Large vs. small: How much does size matter in corpus-based data extraction?

Ulrich Heid

University of Hildesheim
Institute for Information Science and NLP

Valladolid: AELINCO congress 2015

Overview

- Reminder:
  The scenario of data extraction from corpora
- Quantitative issues with tasks and tools:
  - Sampling of texts to build a corpus
  - Linguistic preprocessing:
    Tokenizing – Lemmatization – Parsing
- Quantitative issues in applications of corpus processing
  - Lexicography:
    Profiting from large corpora
  - Terminology:
    Counterbalancing a lack of data
- Conclusions

Preface

- Caveat
  - Very general overview:
    Not my research focus
- Partial:
  (a) selective
  (b) preferring corpus-based work
- Contributions from many colleagues
  - Prof. Prinsloo and Bothma: Univ. of Pretoria
    e-dictionaries Project SeLA
  - M. Dick and A. Blessing: U. Hildesheim/Stuttgart
    corpus sampling Project e-Identity
  - I. Rösiger, S. Tannert, T. George, J. Schäfer: U. Stuttgart
    term candidate extraction (industry project)

Corpus-based work

Basic scenario: applications

- Research question that can be answered on the basis of textual material, e.g. on

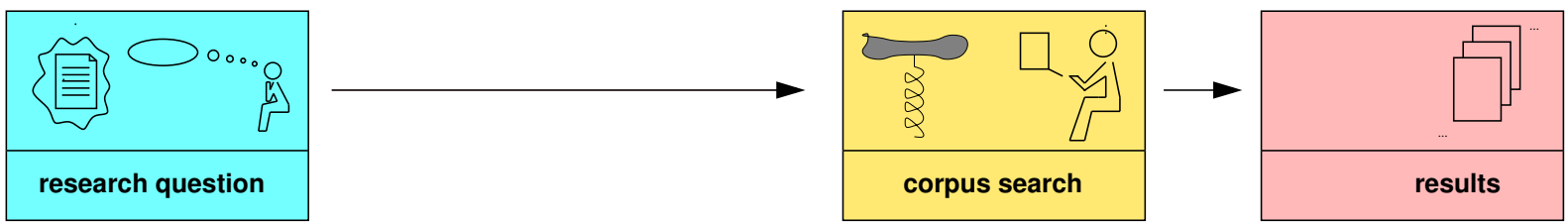
  – *grammatical phenomena*
  – *lexical and/or terminological items* } linguistics
  – *discourse phenomena*

  – *style, topics, ...in literature*
  – *decisions, motivations, ...in politics* } digital humanities
  – *facts, motivations, ...in history*

Corpus-based work

Basic scenario: corpus exploration

- Search and retrieval:
  - Interactive
  - Automatic
- Objectives:
  - Find evidence relevant for the research question
  - Possibly classify and quantify the evidence

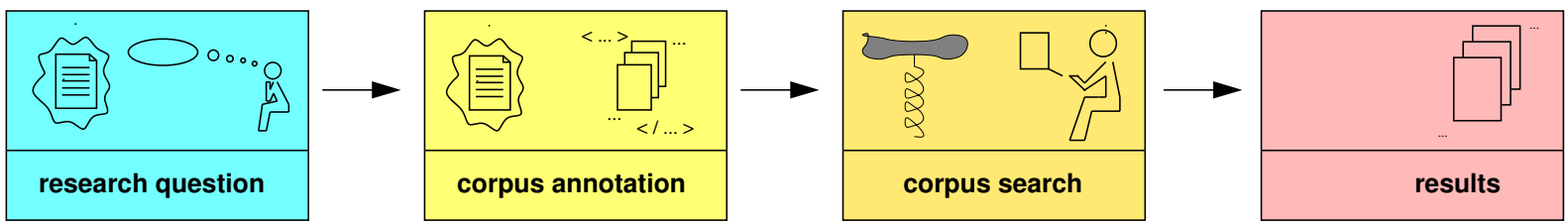


Corpus-based work

Basic scenario: tools for corpus preprocessing

- Search:
  - On "raw" text
  - On annotated text:
    Annotation as a generalization device:
    - Lemmatization — abstracts away from word forms
    - POS-tagging — abstracts away from lemmas
    - Parsing — abstracts away from phrases or grammatical functions

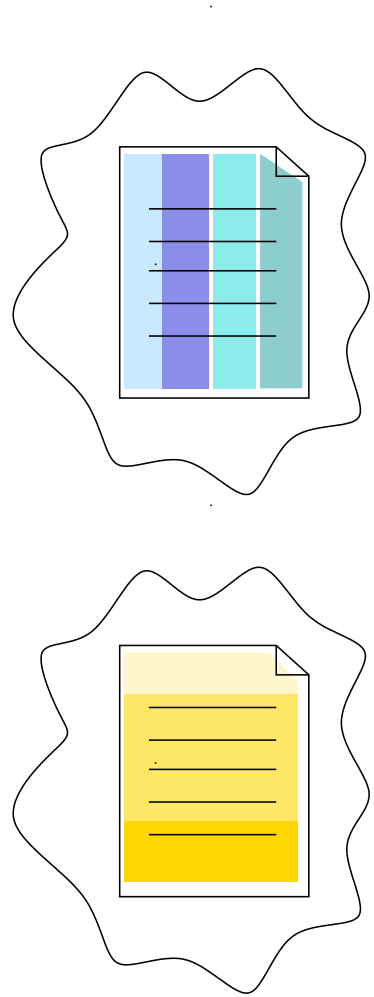
$\Rightarrow$ Need for preprocessing



Corpus-based work

Basic scenario: metadata

- Metadata for corpus composition:
  Related with sampling
  - Genres
  - Text types
  - Authorship, addressees
  - Creation date
  - ...
- Metadata for text objects:
  Internal structure of texts
  - Headlines
  - Sectioning
  - Contents
  - Functional text components



"Philosophical" issues of corpus linguistics

Corpora vs. intuition

- Basic assumption underlying this talk:
  - There is some relevance and usefulness in corpus exploration, at least for some tasks
  - Obviously:
    Not all tasks equally profit from corpus-based treatment,
    and end-users may profit more than, e.g. (sub-)language experts
- Size of corpora:
  - Size may influence usefulness of corpus data:
    see later discussion of selected applications
  - Size
    definitely influences the choice of tools and methods to be applied

$\Rightarrow$ to be shown in this talk

"Philosophical" issues of corpus linguistics

Corpus-based vs. corpus-driven methodologies (1/2)

- Simplistic account:
  Relationship between theory and data
  - Corpus-based:
    Start from theoretical assumptions and verify these on corpus data
  - Corpus-driven:
    Start from (statistical) account of data and draw conclusions:
    hypothesis building by use of
    aggregation, generalization, etc.

"Philosophical" issues of corpus linguistics

Corpus-based vs. corpus-driven methodologies (2/2) Dipper 2009

- Impact of size of corpora
  - Corpus-driven approach:
    More text → more data points
    → clearer tendencies, for more items
  - Corpus-based approach:
    * Quantitative needs may depend on task
    * Certain properties of lexical and grammatical phenomena can be seen also in small samples
- To be considered:
  - Languages with fixed corpus size (e.g. in diachronic studies)
  - Both approaches tend to be combined in recent work
- In this talk: emphasis on corpus-based approach

Quantitative issues in corpus analysis steps

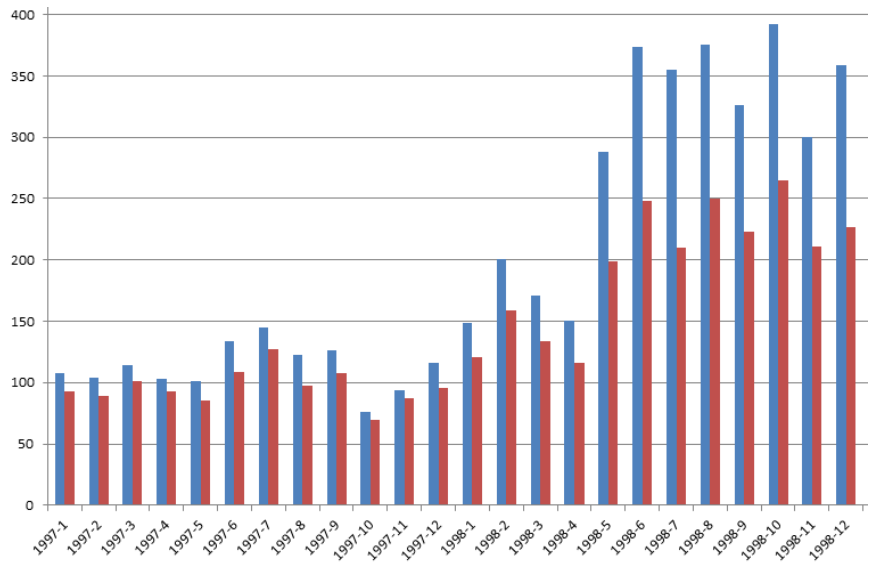
Overview

- Typical (idealized) chain of processing:
  - Sampling — finding the right texts
  - Tokenizing — sentence splitting, word form identification
  - Tagging, lemmatization, parsing — linguistic annotation
- Techniques for each may need to be assessed individually
  - Typically statistical tools: for tagging, lemmatization, parsing
  - Symbolic tools: tokenizing, parsing

Quantitative issues: sampling

Situation

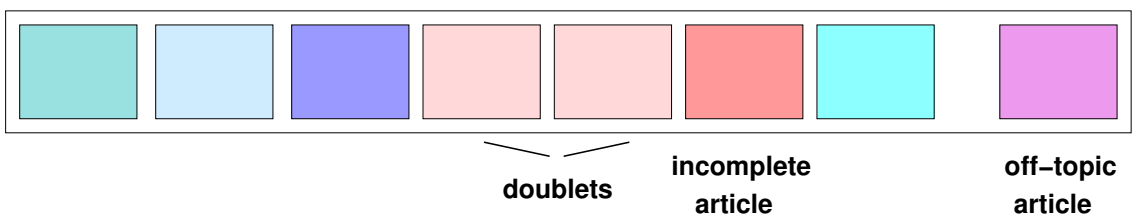
- Research question with quantitative aspects,
  e.g. in political science:
  - Issue cycles:
    media attention over time
  - Examples:
    * Using terrorism as a
      motivation for new laws
    * Ways of talking
      about an issue,
      e.g. “sustainability”
    * ...



Quantitative issues: sampling

Problems

- News corpus may contain redundancies:
  - Double occurrences of articles
  - Variants of articles: morning vs. evening issue...
  - ...
- News corpus may contain off-topic articles
  - Due to way of extracting data,
    e.g. from an archive: by keywords only
    Lexis-Nexis, Factiva,...
  - Due to (related) metadiscourse:
    book reviews, film titles,...



Quantitative issues: sampling

Sampling strategies – sampling error removal

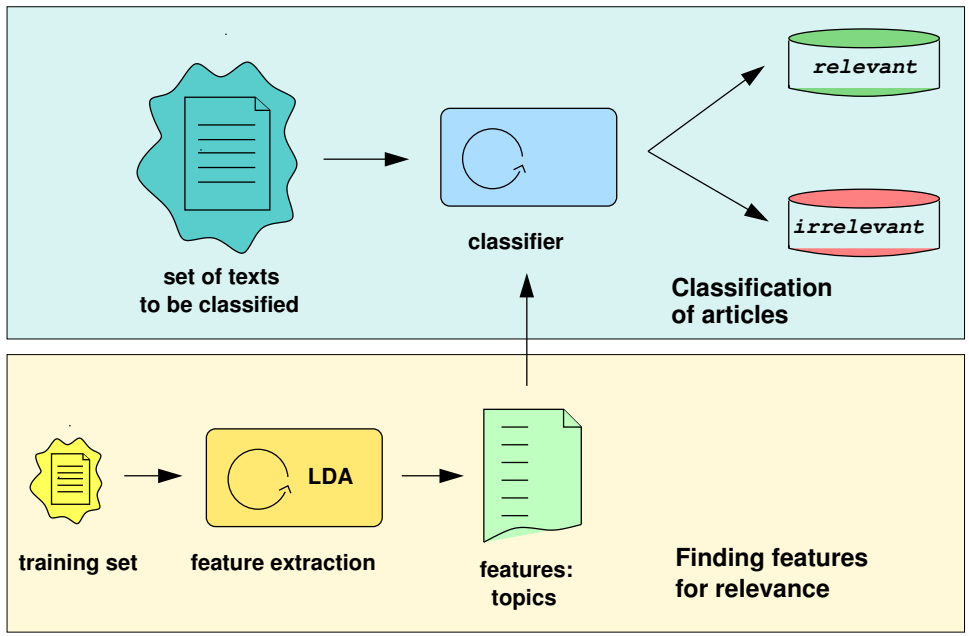
- Avoiding doublets:
  Comparison of text pairs by similarity:
  ⇒ computationally dependent on corpus size,
  not wrt the results
- Removal of incomplete texts:
  Based on regular expressions –
  E.g.: does last paragraph end on a full sentence?
  ⇒ Independent from corpus size
- Removal of off-topic articles: Dick, Blessing, Heid 2015
  Recent work based on classification and machine learning
  ⇒ Requires training material prepared manually
  ⇒ Can be manually adjusted

Quantitative issues: sampling

Removing off-topic articles

Dick, Blessing, Heid 2015

- A two-step approach:
  1. manual annotation of training data
  2. automatic classification of a set of texts
- Possibility to adjust training set interactively



⇒ Small training sets

⇒ Application to large set of texts

Blei et al. 2003

Based on "topics"

identified by means of Latent Dirichlet Allocation

Quantitative issues: Preprocessing

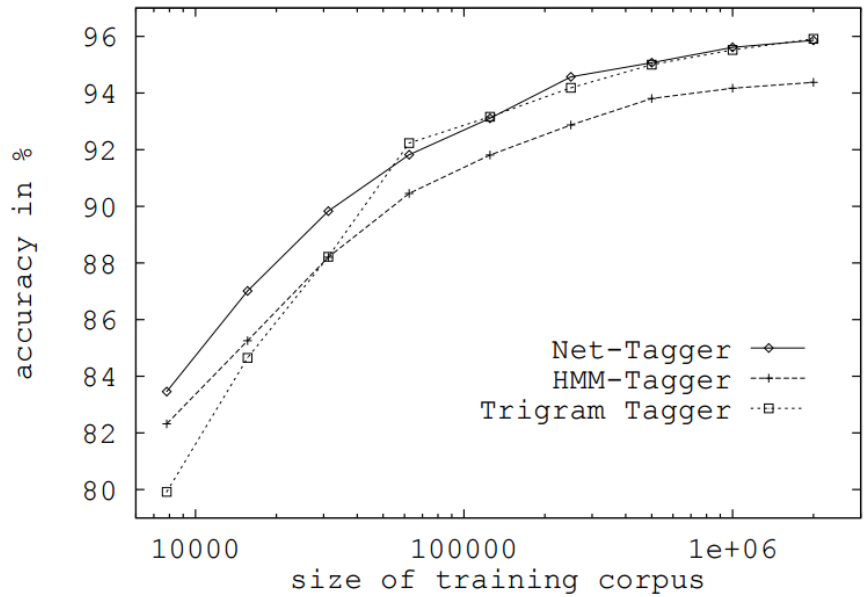
Tokenizing

- Tokenizing: Identification of
  - Sentence boundaries
  - Word form boundaries
- Problem cases:
  - Fused forms: IT *comprandolo*, FR *des, du,...*
  - Multiwords: *New York*, FR *pour que* ("in order to"),...
  - Non-word forms: *system FS-210, on 17-3-2015*,...
- Size of texts:
  - May matter for trainable systems Schmid 2000
  - Does not matter for rule-based systems: operating on regular expressions

Quantitative issues: Preprocessing

POS-tagging and lemmatization

- Tools are typically training-based:
  need big enough training corpus



- Training sets: size depends on:
  - Number of different tags in tagset
  - Complexity of tagset: hierarchical vs. flat
  - Type of tag definition criteria:
    lexical, morphosyntactic, distributional, ...
- Standard tagger: TreeTagger (=Trigram Tagger): Schmid 1994
  On a tagset of 60-65 tags:
  40.000 words of annotated (news) text are the minimum to be used

Quantitative issues: Preprocessing

Tagging and lemmatization quality

- Standard tools: 96 – 98% accuracy:
  - 2 to 4 errors per 100 items,
    2 to 4 million errors per 100 million items
- Lemmatization problems:
  - Frequent in texts from specialized language:
    * *Brandschutzklasse* ("fire protection class")
    * *Kunststoffbürste* ("plastic brush")
    * *Zementsorten* ("types of concrete")
  - To enhance the quality of subsequent treatment steps:
    need for additional procedures

Quantitative issues: Preprocessing

Devices to enhance lemmatization results

Gojun et al. 2012

- Simple morphological rules,
  based on inflectional properties of DE words:
  `N -ungen -ung  # Bohrungen-Bohrung`
- Compound splitting,
  plus check whether head is in the tool's dictionary:
  - *Brandschutzklasse* → *Brand*|*Schutz*|*Klasse*
  - *Klasse* → noun, fem.
- Or: compound splitting plus morphological rule:
  Aufbewahrungsdöschens ("storage container")
  Aufbewahrung|*Döschens* → Döschen → noun, neutr.

Quantitative issues: Preprocessing

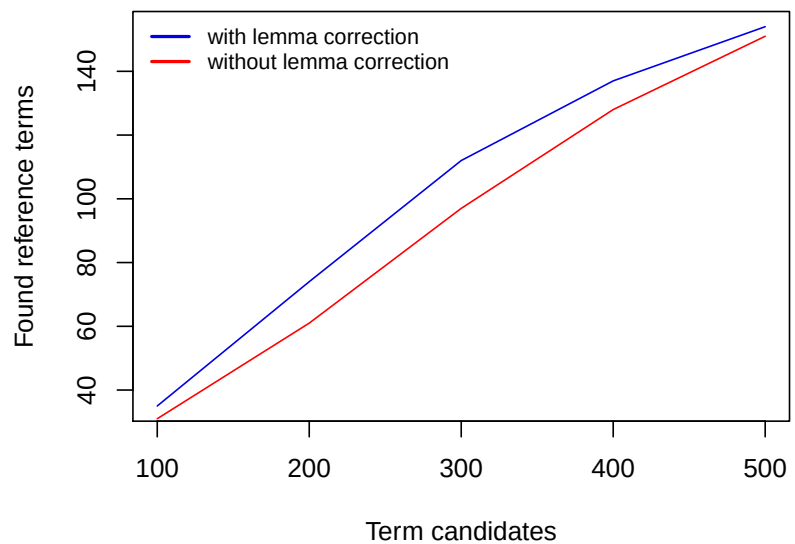
Relevance of lemmatization correction

Gojun et al. 2012

- In very large corpora:
  - Significant reduction of amount of errors
  - As lemmatization confirms pos-tagging, use for re-tagging with higher precision

- In very small corpora:
  - Available data can be exploited better

$\Rightarrow$ Improvement of both:
  - Recall
  - Precision



Quantitative issues: Preprocessing

Parsing

- Example case:
  trainable dependency parser *mate* Bohnet 2010
  - Trained on treebank data
  - The more the better
  - For German; TiGer treebank: 50.000 sentences, 900.000 tokens
- Open issue for ongoing research:
  Domain adaptation:
  how much material is needed in addition to existing training data?

Large vs. small corpora in applications

General aspects

- Statistical tools:
  The more the better!
  - Statistical Machine Translation
  - Distributional Semantics

  ⇒ Typically,
  developement is based on large data sets:
  EUROPARL, UNO texts, etc.
- Applications with philological orientation:
  - Lexicography of general language
  - Terminology work
  - Digital humanities applications

  ⇒ Examples

Lexicography: profiting from large corpora

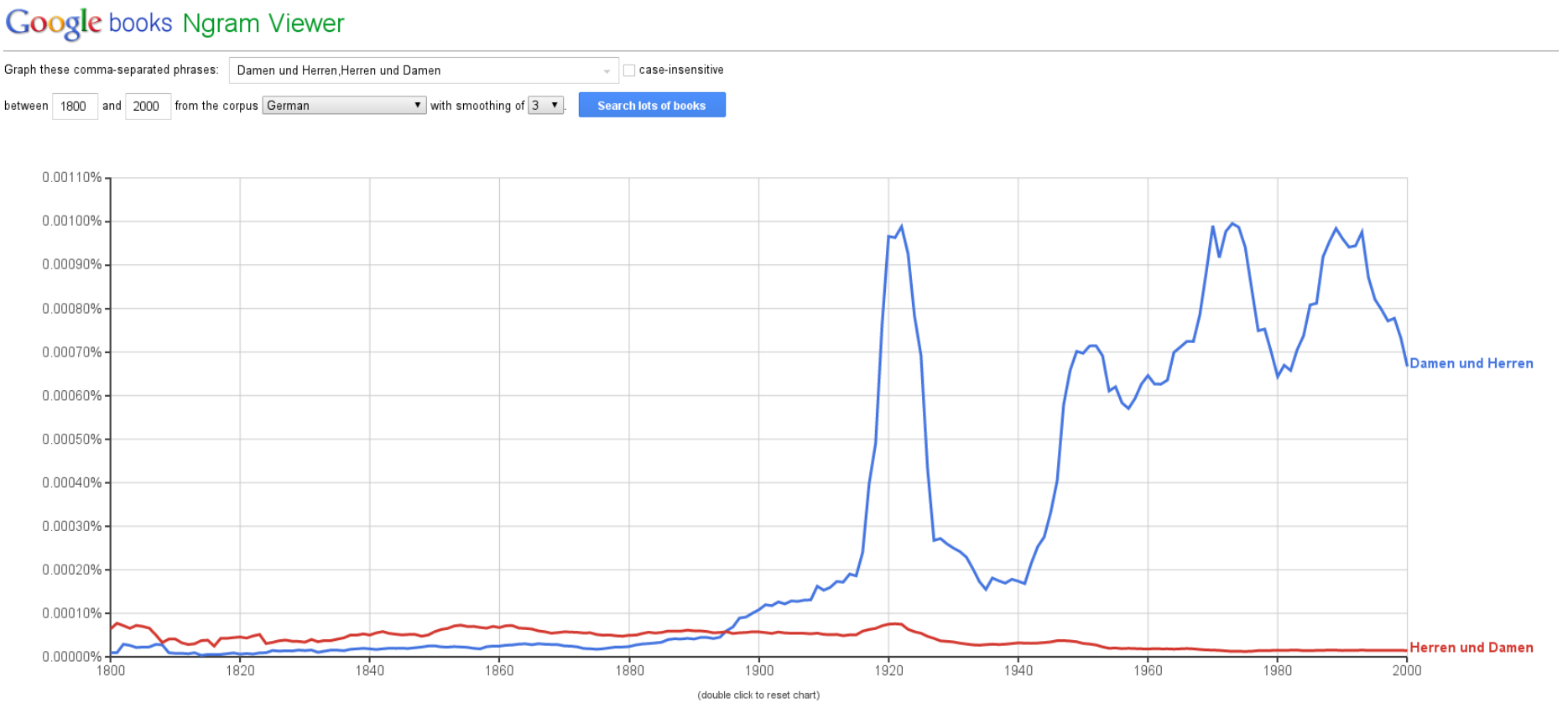
Examples

- National corpora: 100 to 1000 million words
- General tendency:
  Data centers collect as much data as possible:
  - IdS: over 20 billion words,
  - Austrian news corpus (APA, OeAW): over 6 bn
- Added value of (very) large corpora:
  - Possibility to see developements over time:
    cf. Google n-gram viewer
  - Possibility to quantify, e.g. competing phenomena
  - If metadata are available:
    Possibility to compare corpus subsets

Lexicography: profiting from large corpora

Development over time: Google n-gram viewer

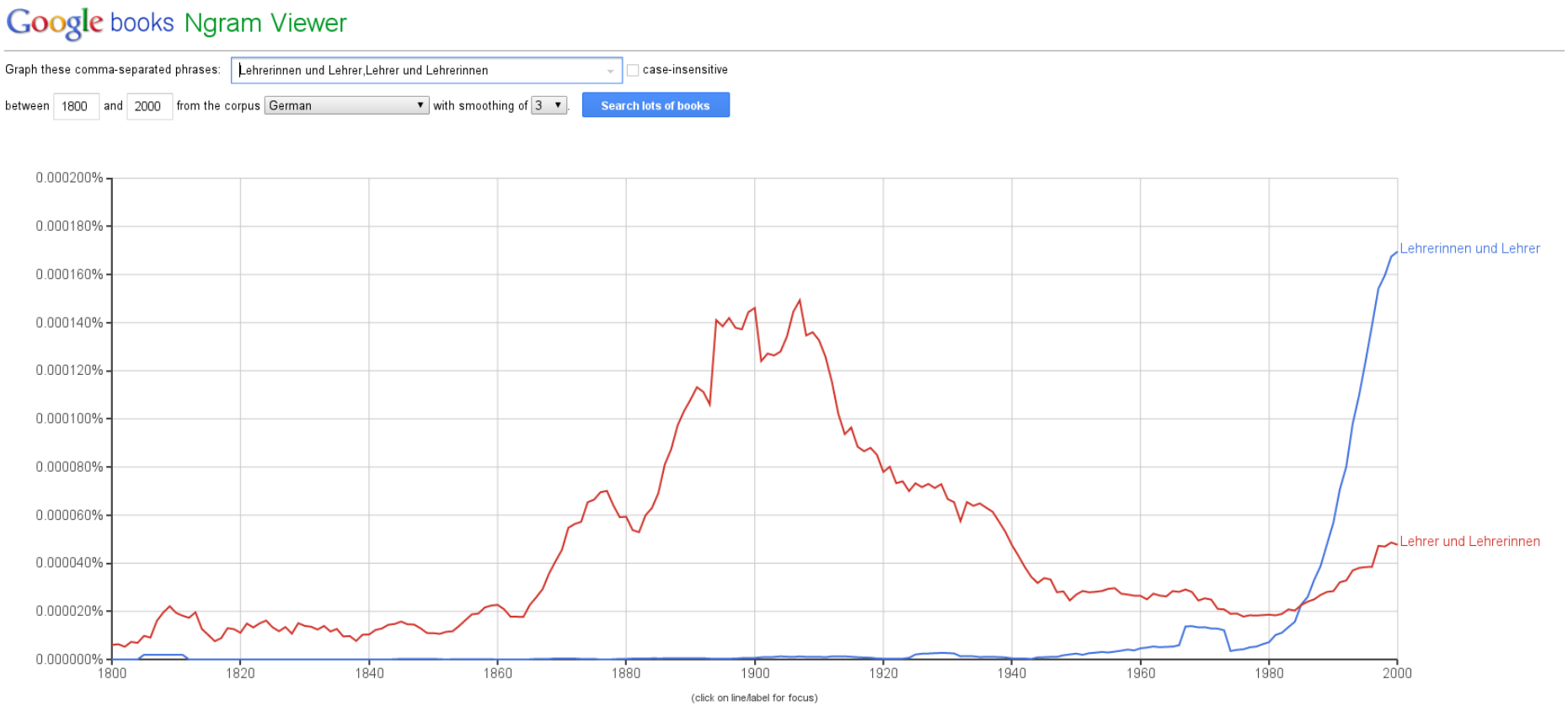
- Based on 5-grams from books at "Google books"
- Example of a (lexical) development over time:
  DE *Herren und Damen* vs. *Damen und Herren*



Lexicography: profiting from large corpora

Development over time: Google n-gram viewer

- Example continued:
  Lehrer und Lehrerinnen vs. *Lehrerinnen und Lehrer*



Lexicography: profiting from large corpora

Particular phenomena in corpus subsets

Engelberg et al. 2012

- Objective:
  - Identification of relative importance
    of certain verb valency patterns across different "genres"
  - Genres identified:
    1. newspaper text
    2. fiction: novels, autobiographies,...
    3. spoken: interviews, media data,...
    4. non-fiction: self-help books, letters, bibliographies,...
    5. science: scientific articles
- Example case: *widersprechen* ('contradict')

Genre	1	2	3	4	5	Total
Occurrences	200	291	61	91	96	739

Lexicography: profiting from large corpora

Particular phenomena in corpus subsets

Engelberg et al. 2012

- Example case *widersprechen*: patterns analyzed

1.	[Er]$_{Arg2}$ [he.NOM]$_{Arg2}$	widersprach contradicted	[dem Bericht]$_{Arg3}$. [the report.DAT]$_{Arg3}$	
2.	[“Das ist unmöglich,”]$_{Arg1}$ [“That is impossible”,]$_{Arg1}$	widersprach contradicted	[sie]$_{Arg2}$ [she.NOM]$_{Arg2}$	[ihm]$_{Arg4}$. [him.DAT]$_{Arg4}$
3.	[Die Berichte]$_{Arg1}$ [the reports.NOM]$_{Arg1}$	widersprechen contradicted	[sich]$_{Arg3}$. [themselves.DAT]$_{Arg3}$	

Genre	1	2	3	4	5
Type 1	18	77	14	8	5
Type 2	1	3	0	0	0
Type 3	5	4	5	3	3

Lexicography: profiting from large corpora

Using analyses of corpus subsets

- For reading distinctions: Prinsloo/Bothma/Heid 2013
  Example: DE loanword *Performance*
  - 500 occurrences in ca. 60 million words of *die Zeit*
  - Distribution over domains:

Domain	Total	Percent
Literature/Culture	235	47.0
Economy	106	21.2
Politics	47	9.4
Science	44	8.8

- To support contrastive synonym lexicography:
  Quasi-synonyms denoting "importance" in DE:

Domain	# N tokens	Relevanz	Signifikanz	Wichtigkeit
Literature	3.4 M	2.9	1.3	1.6
Politics	3.7 M	1.8	2.5	1.3
Sciences	2.0 M	3.3	6.9	1.1
Economy	2.5 M	1.7	4.2	0.4
History	0.15 M	1.3	1.3	2.7

Lexicography: profiting from large corpora

Open issue in e-lexicography: corpora as a knowledge source for end users (1/2)

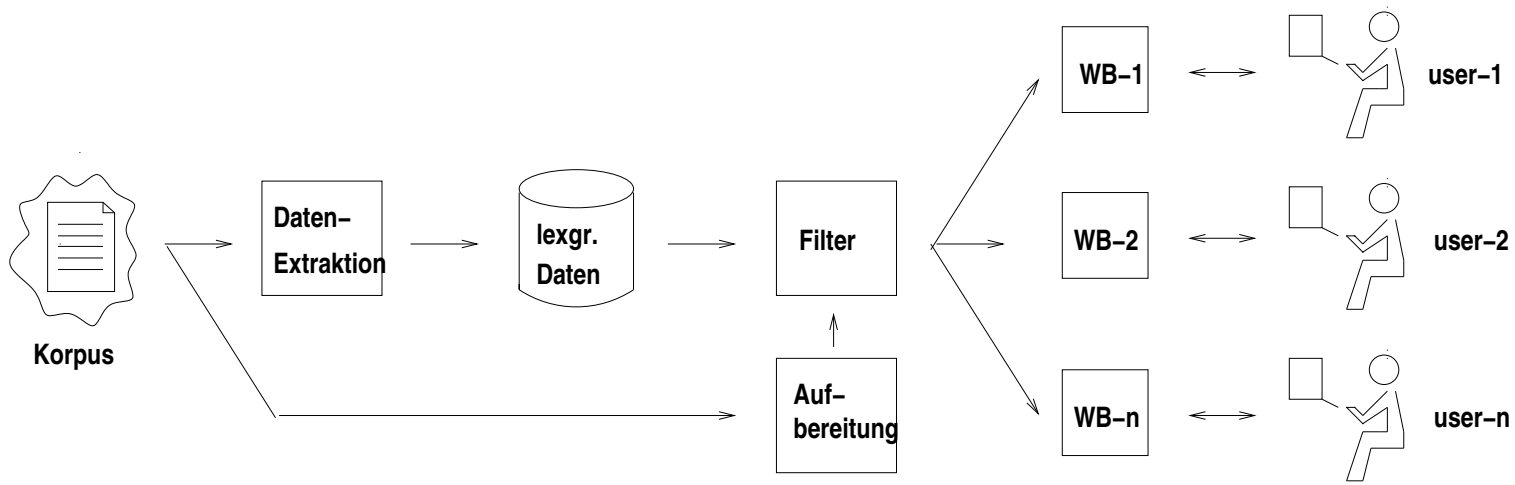
Requirements from lexicgraphic theory — Tarp 2012

Type	Contents	User orientation	Extra data offer
1	as in a print dictionary	- -	- -
2	follows print model, quick access	-	links within dictionary
3	"dynamic articles with dynamic data"	by user profiles	links to relevant websites: cognitive information
4	"dynamic articles with dynamic data"	individualized	"re-create and re-present relevant data"

Lexicography: profiting from large corpora

Open issue in e-lexicography: corpora as a knowledge source for end users (2/2)

Corpus data as an additional data offer for the end user



Required:

- Advanced tools for
  preparation, generalization and aggregation of data

Ways of counterbalancing a lack of data

The example of terminology

- Specialized languages:
  typically comparatively small corpora
  - User generated texts: forum contributions
  - Popular science
  - Didactic material: handbooks, manuals, ...
  - Expert-to-expert communication

  $\Rightarrow$ For increasing level of expertise
  and for increasing specialization:
  typically less text data available
- Examples:
  - Electrochemical gas sensor technology: $<$ 100.000 words
  - DIY instructions: 2,7 million words
  - Juridical journals on IPR: 1946-2006: 75 M words

Ways of counterbalancing a lack of data

Term candidate extraction – general observations

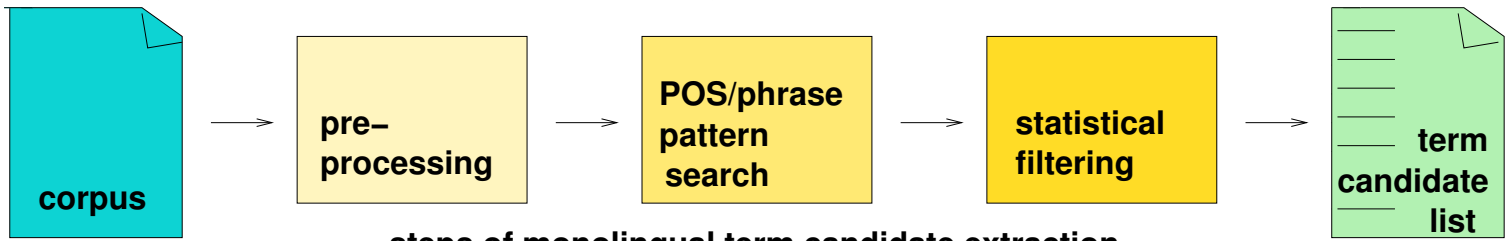
- Specialized lexicography:
  Domain experts: may "know" the domain:
  likely less profit from corpus-based term extraction
- Translation services and freelance translators:
  Need for updates of existing term collections
- Analysis of user-generated contents:
  - Not with lexicographic aims,
  - But with a view to domain mapping
    and to the analysis of topics discussed

Ways of counterbalancing a lack of data

Term candidate extraction: scenario

- Hybrid system: from EU-project TTC

Gojun et al. 2012



steps of monolingual term candidate extraction

- Use of 2.7 M DIY instructions corpus
- Examples of strategies:
  - Identification of term variants
  - Analysis of DE compounds, in order to relate them with phrasal constructs

Ways of counterbalancing a lack of data

Term candidate extraction: Variant identification (1/2)

- Definition of "term variant": Daille 2007

  "A variant of a term is

  an utterance → text occurrence

  which is

  semantically and conceptually related → synonymous/related

  with an original term." → base term

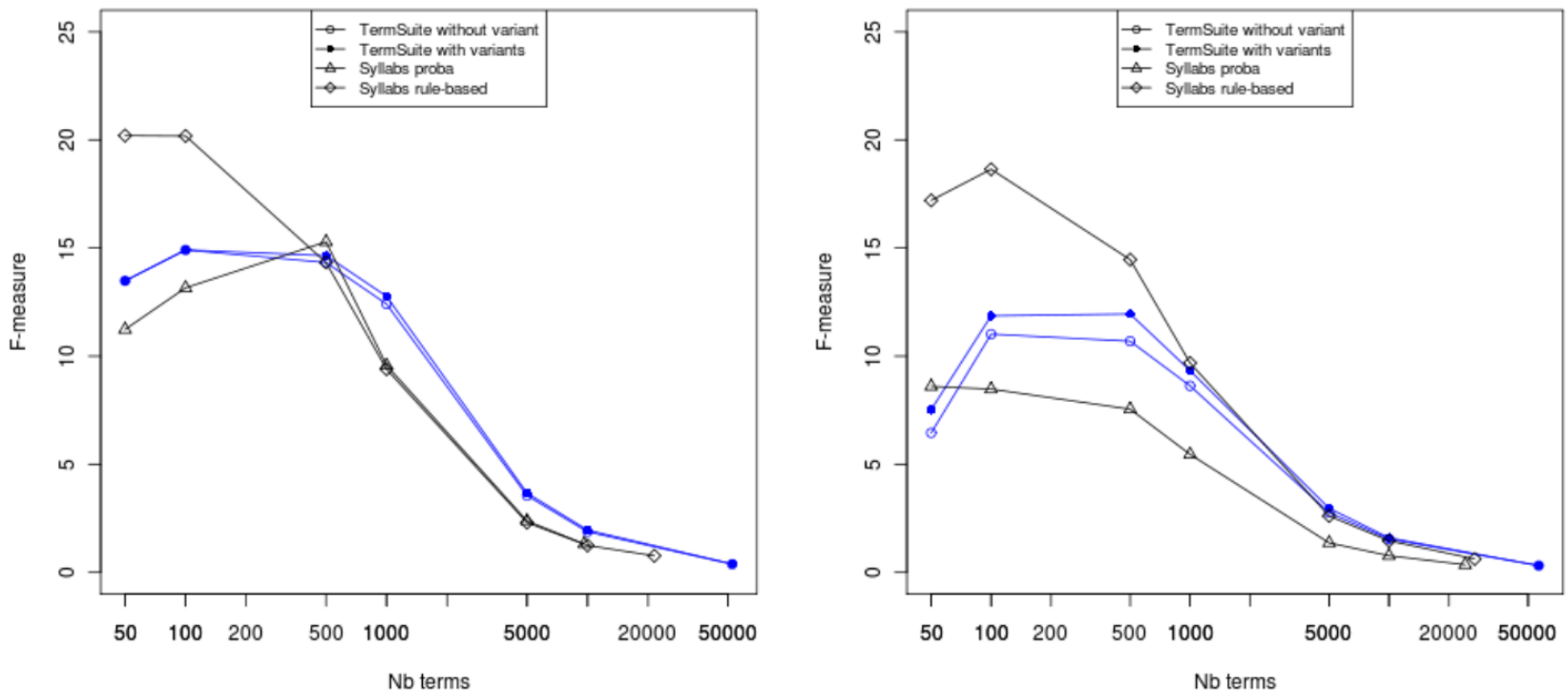
- Examples: domain of wind energy

	Base	**Variant**	**Example**
DE	N1 N2	N2 Sp N1	Wärmeproduktion ↔ Produktion von Wärme
EN	N1 N2	N2 Sp N1	energy production ↔ production of energy
ES	N1	A1	rotor ↔ rotórico
FR	N1 Art N2	N1 A2	tension du stator ↔ tension statorique

Ways of counterbalancing a lack of data

Term candidate extraction: Variant identification (2/2)

- Approach: based on variation patterns:
  - correspondence rules: variant pattern → base pattern
  - use of rules for derivation (suffixes)
  - splitting around hyphen
  - edit distance to identify orthographic variants
- Evaluation: Wind energy: EN (left), ES (right)



Ways of counterbalancing a lack of data

Term candidate extraction: compounds and phrases (1/2)

- Extracting verb+object pairs from dependency parses:
  - *Nun konnte der neue Laminatboden verlegt werden*
  - "Now, the new laminate flooring could be installed"

word	POS	lemma	gram. fct.	metadata
Nun	R	nun	NULL	project
konnte	Vois3s	können	NULL	project
der	T-dmsn	die	OBJ-EMB-45627	project
neue	A-pmsn	neu	OBJ-EMB-45627	project
Laminatboden	Ncmsn	Laminatboden	OBJ-HD-45627	project
verlegt	Vmp	verlegen	V-M-P-45627	project
werden	Van-	werden	NULL	project
.	S	.	NULL	project

Ways of counterbalancing a lack of data

Term candidate extraction: compounds and phrases (2/2)

- Compound splitting
  - Use of COMPOST, Cap 2015
    a hybrid tool combining
    a rule-based morphology and corpus data Schmid et al. 2004
  - *Laminat|boden*
  - *Boden|verlegung*
- Objective:
  finding text occurrences that "speak about the same thing",
  but in different ways:
  - *Holz|bohrer* (wood drill) $\leftrightarrow$ *Holz bohren*
  - *Stein|bohrer* (stone drill) $\leftrightarrow$ *Stein bohren*
  - *Diamant|bohrer* (diamond drill) $\not\leftrightarrow$ *Diamant bohren*

Conclusions

...has been shown

- Impact of corpus size
  - In the standard workflow of corpus linguistics:
    Sampling – Preprocessing – data extraction
  - In selected applications:
    * Lexicography
    * Terminology work
  - Possibilities to profit from very large amounts of data
  - Ways to counterbalance a lack of data

    $\Rightarrow$ In either case:
    Use of computational linguistic tools improves the results

Conclusions

Directions for future work

- In the field in general:
  - Domain adaptation techniques for statistical tools
  - Enrichment of statistical tools with linguistic knowledge
  - Tools and methods for dealing with small corpora:
    a necessity for some applications in the digital humanities
- Our own work:
  - Exploiting small corpora,
    e.g. by combining compound splitting and syntax
  - Exploring corpus use and generalization
    as a knowledge source for end users in e-dictionaries