German Text Archive / DWDS & Case Studies

Plan for today

- Examples for corpus-based research on the history of (more recent stages of) German

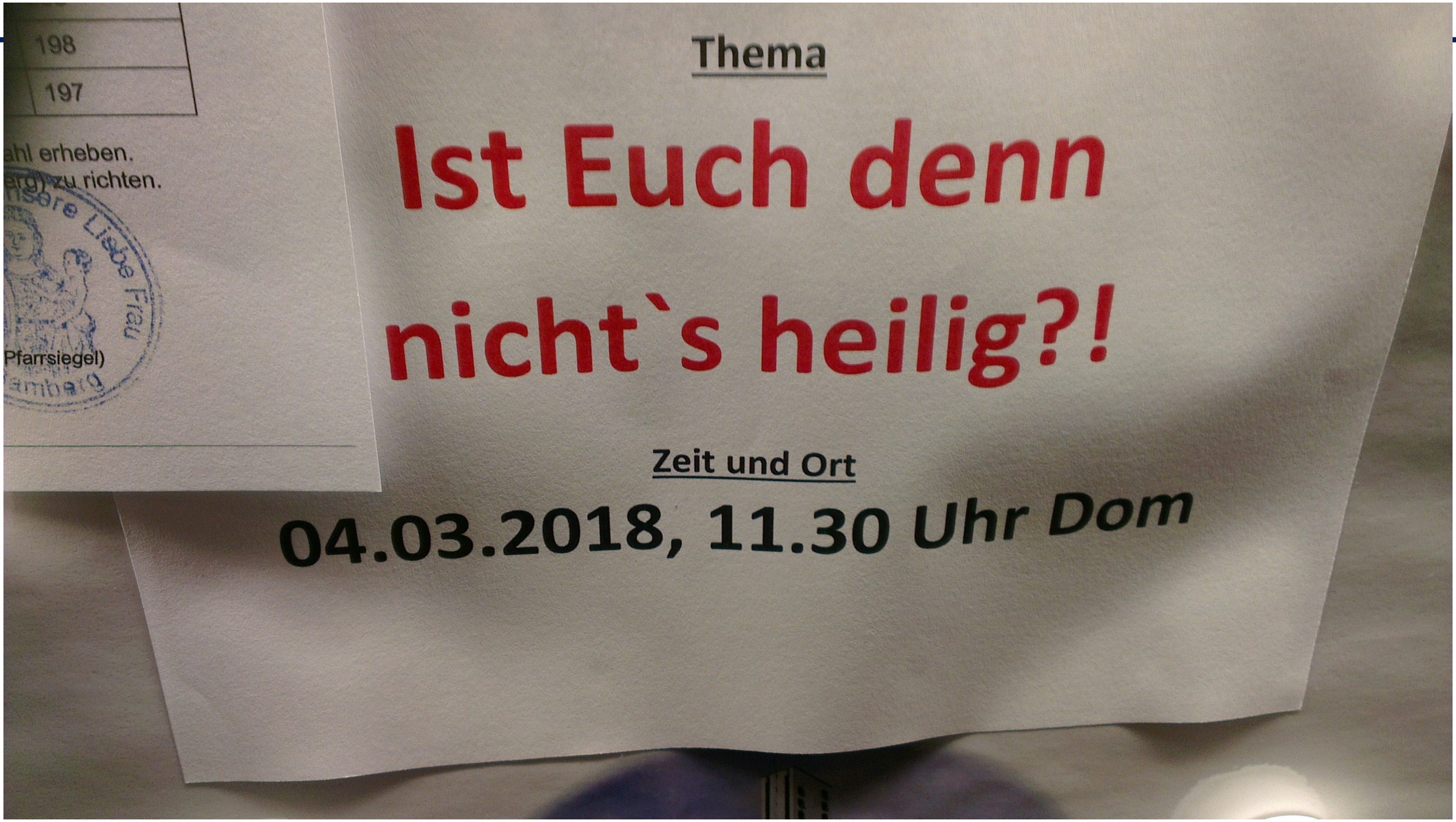
- DWDS and the German Text Archive

- Hands-on examples

Some example studies

Example studies

1. Graphemics: Functional expansion of the apostrophe <'>
2. Morphology: Word formation change
3. Morphosyntax: the x-er the y-er



198

197

ahl erheben.

erg) zu richten.

nsere Liebe Frau

(Pfarrsiegel)

amberg

Thema

Ist Euch denn nicht`s heilig?!

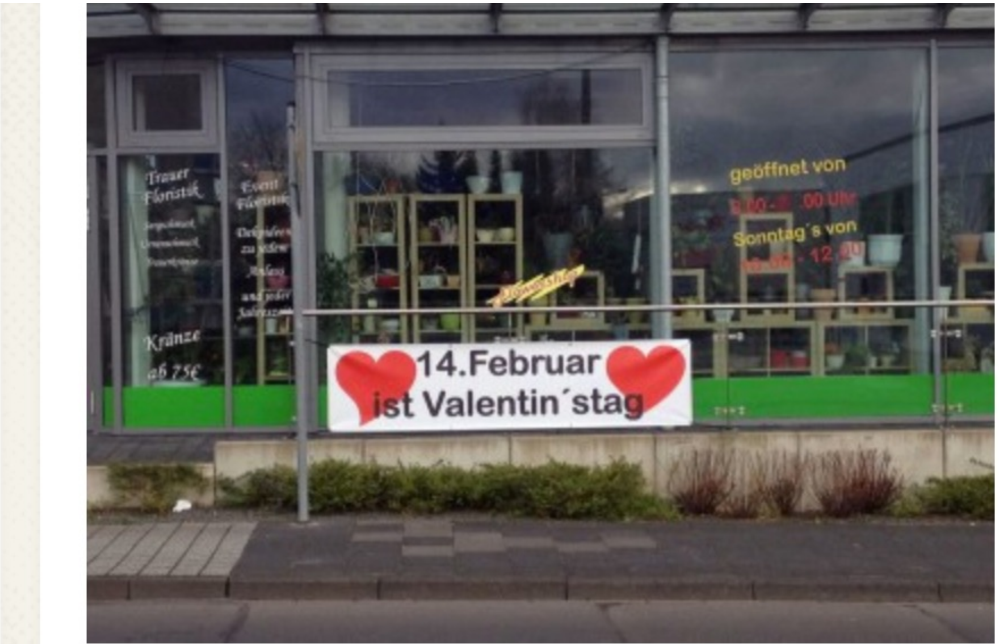
Zeit und Ort

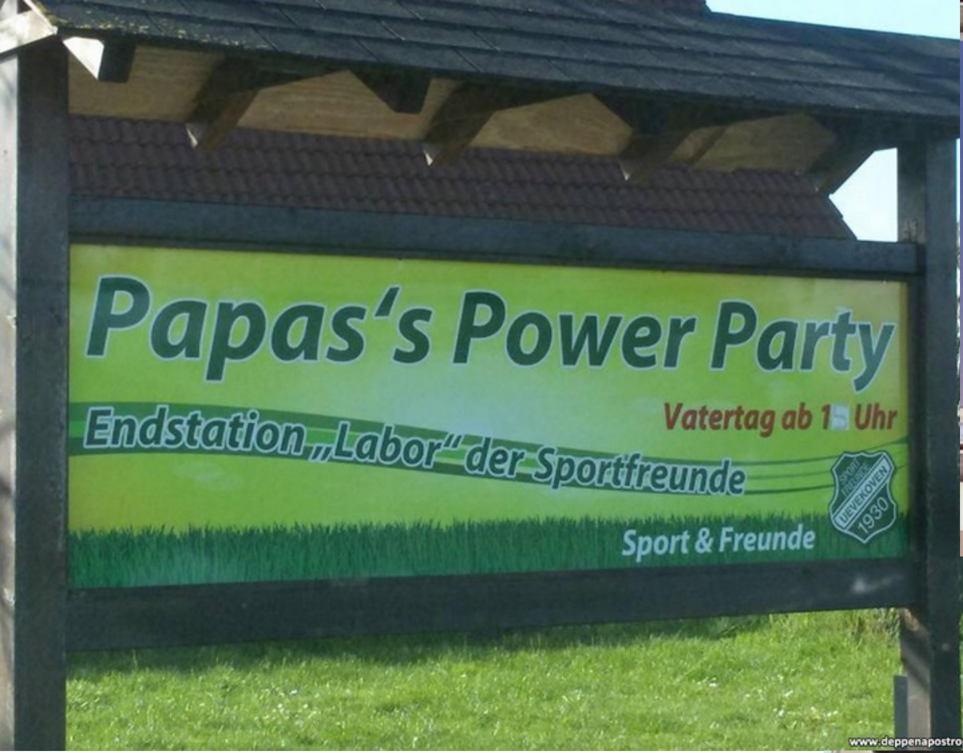
04.03.2018, 11.30 Uhr Dom

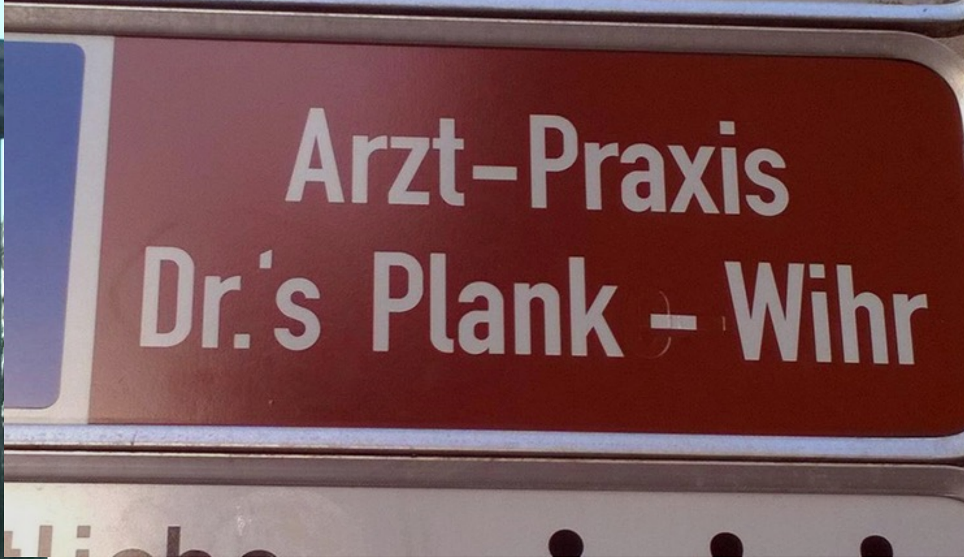
(Vortragsankündigung im Bamberger Dom)

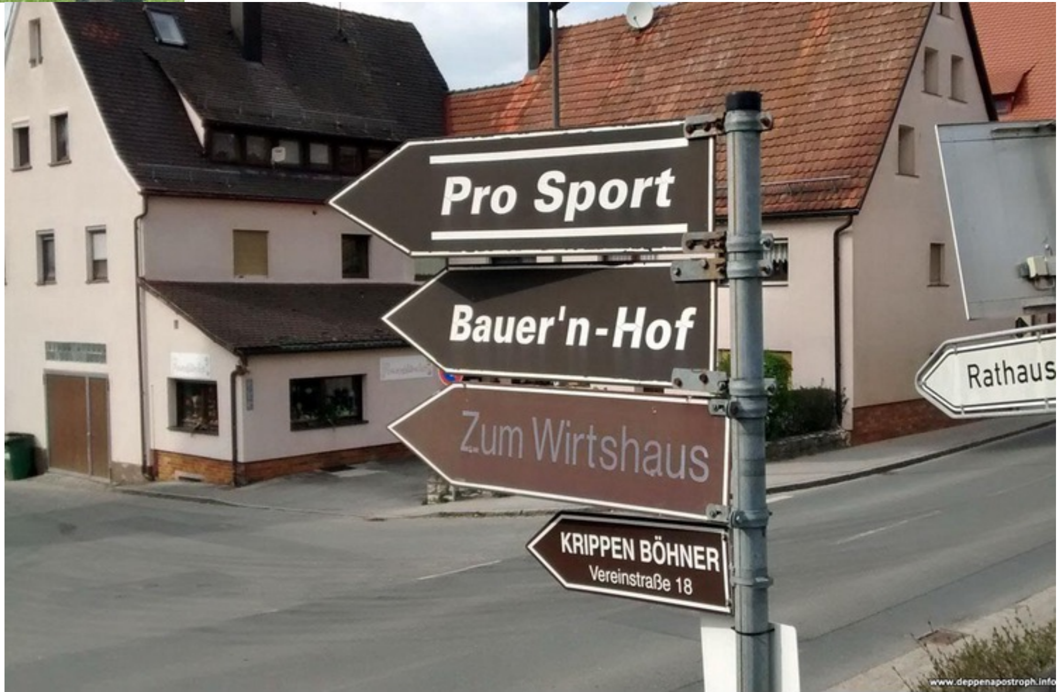
Apostrophe (Scherer 2013)

- phonographic apostrophe: *habe es > hab's, gibt es > gibt's*
- morphographic apostrophie: *Moni's Friseursalon, Dienstag's Schnitzeltag*




Papas‘s Power Party
Vatertag ab 1[illegible] Uhr
Endstation „Labor“ der Sportfreunde
Sportfreunde Uevekoven 1930
Sport & Freunde
www.deppenapostroph.info


Arzt-Praxis
Dr.´s Plank - Wihr


Pro Sport
Bauer'n-Hof
Zum Wirtshaus
KRIPPEN BÖHNER
Vereinstraße 18
Rathaus
www.deppenapostroph.info

Elision apostrophe

- The elision apostrophe is a so-called syngrapheme → marks omission of elements in a word
- Usually vowels are subject to elision
- consonants are omitted rarely – if so, usually in combination with vowels: *für den* > *für'n*
- Elisions can be word-initial, word-medial, or word-final

(aus Klein 2006)

Elision apostrophe

- Elision is, first and foremost, a phonological phenomenon
- Elision apostrophes stand for omitted sounds
  → phonographic representation

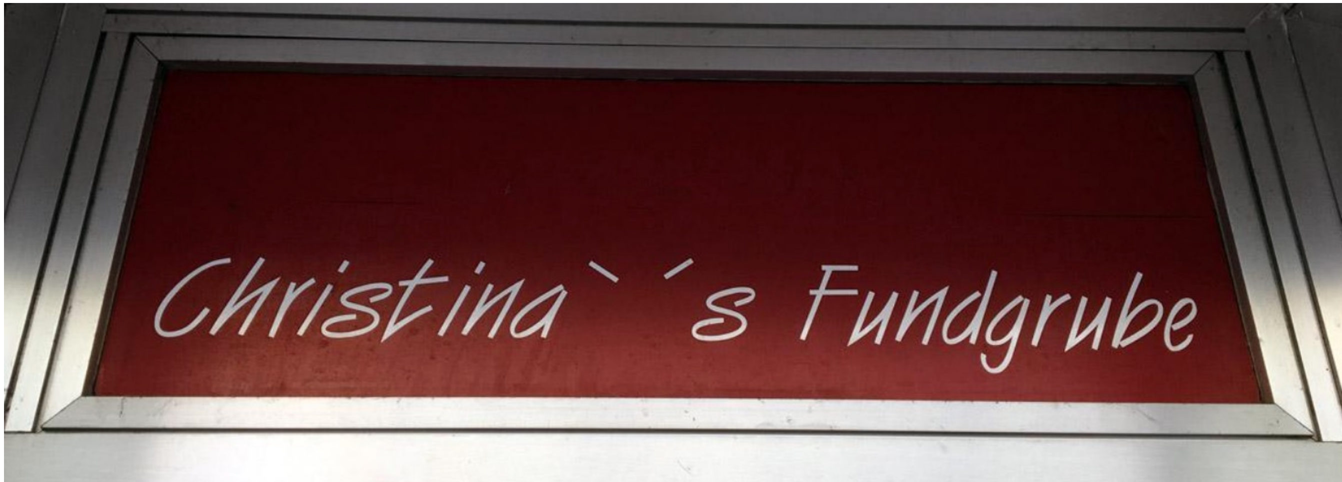
(vgl. Klein 2006)

From phonographic to morphographic apostrophe

- Especially in non-standard writing, the apostrophe can also signal morpheme boundaries
- Scherer (2013): Howe frequent is the morphographic apostrophe in written German, and which factors determine its use?

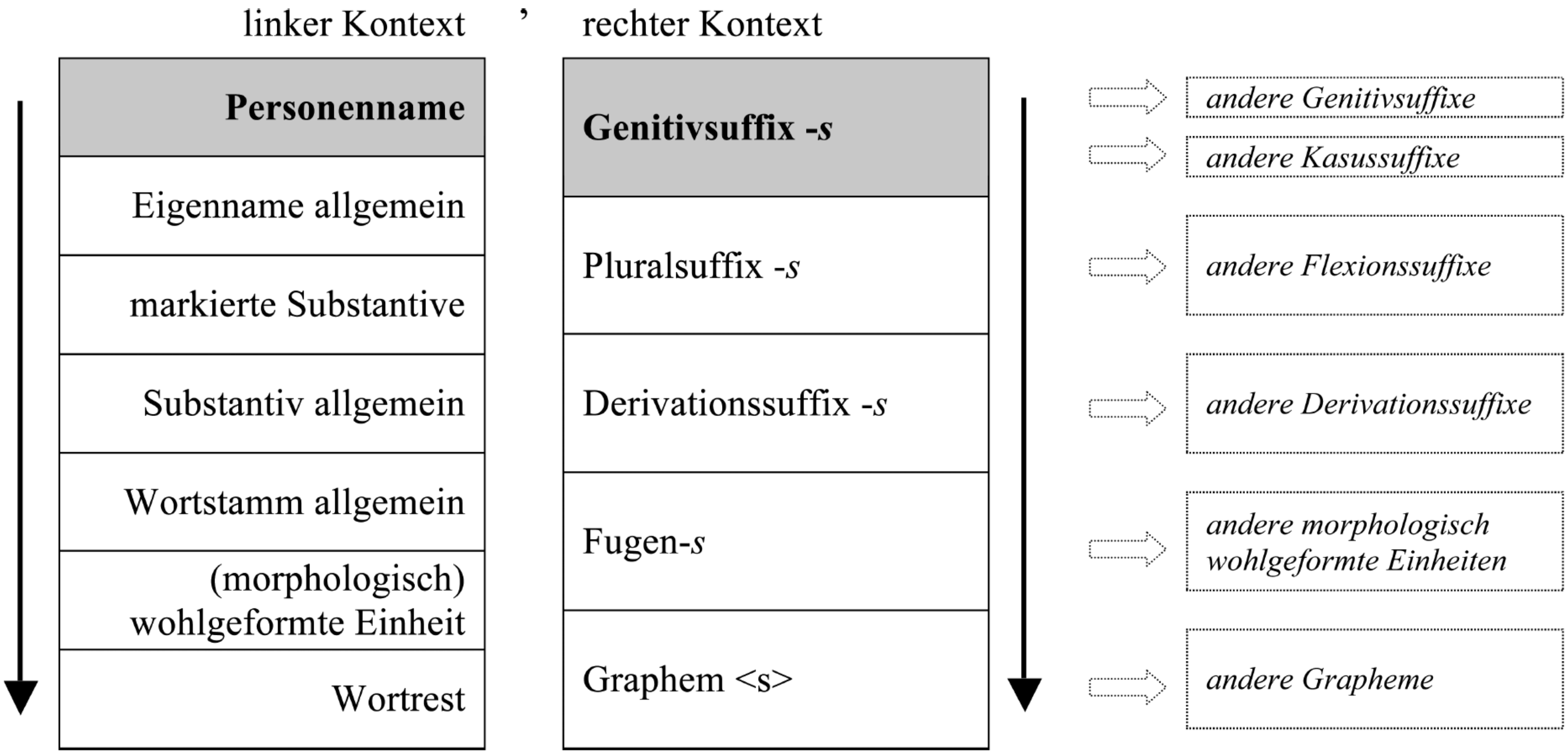
Results

- 10-20 % of apostrophes in her corpus are morphographic
- morphographic apostrophe especially for marking genitives with person names → prototypical context



Results

- functional expansion of apostrophe



linker Kontext ’ rechter Kontext

Personenname

Eigenname allgemein

markierte Substantive

Substantiv allgemein

Wortstamm allgemein

(morphologisch) wohlgeformte Einheit

Wortrest

Genitivsuffix -*s*

Pluralsuffix -*s*

Derivationssuffix -*s*

Fugen-*s*

Graphem <s>

andere Genitivsuffixe

andere Kasussuffixe

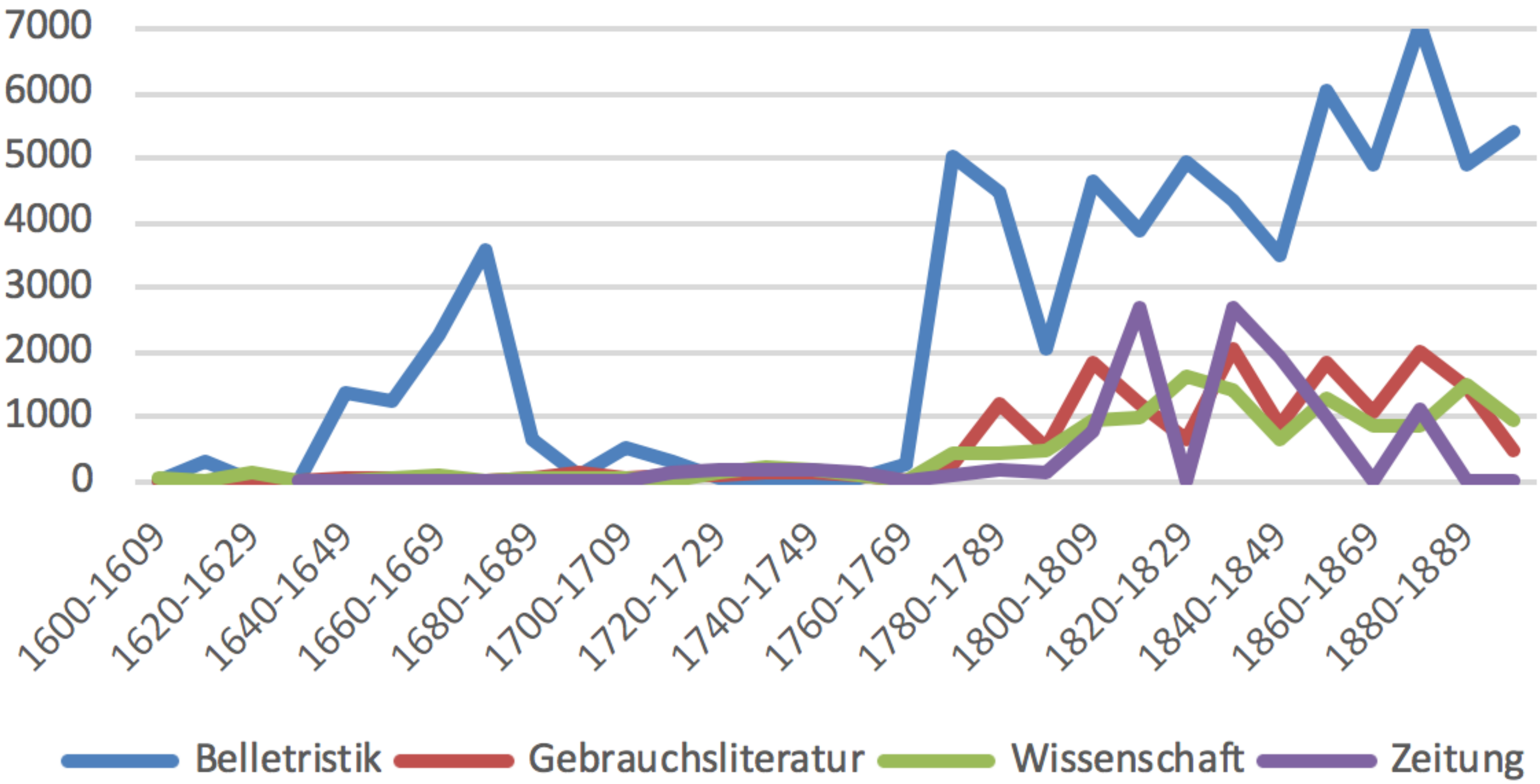
andere Flexionssuffixe

andere Derivationssuffixe

andere morphologisch wohlgeformte Einheiten

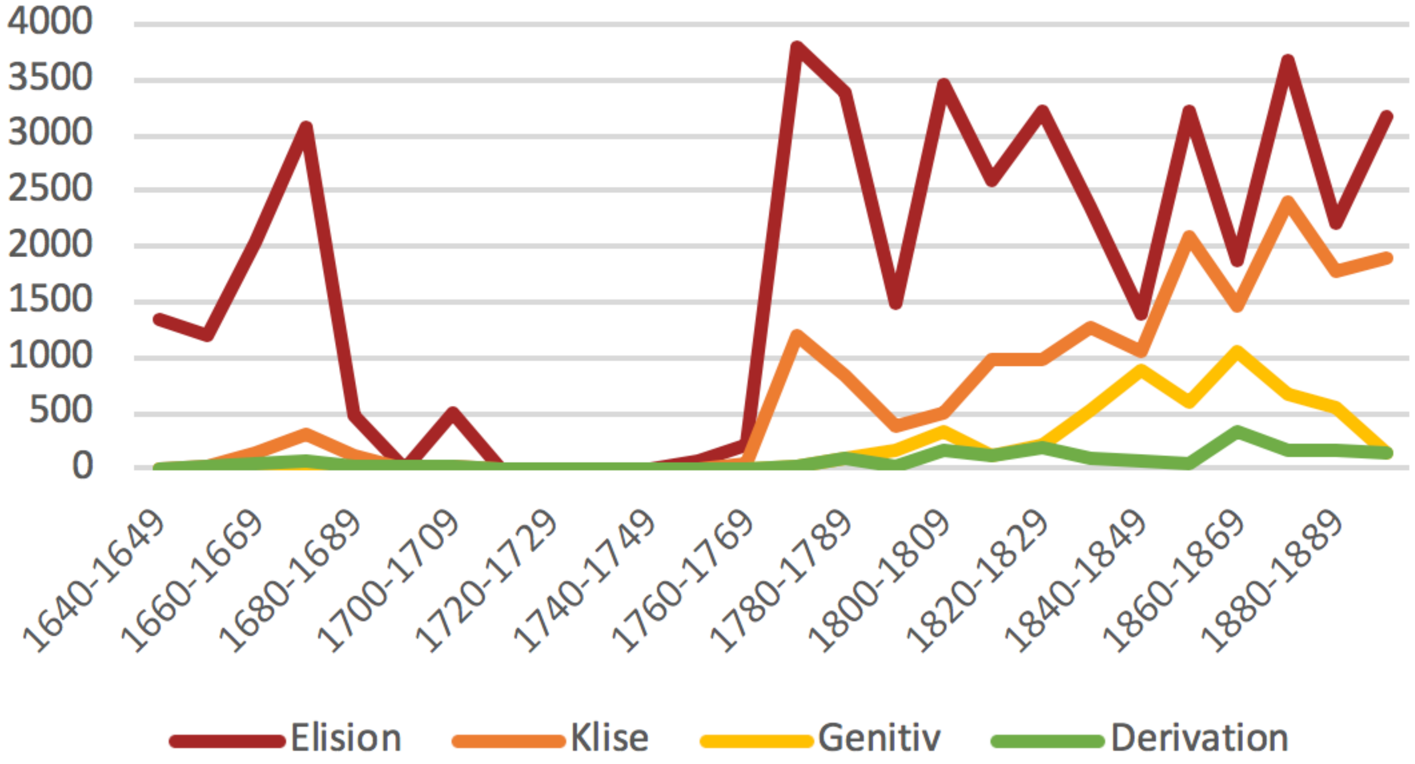
andere Grapheme

Apostrophe in the DTA



(Frequency per million words, Kempf 2019)

Apostrophe in the DTA



schöneren > schön'ren	*in das > ins*	*Stefan's*	*Angesicht's*
es ist > 's ist	*gibt es > gibt's*	*Luise's*	*Wolff'sche*

(only fiction texts, from Kempf 2019)

Morphographic apostrophe in the DTA

- apparently reanalysis: phonographic elision marker > morphographic boundary marker

*Gott'***s** *Wahrheit* 'God's truth' (< Gottes)

>

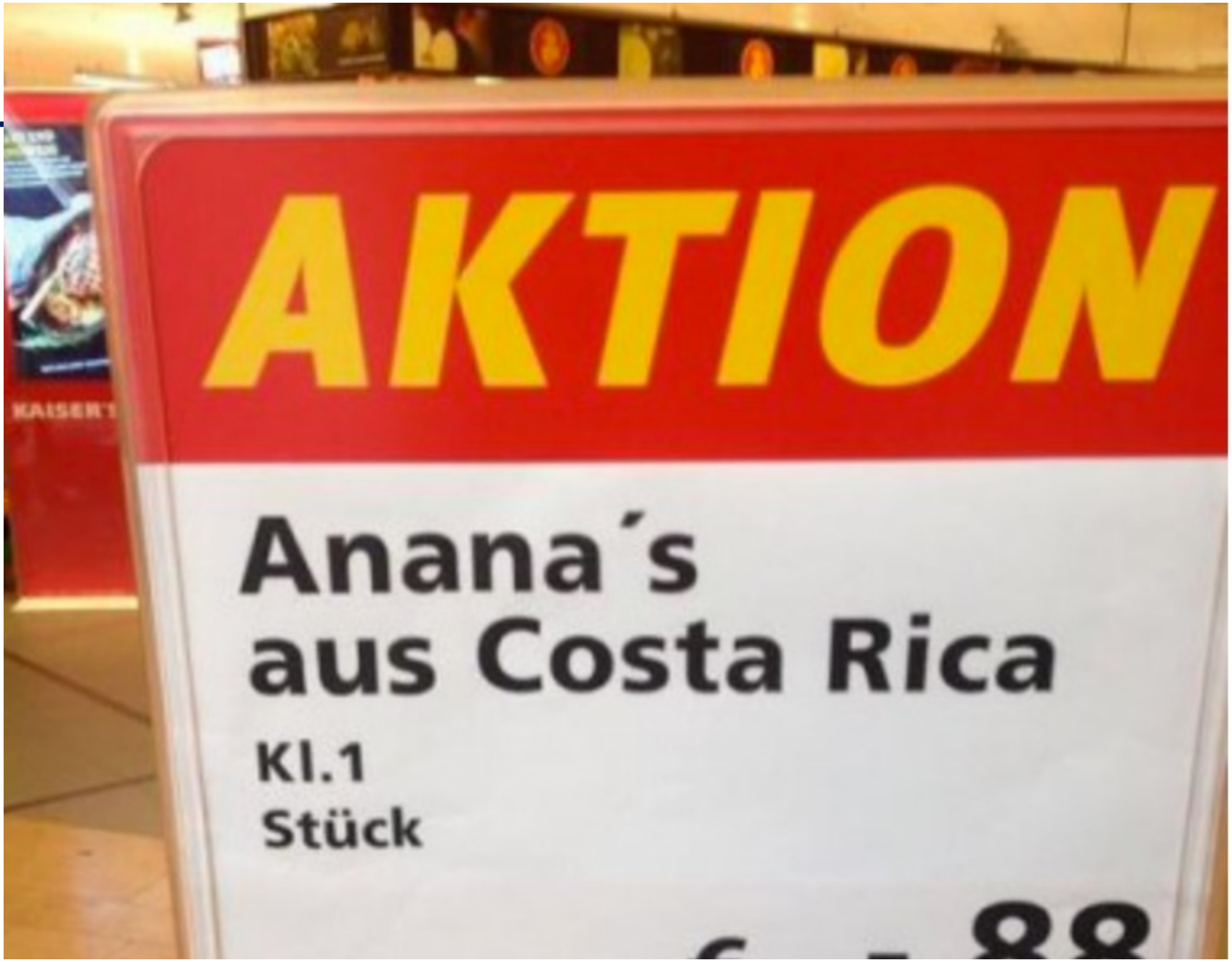
Bonaparte's Benehmen 'Bonaparte's behavior'

Development of genitive apostrophe

- Kempf (2019): zwei "haydays" of genitive apostrophe, first in the 17th century, then in 18th/19th century
- probably no continuity between these high-frequency phases:
  - In the first phase, the genitive apostrophe predominantly combines with native appellatives, in the second one with proper names (in the beginning, mostly non-native ones)
  - Declins in morphographic apostrophe types towards the end of the 19th century

Wrap-up on genitives

- Apostrophe can contribute to making morphology "visible"
- in present-day German standard orthography, it can only function as an elision marker
- in non-standard writing its morphographic function is retained (or is coming back)
- … even in cases where the apparent morphological structure is the result of reanalysis!


AKTION
Anana´s
aus Costa Rica
Kl.1
Stück
88

Word-formation change: -gate

The most devastating detective story of this century.
REDFORD/HOFFMAN
"ALL THE PRESIDENT'S MEN"

Watergate

Nipplegate

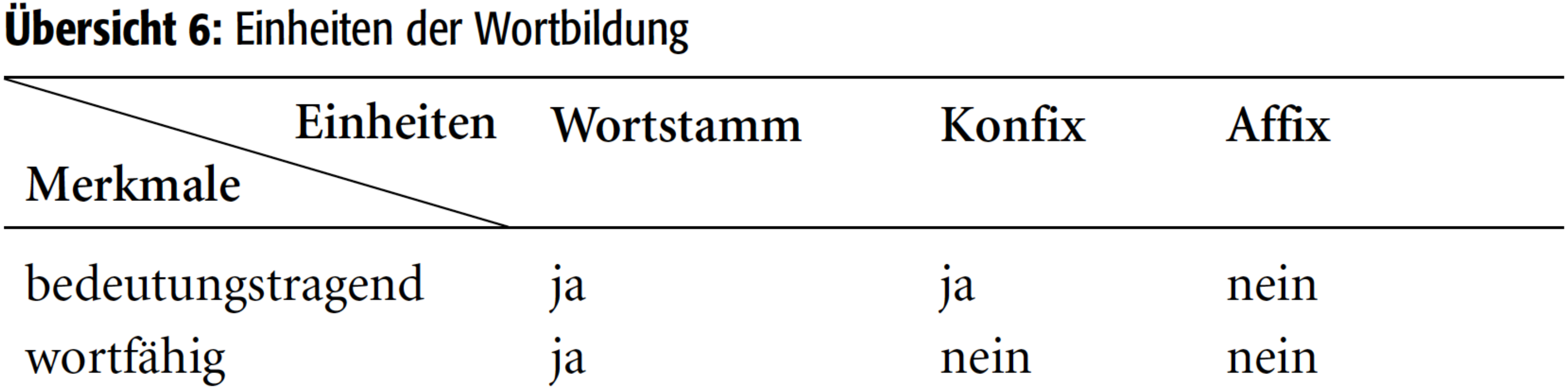
Hosen-Gate

-gate as an onymic confix

- *-gate* as an onymic confix: it is used to derive proper names
- unlike common nouns, proper names are characterized by
  - monoreference: referring to exactly one entity (e.g. *Alexander Bergs*)
  - direct reference: no "detour" via potential / prototypical meaning

-gate as a confix

- Confixes share properties with affixes and with free words:
  - like affixes, they are bound to a stem;
  - like free words, they carry lexical meaning.



Übersicht 6: Einheiten der Wortbildung

Einheiten Merkmale	Wortstamm	Konfix	Affix
bedeutungstragend	ja	ja	nein
wortfähig	ja	nein	nein

(Fleischer & Barz 2012: 64)

Confix vs. affixoid

- Affixoid as a unit between word and affix
- unlike confixes, affixoids are characterized by semantic bleaching
- e.g.: ***Riesen****krach* 'giant noise' (not *'nouse of a giant'), *Laub****werk*** (not a 'work', but a collective noun for fallen leaves)
- disputed concept (vgl. z.B. Schmidt 1987, Stevens 2005)

-gate as an onymic concept

- reanalyzed from *Watergate*
- first –*gate* formations in English as early as 1972/73 (time of Watergate affair)
- became productive in German in the last few years as well, e.g. *Hosen-Gate*

(vgl. Flach, Kopf & Stefanowitsch im Ersch.)

Examples (from Wortwarte)

- "Schnell war von " **Guacamole-Gate** " die Rede . Die Debatte nahm beinahe Loriot'sche Dimensionen an , frei nach dem Motto : Die Erbse bleibt draußen !"
- Falls hier eine Trennwand geplant war, fehlt für ihre Installation der nötige Platz. Unter Mitarbeitern des russischen Außenministeriums kursiert noch eine zweite Erklärung, wie es zum "**Toiletten-Gate**" kommen konnte.

Development

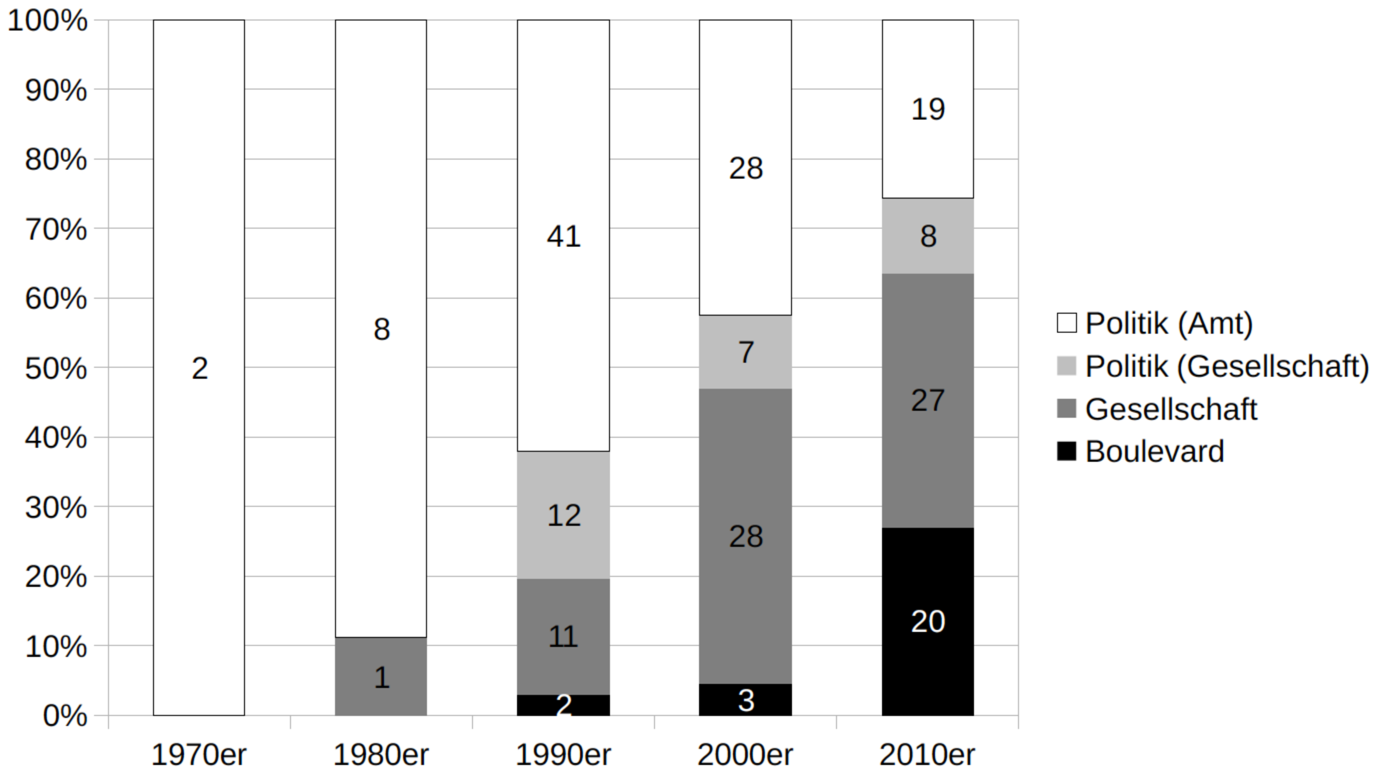


Abb. 5a. Deutsche Erstbelege (Entlehnungen und dt. Bildungen) in DeReKo/ZEIT nach Skandalfeld (n = 217).

aus Flach, Kopf & Stefanowitsch (im Ersch.)

Example

- How can we search for -*gate* formations in a corpus of present-day German?

- How can we make sure to exclude expectable false positives?

Meaning shift of *-gate*

- Flach et al. (2018) show that *-gate* in German is subject to "trivialization": from big political scandals to smaller boulevard affairs
- Are there similar developments in other domains?

"X-phemism mill"

- Expressive meanings tend to show signs of attrition over time
- e.g. *scheiße* 'shit, crap': de-tabooization over the course of the (late) 20th century
- Allan & Burridge (2006): "X-phemism mill" – euphemysms and dysphemisms tend to loose their expressive meaning and become replaced by new ones
- cf. e.g. MHG *kranc* 'weak' > NHG *krank* 'ill'

Expressivity

- concept often used quite vaguely
- Traugott & Dasher (2002: 94) attribute it to Traugott (1982)
- There it predominantly refers to the "interpersonal" component of language in the sense of Halliday & Hasan (1976)

Halliday, M.A.K. & Ruqaia Hasan. 1976. *Cohesion in English*. London: Longman.

Traugott, Eliabath Closs. 1982. From propositional to textual and expressive meanings; some semantic–pragmatic aspects of grammaticalization. In Winfred P. Lehmann and Yakov Malkiel, eds., Perspectives on Historical Linguistics, 245–271. Amsterdam: Benjamins

Traugott, Elizabeth Closs & Richard B. Dasher. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Expressivity

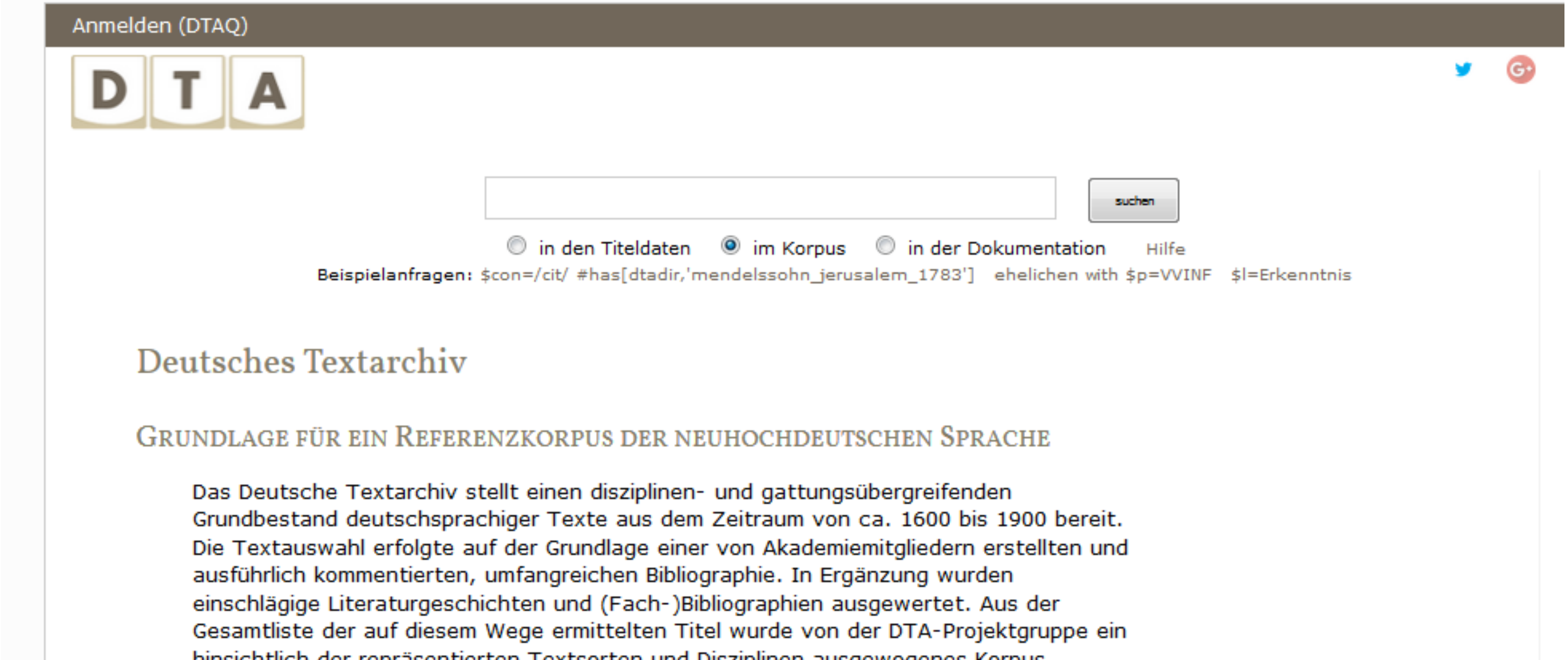
- Traugott & Dasher (2002) use *expressivity* quasi-synonymously with *subjectivity*
- diachronic emergence of subjective meaning as *subjectification*

Subjectification: "the development of a grammatically identifiable expression of speaker **belief** or speaker **attitude** to what is said" (Traugott 1995)

DWDS & DTA

German text archive

- available via https://deutschestextarchiv.de/ OR https://dwds.de OR https://kaskade.dwds.de/dstar/



○ in den Titeldaten ● im Korpus ○ in der Dokumentation

Beispielanfragen: $con=/cit/ #has[dtadir,'mendelssohn_jerusalem_1783'] ehelichen with $p=VVINF $l=Erkenntnis

Deutsches Textarchiv

Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache

Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern erstellten und ausführlich kommentierten, umfangreichen Bibliographie. In Ergänzung wurden einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Aus der Gesamtliste der auf diesem Wege ermittelten Titel wurde von der DTA-Projektgruppe ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus

German Text Archive

- www.deutschestextarchiv.de

  → basic search functions, good for simple searches that require lots of context and perhaps even faksimiles of the original print; data for download

- www.dwds.de

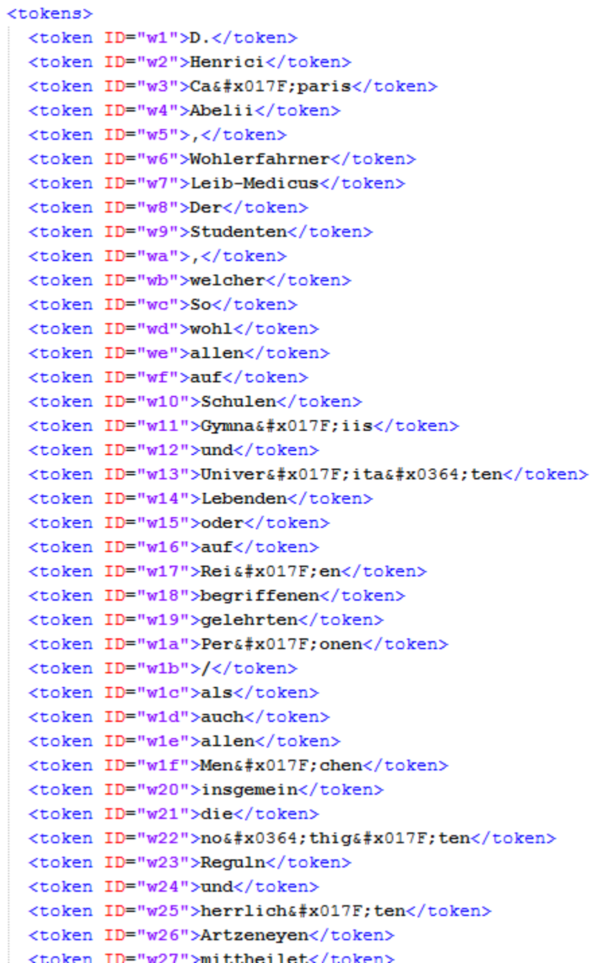
  → advanced search functions, best choice for most corpus queries; useful export functions

- https://kaskade.dwds.de/dstar/

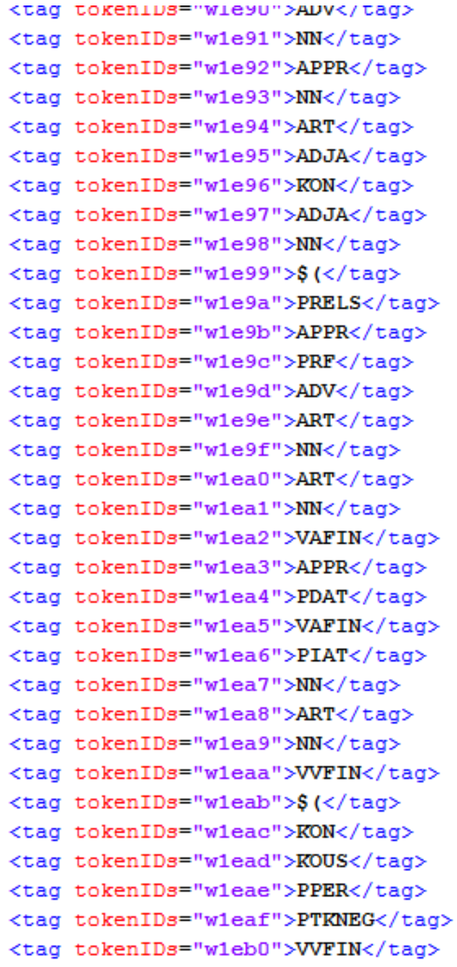
  → expert search functions, less limited than the dwds.de search interface; less ideal export functions; very good for advanced count operations

German Text Archive

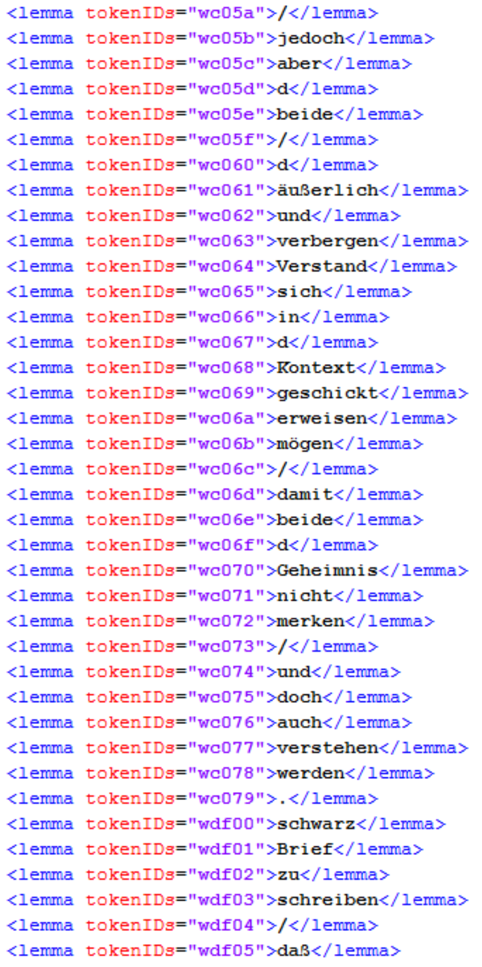
- DTA is **tagged** and **lemmatized**



```
<tokens>
  <token ID="w1">D.</token>
  <token ID="w2">Henrici</token>
  <token ID="w3">Ca&#x017F;paris</token>
  <token ID="w4">Abelii</token>
  <token ID="w5">,</token>
  <token ID="w6">Wohlerfahrner</token>
  <token ID="w7">Leib-Medicus</token>
  <token ID="w8">Der</token>
  <token ID="w9">Studenten</token>
  <token ID="wa">,</token>
  <token ID="wb">welcher</token>
  <token ID="wc">So</token>
  <token ID="wd">wohl</token>
  <token ID="we">allen</token>
  <token ID="wf">auf</token>
  <token ID="w10">Schulen</token>
  <token ID="w11">Gymna&#x017F;iis</token>
  <token ID="w12">und</token>
  <token ID="w13">Univer&#x017F;ita&#x0364;ten</token>
  <token ID="w14">Lebenden</token>
  <token ID="w15">oder</token>
  <token ID="w16">auf</token>
  <token ID="w17">Rei&#x017F;en</token>
  <token ID="w18">begriffenen</token>
  <token ID="w19">gelehrten</token>
  <token ID="w1a">Per&#x017F;onen</token>
  <token ID="w1b">/</token>
  <token ID="w1c">als</token>
  <token ID="w1d">auch</token>
  <token ID="w1e">allen</token>
  <token ID="w1f">Men&#x017F;chen</token>
  <token ID="w20">insgemein</token>
  <token ID="w21">die</token>
  <token ID="w22">no&#x0364;thig&#x017F;ten</token>
  <token ID="w23">Reguln</token>
  <token ID="w24">und</token>
  <token ID="w25">herrlich&#x017F;ten</token>
  <token ID="w26">Artzeneyen</token>
  <token ID="w27">mittheilet</token>
```



```
<tag tokenIDs="w1e90">ADV</tag>
<tag tokenIDs="w1e91">NN</tag>
<tag tokenIDs="w1e92">APPR</tag>
<tag tokenIDs="w1e93">NN</tag>
<tag tokenIDs="w1e94">ART</tag>
<tag tokenIDs="w1e95">ADJA</tag>
<tag tokenIDs="w1e96">KON</tag>
<tag tokenIDs="w1e97">ADJA</tag>
<tag tokenIDs="w1e98">NN</tag>
<tag tokenIDs="w1e99">$(</tag>
<tag tokenIDs="w1e9a">PRELS</tag>
<tag tokenIDs="w1e9b">APPR</tag>
<tag tokenIDs="w1e9c">PRF</tag>
<tag tokenIDs="w1e9d">ADV</tag>
<tag tokenIDs="w1e9e">ART</tag>
<tag tokenIDs="w1e9f">NN</tag>
<tag tokenIDs="w1ea0">ART</tag>
<tag tokenIDs="w1ea1">NN</tag>
<tag tokenIDs="w1ea2">VAFIN</tag>
<tag tokenIDs="w1ea3">APPR</tag>
<tag tokenIDs="w1ea4">PDAT</tag>
<tag tokenIDs="w1ea5">VAFIN</tag>
<tag tokenIDs="w1ea6">PIAT</tag>
<tag tokenIDs="w1ea7">NN</tag>
<tag tokenIDs="w1ea8">ART</tag>
<tag tokenIDs="w1ea9">NN</tag>
<tag tokenIDs="w1eaa">VVFIN</tag>
<tag tokenIDs="w1eab">$(</tag>
<tag tokenIDs="w1eac">KON</tag>
<tag tokenIDs="w1ead">KOUS</tag>
<tag tokenIDs="w1eae">PPER</tag>
<tag tokenIDs="w1eaf">PTKNEG</tag>
<tag tokenIDs="w1eb0">VVFIN</tag>
```



```
<lemma tokenIDs="wc05a">/</lemma>
<lemma tokenIDs="wc05b">jedoch</lemma>
<lemma tokenIDs="wc05c">aber</lemma>
<lemma tokenIDs="wc05d">d</lemma>
<lemma tokenIDs="wc05e">beide</lemma>
<lemma tokenIDs="wc05f">/</lemma>
<lemma tokenIDs="wc060">d</lemma>
<lemma tokenIDs="wc061">äußerlich</lemma>
<lemma tokenIDs="wc062">und</lemma>
<lemma tokenIDs="wc063">verbergen</lemma>
<lemma tokenIDs="wc064">Verstand</lemma>
<lemma tokenIDs="wc065">sich</lemma>
<lemma tokenIDs="wc066">in</lemma>
<lemma tokenIDs="wc067">d</lemma>
<lemma tokenIDs="wc068">Kontext</lemma>
<lemma tokenIDs="wc069">geschickt</lemma>
<lemma tokenIDs="wc06a">erweisen</lemma>
<lemma tokenIDs="wc06b">mögen</lemma>
<lemma tokenIDs="wc06c">/</lemma>
<lemma tokenIDs="wc06d">damit</lemma>
<lemma tokenIDs="wc06e">beide</lemma>
<lemma tokenIDs="wc06f">d</lemma>
<lemma tokenIDs="wc070">Geheimnis</lemma>
<lemma tokenIDs="wc071">nicht</lemma>
<lemma tokenIDs="wc072">merken</lemma>
<lemma tokenIDs="wc073">/</lemma>
<lemma tokenIDs="wc074">und</lemma>
<lemma tokenIDs="wc075">doch</lemma>
<lemma tokenIDs="wc076">auch</lemma>
<lemma tokenIDs="wc077">verstehen</lemma>
<lemma tokenIDs="wc078">werden</lemma>
<lemma tokenIDs="wc079">.</lemma>
<lemma tokenIDs="wdf00">schwarz</lemma>
<lemma tokenIDs="wdf01">Brief</lemma>
<lemma tokenIDs="wdf02">zu</lemma>
<lemma tokenIDs="wdf03">schreiben</lemma>
<lemma tokenIDs="wdf04">/</lemma>
<lemma tokenIDs="wdf05">daß</lemma>
```

German Text Archive

- Search Syntax: DDC; see https://www.dwds.de/d/korpussuche (German) or https://www.cudmuncher.de/~moocow/software/ddc/querydoc.html (English)

German Text Archive / DWDS

- annotation layers:

$w	words/tokens (in DTA: Latin-1 text)
$l	Lemma
$p	part of speech

additionally in DTA:

$u	original text in DTA
$v	normalized word form

Examples

How can we search for...

1. the exact word **form** *König* 'king' (i.e. not *Könige, Königs ...*)
2. The **lemma** *laufen* 'go,run'
3. the plural forms *Wagen* vs. *Wägen* 'cars/waggons'
4. the construction ADJ *werden* 'become ADJ' (e.g. *verrückt werden* 'go crazy')
5. The sequence *weil* + personal pronoun + verb (e.g. *weil ich sag das halt so*)
6. Apostrophe with genitives of words ending in -s, e.g. *des Korpus'*
7. Infinitives without *zu* and *zu* infinitives
8. Frequency of *ward* vs. *wurde* across centuries

Dstar

D* (Dstar)

https://kaskade.dwds.de/dstar/

- alternative interface for the BBAW corpora
- particularly suitable for frequency counts
- documentation is a bit suboptimal

D* (Dstar)

- Useful hints in this tutorial by Andreas Blombach http://sprachwissenschaft.fau.de/personen/daten/blombach/korpora.pdf
- and in this blog post by Frank Wiegand: https://sprache.hypotheses.org/723

D* (Dstar)

- Basic pattern for count queries:

COUNT (insert normal DDC query here)

- Example:

COUNT($p = /NN/g) #sep

(counts all common nouns)

D* (Dstar)

- by-Operator: count by $l (Lemma), $p (POS) etc.

```
COUNT ( insert ddc query here )
```

- Example:

```
COUNT( $p = /NN/g ) #BY($l) #sep
```

(counts all common nouns by lemma)

D* (Dstar)

- Basic pattern for frequency counts:

```
COUNT ( insert normal ddc query here )
```

- More complex example:

```
count( "$w=/[Jj]e/g $w=/.*er/g=1" &&
"$w=/desto/g $w=/.*er/g=2" )
```

The mother of all (German) corpora: DeReKo

DeReKo

- since 1964
- biggest collection of corpora of present-day German
- not a balanced corpus – instead, it is a collection of "archives" designed in such a way that one can create "virtual corpora" balanced for aspects relevant for the current research question

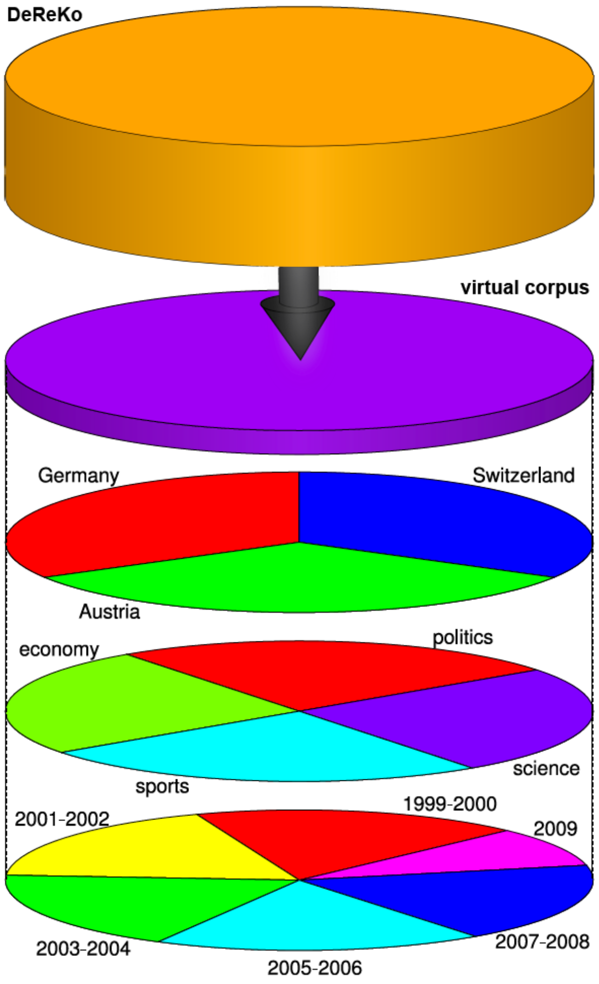
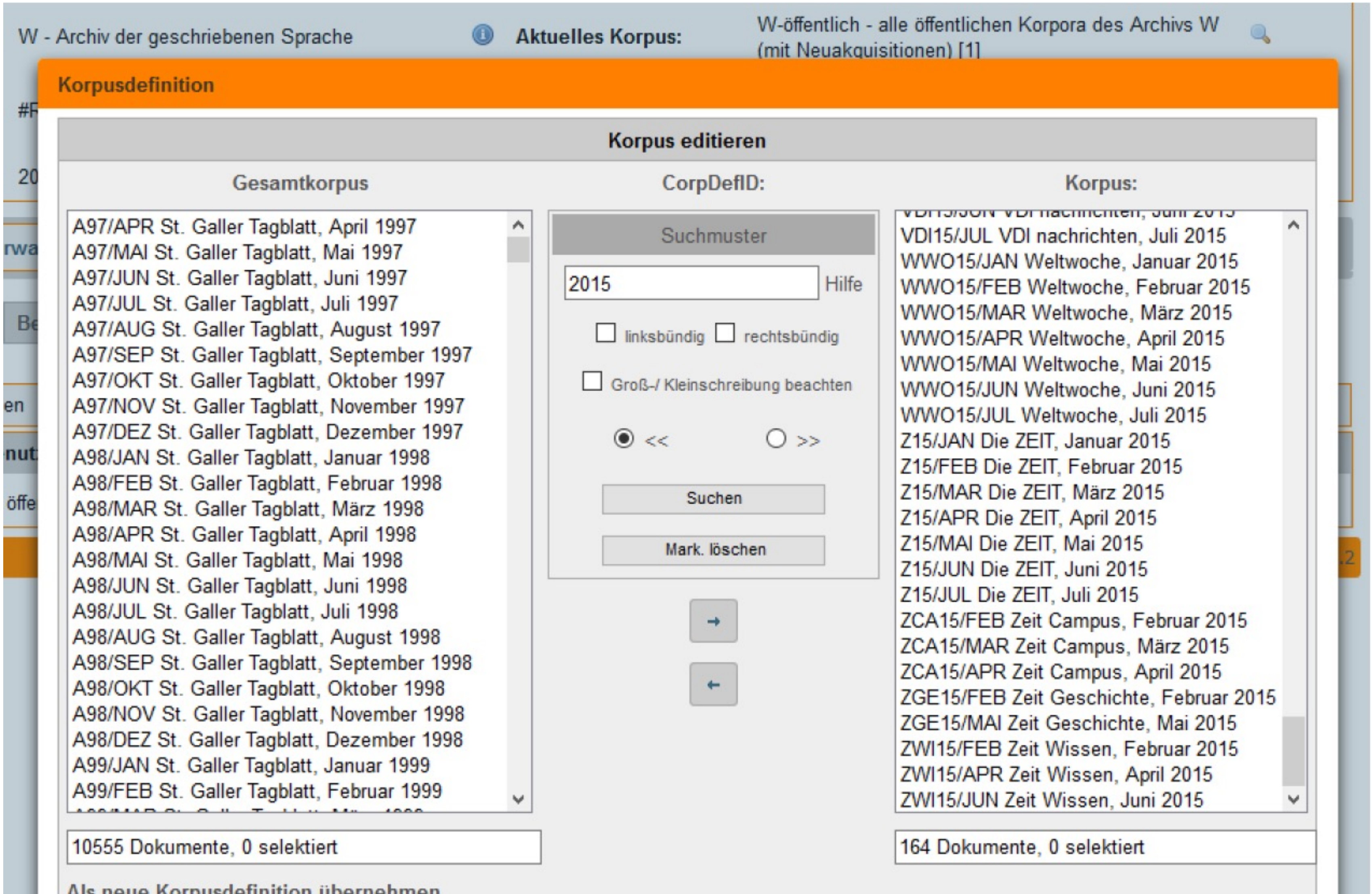


Figure 1: Defining a virtual corpus by specifying its distribution across the metadata dimensions *country of origin* (top), *topic* (center), and *time* (bottom).

(from Kupietz et al. 2010)

Create virtual corpora



W - Archiv der geschriebenen Sprache

Aktuelles Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W (mit Neuakquisitionen) [1]

Korpusdefinition

Korpus editieren

Gesamtkorpus

A97/APR St. Galler Tagblatt, April 1997
A97/MAI St. Galler Tagblatt, Mai 1997
A97/JUN St. Galler Tagblatt, Juni 1997
A97/JUL St. Galler Tagblatt, Juli 1997
A97/AUG St. Galler Tagblatt, August 1997
A97/SEP St. Galler Tagblatt, September 1997
A97/OKT St. Galler Tagblatt, Oktober 1997
A97/NOV St. Galler Tagblatt, November 1997
A97/DEZ St. Galler Tagblatt, Dezember 1997
A98/JAN St. Galler Tagblatt, Januar 1998
A98/FEB St. Galler Tagblatt, Februar 1998
A98/MAR St. Galler Tagblatt, März 1998
A98/APR St. Galler Tagblatt, April 1998
A98/MAI St. Galler Tagblatt, Mai 1998
A98/JUN St. Galler Tagblatt, Juni 1998
A98/JUL St. Galler Tagblatt, Juli 1998
A98/AUG St. Galler Tagblatt, August 1998
A98/SEP St. Galler Tagblatt, September 1998
A98/OKT St. Galler Tagblatt, Oktober 1998
A98/NOV St. Galler Tagblatt, November 1998
A98/DEZ St. Galler Tagblatt, Dezember 1998
A99/JAN St. Galler Tagblatt, Januar 1999
A99/FEB St. Galler Tagblatt, Februar 1999
A99/MAR St. Galler Tagblatt, März 1999

10555 Dokumente, 0 selektiert

CorpDefID:

Suchmuster

2015 Hilfe

□ linksbündig □ rechtsbündig

□ Groß-/ Kleinschreibung beachten

● << ○ >>

Korpus:

VDI15/JUN VDI nachrichten, Juni 2015
VDI15/JUL VDI nachrichten, Juli 2015
WWO15/JAN Weltwoche, Januar 2015
WWO15/FEB Weltwoche, Februar 2015
WWO15/MAR Weltwoche, März 2015
WWO15/APR Weltwoche, April 2015
WWO15/MAI Weltwoche, Mai 2015
WWO15/JUN Weltwoche, Juni 2015
WWO15/JUL Weltwoche, Juli 2015
Z15/JAN Die ZEIT, Januar 2015
Z15/FEB Die ZEIT, Februar 2015
Z15/MAR Die ZEIT, März 2015
Z15/APR Die ZEIT, April 2015
Z15/MAI Die ZEIT, Mai 2015
Z15/JUN Die ZEIT, Juni 2015
Z15/JUL Die ZEIT, Juli 2015
ZCA15/FEB Zeit Campus, Februar 2015
ZCA15/MAR Zeit Campus, März 2015
ZCA15/APR Zeit Campus, April 2015
ZGE15/FEB Zeit Geschichte, Februar 2015
ZGE15/MAI Zeit Geschichte, Mai 2015
ZWI15/FEB Zeit Wissen, Februar 2015
ZWI15/APR Zeit Wissen, April 2015
ZWI15/JUN Zeit Wissen, Juni 2015

164 Dokumente, 0 selektiert

Als neue Korpusdefinition übernehmen

The concept of DeReKo

- DeReKo as "Urstichprobe" 'original sample' (Kupietz 2010)
- i.e. no ready-to-use sample but a data pool from which the user can create a sample balanced for relevant criteria

Terminology

DeReKo uses the following terms:

- **Dokument 'document:** contains at least one text (e.g. a novel) or multiple text (e.g. one month of newspaper articles from the St. Galler Tagblatt)
- **Korpus 'corpus':** contains multiple documents, e.g. all documents of the St. Galler Tagblatt
- **Virtuelles Korpus 'virtual corpus':** user-defined collection of multiple documents or corpora.
- **Archiv 'archive':** uppermost level of DeReKo. E.g. "archive of written language" contains all corpora of written language in DeReKo (e.g. the St. Galler Tagblatt corpus).

Quelle: https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_www_erste.html

DeReKo

- accessible via COSMAS II interface
- advantage: relatively flexible search options
- disadvantage: limited export options (max. 10,000 hits)

Regular expressions in COSMAS

Wildcards (Source: http://www.ids-mannheim.de/cosmas2//win-app/hilfe/suchanfrage/eingabe-grafisch/syntax/WORT.html)

- * 0, 1, 2, … characters.
- **+** 0 or 1 character
- **?** 1 character
- The placeholders can be used multiple times within one word form.
- They can be placed anywhere within a word form.
- When using * at least two characters have to be speficfied.
- Wildcard function can be escaped with \

Annotation of COSMAS

- Largest archive W not pos-tagged
- However, there is a subcorpus with tagged texts:
  - Tagged-C, Tagged-C2 (from 2010): tagged with Connexor
  - Tagged-T, Tagged-T2 (from 2010): tagged with TreeTagger
- The tagged archives still have > 1 billion tokens.

Example

How can we find

- *wegen* + NP
- *weil* + personal pronoun + Verb (*weil ich sag das halt so)*
- Frequency of *ward* vs. *wurde* in historical texts
- Usage variants of the verb *kommunizieren*