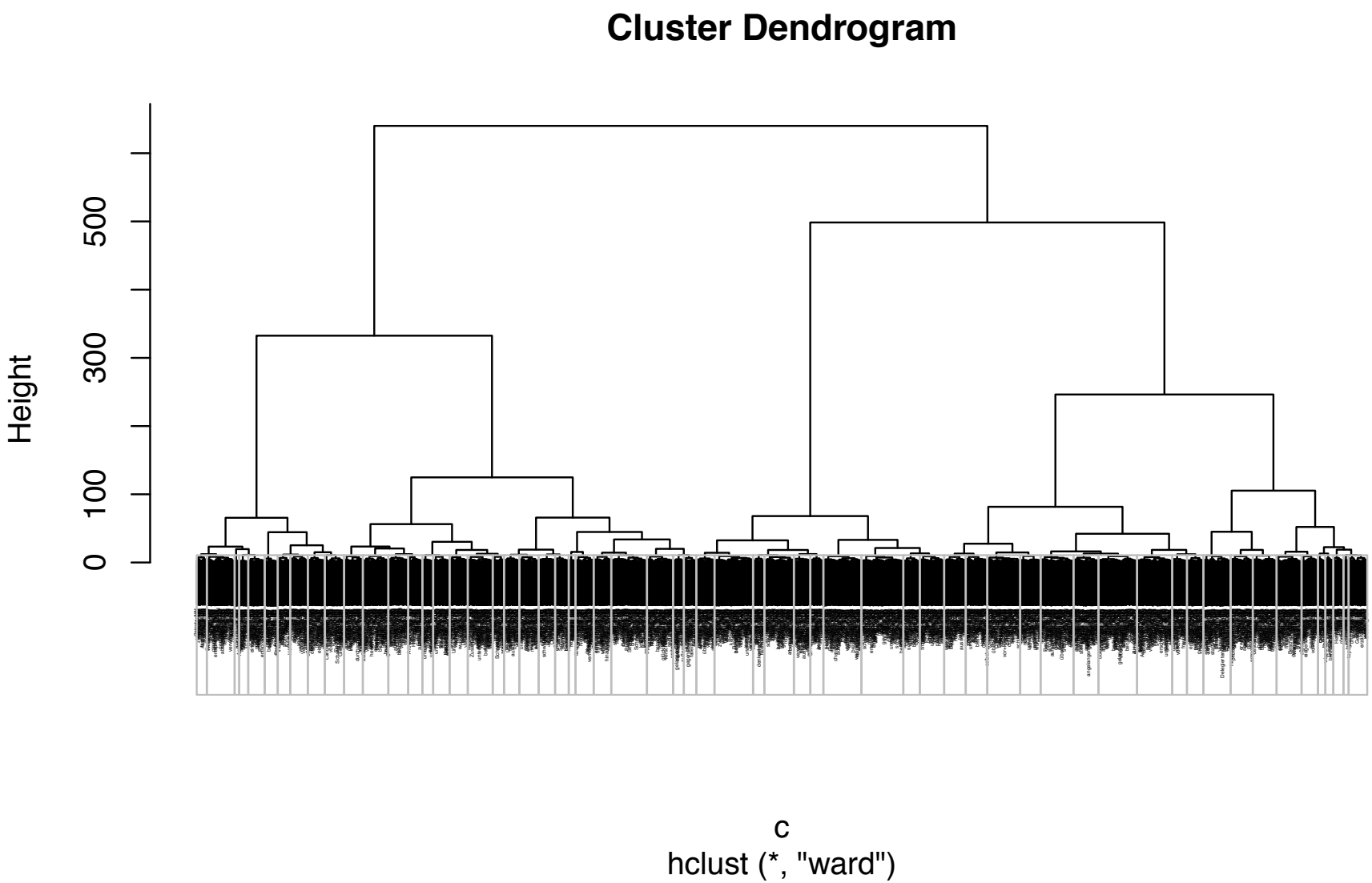


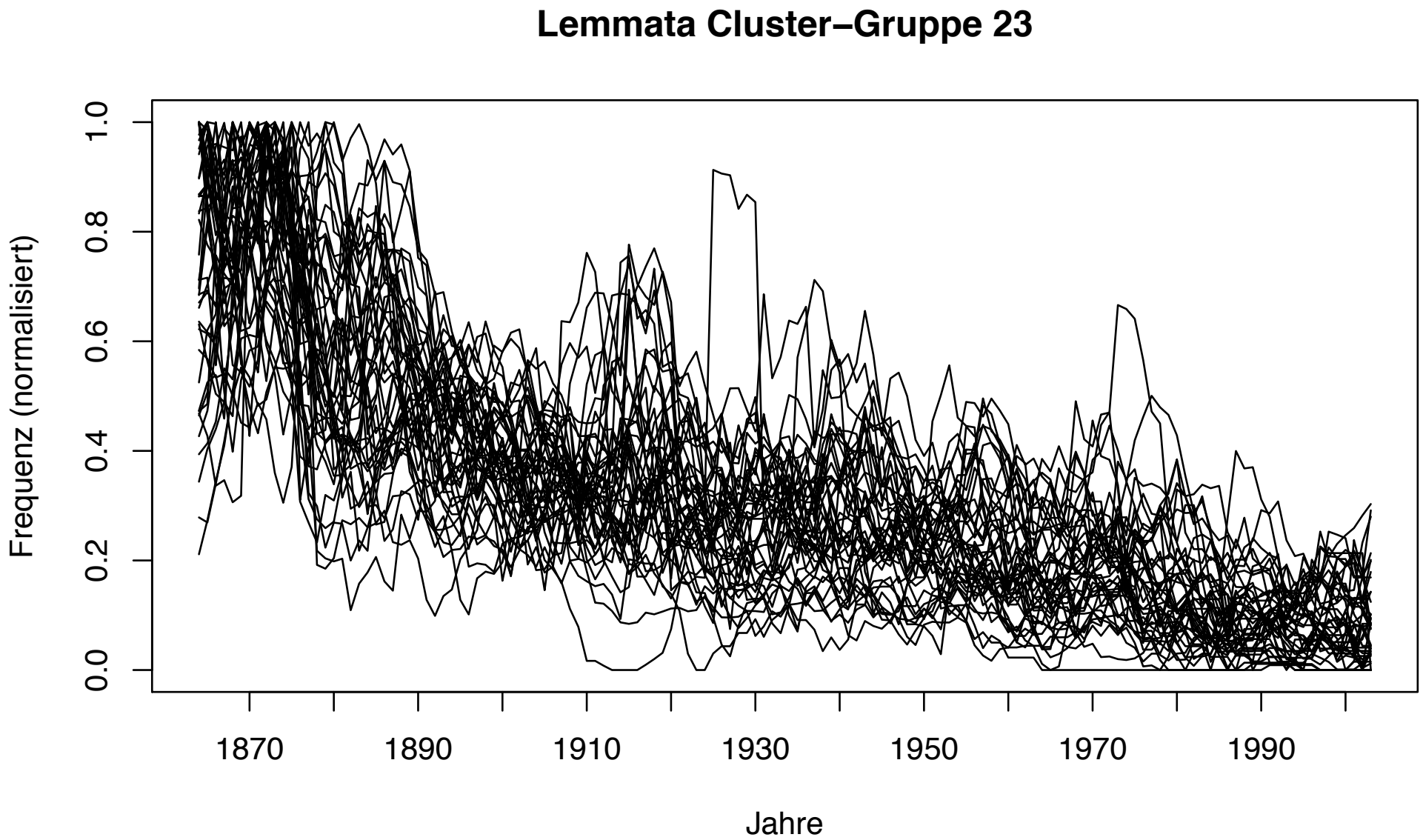
Typisches Vokabular
2000-2011

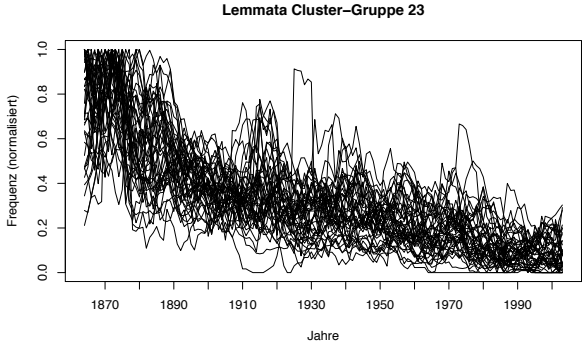
Data Driven Calculation of Time Specific Vocabulary

- no
- per
  (no
- sel
  me
  200
- clu
  sim
  - I
  - 60




Cluster Dendrogram
Height
0
100
300
500
c
hclust (*, "ward")

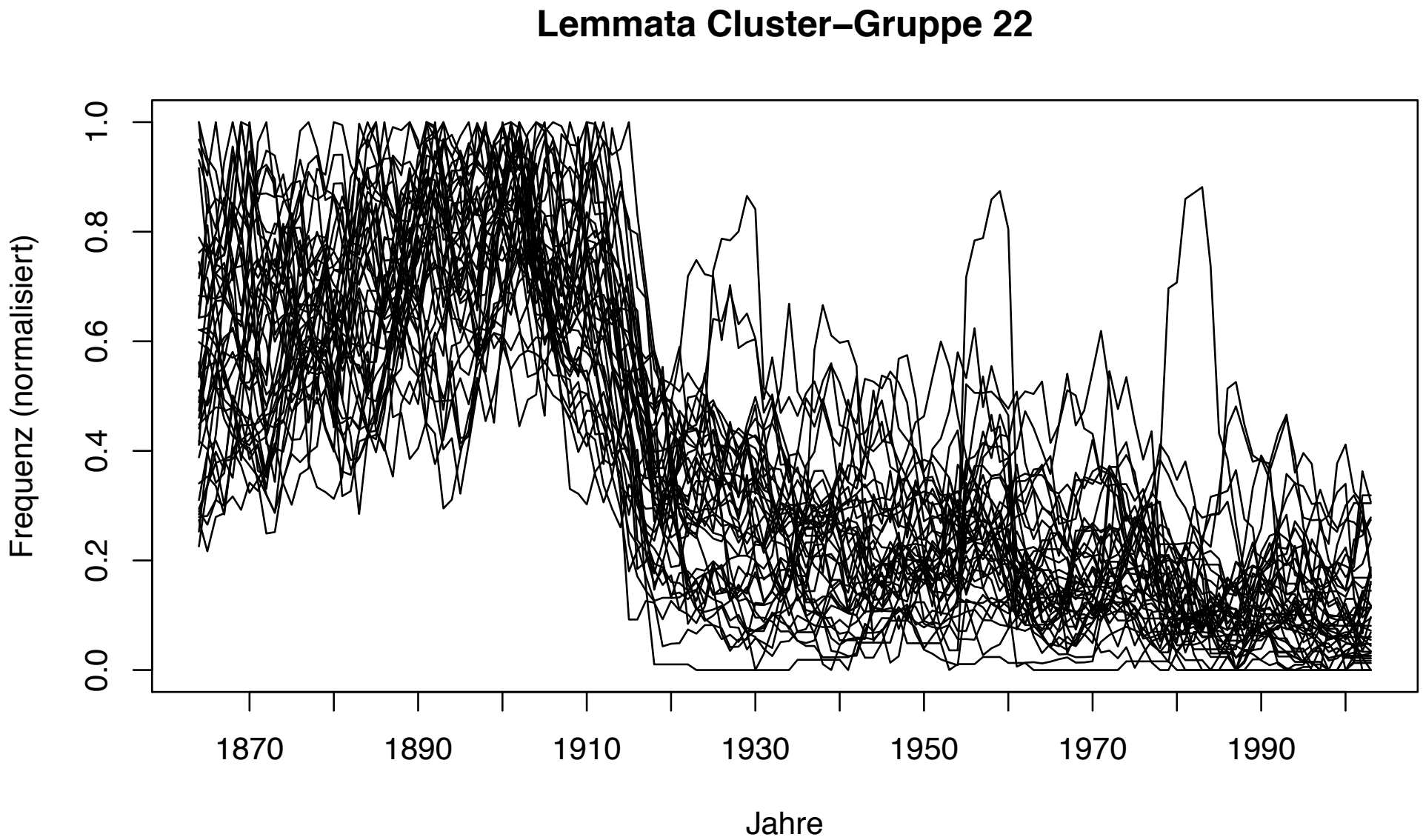

Lemmata Cluster−Gruppe 23
Frequenz (normalisiert)
0.0
0.2
0.4
0.6
0.8
1.0
1870
1890
1910
1930
1950
1970
1990
Jahre

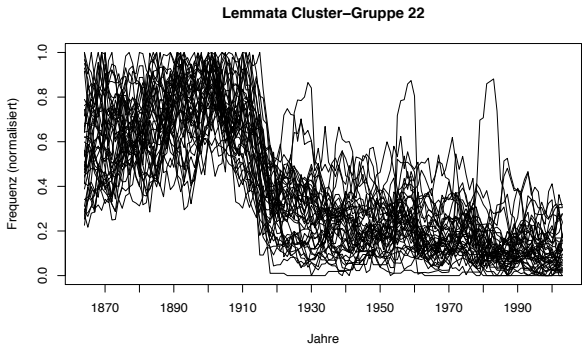


Wortwolke Gruppe 23

passieren
Einschnitt
Standpunkt
Joch
an+langen
hochöchst
namentlich
Besitz
zerrissen
bekanntlich
Sitte
freundlich
jetzig
teils
Spitze
jenseitig
gleichsam
Horn
nackt
Notiz
Ausläufer
kahl
grün
lagern
vorig
irgend
reich
reich
Stellung
Haupt
Ganze
Alpenrose
hie
nieder
Ortschaft
Strom
Gräte
Trümmer
Entfernung
daselbst
Kette
Wasserscheide
Beschaffenheit
österreichisch

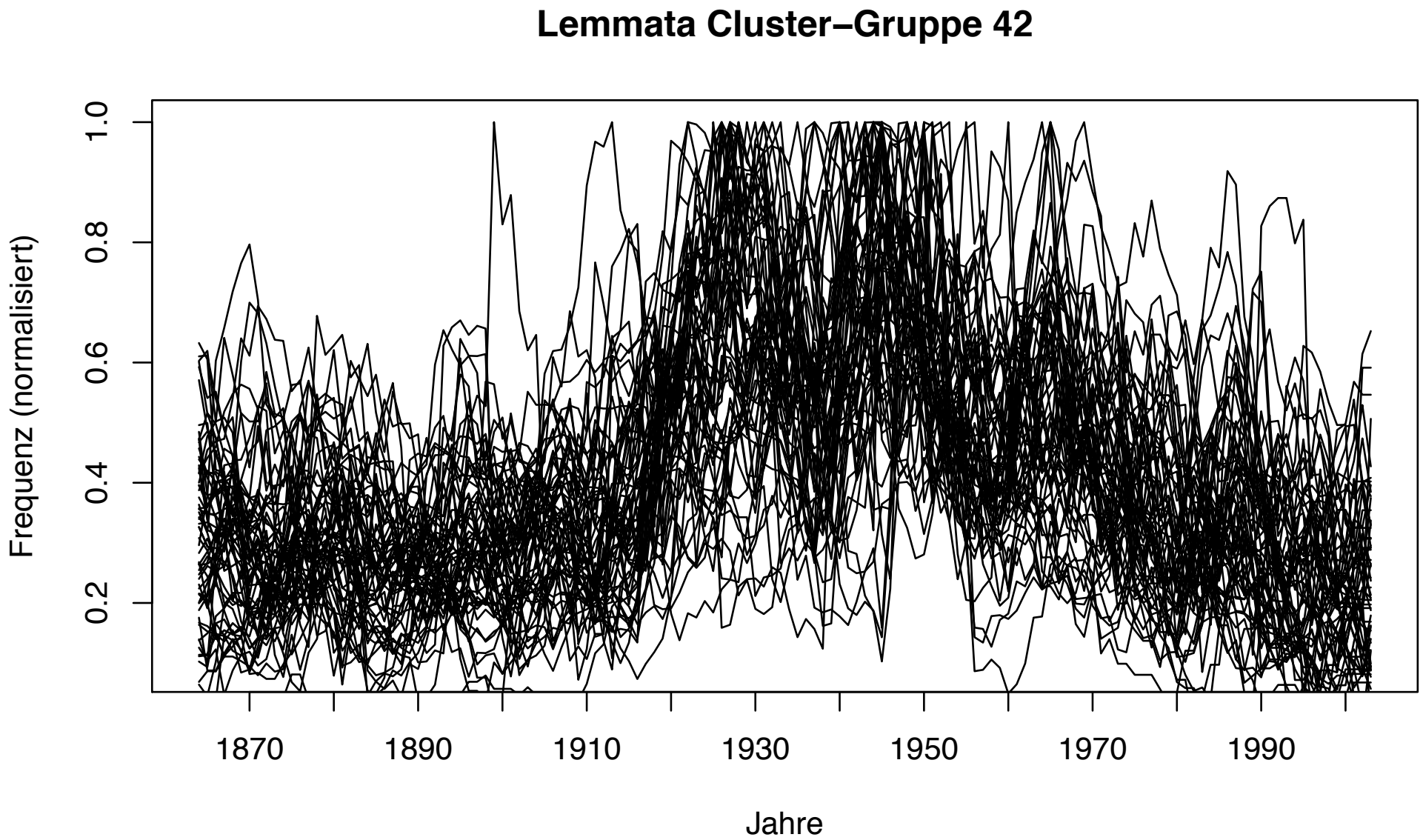
Nomen (21) Adjektive (20) Verben (3) Personalpronomen (0)

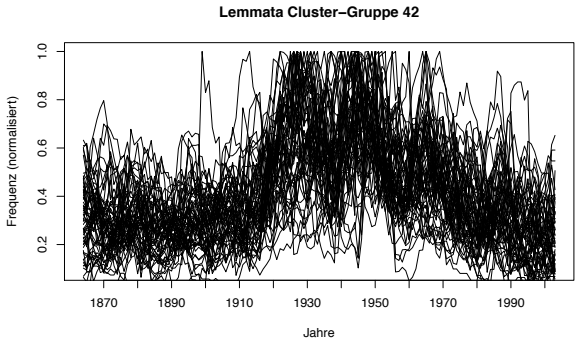

Lemmata Cluster−Gruppe 22
Frequenz (normalisiert)
0.0
0.2
0.4
0.6
0.8
1.0
1870
1890
1910
1930
1950
1970
1990
Jahre



Wortwolke Gruppe 22

Nomen (19) Adjektive (14) Verben (9) Personalpronomen (0)

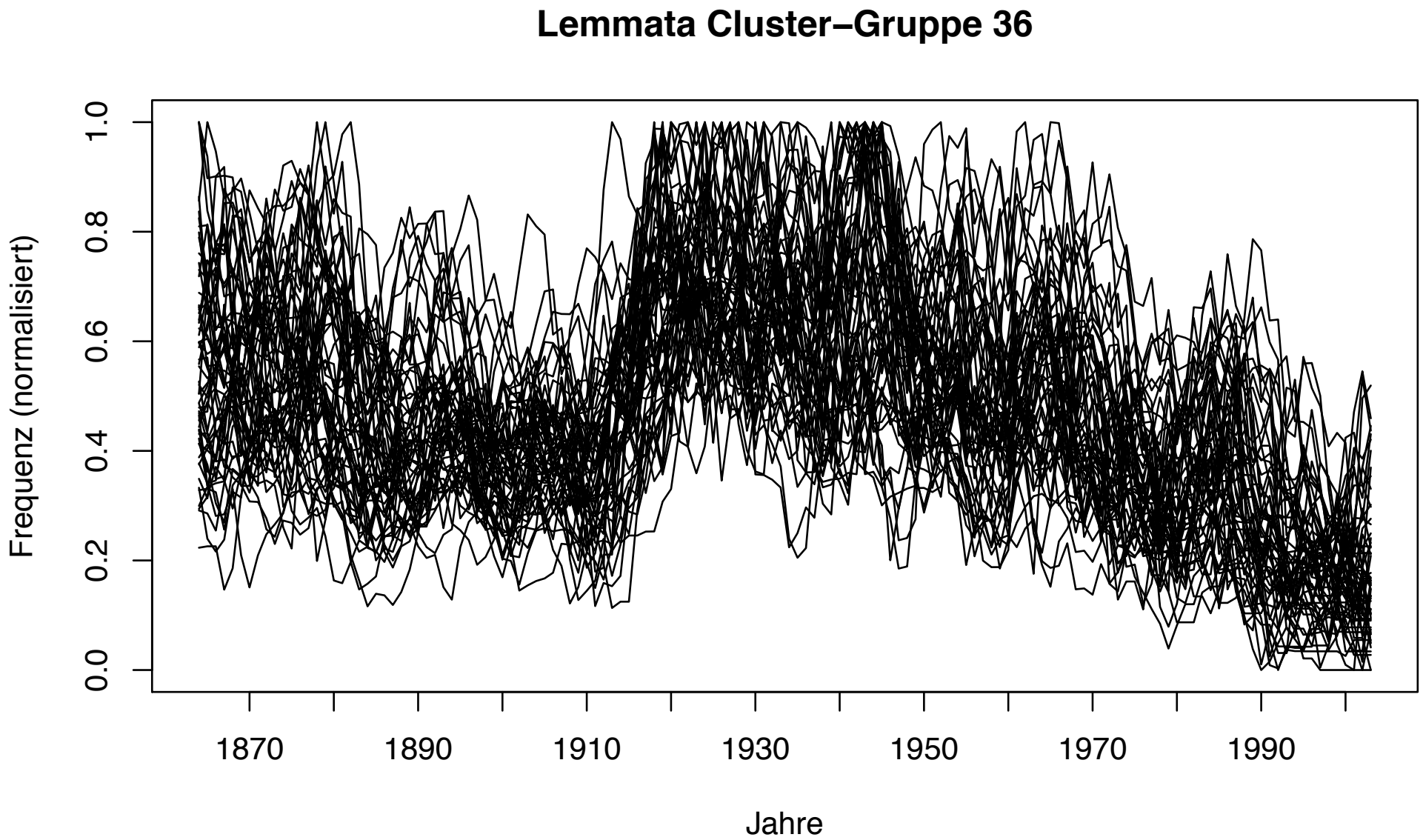

Lemmata Cluster−Gruppe 42
Frequenz (normalisiert)
0.2
0.4
0.6
0.8
1.0
1870
1890
1910
1930
1950
1970
1990
Jahre

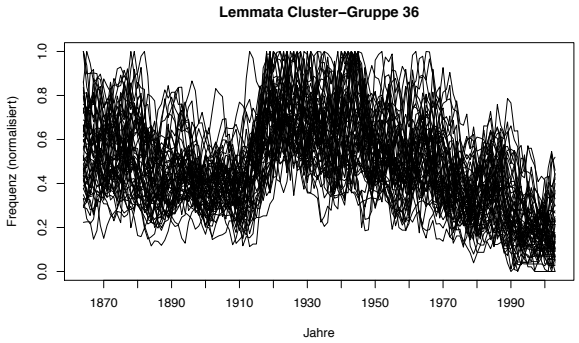


Wortwolke Gruppe 42

greifen
Liebe brennen
Schneesturm
Blume Bewußtsein gesund
dunkel Wächte Tief
grüßen hängen Schein Glas
Kampf Ohr Forschung Wurzel
Einzelheit Fahrt
Führung ragen Licht Bergfahrt
Höhle stecken luftig fein still unbekannt
Kälte schauen inner
geistig eisig Fortsetzung Gott
Finger Herz Sturm Blau golden
Fenster lesen Stern leuchtend
Unterlage müde
kriechen eng sinken
gleiten Lawine Geist
Wesen Tempo Bruch
reißen lösen Aufnahme
Schneemasse geschlossen
Pulverschnee

Nomen (36) Adjektive (14) Verben (13) Personalpronomen (0)


Lemmata Cluster−Gruppe 36
Frequenz (normalisiert)
0.0
0.2
0.4
0.6
0.8
1.0
1870
1890
1910
1930
1950
1970
1990
Jahre



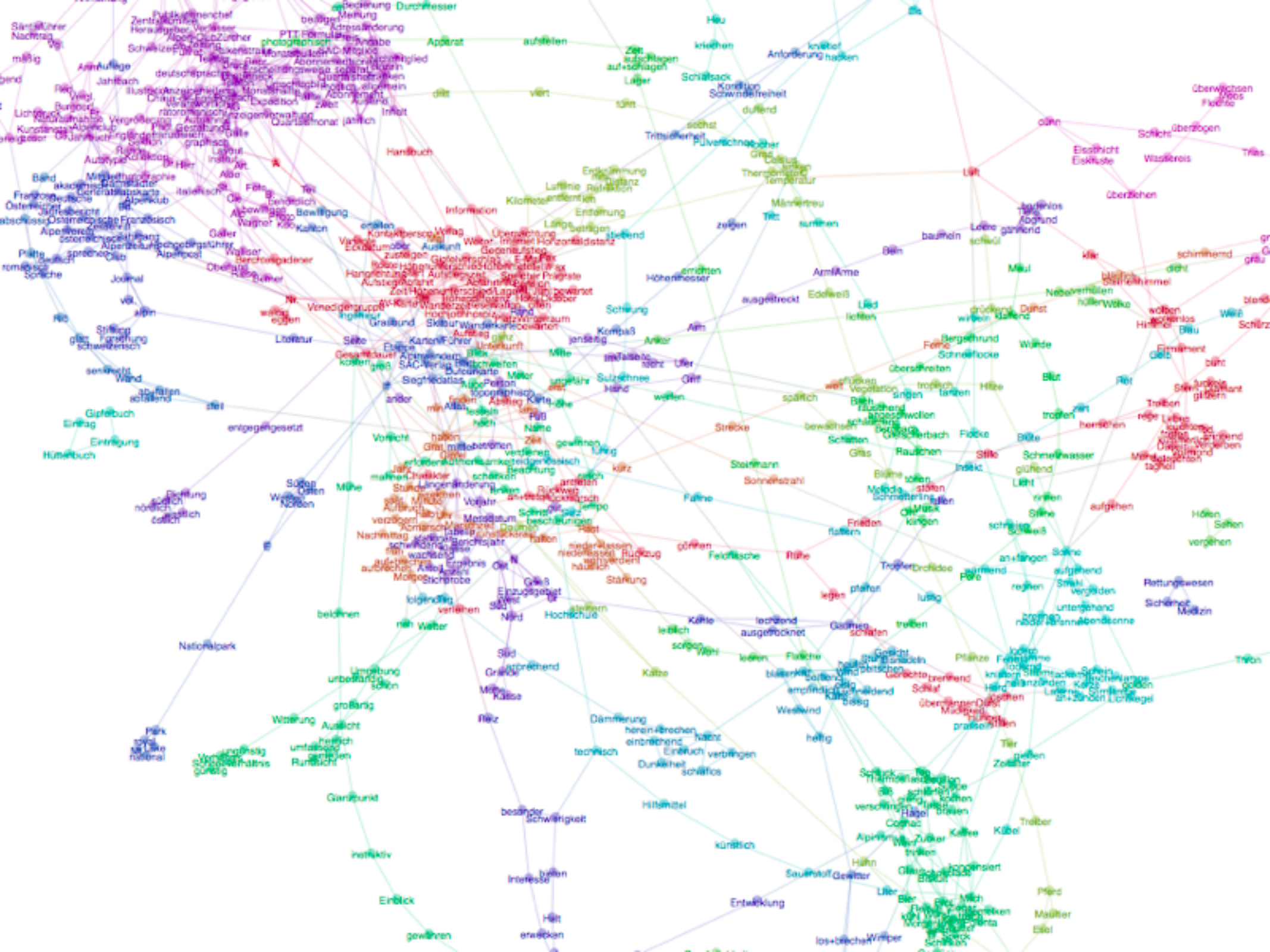
Wortwolke Gruppe 36

Schönheit Schlag
braun licht
Skizze
Lust heben
überwinden
Flucht
sagen blank fassen Luft
Schatten selbstverständlich
köstlich lösen treiben Sprung
Arm dringen schlagen
drängen schlagen
bloß ruhig Sieg
eilen Tat hell ziehen
Feuer Betrachtung Spalte Bewegung
Mond arm vermag dunkel hell nächtlich
Kante hinweg ewig Kreuz
weich Glied Dämmerung vorn
weisen Wirkung Körper flach
vor+kommen Eile
wolkenlos wiederholt
Entschluß

Nomen (24) Adjektive (19) Verben (15) Personalpronomen (0)

Collocation Networks

- calculation of collocates to all words which are typical for the corpus (compared to reference corpus)
- calculation of the 2nd order collocates
- visualization as collocation network
- detection of node clusters with significant interconnections



Complex Patterns (n-Grams)

- Aim: Finding typical formulations
- Operationalization:
  - n-grams:
    as you know
  - complex n-grams:
    as [personal pronoun] [verb]
  - frequency calculation of all combinatory possible complex n-grams of length *n* (3-5)
  - comparing frequencies of n-grams between time periods → getting time specific n-grams

Complex Patterns (n-Grams): Corpus A (1880–1899)

- **ADV erreichten PPER ART NN**
  endlich erreichten wir den Aaresattel
  (Bald) nachher erreichten wir den Guggistafel
  Nun erreichten wir das Gebiet (des Kalkfelsens)
- **VVFIN APPR CARD Uhr**
  stiegen um 12 Uhr
  erreichten um 2 Uhr
  verließen um 3 Uhr
- **an der ADJA NN des**
  an der linken Seite des
  an der rechten Seite des
  an der anderen Seite des
  an der breiten Wand des

Complex Patterns (n-Grams): Corpus A (1880–1899)

- **APPR ART Nähe ART NN**
  in der Nähe des Gipfels
  in der Nähe der Grenze
  in der Nähe des Muttensees
- **ADJA Weg APPR ART NN**
  alten Weg über den Feegletscher
  anderen Weg auf das Gabelhorn
  ausgetretenen Weg durch den Moränenschutt
- **APPR ART Führer NE NE**
  mit den Führern Alois Pinggera
  neben dem Führer Alphons Supersaxo
- **Dr. NE NE // APPR Prof. Dr. NE NE**
  Dr. Emil Burckhardt
  von Prof. Dr. K. Schulz
  von Prof. Dr. G. Meyer

Complex Patterns (n-Grams): Corpus B (1930–1949)

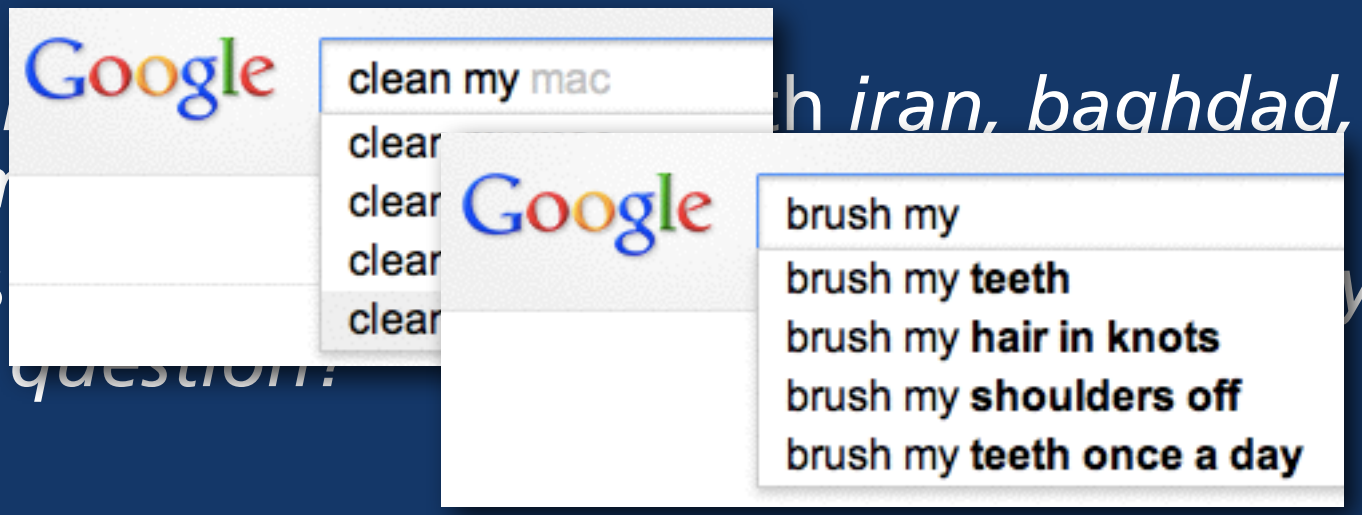
- **ADV VVFIN ART ADJA NN .**
  Draussen erwachte ein neuer Tag .
  Dann kam ein trüber Tag .
  Nun naht das schwierigste Stück .
- **ADV VVFIN PPER VVINF**
  So lasst uns eilen
  jetzt heisst es handeln

Complex Patterns (n-Grams): Corpus B (1930–1949)

- **dann VVFIN ART NN**
  dann VVFIN ART ADJA
  dann VVFIN PPER auf
  dann kündete die Gipfelglocke
  dann geschieht das Wunder
  dann folgt ein heikler
  dann standen wir auf (dem kühnen Gipfel)
- **KOUS PPER APPR ART NN VVFIN**
  KOUS PPER ART NN VVFIN
  Wie ich in den Riss einstieg
  als wir in der Gabel anlangten
  Bevor wir in das Couloir hinübersteigen
  während wir der Hütte zustrebten
  während wir die Steigeisen ablegten
  Als wir die Passhöhe erreichten

Advanced Corpus Linguistics

- Data mining and corpus linguistics are able to find patterns in big data.
- What are the reasons for the existence of these patterns?
  - grammar: ADJ before NOUN
  - idiomaticity: *brush my teeth* (instead of *clean my teeth*)
  - semantics: *…h iran, baghdad, mesopotam…*
  - pragmatics: *…y I ask you a question?*



Advanced Corpus Linguistics

- What are the reasons for the existence of these patterns? (continued)
  - social/cultural/historical settings:
    - in the U.S. presidential debates 2008, *iraq* goes together with *afghanistan, war, costs, win* etc.
    - they way we write about climbing mountains

Universität Zürich[UZH]

Noah Bubenhofer

Working with Specialized Corpora

Conclusion: Specialized Corpora

- Language use sometimes greatly differ between text types.
  - Example: Use of „I“ in english texts (press vs. informal letters)
  - Use of expressions like „friendly fire“ in time
- Differences sometimes can only be seen, if frequencies of the phenomena are compared in different corpora.
- So:
  - either compile a corpus containing different text types, genres etc. → balanced corpus
  - or just work ad hoc with different corpora
- Metadata about each match is important!

Balanced Corpus

- Aim: Representing different genres, text types, time periods etc. in one corpus.
  - BNC: 90% written, 10% spoken
  - DWDS: 21% functional literature, 28% poetry and fiction, 23% science, 27% press
- Typical for general purpose „national corpora“.
- Problem: Which genres, text types etc. should be included? How to define the mix?
  - No general rule; what should it be derived from?
  - Highly dependent on the research questions!

„Ad hoc“ Corpus Composition

- Idea:
  - Work with different corpora.
  - Start with your research needs.
  - Compose your corpus collection in dependency of your research questions.
- Example DeReKo Deutsches Referenzkorpus (IDS Mannheim):
  - collection of a great variety of corpora
  - user should build a „virtual corpus“ containing the desired corpora

Research- / Reference Corpus Design

- Idea of typical settings in social science: research group and control group
- Research corpus: Specialized corpus representing language use of text types, genres, time span etc. you are interested in.
- Reference corpus: Corpus unspecific concerning the language use you are interested in.
- Example – RQ: linguistic features of informal letters.
  - research corpus: informal letters
  - reference corpus: formal letters; letters in general; mix of letters, prose, fiction…

Tutorial Session 1

- Doing corpus research, mainly comparing vocabulary use over time.
- We work with the english Text+Berg corpus of the Alpine Journal.
- Please go to www.bubenhofer.com/tbilisi/ ➔ Part 1
- Username: guest[your number] e.g.: guest03
- Password: guest[your number] e.g.: guest03

Noah Bubenhofer

Building (Specialized) Corpora

Turning Text into a Corpus

No	Filename	Solution 451 to 500	Page 10 / 106
451	1968_de_a28.a28-s188	suchen den Zugang zum verwunschenen Schloss ... Der Tarodalt ist einer der_ART	Berge , wie es viele gibt im Hogg
452	1905_mul_a17.a17-s207	Gehänge , aber jenseits meldet sich im Larkiowald schon der Einfluß der_ART	Berge . Wo der Fluß eine kleine T
453	1997_de_a118.a118-s48	Gegend des Wadi Rum und wurden von Sheik Hamdan auf einige der_ART	Berge geführt . Im Herbst 1984 w
454	1899_mul_a11.a11-s45	auf das Fogarascher Gebirge angewandt . Wir greifen demnach speciell einige der_ART	Berge des Burzenlandes und den
455	2002_de_a280.a280-s59	der Entwicklung ein Profil zu geben , die weitere Einrichtung der_ART	Berge gemeinsam und zurückhalt
456	1949_mul_a19.a19-s143	Blütenlaube klingt leise aus Frauenmund ein altes Liebeslied . O Einsamkeit der_ART	Berge , was bist du gegen die Eins
457	1952_mul_a70.a70-s14	solcher Stunde waren wir schweigsam , umsponnen vom Gefühl der Einsamkeit der_ART	Berge . Wir allein bewegten uns o
458	1987_de_a13.a13-s126	einigen Umwegen entdecken wir Klöster , die , in der Einsamkeit der_ART	Berge verborgen , von der Kultur
459	1961_de_a13.a13-s103	Entwicklungen aber kann man nie zurückschrauben , wie auch die Einsamkeit der_ART	Berge unwiederbringlich ist . Dies
460	1978_de_a17.a17-s368	aus längst vergessenen Schauerromanen . Drei Männer gehen in die Einsamkeit der_ART	Berge , zwei kommen zurück , ein
461	1980_de_a6.a6-s83	Künder des extremen Alpinismus , Leo Maduschka : « O Einsamkeit der_ART	Berge , was bist du gegen die Eins
462	1968_de_a35.a35-s15	allein direkt zum Gipfel aufzusteigen versuchen . In der stillen Einsamkeit der_ART	Berge beginnt er die Kletterei , do
463	1889_mul_a11.a11-s4	die Stille der Wüste zu begeben und uns in die Einsamkeiten der_ART	Berge und Thäler des Sinai zu ver
464	1913_mul_a20.a20-s9	berührt sympathisch und ist vertrauenerweckend . Neu ist die Einteilung der_ART	Berge und Pässe , nicht in orograp
465	1889_mul_a46.a46-s6	Alter , der petrographischen Beschaffenheit , sowie der Schnee- und Eisbedeckung der_ART	Berge , und die Gletscherbeobach
466	1978_de_a9.a9-s184	zu wollen . Der herrlichste Sternenhimmel wölbte sich über den Eiskronen der_ART	Berge . Kein Wölkchen war am H
467	1930_mul_a36.a36-s217	nach einigen Minuten Ansteigens wieder mitten im Kampf mit den Elementen der_ART	Berge , der Kälte , dem Sturm , de

Data Processing

- Preprocessing: Scanning, OCR text recognition, transcribing…
- Processing: Turning the digital data in a useful format.
- Different data types:
  - unstructured data:
    text without any (or with no useful) meta data
  - semistructured data:
    text with xml, sgml, html annotations
  - (structured data:
    text in data bases)

Noah Rubenhofer

Unstructured Dat

Dr. **Heinrich Dübi.**

Die Vispertaler Sonnenberge.

Von *F. G. Stehler,*
Zürich.

(Nachdruck verboten.)

Die Vispertaler Sonnenberge.

Das Wallis. •— Lötschbergbahn. — Die Rarner Schattenberge.

as Thema führt uns wieder in das gottgesegnete Wallis. Das' Wallis ist ein ganz eigenartiges Land; es hat ein eigenes Klima, eigene Bodenverhältnisse, eigene Flora und eine eigenartige Pflanzenkultur. Aber auch der Viehstand und das Volk sind eigenartig; dieses hat eine eigene Geschichte, die an Ruhm und Heldenhaftigkeit derjenigen der Waldstätte nicht nachsteht. Bis vor kurzem dem Verkehr noch wenig erschlossen, haben sich die alten Einrichtungen, Sitten und Gebräuche noch ziemlich allgemein erhalten. Ganz' besonders ist dies in den vom grossen Fremdenstrom ab.seits liegenden Bergdörfern der Fall.

Im östlichen Teil des Kantons spricht man deutsch; im westlichen Abschnitt hat man einen welschen Dialekt. Die einzelnen Landesteile haben aber unter sich viel Gemeinsames; dies ist namentlich im deutschen Teil, dem Oberwallis, der Fall. Vergleicht man z. B. das Goms mit Lötschen, das Vispertal mit den Bergdörfern des Haupttales, die Dorfschaften von Salgesch bis Brig, so findet man viel Gleichartiges, Übereinstimmendes. So wird man auch in dieser Arbeit manche Anklänge an meine frühern Beschreibungen «Ob den Heidenreben», «Das Goms und die Gomser», «Lötschen und die Lötscher» und «Sonnige Halden am Lötschberg» herausfinden. Trotzdem hat jeder

Semistructured Data

```
<?xml version="1.0" encoding="utf-8"?>
<book id="1929_mul">
  <article n="1">
    <tocEntry author="Hektor Küffer" authorID="1741" category="Gedicht" lang="de
    <div>
      <s lang="de" n="1-1">
        <w lemma="@card@" n="1-1-1" pos="CARD">1</w>
      </s>
    </div>
    <div>
      <s lang="de" n="1-2">
        <w lemma="bergkreuz" n="1-2-1" pos="ADJA">Bergkreuz</w>
      </s>
    </div>
    <div>
      <s lang="de" n="1-3">
        <w lemma="von" n="1-3-1" pos="APPR">Von</w>
        <w lemma="Hektor" n="1-3-2" pos="NE">Hektor</w>
        <w lemma="unk" n="1-3-3" pos="NN">Küffer</w>
        <w lemma="." n="1-3-4" pos="$.">.</w>
      </s>
    </div>
    <div>
```

Semistructured Data: XML

```
<?xml version="1.0" encoding="utf-8"?>
<corpus>
        <article id="4673">
                <metadata>
                        <title>My Title</title>
                        <author>Noah Bubenhofer</author>
                </metadata>
                <text>
                        Here comes the text of my article! Bla bla…
                </text>
        </article>
        <article id="5783">
            …
        </article>
</corpus>
```

Flat XML Format

```
<?xml version="1.0" encoding="utf-8"?>
<corpus>
        <article id="1" title="Titel des Beitrags"
author="Hanspeter Meier">
            Hier kommt der Text. Hier kommt der Text. Hier
kommt der Text. Hier kommt der Text. Hier kommt der Text.
Hier kommt der Text. Hier kommt der Text. Hier kommt der
Text.
        </article>
        <article …

        </article>
        …
</corpus>
```

Semistructured Data: XML

- **Ex**tensible **M**arkup **L**anguage
- Meta language: DTD, XML-Schema
  → defines the elements and there position in the hierarchy
- well-formed: document consistent with xml rules:
  - one root element
  - all elements with begin and end tag
  - elements properly nested
- valid: document consistent with the rules defined in the DTD / in the schema

XML Formats

- You can invent your own xml format! If you do so: Try to find a generic format (suitable not only for one specific corpus).
- Often it is useful to use XML standards:
  - TEI (Text Encoding Initiative): http://www.tei-c.org
  - xces (Corpus Encoding Standard for XML): http://www.xces.org/
- Transformation between different XML formats is possible → ideal not only for working with data but also for archiving
- Important: include as much metadata as possible!

TEI P5 (Text Encoding Initiative)

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href=„http://www.tei-c.org/release/xml/tei/custom/schema/…
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Title</title>
      </titleStmt>
      <publicationStmt>
        <p>Publication Information</p>
      </publicationStmt>
      <sourceDesc>
        <p>Information about the source</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <p>Some text here.</p><p>And another paragraph.</p>
    </body>
  </text>
</TEI>
```

TEI P5 (Text Encoding Initiative)

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href=„http://www.tei-c.org/…
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>title of corpus</title>
                <author>author</author>
            </titleStmt>
                    <publicationStmt>
                        <p>Publication Information</p>
                    </publicationStmt>
                    <sourceDesc>
                        <p>Information about the source</p>
                    </sourceDesc>
        </fileDesc>
    </teiHeader>
    <TEI xml:id="MyTextNumber1">
        <teiHeader>
            <fileDesc>
                <titleStmt>
                    <title/>
```

Tutorial Session 2

- www.bubenhofer.com/tbilisi/ → Part 2
- Download the „All.xml“-template
- Open it in your favorite text editor; recommendations:
  - Mac: Textwrangler
  - Linux: gedit (already installed)
  - Windows: Notepad++, Textpad
  - All: Sublime Text, jEdit
- Fill in your texts.
- Open it in a web browser; if no error message appears, the file is well formed!

Tutorial Session 2: Advanced

- Try to define an XML schema which suits your needs!
- Define the elements you need.
- Define the hierarchical structure.

Tutorial Session 2: Advanced

- Use an XML editor to write your XML files:
  - Oxygene (commercial, trial version available)
  - Syntext Serna Free
- Try to automate the conversion of a non XML file in a XML file using regular expressions:
  - intro regular expressions: http://regexone.com/
  - more detailed: http://www.regular-expressions.info/

Corpus Linguistics: Sources

- Use existing corpora:
  - English
    - British National Corpus (BNC)
    - Corpus of Contemporary American English (COCA)
    - Corpus of Historical American English (COHA)
  - German
    - Deutsches Referenzkorpus (IDS-Korpora)
    - DWDS-Korpus
  - Other languages: Search for ‚national corpus‘

Corpus Linguistics: Sources

- Build your own corpus!
  Consider the following questions:
  - What language do you want to analyze?
  - What texts represent this language?
  - What sample of texts representing this language can be collected?
  - How do you want to examine the text data?
  - What kind of annotations are needed?
  - What data management platform suits your needs?