What is a good corpus?

Anke Lüdeling, Roland Schäfer, Elizabeth Pankratz,
Thomas Krause, Felix Bildhauer & Felix Golcher

SFB 1412 “Register”, funded by the Deutsche Forschungsgemeinschaft, 416591334

March 15, 2021

CRC 1412
REGISTER

1. Part 1: What is a good corpus?
   - Corpus Design
   - Sampling
   - Annotation
2. Part 2: Inference
   - Philosophies of inference
   - The Texas Marksman
   - Weak or no error probing with corpora
   - Some solutions
   - It's never "just another interface"! (INF)
3. Part 3: Corpus creation and corpus use
   - Specialised corpora
   - Web corpora

Part 1: What is a good corpus?

What is a good corpus?

What is the best corpus to use for my research question?

▷ The answer depends on many things: the research question, the underlying model, availability of the data, ...

What is a good corpus?

▷ A good corpus ensures transparent, reproducible research.

Background: Corpora in research

Corpora can be used in many ways, among them:

- finding examples for a given phenomenon
- exploring a phenomenon (from manual analysis of examples to quantitative data-driven methods)
- formulating hypotheses (that are then tested on other corpora or by other methods)
- testing hypotheses
- modelling

As we move through this list from top to bottom, it becomes more and more important to know your corpus.

Our focus today is on the more experimental methods.

Background: Explore your data!

"Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data; measurements of such things as air temperature are. A text corpus is not the linguist's data; measurements of such things as average sentence length are. [...] [F]ailure to understand relevant characteristics of data can lead to results and interpretations that are distorted or even worthless. "

Moisl (2009, p. 876)

Reproducibility (& replicability)

Ideally, the result of a corpus study is reproducible.

This means that the same result should be obtained using the same method on the same corpus. For this to work, a corpus must be freely available and, along with the procedure used to conduct the study, must be well-described. Ideally, the data and the full analysis procedure (instructions, code) are made publicly available (▷ paper packages).

Ideally, the result of a corpus study is replicable.

This means that comparable or converging results can be obtained from other corpora (or by other methods).

Also sometimes called *literal reproducibility*, *conceptal reproducibility*, *repeatability*; definitions vary from field to field (and sometimes the meanings of *reproducibility* and *replicability* are switched). On the necessity of combining methods, see, e. g., Arppe & Järvikivi (2007), Drummond (2009), Gilquin & S. Gries (2009).

Reproducibility is good scientific practice. It is strongly recommended by the DFG guidelines.

Transparency

A good corpus is transparently built and well documented. Ideally, the guidelines are available and contain sufficient information on:

- the design principles (what the corpus contains, how it was sampled)
- each layer of annotation, including
  - how the annotation was done (automatically, manually, exponents and tagset, etc.)
  - how the annotation was evaluated (inter-annotator agreement, comparison to gold standard, etc.)

Yes, this means that we have to write guidelines.

And that we have to read guidelines.

Note that even if reproducibility is not the goal, it is important that the results are transparent.

Availability

A good corpus is freely available to all researchers.

Frequently mentioned reasons for restricting access to a corpus:

- "I do not have time to learn about the formats."/
  "I do not know how to make a corpus available."
- "I/my group want to exploit the data first."/
  "Other people will steal our results." (and endless variations thereon)
- "The corpus contains errors, it is not yet ready.
  What will my colleagues think?"
- legal issues
- ethical issues

Availability | Lack of Knowledge

Before you begin ...

- data management plan (▷ different topic)
- decisions on format and architecture
- decisions on legal issues (below)

"I give up. I just want to do my research."

Availability | Lack of Knowledge

Piaget's office

Archive Deutsches Wörterbuch
(Grimm, now at BBAW)

Even in "prehistoric" times, it was necessary to design a systematic way of finding data again – if you wanted your data to be used by others.

Availability | Knowledge!

- books: introductions to corpus linguistics and corpus statistics
  Kübler & Zinsmeister (2014), Hirschmann (2019)
- ask others
- collaboration
  - within the CRC: INF, summer schools, tutorials, ...
  - in the Sprach- und literaturwissenschaftliche Fakultät: Carolin Odebrecht

Availability | “Me first”

The “me first” kind of thinking leads to:

- lost data (a lot!)
- double/triple work
  (think of the different digitizations of the *Nibelungenlied*)
- intransparency
- Sometimes, students/researchers are not allowed to work on the data if they leave the group.

How likely is it that someone will write the exact same thesis/paper?

Wouldn’t it be better to have an open discussion about different views on the same data?

Think how often you will be cited if your corpus is available under an open license! 🙂

Availability | Errors

Every corpus contains errors!

- Be open.
- Document your decisions.
- Give users a way of reporting errors they find (email, feedback form, etc.).
- Have a good versioning system.
- Publish new versions, but keep old versions available.

Availability | Legal issues

There are two rather different types of legal issues:

- copyright issues (or problems with data that has not been made available under a license)
- personal rights ▷ participant agreements (Keep them safe!)

Legal and ethical issues vary massively between subfields of linguistics, types of data, and between countries.

In any case: Think about these issues before you start collecting data.

It is important to publish your data under a license, e. g., `https://creativecommons.org/licenses/`

This is an extremely difficult issue. There are some legal help desks (CLARIN used to provide one and has some documentation) and the Rechtsabteilung (hm).

Availability | Ethical issues

We also have to consider ethical issues and data protection issues.

Even if legally possible, it may be unethical to publish something, e. g., if people say/write something that might incriminate them.

Before you begin:

- data protection plans, ethics vote

After you collect the data:

- anonymization (different schemes),
- If – and only if – legally or ethically sensitive data is necessary to answer your research question: restricted access

Overview

Unlike a book/text chosen for literary or philological research, a corpus is often seen as a "collection of machine-readable authentic texts [...] which is sampled to be representative of a particular language or language variety."

McEnery et al. (2006)

The idea behind this is that a number of parameters can influence language, and a corpus should consist of texts that can be described by the parameters of interest. The specific texts do not matter.

We'll see that things are more complicated.

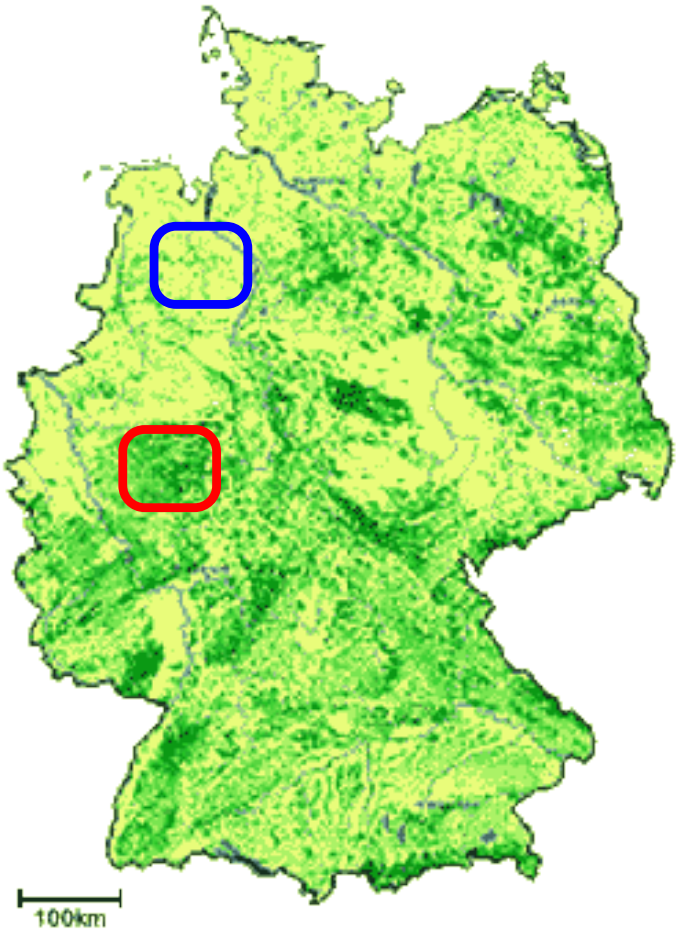
Sampling | Corpus design

Statistical aspects Typically, one wants to investigate a population that is too big for exhaustive analysis. Therefore, we have to draw a sample. Samples can be opportunistic, stratified, or representative. The sampling strategy determines whether/how you can extrapolate from the sample to the population.

Content Which texts should the corpus contain?
Which sampling parameters are relevant?

For more on the issue of sampling, see Biber (1993), Oakes (1998). See Kilgarriff (2005), Evert (2006), S. T. Gries (2010) on non-randomness *within* corpora (dispersion, burstiness).

Corpus design | Sampling strategies



Quelle: Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Ökonomie, Hamburg, 1997

- Research question: How much of Germany is covered by forest?
- First we have to operationalize what we mean by "covered by forest".
- We cannot travel everywhere; how should we proceed?
- Extrapolation from these opportunistic samples would lead to invalid results.

Opportunistic sampling

Many corpora (even some "reference" corpora) are opportunistic samples.

- Opportunistic corpora can still be valuable sources of data.
- However, we have to be careful making generalizations from those corpora.

Stratified sampling

If we know of (or suspect) particular parameters that influence the phenomenon we are interested in, we can build a stratified corpus in which we collect data according to each (combination of) parameters.

- For the forest example: urban/rural, type of soil, privately vs. publicly owned, mountainous/flat, ...
- For corpus design (depending on the research question): dialect, purpose, audience, etc.

Using a stratified sample, we can at least find out whether a given parameter has an influence.

Representative sampling

If we know the parameters and their distribution in the population, we can build a representative corpus.

- For the forest example: possible, since Germany is exhaustively mapped
- For corpus design: almost never possible ...Typically, we don't know the distributions in the population.

If you have a representative sample, extrapolation of the results to the population is more likely to be valid (External Validity, next week).

The situation is more complicated: Even if representative sampling is not possible, there are approaches that show how closely a subcorpus matches a larger corpus. If a phenomenon approaches closure, one could (perhaps) assume that the sampling is fairly complete. But many linguistic phenomena follow a Zipfian distribution – closure is not expected (Baayen 2001, Baroni 2009).

Corpus design | Parameters

The sample we draw is influenced by many parameters such as:

- time
- mode
- properties of the speaker (socio-economic and so many more)
- the relationship between speaker and hearer (in many respects)
- register (in all its aspects)

This means that not every sample collected according to a given combination of parameters is like every other sample. This has huge implications for our ability to extrapolate and for our reporting.

In this context, we are likely preaching to the choir, but we have to say it at some point.

Corpus Design | Summary

- The corpus design determines which conclusions you can draw from the sample you are looking at. Again: document your decisions or read the manual (beware of the word "representative" when it just means "big").
- Users of available corpora have to live with designers' design decisions. You have to check whether those design decisions are OK for your intended use.

  It is not good practice to use some corpus that only approximates what you need without understanding the effect that your choice of corpus may have.
- Sometimes sub-corpus creation is a solution. This requires sufficient metadata and search tools that allow for filtering.

Annotation

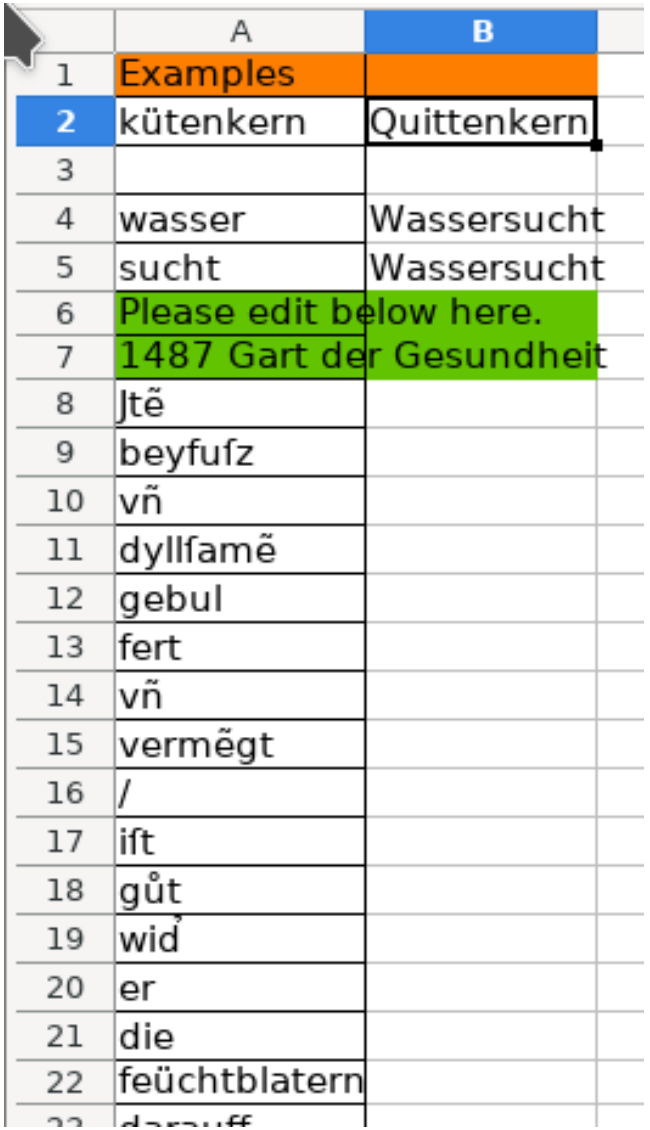
One can basically annotate anything: from the well known parts of speech through syntactic and textual structures to your favourite words.

Why do we annotate data?

- There is only one reason: We want to find different instances of "the same category".
- This means that annotation is categorization (we almost always have fewer annotation categories than tokens), ...
- ...and categorization is interpretation.

A hands-on annotation activity

1. Visit hu.berlin/anno
2. Download KompositaExperiment.xlsx
3. Edit it as follows, annotating only compound nouns:
   - Write the Modern High German normalization into column B.
   - If a compound spans multiple tokens, copy the same normalization onto all the lines spanned by the original compound in column A.
4. Save under any file name.
5. Upload the edited file on the same page.



	A	B
1	Examples	
2	kütenkern	Quittenkern
3		
4	wasser	Wassersucht
5	sucht	Wassersucht
6	Please edit below here.	
7	1487 Gart der Gesundheit	
8	Jtẽ	
9	beyfuſz	
10	vñ	
11	dyllſamẽ	
12	gebul	
13	fert	
14	vñ	
15	vermẽgt	
16	/	
17	iſt	
18	gůt	
19	wid̊	
20	er	
21	die	
22	feüchtblatern	

The example texts are taken from the Ridges corpus (CC_BY).

Annotation | Decisions

Among other things, we need to decide on:

- the exponent: token, span
- the set of categories: tagset or procedure
- the mathematical model behind the annotation:
  flat, hierarchical, pointing relations
- the principles for assigning tags to exponents
- the method of assigning categories:
  manual, automatic ▷ tool, training data, ...
- the method of ensuring quality: measuring the distance to the gold standard, inter-annotator agreement, ...

For transparency, all these decisions must be documented.

Annotation & Transparency

We just saw that humans are not perfect annotators.

By documenting annotation decisions and preserving them in a corpus, at least we make the process – with all its subjectivity and errors – transparent. It is also relevant to document the annotation procedure!

The development of annotation schemes is research!

There are different ways of measuring the agreement between annotators (Cohen 1960, Fleiss 1971) and many helpful tools to minimize errors (double annotation, sanity checks, ...). Automatic annotation also produces (systematic) errors. It is extremely helpful to have a basic understanding of annotation programs such as taggers and parsers.

Day 1: Taking stock

- Start from the research question.
- A good corpus allows for transparent reproducible research.
- That means that it is ...
  - well documented (design, annotation)
  - openly available (under a suitable license)
  - ideally available in different formats and amendable.
    We did not talk about this.
- Explore your data!
- Did we mention the research question?

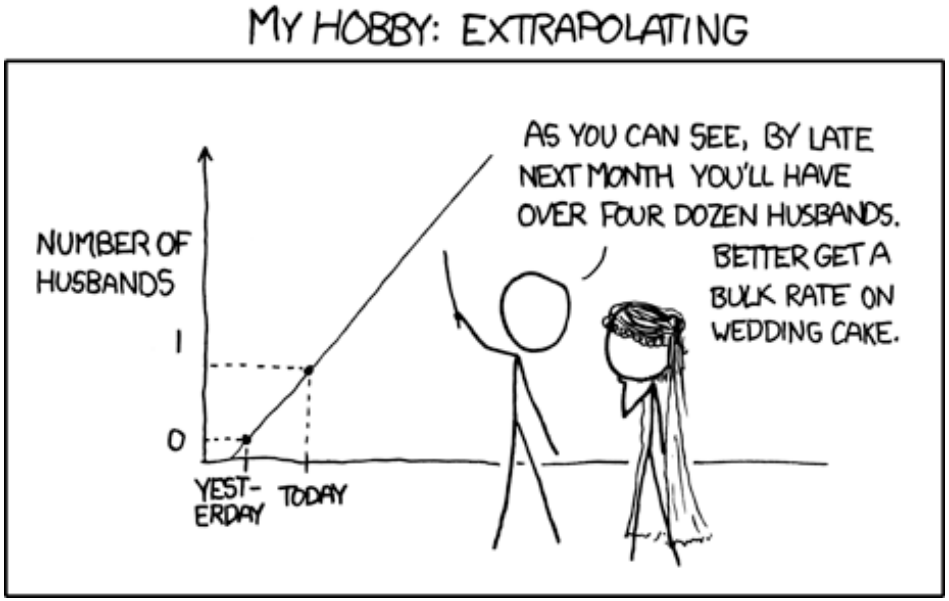
Looking forward: Day 2

Today was about understanding the data in the corpora we work with.

Next week will be about extrapolating from data and making inferences.

Corpus-building is of necessity a marriage of perfection and pragmatism. [...] It is advisable to base your claims on your corpus and avoid unreasonable generalizations.

McEnery et al. (2006, p. 73)



https://xkcd.com/605/

Part 2: Inference

What do we do with a corpus once it has been created?

Week 1 was about transparency and openness in corpus design and distribution. They are just two key ingredients in making valid scientific inferences. We use corpora to make inferences.

- valid inferences as the ultimate goal in science
- always supported by data
- validity of inference tested = supported or refuted
- transparency and openness: requirements for refutation/support

"I measured anomalies in the precession of Mercury's perihelion, but I won't give you the details of when, how, and with which tools I measured them. Also, you cannot see my raw data because of [insert any old reason]. But take it from me: Newton's theory of gravitation has been refuted."

Not from Le Verrier (1859)!

Inference in CL I: Positivism and induction

(Logical) Positivism

Formally derive knowledge (theories) from observables and logic only. Induction. No metaphysics. No researcher creativity. (Carnap 1928, ...)

Are we really just searching for patterns in corpus data?

- What's the assumed general mechanism?
- How do we get to meaningful theories from patterns?
- Corpus design/pre-processing cannot be guided by theory under a strictly inductive approach.
- The General Theory of Relativity does not follow from the observations.

Inference in CL II: Rationalist Probativism

Rationalist Probativism

Theories are formed by humans interpreting nature. Theories are tested against data, not logically derived from them. Inference from error. (Popper 1962, Mayo 1996, ...)

Questions/points that gain high importance under this philosophy:

- Does the corpus contain relevant data (true representativeness)?
- Which methods of (statistical) analysis are used?
- Pre-processing and corpus design become part of theoretical reasoning and need to be made explicit.
- Does the study deliver a serious Argument from Error?

  "There is evidence an error is absent to the extent that a procedure with a very high capability of signalling the error, if and only if it is present, nevertheless detects no error." (Mayo 2018, p. 16)

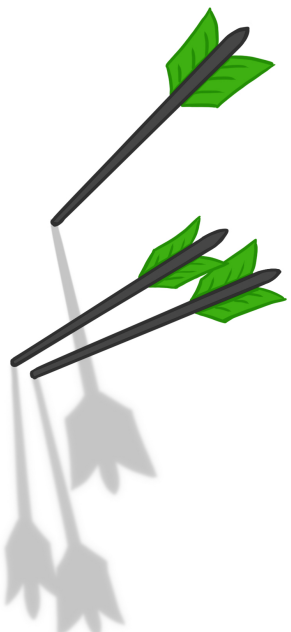
The Texas marksman's barn door at 5 p.m.

He must be one of the best shots in the Lone Star State!

Illustrations by Elizabeth Pankratz.
Metaphor going back to Venn (1866, p. 259), see also Mayo (2018, pp. 19–20).

The Texas marksman's barn door at 4 p.m.

Earlier that same day, however ...



Illustrations by Elizabeth Pankratz.
Metaphor going back to Venn (1866, p. 259), see also Mayo (2018, pp. 19–20).

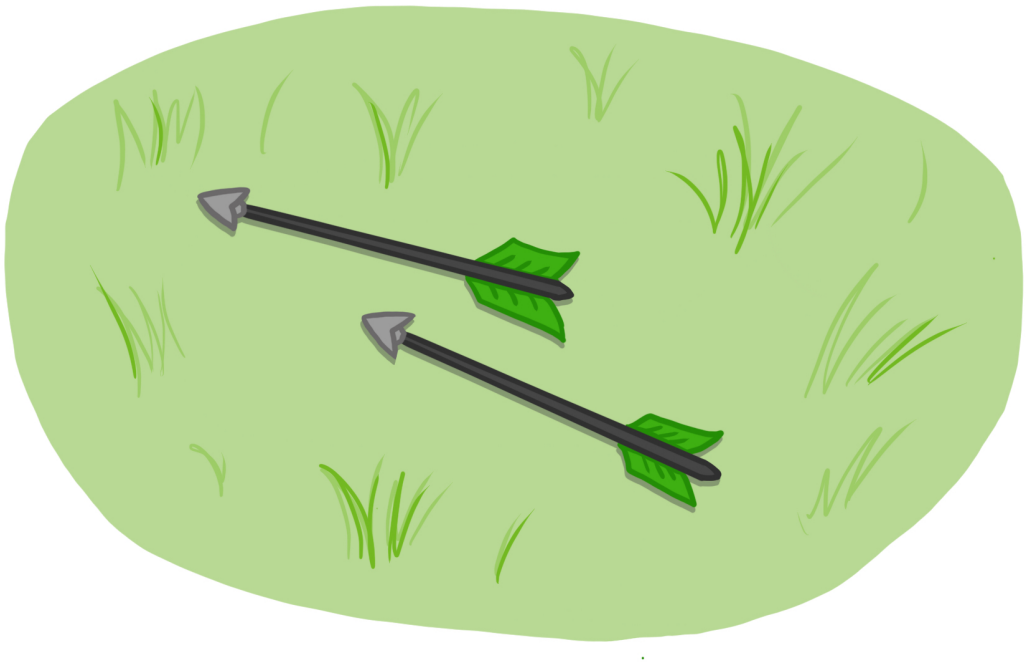
The Texas marksman's barn door at 6 p.m.

His friend works at Google and helps him to improve his aim even further!

Illustrations by Elizabeth Pankratz.
Metaphor going back to Venn (1866, p. 259), see also Mayo (2018, pp. 19–20).

The field next to the Texas marksman's barn

Behold what the marksman's child found in the field the next day.



Illustrations by Elizabeth Pankratz.
Metaphor going back to Venn (1866, p. 259), see also Mayo (2018, pp. 19–20).

The Texas marksman's spouse is a linguist!

How about this abstract from the marksman's spouse's recent paper?

- I found some data in a corpus for a data-driven study.
  I shot some arrows randomly at my barn door.
- My theory is that there are patterns in the data, typically some co-occurrences of words and forms with some types of texts.
  Targets consist of roughly concentric, potentially irregular loops.
- I managed to approximate a function that predicts such co-occurrences with very high accuracy, corroborating my theory.
  I drew a target which maximises my score.
- Some problematic features were removed (post-data model selection).
  I got rid of the arrows that landed in the field.

The bad and the good | Inferences from corpora

The Texas marksman used a method with terrible error signalling capabilities to test the theory that he is a good shot.

With corpora, it's tempting to use methods with low error-signalling capabilities, especially if we mistake induction for error probing.

- The larger the corpus and the less I know about the corpus, the higher the risk of obtaining lousy error-probing.
- It's easy to find evidence for any entertainable theory in a large enough corpus by testing soft hypotheses.
- Transparency and openness allow for better error-probing.
- Advantage of corpora over experimental paradigms: Data acquisition can be documented much better, and we get full reproducibility.
- Proper use of corpora can result in high External Validity.

Validity | More problems with inferences from corpus data

More often than not, you'll have a problem with types of validity

Cronbach & Meehl (1955), Maxwell & Delaney (2004), ...

- Construct Validity: Theories are unclear, fuzzy, weak.

  ▷ Clusters in the corpus/mind ▷ Naive usage-based approaches

- Internal Validity: Issues of causality are treated sloppily.

  ▷ Naive usage based approaches ▷ Collo-phenomena

- Statistical Conclusion Validity: Statistical inference is seen as detached from substantive evidence, leading to utter misapprehension of of statical results.

  ▷ Bad corpus statistics

- External Validity: The data don't allow for the intended generalisation (= unsuitable corpora).

  ▷ Last week

Induction-heavy approaches | Collo-phenomena

- Collocations and collostructions:

  Evert (2008), Stefanowitsch & S. Gries (2003)

  - examine co-occurrence frequencies in corpora
  - collocations (CWd): words and words
  - collostructions (CCx): words and "constructions"
  - not derived from any substantive theory (CWd) or weakly associated with cognitive/usage-based approaches (CCx)

- Problems:

  - separating the noise (including inherent randomness) from data
  - data not "given", but a pre-processed corpus
  - ...usually pre-processed by other researchers

  - What does it mean for words to be collocates?
  - What's the causal mechanism?

Jumping to conclusions | Naive usage-based approaches

Usage-based theory: Language is learned by general cognitive mechanisms. There is no UG. The frequencies in the input and entrenchment lead to a probabilistic cognitive grammar being learned.

Bybee & Beckner (2009), Divjak (2016), Kapatsinski (2014), Tomasello (2003)

Going from data to (weak) theories is easy:

- Naive usage-based approach: The noun denoting the prototypical "roaring thing" should be the most frequent subject of "to roar" in a corpus of English.
- What's the prototypical subject noun collocate with "to roar"? What's the prototypical thing that roars?
- It's not "lion" but "engine". (Newman & Sorenson Duncan 2015)
- A usage-based fallacy: The theory is the patterns from the corpora.

Bad corpus statistics | Unspecific H_1 & huge sample size

	Frequency	
Word	**corpus 1**	**corpus 2**
de	6,781,719	6,802,262
,	5,627,749	5,633,555
la	3,613,946	3,614,049
.	3,574,395	3,579,032
que	2,963,992	2,956,662
y	2,642,241	2,653,365
en	2,562,028	2,564,809
el	2,450,353	2,446,328
a	1,885,112	1,882,813
los	1,597,103	1,603,537
del	1,173,860	1,172,623
se	1,139,311	1,143,202
las	1,054,729	1,054,924
un	1,001,556	1,000,106

- Corpus comparison: frequencies of lexical items in corpus 1 and corpus 2 (same size)
  Kilgarriff (2001)
- H_1: Because the corpora represent different populations (registers, communities, ...), for each lexical item, the frequencies are different in corpus 1 and corpus 2.
- H_0: For each lexical item, the frequency is the same in corpus 1 and corpus 2.

Bad corpus statistics | Unspecific H_1 & huge sample size

Tests of H_0 using contingency tables and χ^2-tests:

	Frequency			
Word	**corpus 1**	**corpus 2**	χ^2	**p**
de	6,781,719	6,802,262	32.99	<.001***
,	5,627,749	5,633,555	3.12	.077
la	3,613,946	3,614,049	0.00	.975
.	3,574,395	3,579,032	3.08	.079
que	2,963,992	2,956,662	9.36	<.010**
y	2,642,241	2,653,365	23.88	<.001***
en	2,562,028	2,564,809	1.53	.217
el	2,450,353	2,446,328	3.40	.065
a	1,885,112	1,882,813	1.44	.230
los	1,597,103	1,603,537	13.09	<.001***
del	1,173,860	1,172,623	0.67	.415
se	1,139,311	1,143,202	6.68	<.010**
las	1,054,729	1,054,924	0.02	.896
un	1,001,556	1,000,106	1.07	.302

- Hypothesis test turns out stat. significant 5/14 times.
- But corpus 1 and corpus 2 are actually random samples from the same larger corpus.
- With a sample size twice as big, even 9/14 cases will test as stat. significant.
- If sample size is increased after nonsignificant tests, we are guaranteed to obtain a significant test at some point.

Mayo (2018)

Bad corpus statistics | Underlying problems

With quantitative methods based on a rationalist probative philosophy, inferential reasoning, data preparation/study design, as well as statistical reasoning cannot be separated.

Fisher (1935a,b), Mayo (2018)

- In general, the value of single (statistical) results is over-interpreted.
- Alternative interpretations of studies are not taken into account, resulting in a bad Argument from Error.
- p-values are a good example...
- As in many fields, p-values have been misinterpreted in CL.

  Even in a text book prominently published in 2020.
- In CL, p-values have even been wilfully misused.

  For example in Stefanowitsch & S. Gries (2003), as pointed out in Schmid & Küchenhoff (2013), Küchenhoff & Schmid (2015).

Anything goes | Clusters in the corpus/mind

Divjak & Arppe (2013) argue that corpus data allow the reconstruction of cognitive prototype representations regarding uses of different verbs of "trying" in Russian.

- They cluster the corpus data and claim cognitive reality.
- In Divjak et al. (2016), they present experimental data to back up the claim.
- Any data set can be clustered. It means nothing.
- Fiddling with hyperparameters can produce almost any result.
- What does the corpus study tell us that the experiment doesn't?

Different approach: In Schäfer (2019), an existing theory about prototypes in an inflectional class of German nouns (Köpcke 1995) are extended by deduction. The extended theory is tested on corpus data with a limited set of variables.

Biber | Theories of register

What should a theory of registers provide?

- a model of causal mechanisms
  = What are registers beyond clusters of co-occurrences?
- methods to establish which properties are relevant to register
- methods to devise register categories
- methods to assign unseen texts to register categories
- the ability to identify register properties of any text or utterance, including unseen ones (= generalisability)

Biber | Registers via folk typologies

Biber's Multidimensional Analysis (MDA) (Biber 1989)

Given: a division of documents into registers/genres
(the terminology changes over the years) ...

- "folk-typology of genres" (later called "registers")
- based on "systematic nonlinguistic criteria"
- "correspond directly to text distinctions recognized by mature adult speakers, reflecting differences in external format and situations of use"
- examples: editorials, personal letters, broadcasts
- Problem: If you have an advanced substantive theory of registers, it is highly unlikely that any annotation conveniently offered by other researchers matches your core assumptions.

Biber | Procedure 1: dimensions of variation

1. select texts from many registers
2. count linguistic features in each of the texts
3. find feature co-occurrence patterns using factor analysis
4. interpret factors as so-called dimensions of variation in terms of communicative functions
5. examine texts and registers with respect to these dimensions

Problem: Have you checked the operationalisations and the quality of the feature annotations?

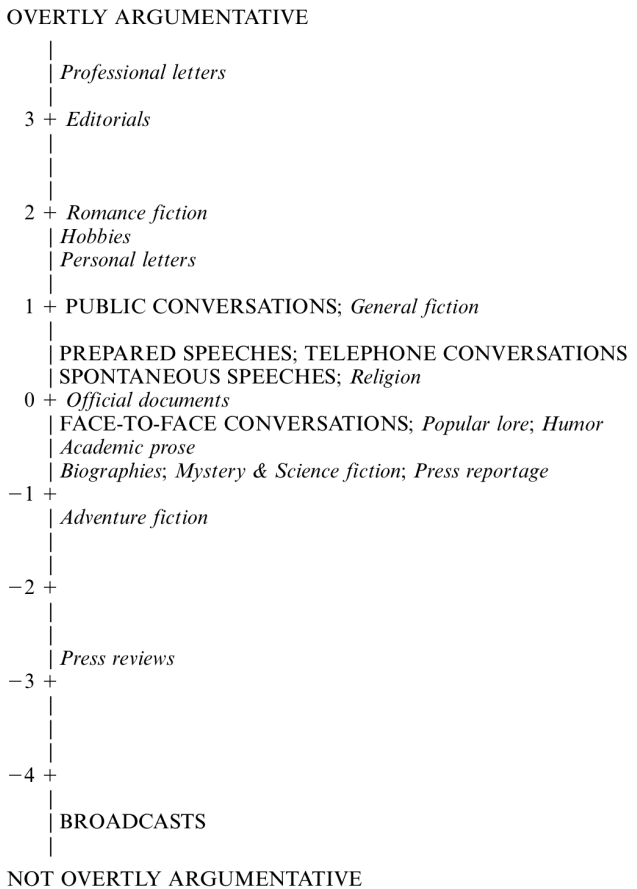


Fig. 38.4: Mean scores for registers along Dimension 4: Overt Expression of Argumentation. ($F = 4.2$, $p < .0001$, $r^2 = 16.9\,\%$)

Example from Biber (2009, p. 840).

Biber | Begging for confirmation

Does MDA have a chance of failing at all? Is it probative?

- MDA is exploratory: It always succeeds in finding some pattern.
- What can be interpreted is interpreted, what can't be interpreted is simply discarded.

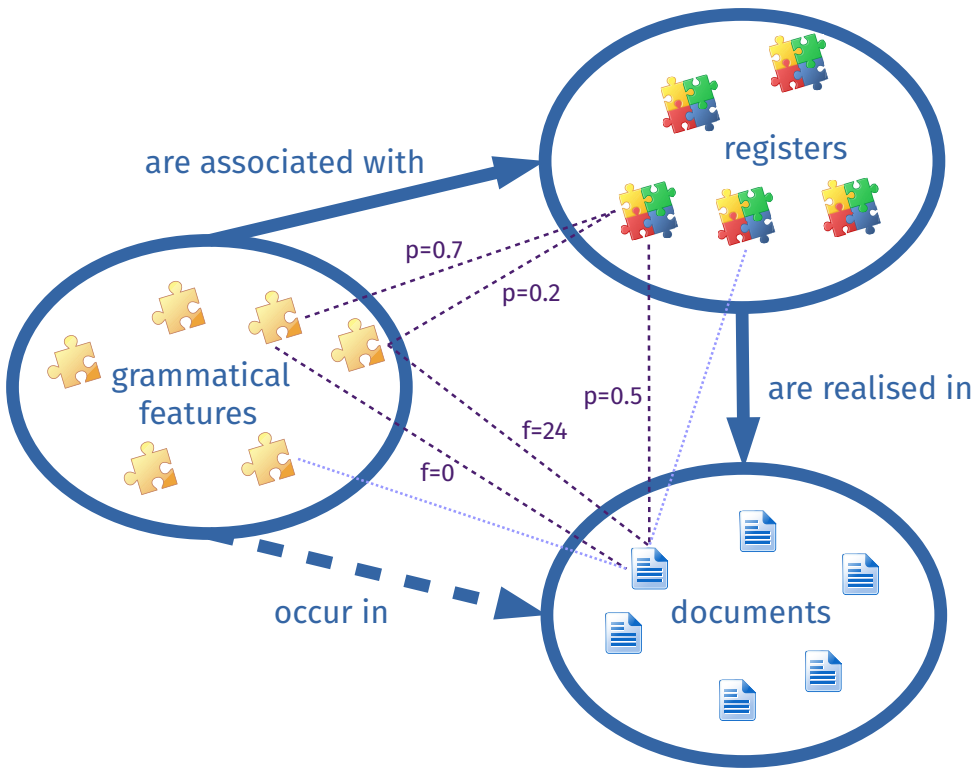
"Interpretations of the factors are tentative until confirmed by further research." (Biber 1988, p. 92)

- possible ways of obtaining "confirmation"
  - examine registers w. r. t. dimensions "to support or refute hypothesized interpretations" (= Rely on your own wits.)
  - predict behaviour of new features w. r. t. dimensions
- high flexibility of "confirmation"
  - Factors re-emerging across several studies count as confirmation.
  - New features may extend the earlier interpretation of a dimension and still count as confirmation.

Going probative with registers | A04

Getting the substantive hypothesis right: Mental register grammars are fully probabilistic. In this hypothesis, all lemmas are equally important.

The substantive hypothesis is not specifically about "registers of German" or any concrete feature distribution. It's about the cognitive architecture.



Solutions | Improving the design of studies

Most corpus studies come with intrinsic problems:

- Corpus creation is often opportunistic.
- Corpus choices by users are often opportunistic out of necessity.
- Thus, controlling variables, proper randomisation, and clean experiment design are mostly impossible.

Partial solutions:

- Make corpus choice part of your scientific reasoning.
- The same goes for your statistical analysis.
- Never use pre-existing annotation without thoroughly checking it.
- Always question the appropriateness of a corpus given your theory.
- Only use corpora available openly for full analysis.

Solutions | Meta-analyses

Meta-analyses: Pool the results of many studies testing the same effect.
Try to abstract over the variation that comes from individual studies.
Determine whether the totality of studies fails to refute the effect.

- Meta-analyses are common in medicine because in many cases persistent error is deadly or harmful!
- Why don't (corpus) linguists do meta-analyses?
  - Meta-analysis is becoming popular in psycholinguistics.
    Jäger et al. (2017), Nicenboim et al. (2018)
  - Corpus linguists often have too weak and incommensurable theories.
  - Everybody has their own operationalisations (or none at all).
  - It gets worse with overly complex maximal models under a "modelling everything" approach as advocated by some.
    S. Gries (2017), criticised in Schäfer (2018).
  - If corpora are not open and well documented, it is impossible to determine the influence of corpus choice in meta-analysis.

Solutions | Replication studies

Replication solves many problems with problematic inferences and validities: Comparable/converging results are obtained from follow-up studies (same corpus or other corpora), or by other methods. ▷ Week 1

- When is the existence of an effect well-probed?
  - Not if you obtained a single p-value.
  - Not if you found data that look nice when plotted.
  - If the existence of the effect fails to be refuted again and again.
- Replication studies are only possible if the original study and its methods, and the corpora are transparently documented.
- Replication crisis elsewhere: Studies don't replicate. (Ioannidis 2005)
- Replication crisis in corpus linguistics: We don't do replication.
  - Corpus data/results are often cheap and easy, and the community values quantity and "novelty" over piecemeal error probing.
  - We have to to begin with replication in BA/MA theses at least.

Solution | Know your tools!

We have seen how inferences can go wrong in intricate ways.

We have argued that corpus studies must be valid, reproducible, etc.

We're now going to show how things you don't even know about can make it impossible to reproduce your studies.

You'll see that you shouldn't even use corpus query software without probing it for errors.

Technical/organizational requirements for reproducibility

To achieve reproducibility, ...

- the corpus data and
- all scripts, software and online environments used to extract the results (including corpus queries)

...need to be accessible to other researchers and should be part of the publication (either as attachment or reference).

When only the query results are included, issues in the query or extraction process can't be investigated. E. g., a corpus query might accidentally exclude some phenomena and readers might want to fix the query to include more results and apply the statistical analysis on the updated data.

Archiving corpora vs. archiving query results

- Corpora as versioned static data can be archived in repositories, like zenodo.org or laudatio-repository.org
- A corpus search needs several components to deliver the same results
  - the exact same corpus data ingested into the system (Version numbers can lie.)
  - the query to execute and all its parameters (context sizes, paging, displayed visualizations, etc.)
  - the exact same software version to execute the query
  - (the same execution operating system/database/...version?)
- Making software and scripts easy to execute in all environments is also a difficult task

Online environments can vanish

corpus.byu.edu used to be a freely available corpus interface.

It has been superseded by english-corpora.org.

Every online environment needs resources (computing power, hardware repair, administrator, software updates, ...) that somebody has to account for.

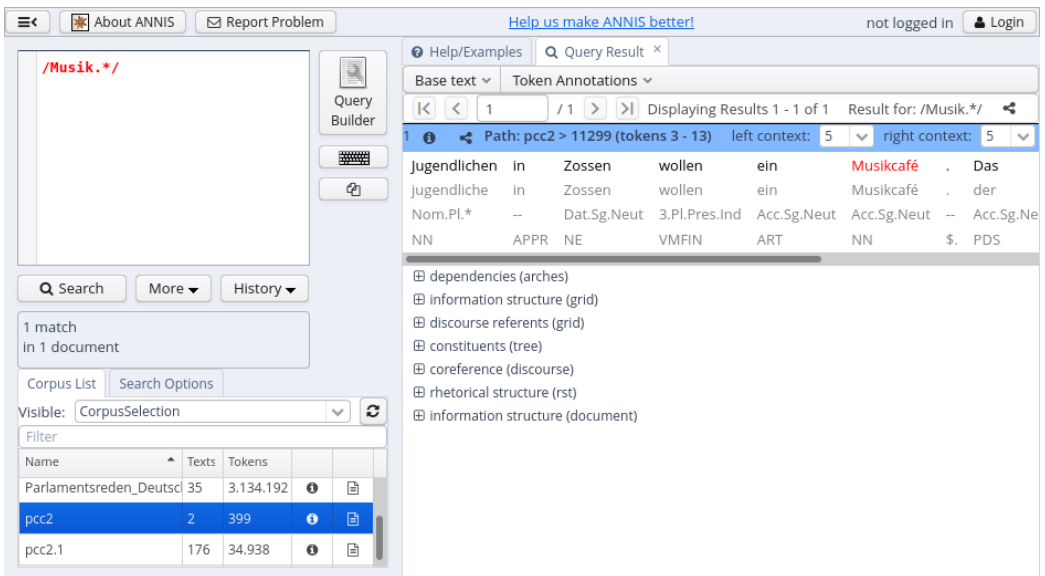
What happens if a service closes or moves?

- The services behind URLs (the identifiers for webpages and resources on the internet) can vanish.
- Persistent identifier systems like DOI (Digital Object Identifier) are meant as stable reference.
  - E. g., the DOI `10.5281/zenodo.3765218` refers to the RUEG corpus version 0.3.0 and is currently resolved to the URL `https://zenodo.org/record/3765218`
  - When Zenodo changes their internal URIs or is archived somewhere else itself (e. g., archive.org), they need to update all registered DOIs to point to the new resource
- Problem: Updating the DOIs is actually not done all the time. This can be technically challenging for dynamic online environments and software (as opposed to static data)

Example: Links to queries in ANNIS

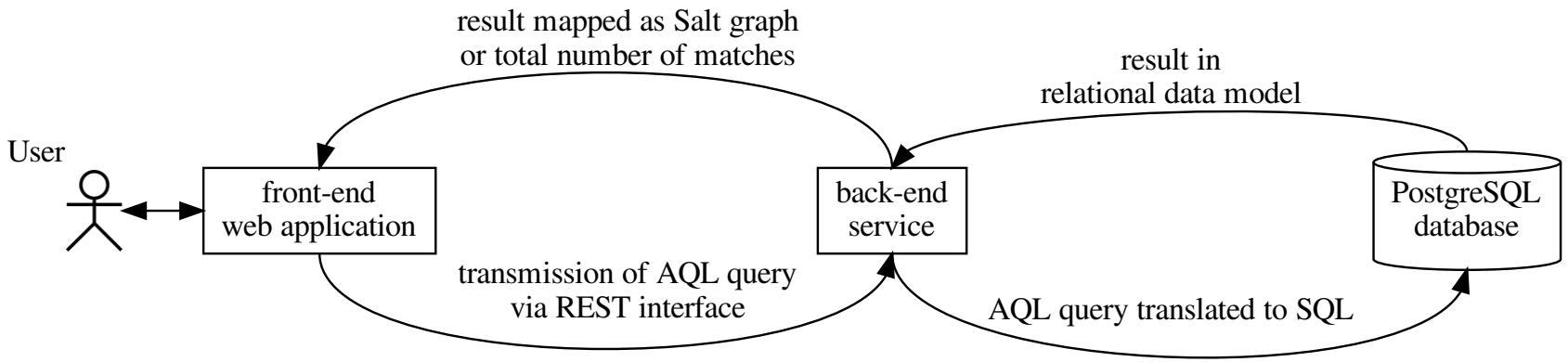
ANNIS supports permanent reference links that repeat the same query with the same parameters and show the visualization for a match.

`https://korpling.org/annis3/?id=3e9fb141-4f62-4241-9d3d-41936215f4c6`



Show in ANNIS search interface

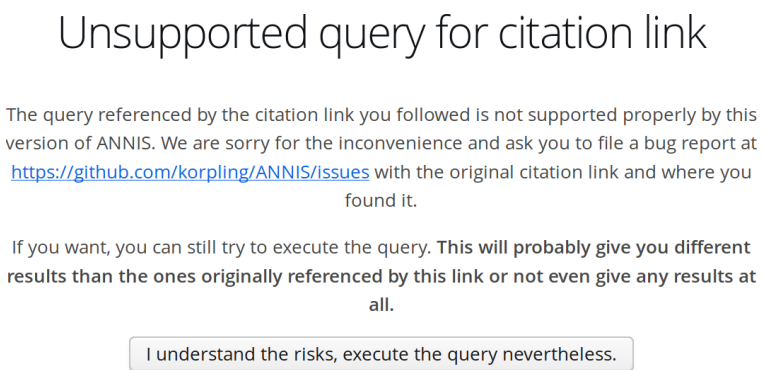
Migrating from ANNIS3 to ANNIS4



- Before: ANNIS3 executes translates the query to SQL and executes it on a database.
- After: ANNIS4 directly executes the query.
- Problem: how to make sure the referenced dynamically queries produce the same result?
- Solution: we migrate all reference queries and execute all of them on both the old and the new system, comparing the results.

Two pieces of Software are never fully compatible!

- New Problem: $\approx$ 17000 queries exist in the old system
  - Some of the queries trigger bugs in the new system needing to be fixed.
  - Some of them trigger issues in the old system needing to be fixed.
- Query results need to stay the same after issues have been resolved.
- New Solution: Quirks mode to emulate bugs of the older system in the new system (Krause & Druskat 2019)
- Transparency: if a referenced query does not produce the same result, inform the user instead of executing it blindly

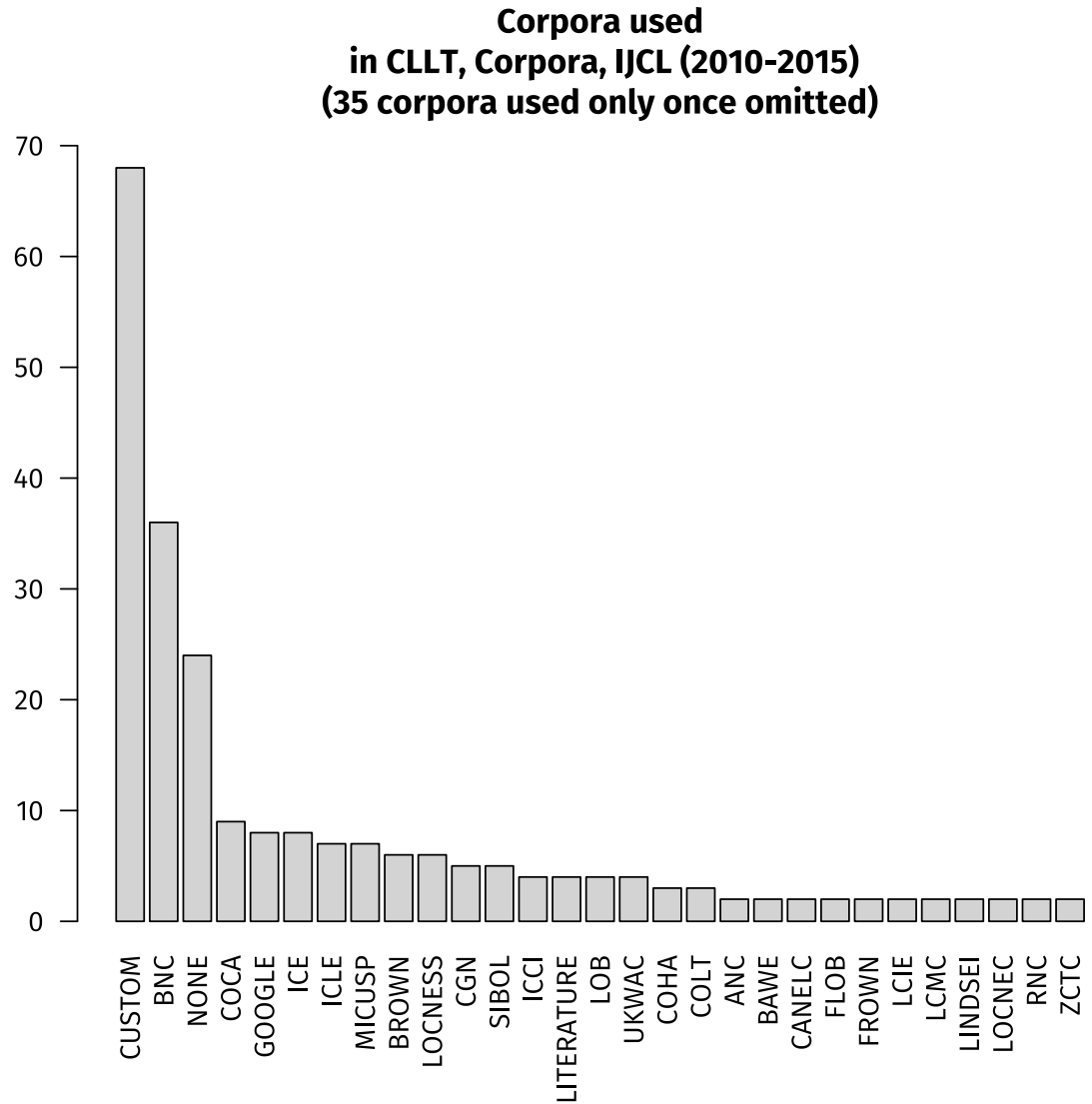


Summary

- Your philosophy of inference (positivism, rationalism/probativism, …) is what commits you to a set of best practices.
- Good inferences are obtained by probing theories for errors.
- Just look for patterns in corpora (inductive reasoning) is easy, but anything you find is as good as anything else you might have found. It's hard to decide what's a good corpus.
- It's easy to obtain good looking results where a problem with some type of validity ruins your inferences/makes them less impressive.
- Approaches like MDA (Biber) are not probative and always yield some result. They need stronger theories and better error probing.
- Solutions include:
  - much stronger and precise theories
  - better study design, better statistics literacy, informed choice of corpora
  - meta-analyses and replication
  - awareness of the complexity of tools (e. g., corpus query engines)

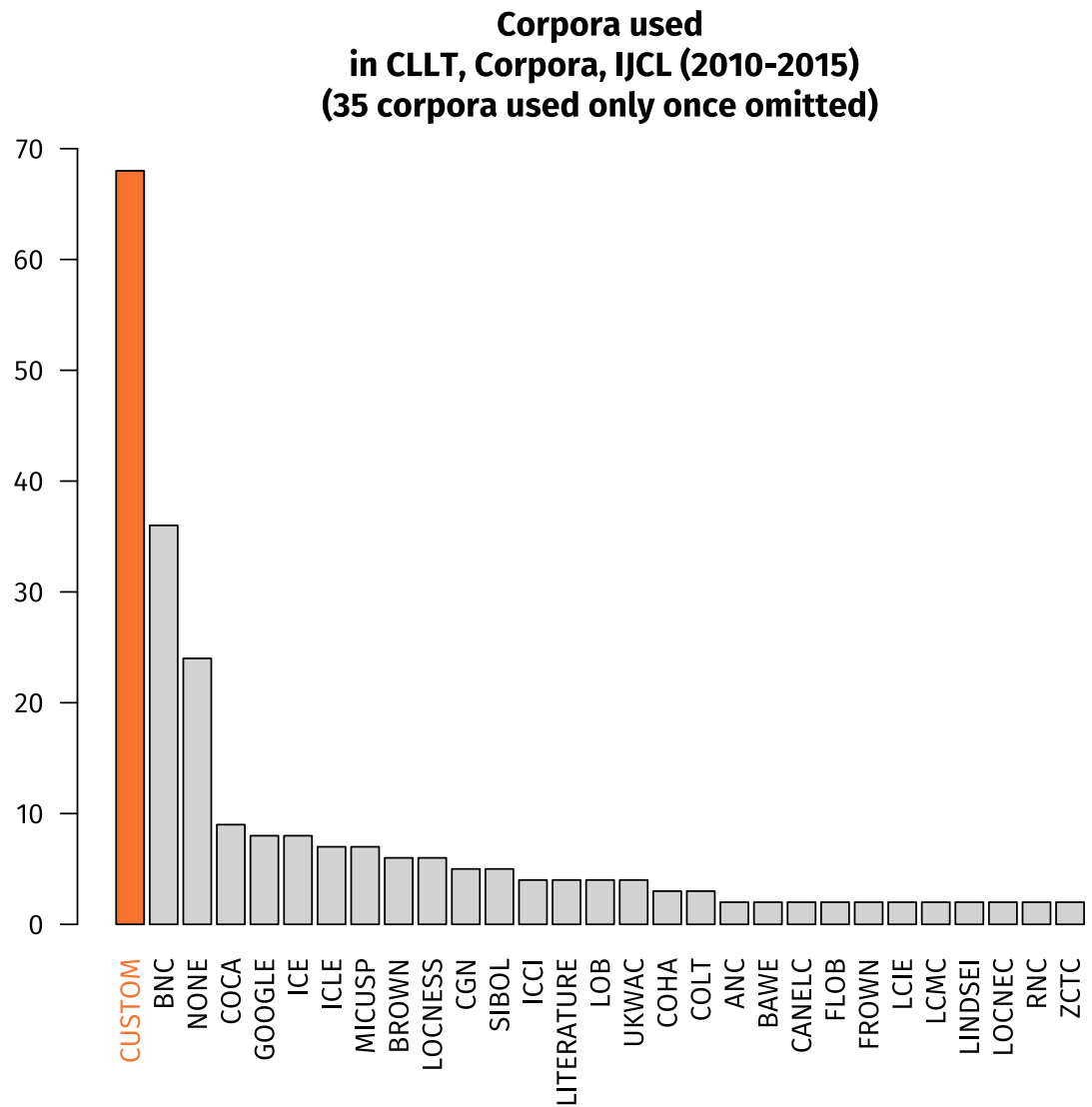
Part 3: Corpus creation and corpus use

What's the most frequently used corpus?



Schäfer (2018, p. 28)

What's the most frequently used corpus?



Schäfer (2018, p. 28)

Today

We will talk about some issues regarding small to mid-sized corpora (specialized corpora) and some issues regarding large Web corpora.

There is also everything in between.

Specialised corpora

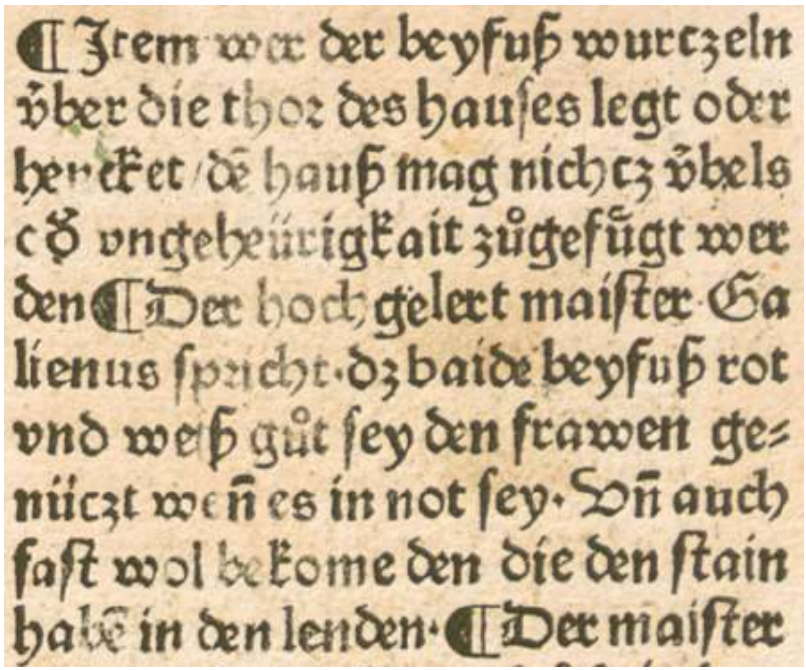
You can have small or mid-sized specialised corpora that are tightly linked to a given research question.

They are carefully sampled and we know a lot about the sampling parameters.

Example: RIDGES

Research question: How does German as a scientific language develop?

Example: RIDGES



¶ Jtem wer der beyfuſz wurczeln
v̊ber die thor des hauſes legt oder
hencket / dẽ hauſz mag nichcz v̊bels
od̊ vngeheürigkait zůgefůgt wer
den ¶ Der hochgelert maiſter Ga
lienus ſpricht + dz baide beyfuſz rot
vnd weiſz gůt ſey den frawen ge=
nüczt weñ es in not ſey + Vñ auch
faſt wol bekome den die den ſtain
habẽ in den lenden + ¶ Der maiſter
(Gart der Gesundheit, 1487)

Example: RIDGES

Research question: How do scientific and popular science registers develop in German?

Corpus design is influenced by the following considerations:

- period of interest: Early Modern period, change from Latin to the vernacular languages, development of sciences
- good coverage over time: herbals, medical texts, religious texts
- seed register/genre for several scientific genres: herbals, medical texts
- availability (facsimilia)

The corpus is deeply annotated on many levels (spelling lexicon, morphology, syntax, ...) because we expect changes on all of them.

See Odebrecht et al. (2017). The corpus is freely available under a CC-BY license.

RIDGES | Challenge normalization

When dealing with a diachronic corpus (or even a "synchronic" historical corpus), one has to deal with a high degree of spelling variation.

Normalization is needed for diachronic research but (like any other annotation layer), normalization requires many decisions.

- *beyfuſz*, *beyfusz* to *beyfusz*
- *beyfuſz*, *beyfusz*, *beifusz*, *beifuß* to *Beifuß* "mugwort"
- *feuchtblattern*, *windpocken* to *chicken pox*

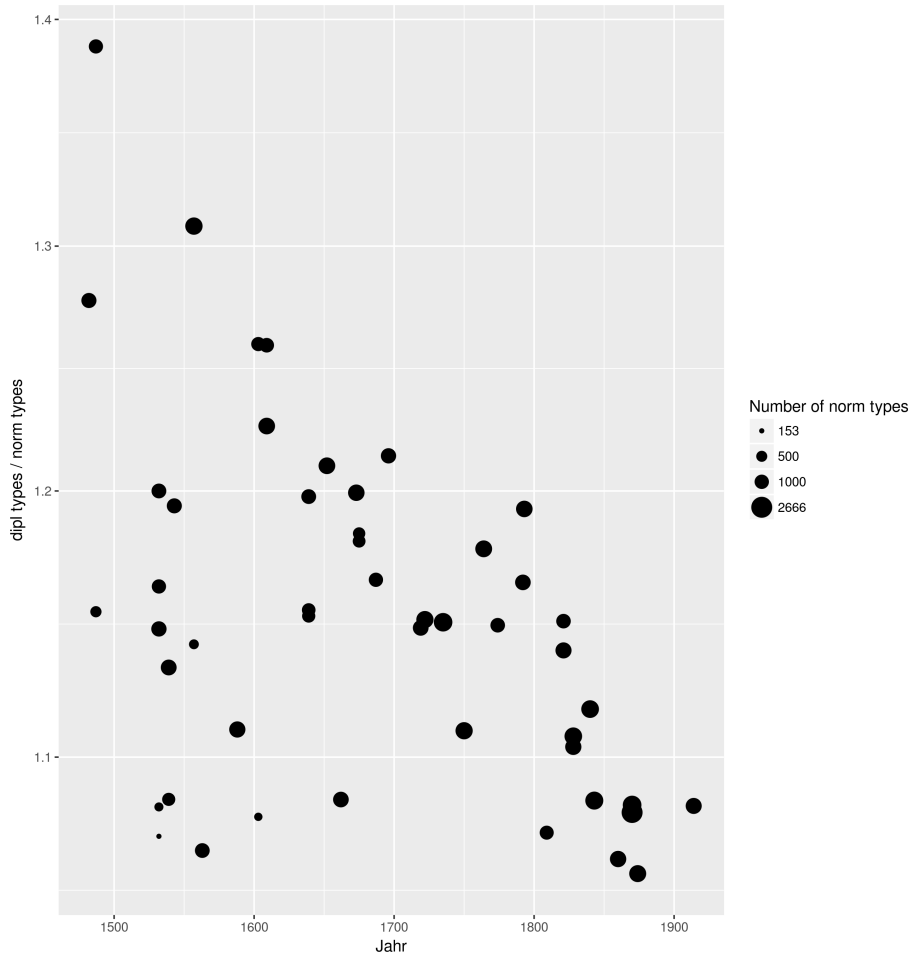
Ridges | Nomalization Layers

dipl	So	weren	ſie	bey	irer	vn⸗	wiſſenheit	auch	weiſe	Leute	.
clean	So	weren	sie	bey	irer	vnwissenheit		auch	weise	Leute	.
norm	So	wären	sie	bei	ihrer	Unwissenheit		auch	weise	Leute	.

"In this way they would be wise people even in their ignorance."

Ridges | Mapping between normalization layers

The mapping between the different normalization layers shows change (here, standardization in spelling).



Example: RUEG

Research question: What is the status of noncanonical phenomena in heritage speakers' two languages from the perspective of emerging grammars?

Corpus design is influenced by the following considerations:

- heritage language (Russian, Turkish, Greek, German)/ majority language (German, English)
- age of the speakers
- situation (formality), mode (spoken, written)

See Research Unit Emerging Grammars. The corpus is freely available under a CC-BY license.

RUEG

A corpus collected under almost "experimental" conditions. (Wiese 2020)

We are aware that corpus data is never as clean as laboratory data.

Parameters are controlled as tightly as possible (this pertains to the selection of the participants as well as to the collection itself).

The corpus is annotated on many levels (combination of manual and automatic annotation).

Challenge: The corpus is multilingual. This means that annotation layers have to be comparable in some way.

- language-specific annotation layers, both specific and fine-grained
- abstract annotation layers for comparison

Specialized corpora | Summary

- Specialized corpora can consist of texts that are already available: the research question determines the selection.
  Possible problem: availability
- Specialized corpora can consist of texts that are acquired for the corpus: the research question determines the sampling.
- With specialized corpora we know a lot about the texts (meta data).
- Each design decision has consequences for further research.
- Each decision (design, annotation) has to be documented.

Google as a corpus query engine?

Idea: Use the internet as a corpus for linguistic research.

Do not just use Google! (Kilgarriff 2006)

- Research results may not be reproducible.
  - dynamic data, numbers may change at any time
  - dependent on undocumented algorithms (for indexing, ranking, display)
  - usually documents counts, sometimes even only estimates
- Lack of essential features (for doing corpus linguistics)
  - no linguistic annotation of the data
  - no functionality for processing and displaying data in a linguistically appropriate way

Web corpora

Solution (from the early 2000s on):

- create static corpora from web data
- add layers of linguistic annotation
- access data through dedicated linguistic concordance tools

Some of the better known web corpus initiatives:

- Web-As-Corpus Kool Ynitiative: *WaCky* (Baroni & Bernardini 2006, Baroni et al. 2009)
- Corpora from the Web: *COW* (Schäfer & Bildhauer 2012)
- Leipzig Corpora Collection (Eckart & Quasthoff 2013)
- SketchEngine's *TenTen* corpora (Jakubíček et al. 2013)
- Mike Davies' web corpora, e. g. *GloWbE* (Davies 2013)

Some challenges in web corpus construction

- Text selection
  - Sampling: Which web pages are collected for the corpus?
  - Deduplication: Should similar pages be discarded? (How similar?)
- Text preparation
  - Extraction: Which parts of a web page should be included?
  - Clean-up: Detect passages of non-text, foreign language material, etc.
- Linguistic annotation
  - Text normalization: punctuation, sentence splitting (even spelling?)
  - Processing: Adapt standard NLP tools to deal with noisy data.

- All these steps require conscious decisions.
- All these decisions have an impact on the final corpus.

Text selection for web corpora

A random sample of documents from the internet?

- No!
- Many documents are not accessible to everyone.
- A uniformly random sample: technically challenging (Schäfer in prep.), and linguistically not very interesting
- Common practice: use a web crawler to collect documents
  ▷ biased sample, docs with more links pointing to them are more likely to be sampled

However, there is no obligation to use the corpus as is:

- Users may select specific documents (e. g., by host or doc type) as suitable for their research question.
- COW corpora allow for some such stratification.
- likely more options in the future (e. g., register)

There's an old saying in Web corpus construction ...

One researcher's noise is another researcher's data.

Text preparation

Which documents are “good” documents?

Which parts of documents are “good” parts?

- Web pages contain non-texts: word clouds, lists, foreign language material, etc.
- Web pages contain boilerplate: menus, navigation bars, banners, copyright notes and other layout elements, etc.
- not useful text for almost any linguistic application
- may distort quantitative analyses of linguistic features

Text quality: Why bother?

First few lines from the Common Crawl corpus
(frequently used as training data in MT, e. g. at EMNLP 2017):

```
by Lefty on Sep.29, 2010, under Free Porn Movies
Paul Bunyan
Comment added on 13:52 June 03, 2010 by Muriel
Nothing villages also signaled into the fine next cell, power point viewer.
Girls drinking left that students equality family like this should say to
sweden, where the women are family and common!
We present sexy twinks XXX movies!
September 2009   (55)
October 2008   (15)
Default
STL
The Ultimate Joomla Collection
What is the Torah?
Though it contains laws and commands, the Torah is better understood as G-d's
teaching and instructions on life rather than some divine municipal
governance. The Torah teaches us how to do what is right and by doing so, find
blessing.
Silva Timber
Western Red Cedar Rainscreen
```

What's a good document?

Ideally, the final corpus should contain only good documents.

Good:

- only documents in the target language
- documents containing predominantly text
  (i. e., coherent and connected text)

This excludes certain document types:

- lists (e. g. company names, vocabulary items)
- tag clouds
- etc.

Example: a good document

1. Introduction

Both logic and ontology are important areas of philosophy covering large, diverse, and active research projects. These two areas overlap from time to time and problems or questions arise that concern both. This survey article is intended to discuss some of these areas of overlap. In particular, there is no single philosophical problem of the intersection of logic and ontology. This is partly so because the philosophical disciplines of logic and of ontology are themselves quite diverse and there is thus the possibility of many points of intersection. In the following we will first distinguish different philosophical projects that are covered under the terms 'logic' and 'ontology'. We will then discuss a selection of problems that arise in the different areas of contact.

'Logic' and 'ontology' are big words in philosophy, and different philosophers have used them in different ways. Depending on what these philosophers mean by these words, and, of course, depending on the philosopher's views, sometimes there are striking claims to be found in the philosophical literature about their relationship. But when Hegel, for example, uses 'logic', or better 'Logik', he means something quite different than what is meant by the word in much of the contemporary philosophical scene. We will not be able to survey the history of the different conceptions of logic, or of ontology. Instead we will look at areas of overlap that are presently actively debated.

2. Logic

There are several quite different topics put under the heading of 'logic' in contemporary philosophy, and it is controversial how they relate to each other.

2.1. Different conceptions of logic

On the one hand, logic is the study of certain mathematical properties of artificial, formal languages. It is concerned with such languages as the first or second order predicate calculus, modal logics, the lambda

Good vs. bad documents: lists

Zutaten für 4 **Portionen** umrechnen

100 g **Marzipan - Rohmasse**
300 ml Milch
3 **Ei(er)**
7 EL Mehl
Für die Füllung:
125 g Mohn - Mischung, backfertig
Butterschmalz zum Ausbacken

Zubereitung

Marzipanrohmasse mit 2 EL Milch geschmeidig rühren. (Am besten mit einem Blitzhacker)
Eier mit Marzipanrohmasse mit dem Schneebesen des Handrührgerätes verquirlen, Die restliche Milch und das Mehl zugeben. Alles zu einem glatten Teig verrühren. 10 Minuten quellen lassen.
Das Butterschmalz in einer Pfanne erhitzen und aus dem Teig darin nacheinander vier goldgelbe Crêpes ausbacken. Die fertig gebackenen Crêpes warm halten.
Die Marzipancrêpes mit der Mohnmasse füllen, zu Dreiecken zusammenfalten.

Arbeitszeit:	ca. 20 Min.
Schwierigkeitsgrad:	normal
Brennwert p. P.:	keine Angabe
Freischaltung:	07.09.06
Rezept-Statistiken:	12.081 (156)* gelesen 148 (0)* gespeichert 439 (5)* gedruckt 14 (0)* verschickt * nur in diesem Monat

Verfasser:

feuermohn

Mitglied seit 09.12.2004
10.466 Beiträge (ø3,68/Tag)


Festliches Eisvergnügen mit wenig Aufwand und großer Wirkung

Schlagworte für dieses Rezept

☐ Dessert, ☐ Mehlspeisen, ☐ Süßspeise

Kombi-Suche

Ähnliche Rezepte

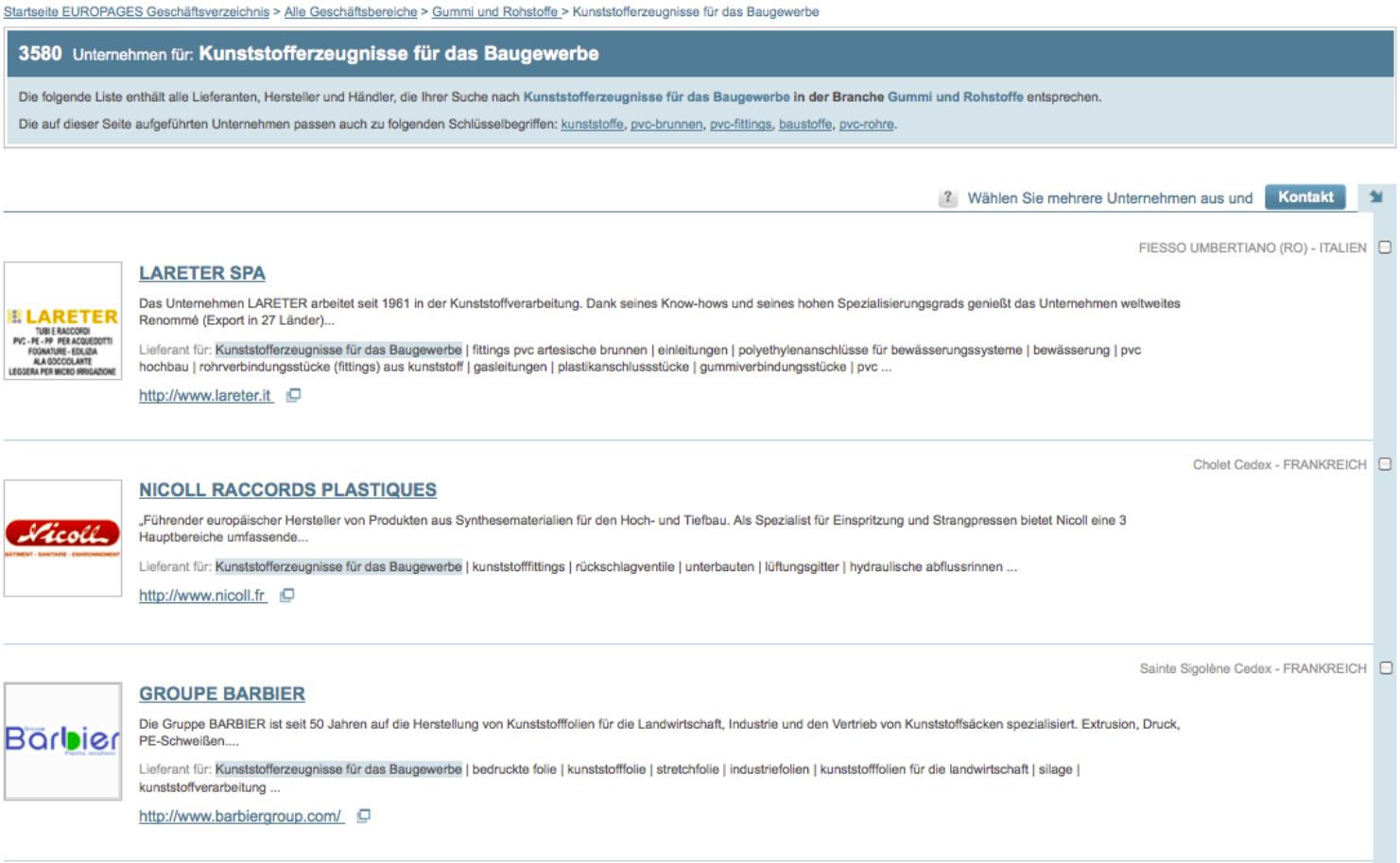
- Semlor
- Pfaffenhüetli-Zitronen
- Spekulatius, gefüllt
- Mandelhippen
- Schoko - Marzipan - Herzen
- Marzipan - Pistazien - Creme
- Pistazieneisparfait mit Nougatsauce
- Schoko - Marzipan - Eis
- Zwetschgennudeln
- Scheiterhaufen mit Äpfeln und Marzipan

Rezeptsammlungen

Dieses Rezept ist in diesen Sammlungen gespeichert:

- Sweeties
- Marzipan
- Kuchen
- Mehlspeisen als Hauptgericht
- Dessert

Good vs. bad documents: lists (II)



Startseite EUROPAGES Geschäftsverzeichnis > Alle Geschäftsbereiche > Gummi und Rohstoffe > Kunststofferzeugnisse für das Baugewerbe

3580 Unternehmen für: **Kunststofferzeugnisse für das Baugewerbe**

Die folgende Liste enthält alle Lieferanten, Hersteller und Händler, die Ihrer Suche nach **Kunststofferzeugnisse für das Baugewerbe** in der Branche **Gummi und Rohstoffe** entsprechen.

Die auf dieser Seite aufgeführten Unternehmen passen auch zu folgenden Schlüsselbegriffen: kunststoffe, pvc-brunnen, pvc-fittings, baustoffe, pvc-rohre.

Wählen Sie mehrere Unternehmen aus und **Kontakt**

FIESSO UMBERTIANO (RO) - ITALIEN

LARETER SPA

Das Unternehmen LARETER arbeitet seit 1961 in der Kunststoffverarbeitung. Dank seines Know-hows und seines hohen Spezialisierungsgrads genießt das Unternehmen weltweites Renommé (Export in 27 Länder)...

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | fittings pvc artesische brunnen | einleitungen | polyethylenanschlüsse für bewässerungssysteme | bewässerung | pvc hochbau | rohrverbindungsstücke (fittings) aus kunststoff | gasleitungen | plastikanschlussstücke | gummiverbindungsstücke | pvc ...

http://www.lareter.it

Cholet Cedex - FRANKREICH

NICOLL RACCORDS PLASTIQUES

„Führender europäischer Hersteller von Produkten aus Synthesematerialien für den Hoch- und Tiefbau. Als Spezialist für Einspritzung und Strangpressen bietet Nicoll eine 3 Hauptbereiche umfassende...

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | kunststofffittings | rückschlagventile | unterbauten | lüftungsgitter | hydraulische abflussrinnen ...

http://www.nicoll.fr

Sainte Sigolène Cedex - FRANKREICH

GROUPE BARBIER

Die Gruppe BARBIER ist seit 50 Jahren auf die Herstellung von Kunststofffolien für die Landwirtschaft, Industrie und den Vertrieb von Kunststoffsäcken spezialisiert. Extrusion, Druck, PE-Schweißen....

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | bedruckte folie | kunststofffolie | stretchfolie | industriefolien | kunststofffolien für die landwirtschaft | silage | kunststoffverarbeitung ...

http://www.barbiergroup.com/

Good vs. bad documents: lists (III)

Home
Blog
Luftfracht Speditionen
Logistik Videoportal
Speditionen
Umzugsspeditionen Deutschland
Logistikzielorte in Deutschland

LOGISTIKJOURNALE

- Cargo Journal
- Umzug Journal
- Internationale Speditionen

LOGISTIK ZIELORTE IN DEUTSCHLAND

- Logistik Zielorte in Baden Württemberg
- Logistik Zielorte in Bayern
- Logistik Zielorte in Berlin
- Logistik Zielorte in Brandenburg
- Logistik Zielorte in Bremen
- Logistik Zielorte in Hamburg
- Logistik Zielorte in Hessen
- Logistik Zielorte in Mecklenburg-Vorpommern
- Logistik Zielorte in Niedersachsen
- Logistik Zielorte in Nordrhein-Westfalen
- Logistik Zielorte in Rheinland-Pfalz
- Logistik Zielorte in Saarland

Liste deutscher Speditionen

Liste internationaler Speditionsunternehmen

Amm Spedition	Amm GmbH & Co KG Spedition
Amm Familie	90451 Nürnberg
Anhalt Logistics	Anhalt Logistics GmbH & Co. KG
Anhalt Familie	25776 Rehm-Flehde-Bargen
Arriva	arriva gmbh
	79115 Freiburg
Beck Gruppe	Beck Spedition+Logistik GmbH
Beck/Kienzler	70794 Filderstadt
Biber Post	Marketing Service Magdeburg GmbH
	39104 Magdeburg
BTG Feldberg	BTG Feldberg & Sohn GmbH & Co. KG
Feldberg Familie	46395 Bocholt
BurSped Gruppe	KG Bursped Speditions-GmbH & Co
	22113 Hamburg
BWG Reimer	BWG Reimer GmbH & Co. KG
	28217 Bremen
Carl Köster & Louis Hapke	Carl Köster & Louis Hapke GmbH & Co. KG

Text quality in COW

- classify documents according to their "textiness"
- metric based on frequent (short) words
  in language identification (Grefenstette 1995)
- does not involve an obviously difficult design decision
- strategy for cleansing: high recall for everyone,
  accept mediocre precision
- For retained documents, use text quality as annotation.
- Let corpus users choose a threshold for text quality
  according to their research question.

The approach is documented and has been evaluated. (Schäfer et al. 2013)

Boilerplate

Boilerplate: text not produced by person on a particular occasion, but:

- generated by content management systems
- similar or identical on many web pages of the same website
- similar or identical across web sites

Why bother?

- BP may bias the frequency of linguistic items in the final corpus.
- One of the most frequent tokens in an experimental German web corpus: *mehr* 'more', as in 'Click to read more ...'
- One of the most frequent sentences in an experimental English web corpus: "You are not allowed to post new content in the forum."

Annotators usually disagree on what is boilerplate to some extent.

Text

taz.de

taz.info | Abo | Anzeigen | Genossenschaft | Stiftung | tazshop | tazcafe | e-Kiosk | Hilfe | Kontakt | taz.die tageszeitung

POLITIK ZUKUNFT NETZ DEBATTE LEBEN SPORT WAHRHEIT BERLIN NORD

NETZPOLITIK NETZÖKONOMIE NETZKULTUR NETZGERÄTE

ARCHIV ZEITUNG BLOGS BEWEGUNG

suchen ...

SCHWERPUNKT ÜBERWACHUNG

09.10.2011 | 15 Kommentare

CHAOS COMPUTER CLUB WARNT

Fieser Geselle Bundestrojaner

Die von den Ermittlungsbehörden genutzte Schnüffelsoftware verursacht Sicherheitslücken bei den betroffenen Computern. Und sie kann mehr als das Bundesverfassungsgericht erlaubt.

Ein Tastendruck in der Ermittlungsbehörde - und Dein PC verliert sein Immunsystem. Bild: Imago/blickwinkel

ONLINE-AKTION KEINE GRENZEN FÜR MENSCHENRECHTE

Die taz wird ermöglicht durch 10.952 Genossinnen

10|10|2011

STICHWORT: SCHWERPUNKT ÜBERWACHUNG

Im Schwerpunkt Überwachung legen wir ein besonderes Augenmerk auf die neuesten Auswüchse der Sammelwut und Kontrollgelüste von Staatsgewalt wie Konzernen. Und natürlich auf Datenpannen aller Art.

Navigation:
Zum Überblick über den Schwerpunkt.

WEITERE SCHLAGZEILEN ...

STREIT UM BUNDESTROJANER
Justizministerin kündigt Aufklärung an

CHAOS COMPUTER CLUB WARNT
Fieser Geselle Bundestrojaner

STREIT UM "GEFÄLLT MIR"-BUTTON
Lobby wettert gegen Weichert

TÜRKISCHE INTERNETZENSUR
Surfen mit dem Kinderpaket

ANTI-NAZI-DEMO IN DRESDEN
Illegale Ermittlungen in Sachsen?

MEISTGELESEN

KAFFEEHAUSCHEFIN ÜBER MACCHIATO-MUTTER

Text (II)

BERLIN *dpa* | Dem Chaos Computer Club (CCC) ist nach eigenen Angaben eine "staatliche Spionagesoftware" zugespielt worden, die von Ermittlern in Deutschland zur Überwachung von Telekommunikationsverbindungen eingesetzt wird. "Die untersuchten Trojaner können nicht nur höchst intime Daten ausleiten, sondern bieten auch eine Fernsteuerungsfunktion zum Nachladen und Ausführen beliebiger weiterer Schadsoftware", teilte der Verein am Samstagabend in Berlin mit.

Der CCC warf den Sicherheitsbehörden vor, aufgrund von groben Design- und Implementierungsfehlern in der Software entstünden "eklatante Sicherheitslücken in den infiltrierten Rechnern, die auch Dritte ausnutzen können". Die Telekommunikationsüberwachung an der Quelle, kurz als Quellen-TKÜ bezeichnet, soll eine Möglichkeit bieten, die Kommunikation über das Internet abzuhören, bevor sie für den Weg durchs Netz verschlüsselt wird.

Ein Sprecher des Bundesinnenministeriums bestätigte auf Anfrage, dass Software-Lösungen für eine Quellen-TKÜ verfügbar seien, sowohl für die Bundesbehörden als auch auf Landesebene. "Für den Einsatz dieser Software gibt es gesetzliche Grundlagen, die beim Einsatz beachtet werden müssen", sagte der Sprecher. Für Ermittlungen auf Bundesebene sei hier etwa das BKA-Gesetz relevant. Außerdem gibt es in einigen Bundesländern Regelungen zum Einsatz der Quellen-TKÜ.

Die Bestrebungen für eine Online-Durchsuchung bei Verdächtigen reichen ins Jahr 2005 zurück, in die Amtszeit des damaligen Bundesinnenministers Otto Schily (SPD). Danach setzte unter dem Schlagwort "Bundestrojaner" eine heftige Debatte über die Zulässigkeit solcher Eingriffe in die Privatsphäre des persönlichen Computers ein.

KAFFEEHAUSCHEFIN ÜBER MACCHIATO-MÜTTER
"Die Weiber denken, sie wären besser"

POLIZEISKANDAL IN DÄNEMARK
Schnüffeln, um zu denunzieren

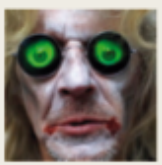
PARLAMENTSWAHL IN POLEN
Donald Tusk kann weitermachen

STUDIE ÜBER BEZAHLSTUDIUM
Uni-Gebühren schrecken nicht ab

UKRAINISCHER SCHACHMEISTER IN BRASILIEN
Pistole auf der Brust

BILDERGALERIE

KRIEG IN AFGHANISTAN
Zehn Jahre dauert der Krieg in Afghanistan an. Nein - viel, viel länger, sagt der Fotograf Simon Norfolk. Er reiste auf den Spuren von John Burke - der 1879 als Erster in Afghanistan fotografierte.



KARIKATUR & TOMS TOUCHÉ

Text (III)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG

bewegung.taz.de

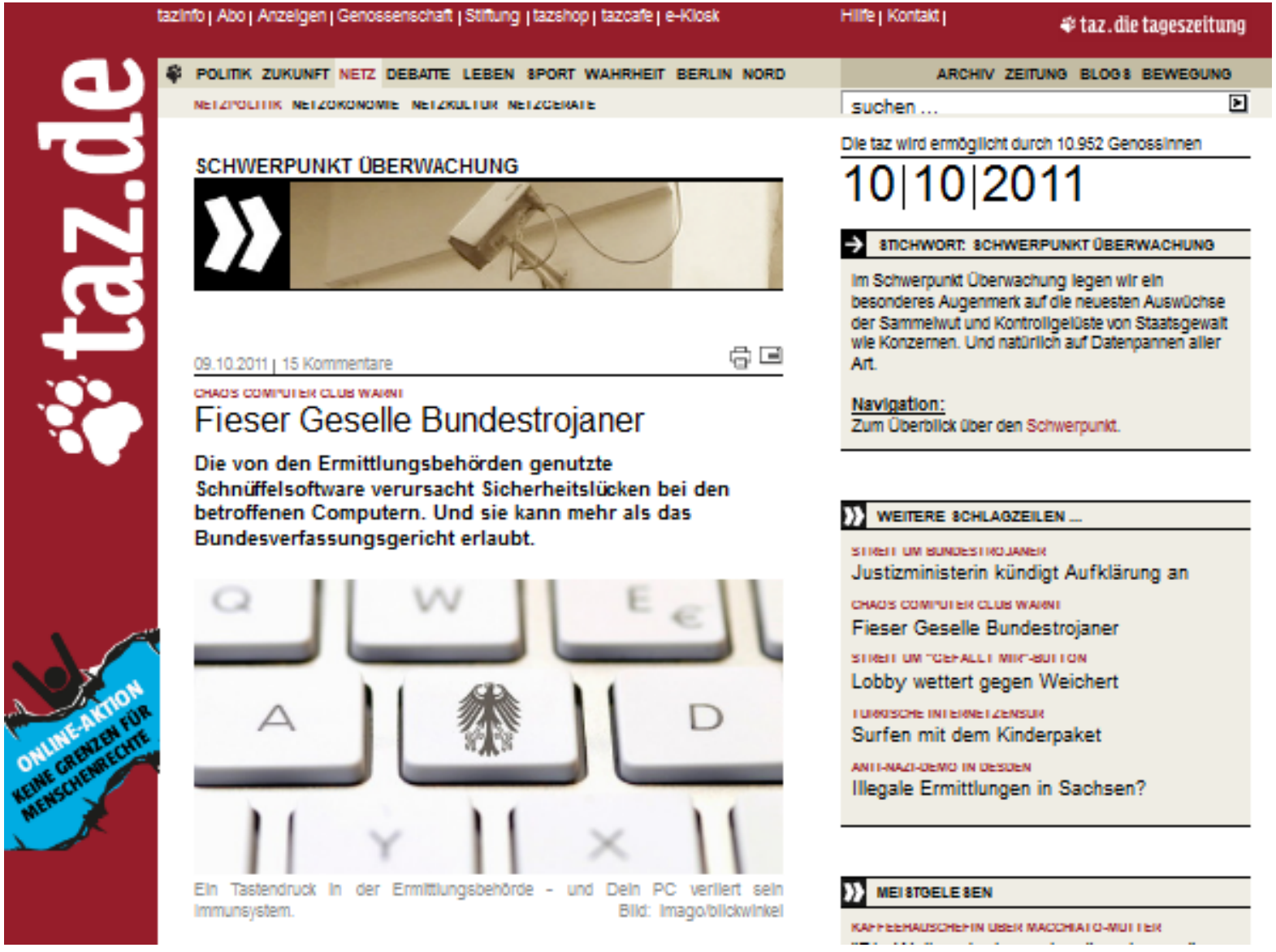
TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

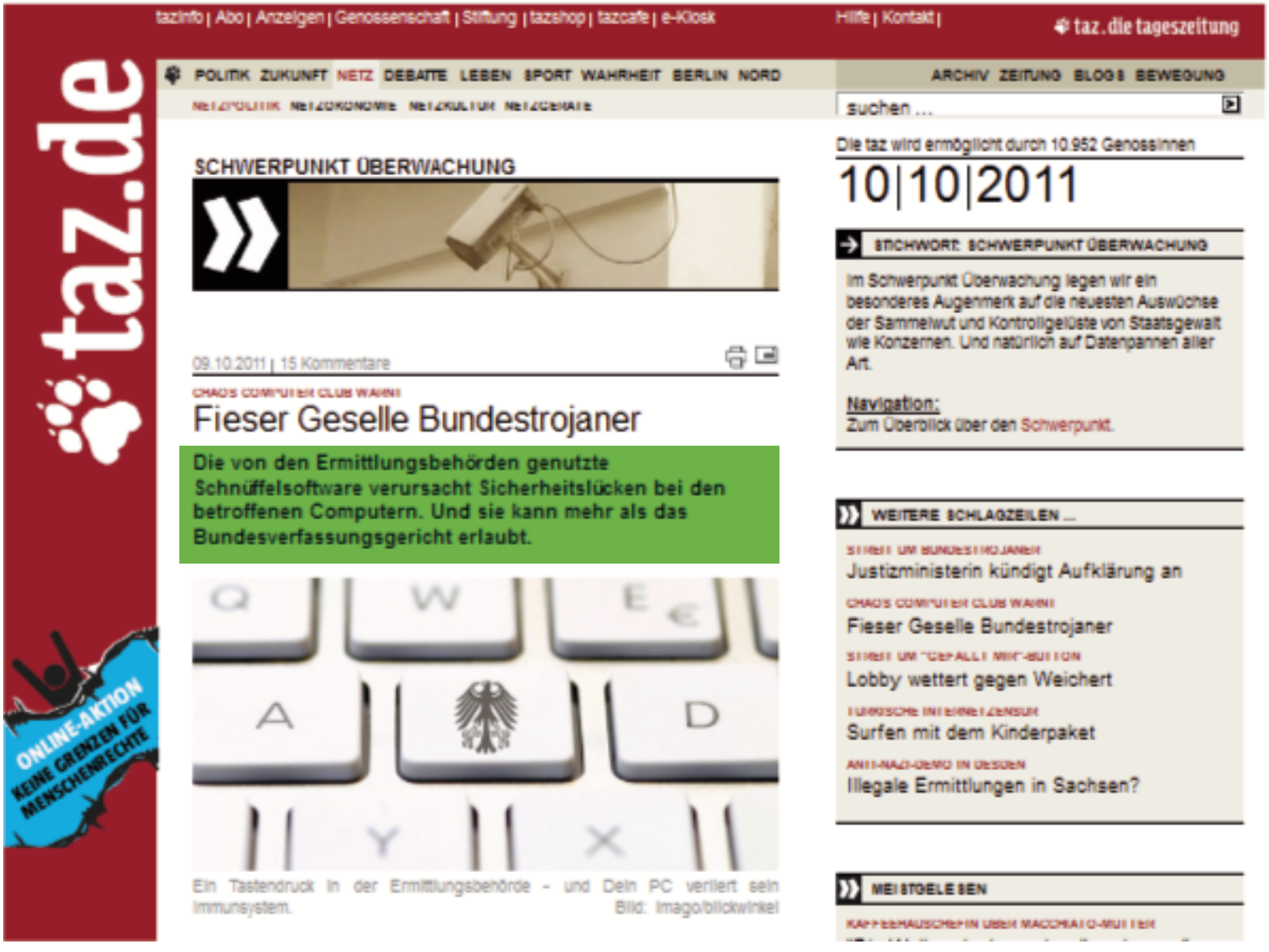
DIE AKTUELLE AUSGABE DER TAZ

ePaper kaufen
PDF-Vorschau

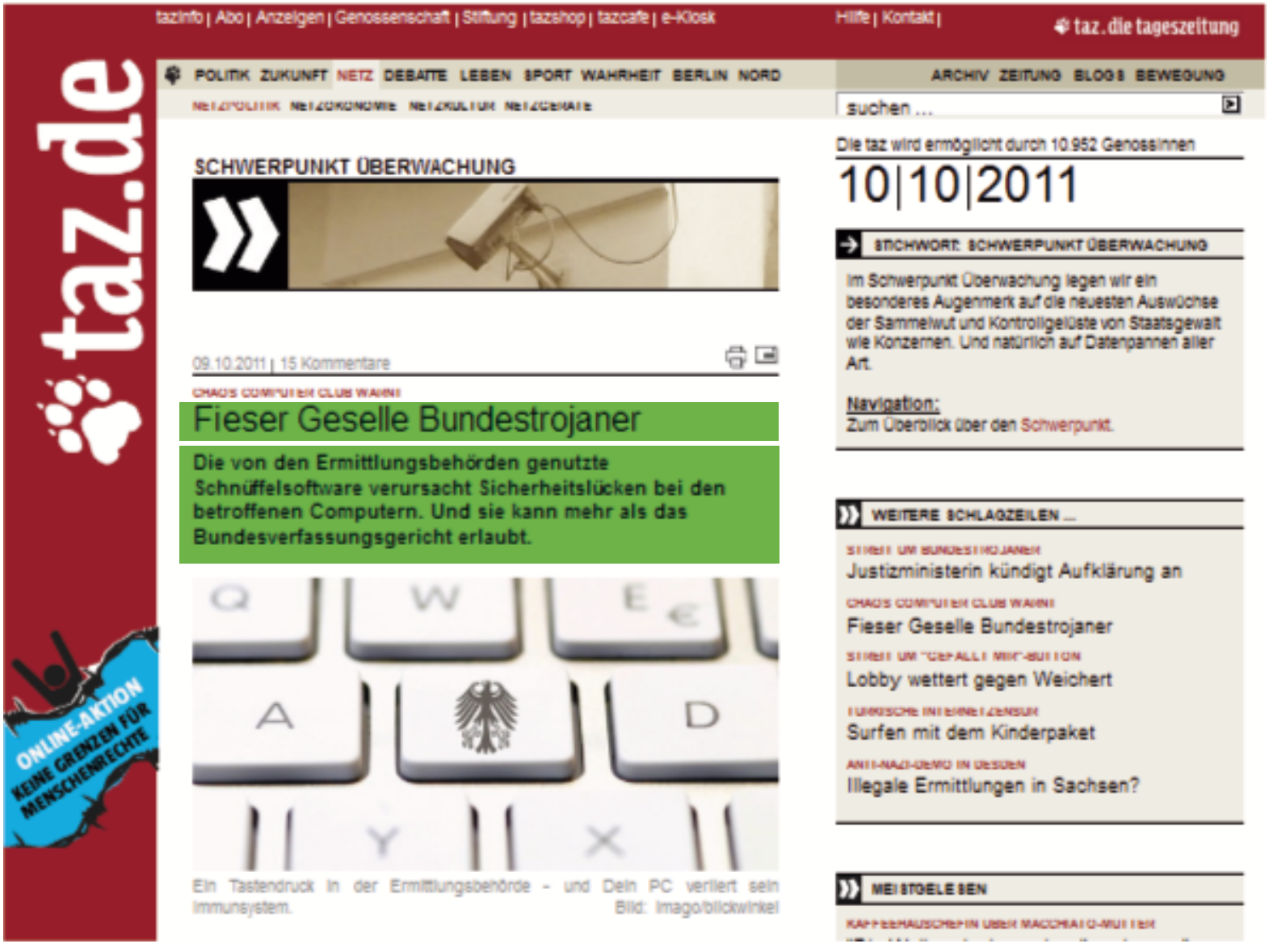
What should count as boilerplate?



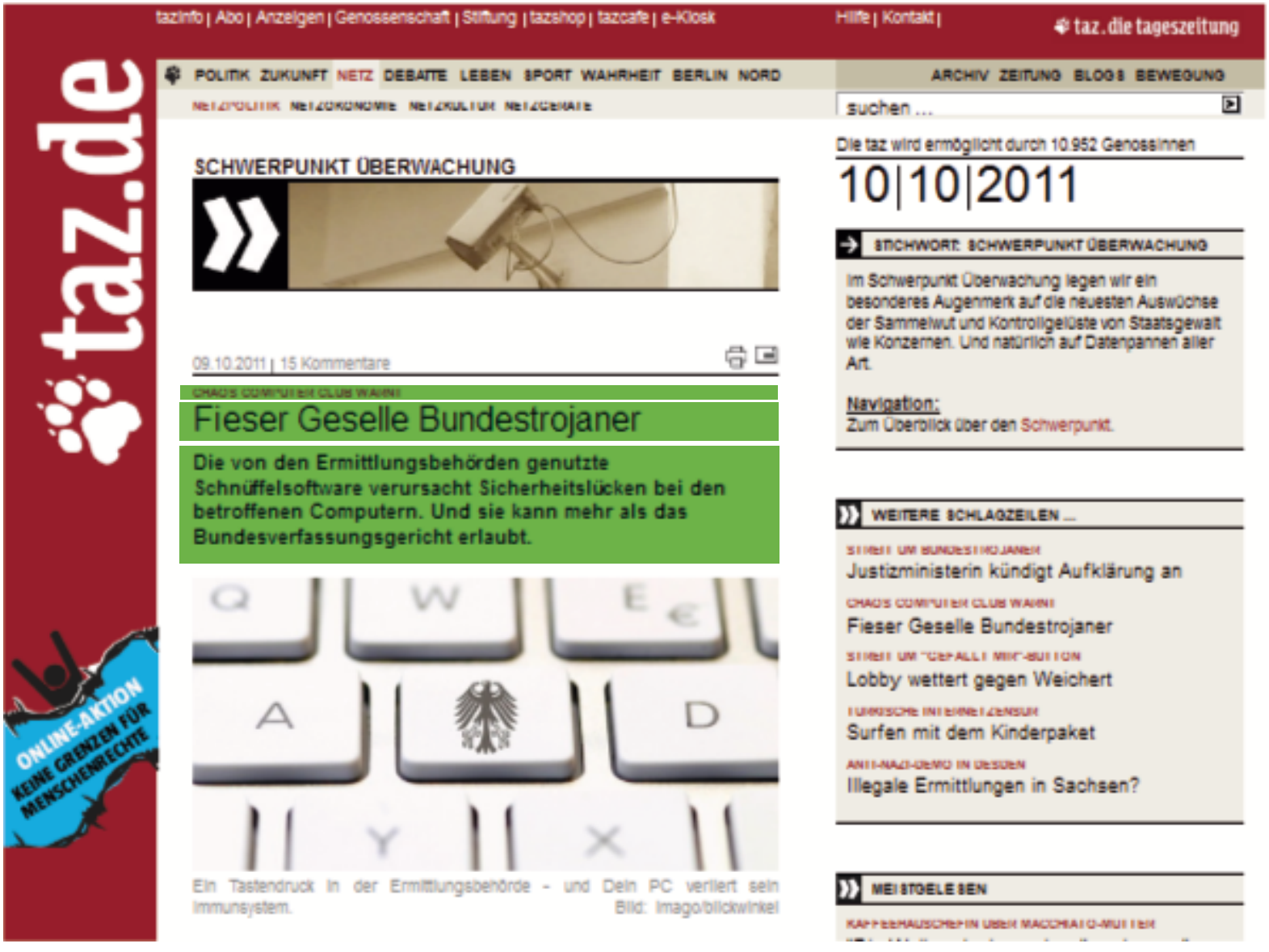
What should count as boilerplate? (II)



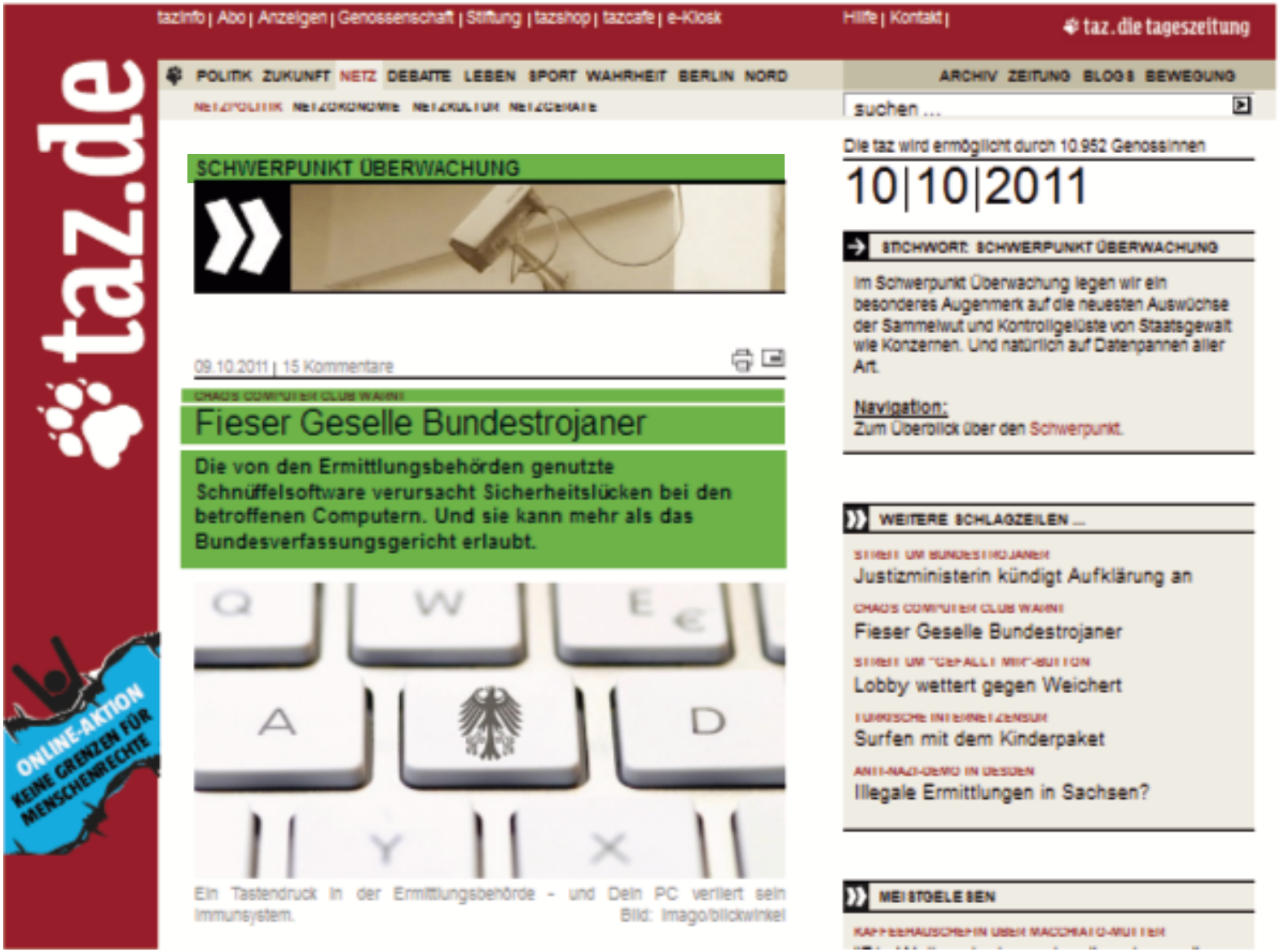
What should count as boilerplate? (III)



What should count as boilerplate? (IV)



What should count as boilerplate? (V)



What should count as boilerplate?(VI)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG
bewegung.taz.de

TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

DIE AKTUELLE AUSGABE DER TAZ

taz.die tageszeitung
Syrische Botschaften gestürmt

ePaper kaufen
PDF-Vorschau

What should count as boilerplate? (VII)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG
bewegung.taz.de

TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

DIE AKTUELLE AUSGABE DER TAZ

taz.die tageszeitung
Syrische Botschaften gestürmt

ePaper kaufen
PDF-Vorschau

What should count as boilerplate? (VIII)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG
bewegung.taz.de

TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

DIE AKTUELLE AUSGABE DER TAZ

taz.die tageszeitung
Syrische Botschaften gestürmt

ePaper kaufen
PDF-Vorschau

What should count as boilerplate? (IX)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG
bewegung.taz.de

TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

DIE AKTUELLE AUSGABE DER TAZ

taz.die tageszeitung
Syrische Botschaften gestürmt

ePaper kaufen
PDF-Vorschau

What should count as boilerplate? (X)

ARTIKEL ZUM THEMA

NETZKONFERENZ RE:PUBLICA
Blogger kontra Offline-Mächtige

Der Trojaner nehme Befehle ohne jegliche Absicherung oder Authentifizierung entgegen. Selbst einfache Absicherungen, wie beim Online-Banking oder bei Flirtportals üblich, gebe es nicht. Es sei für einen beliebigen Angreifer ohne weiteres möglich, die Kontrolle über einen von deutschen Behörden infiltrierten Computer zu übernehmen. "Das Sicherheitsniveau dieses Trojaners ist nicht besser, als würde er auf allen infizierten Rechnern die Passwörter auf "1234" setzen."

DIESER ARTIKEL ... gefällt mir.

eAbo eKiosk mobile Themenalarm

LESERKOMMENTARE

10.10.2011 04:57 | PANHOLD
@Pia

wir das selbe Passwort.

09.10.2011 21:41 | FRAGE
wenn man den Trojaner auf seinem Rechner findet, sollte man dann zum Anwalt gehen und klagen?

09.10.2011 20:01 | GESELLEN SIND NOCH KEINE MEISTER
@Hans: Die NSA will aber nicht, das Syrien und alle anderen ihre Backdoors nutzen.
Und die Zuständigkeiten der Dienste ...

Kommentar schreiben >

Endlagersuche 2012 starten – Gorleben ist verbrannt!

ATTAC
Athen, Madrid, New York: Attac ruft zu Protesten auch in Deutschland auf

ROBIN WOOD
S 21 – Gericht stoppt Bauarbeiten – Erfolg für Klage des BUND!

820 weitere Organisationen stellen sich vor >

PLATTFORM FÜR VERÄNDERUNG
bewegung.taz.de

TAZ SERVICE

DIE ZEITUNG eKiosk
ABO Zeitung | Probeabo | DigiAbo | eBook | iPhone
ANZEIGEN Print | Online | Mediadaten
RECHERCHE Service | Textarchiv | Themenalarm
MITMACHEN Bewegung | Genossenschaft | Akademie
TAZ SHOP
TAZREISEN IN DIE ZIVILGESELLSCHAFT
TAZCAFE tazpressomobil
VERANSTALTUNGEN
TAZ PANTER PREIS
TAZINFO Newsletter | Presse

DIE AKTUELLE AUSGABE DER TAZ

taz.die tageszeitung
Syrische Botschaften gestürmt

ePaper kaufen
PDF-Vorschau

Boilerplate removal

Traditional approach: automatically detect and remove boilerplate.

- Problem 1: Boilerplate status is conceptually unclear in many instances.
- Problem 2: Algorithms trained on human decisions rarely produce 100% correct classifications.

COW approach:

- Automatically detect and classify boilerplate (per paragraph).
- Annotate paragraphs with a BP score generated by the classifier.
- Let corpus users choose the acceptable amount of boilerplate.

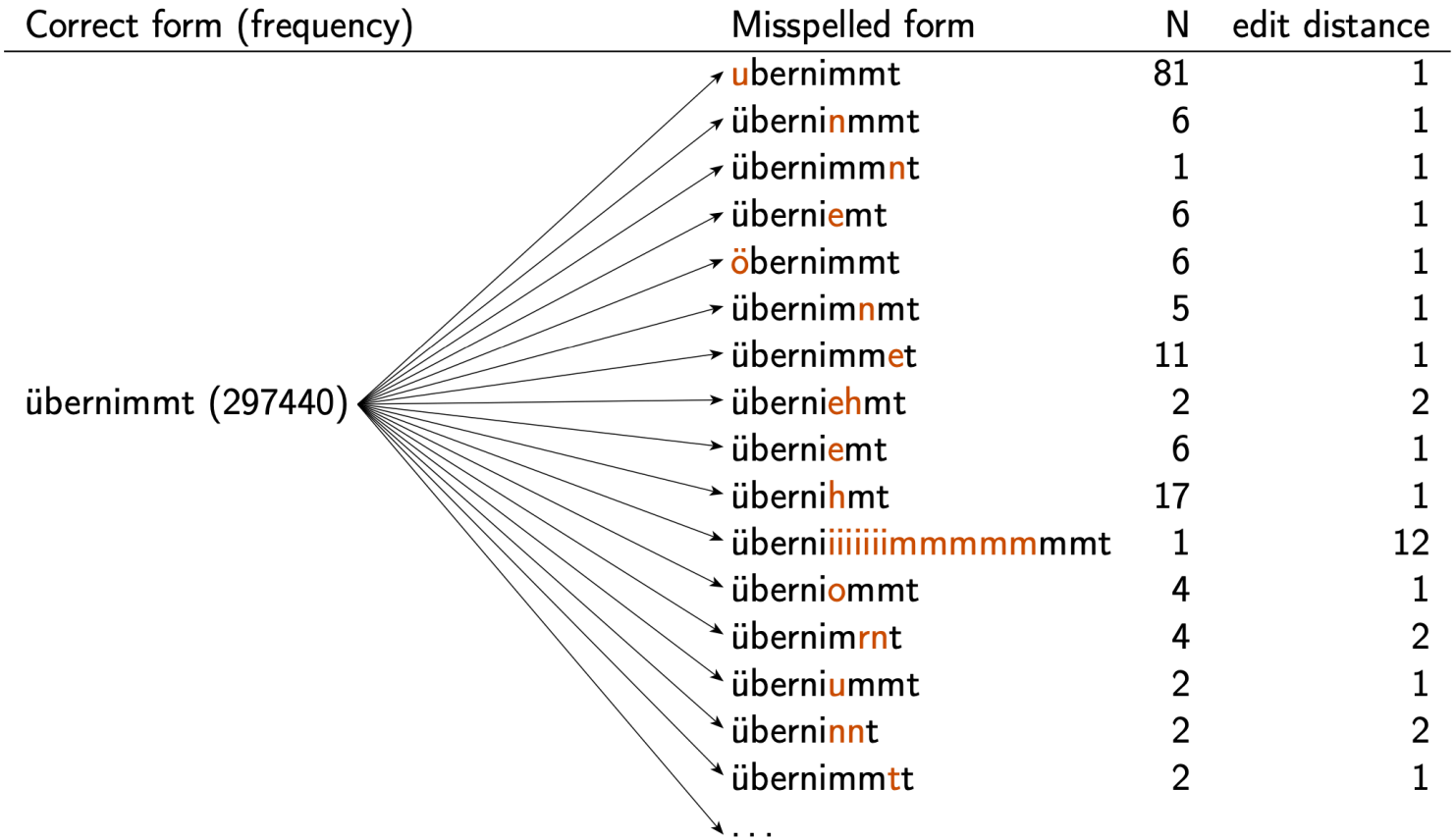
The approach is documented and has been evaluated. (Schäfer 2017)

Linguistic annotation

Two important characteristics of web corpora:

1. Size: Web corpora require automatic linguistic annotation because their size makes manual annotation infeasible.
2. Noise: Off-the-shelf software often under-performs on web data.
   ▷ It needs to be adapted/wrapped.

Example: Spelling variants/errors/noise



Some spelling variants of German *übernimmt* ('takes over') from DECOW2012.

Linguistic annotation for noisy data

- Some amount of annotation error can't be avoided.
- Evaluation is mandatory, e. g. by way of specific shared tasks.

  E. g., GSCL 2015 Shared Task on Automatic Linguistic Annotation of Computer-Mediated Communication/Social Media)

Web corpora | Summary

- Goal: create static corpora from the web for reproducible research.
- Unlike many traditional texts, many (parts of) web documents are seen as noise by corpus users relative to their research questions.
- Web corpus designers cannot foresee users' research questions.
- Instead of doing rigorous cleanups, COW allows users to make transparent decisions to keep or remove potential noise.
- Badness: Is a document a coherent text built from real sentences?
- Boilerplate: Is a paragraph real text or automatically generated?
- Linguistic annotation needs to be adapted to noisy data.

As a consequence, corpus users must actively stratify each query with COW-type corpora.

Come on, just tell us!

What's a good corpus?

What makes a corpus a good corpus?

Answer: A specific theory and a concrete research question!
In other words, you, the corpus user.

- You need a research question.
- You need a theory into which this question is embedded.
- Ideally, the research question should have a high capability of probing your theory for errors.
- It should be specific and precise w. r. t. all substantive and auxiliary connected hypotheses.
- If you just diagnose the existence of patterns without a substantive theory behind you, you cannot decide whether a corpus is good.
- If somebody else has already done something similar, try to replicate, and don't try to be innovative at all costs.

This allows you to make good decisions when it comes to corpus choice and corpus analysis.

Finding a good corpus

When you go looking for a good corpus:

- Be as non-opportunistic as possible.
- Try to find a corpus that was designed guided by a very specific theoretical question compatible to yours (Ridges, RUEG).
- If that's impossible, get to know the corpus very well to avoid noise, ask the creators for guidance (COW).
- Always be sceptical towards linguistic annotation in the corpus.
- Don't use high-level annotation (register) just because it's there. It will ruin your inferences.
- Ensure you and others can probe the corpus and your study for flaws:
  - Use well-documented and openly available corpora.
  - Make transparent your decisions regarding corpus choice.
  - Document everything you did to retrieve, annotate, analyse your data.

We wish you good inferences!

Arppe, Antti & Juhani Järvikivi. (2007). Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.

Baayen, R. Harald. (2001). *Word frequency distributions*. Dordrecht / Boston / London: Kluwer.

Baroni, Marco. (2009). Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. an international handbook*, vol. 2, 803–822. Mouton de Gruyter.

Baroni, Marco & Silvia Bernardini (eds.). (2006). *Wacky! working papers on the web as corpus*. Bologna: GEDIT.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Biber, Douglas. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. (1989). A typology of English texts. *Linguistics* 27(1). 3–43.

Biber, Douglas. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.
Biber, Douglas. (2009). Multi-dimensional approaches. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, vol. 2, 822–855. Berlin: Walter de Gruyter.
Bybee, Joan L. & Clay Beckner. (2009). Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 827–856. Oxford: Oxford University Press.
Carnap, Rudolf. (1928). *Der logische Aufbau der Welt*. Berlin: Weltkreis Verlag.
Cohen, Jacob. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1). 37–46.
Cronbach, Lee J. & Paul E. Meehl. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52. 281–302.
Davies, Mark. (2013). *Corpus of Global Web-Based English*.

Divjak, Dagmar. (2016). Four challenges for usage-based linguistics. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms – new paradoxes – recontextualizing language and linguistics*, 297–309. Berlin/Boston: De Gruyter Mouton.

Divjak, Dagmar & Antti Arppe. (2013). Extracting prototypes from exemplars what can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274.

Divjak, Dagmar, Ewa Dąbrowska & Antti Arppe. (2016). Machine meets man: evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33.

Drummond, Chris. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 1–5.

Eckart, Thomas & Uwe Quasthoff. (2013). Statistical Corpus and Language Comparison on Comparable Corpora. In. Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum & Pascale Fung (eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. 151–165.

Evert, Stefan. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik* 54(2). 177–190.

Evert, Stefan. (2008). Corpora and collocations. In Anke Lüdeling & Maria Kytö (eds.), *Corpus linguistics. an international handbook*, 1212–1248. Berlin: Mouton.

Fisher, Ronald A. (1935a). *The design of experiments*. London: Macmillan.

Fisher, Ronald A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society* 98(1). 39–82.

Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.

Gilquin, Gaëtanelle & Stefan Th. Gries. (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.

Grefenstette, Gregory. (1995). Comparing two language identification schemes. In *Proceedings of the 3rd internation conference on statistical analysis of textual data (jadt 1995)*, 263–268. Rome.

Gries, Stefan Th. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff & Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, 197–212. Amsterdam: Rodopi.

Gries, Stefan Th. (2017). Syntactic alternation research. taking stock and some suggestions for the future. In Ludovic De Cuypere, Clara Vanderschueren & Gert De Sutter (eds.), *Current trends in analyzing syntactic variation*, vol. 31 (Belgian Journal of Lingustics), 7–27. Amsterdam: Benjamins.

Hirschmann, Hagen. (2019). *Korpuslinguistik. Eine Einführung*. Stuttgart: Metzler.

Ioannidis, John P. A. (2005). Why most published research findings are false. *PLos med* 2(8). e124.

Jäger, Lena A., Felix Engelmann & Shravan Vasishth. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language* 94. 316–339.

Jakubíček, Miloš, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý & Vit Suchomel. (2013). The TenTen corpus family. In Andrew Hardie & Robbie Love (eds.), *The 7th international corpus linguistics conference*, 125–127. Lancaster: UCREL.

Kapatsinski, Vsevolod. (2014). What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology* 11. 1–41.

Kilgarriff, Adam. (2001). Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.

Kilgarriff, Adam. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–275.

Kilgarriff, Adam. (2006). Googleology is bad science. *Computational Linguistics* 33(1). 147–151.

Köpcke, Klaus-Michael. (1995). Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache – Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie. *Zeitschrift für Sprachwissenschaft* 14(2). 159–180.

Krause, Thomas & Stephan Druskat. (June 2019). *Die Hard 1.1024.0: backward compatibility of a search engine with persistant IDs*. deRSE19 - Conference for Research Software Engineers in Germany. Potsdam.

Kübler, Sandra & Heike Zinsmeister. (2014). *Corpus linguistics and linguistically annotated corpora*. London: Bloomsbury.

Küchenhoff, Helmut & Hans-Jörg Schmid. (2015). Reply to "More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff" by Stefan Th. Gries. *Cognitive Linguistics* 26(3). 537–547.

Le Verrier, Urbain. (1859). Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 49. 379–383.

Maxwell, Scott E. & Harold D. Delaney. (2004). *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

Mayo, Deborah G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. (2018). *Statistical inference as severe testing: how to get beyond the statistics wars*. Cambridge: Cambridge University Press.
McEnery, Tony, Richard Xiao & Yukio Tono. (2006). *Corpus-based language studies. an advanced resource book*. London: Routledge.
Moisl, Hermann. (2009). Exploratory multivariate analysis. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. an international handbook*, vol. 2, 874–899. Mouton De Gruyter.
Newman, John & Tamara Sorenson Duncan. (2015). *Convergence and divergence in Cognitive Linguistics: Facing up to alternative realities of linguistic catgeories*. Talk given at the 13th international cognitive linguistics conference (ICLC-13).
Nicenboim, Bruno, Timo B Roettger & Shravan Vasishth. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70. 39–55.
Oakes, Michael P. (1998). Statistics for corpus linguistics. In. Edinburgh University Press Edinburgh.

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke LĂ¼deling & Thomas Krause. (2017). RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51(3). 695–725.

Popper, Karl Raimund. (1962). *Conjections and refutations: the growth of scientific knowledge*. New York: Basic Books.

Schäfer, Roland. (September 2017). Accurate and efficient general-purpose boilerplate detection for crawled web corpora. *Language Resources and Evaluation* 51(3). 873–889.

Schäfer, Roland. (2018). *Probabilistic German morphosyntax*. Habilitation thesis. Humboldt-Universität zu Berlin PhD thesis.

Schäfer, Roland. (2019). Prototype-driven alternations: the case of German weak nouns. *Corpus Linguistics and Linguistic Theory* 15(2). 383–418.

Schäfer, Roland, Adrien Barbaresi & Felix Bildhauer. (2013). The good, the bad, and the hazy: design decisions in web corpus construction. In Stefan Evert, Egon Stemle & Paul Rayson (eds.), *Proceedings of the 8th web as corpus workshop (wac-8)*, 7–15.

Schäfer, Roland & Felix Bildhauer. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul: ELRA.

Schmid, Hans-Jörg & Helmut Küchenhoff. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.

Stefanowitsch, Anatol & Stefan Th. Gries. (2003). Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Tomasello, Michael. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard: Harvard University Press.

Venn, John. (1866). *The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its application to moral and social science*. London: Macmillan.

Wiese, Heike. (2020). Language situations: a method for capturing variation within speakers' repertoires. In Yoshiyuki Asahi (ed.), *Methods in dialectology xvi*, 105–117. Frankfurt a. M.: Peter Lang.