

GRK 2839 – Winter School 2023

Corpus linguistics: Fundamentals, corpus compilation & annotation

Prof. Dr. Stephanie Evert

Chair of Computational Corpus Linguistics

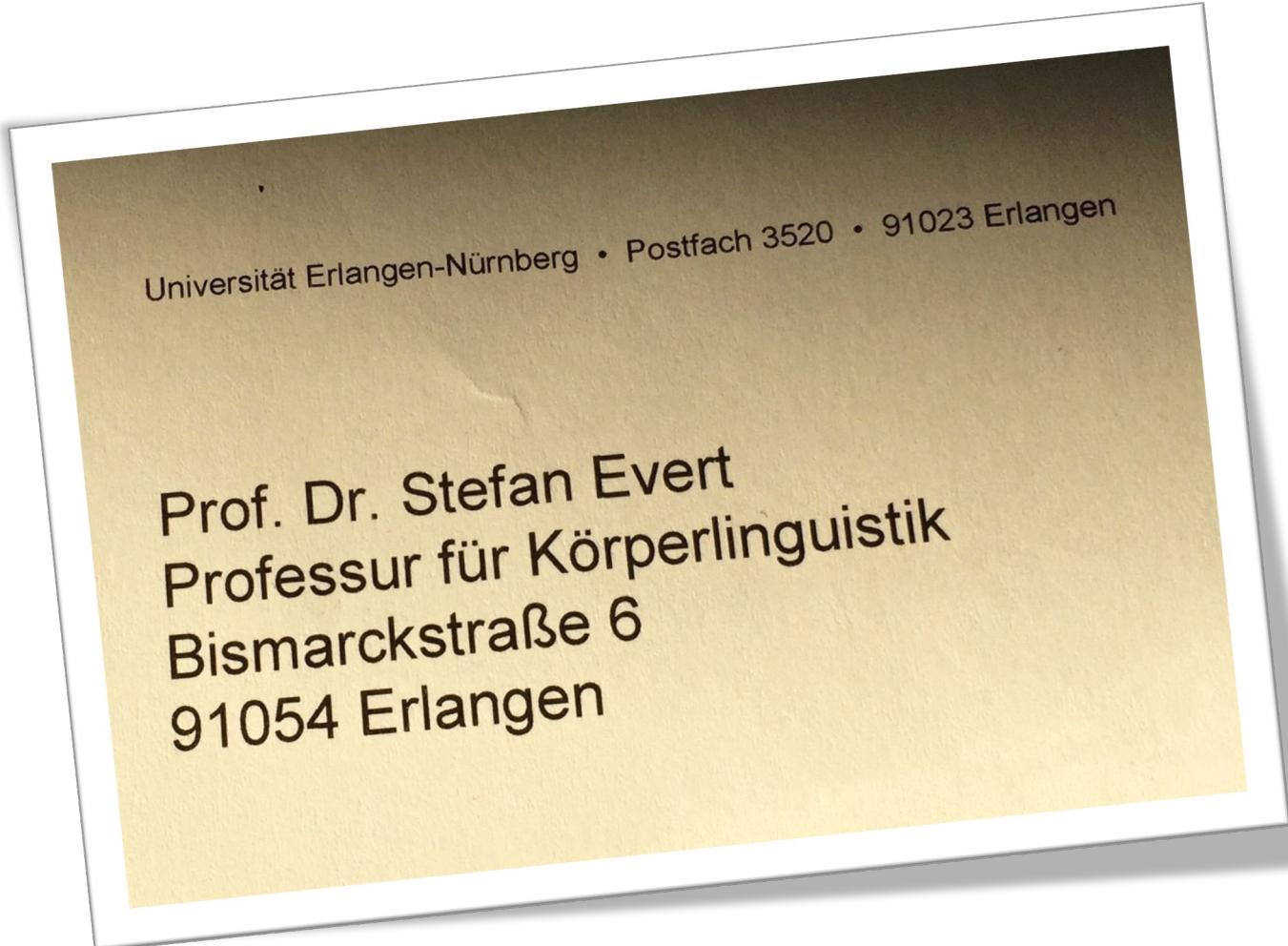
www.linguistik.uni-erlangen.de



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

What is a corpus?



What is a corpus?

corpus | 'kɔ:pəs |

noun (pl. **corpora** | 'kɔ:pərə | or **corpuses**)

1 a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject: *the Darwinian corpus*.

- a collection of written or spoken material in machine-readable form, assembled for the purpose of linguistic research.

2 Anatomy the main body or mass of a structure.

- the central part of the stomach, between the fundus and the antrum.

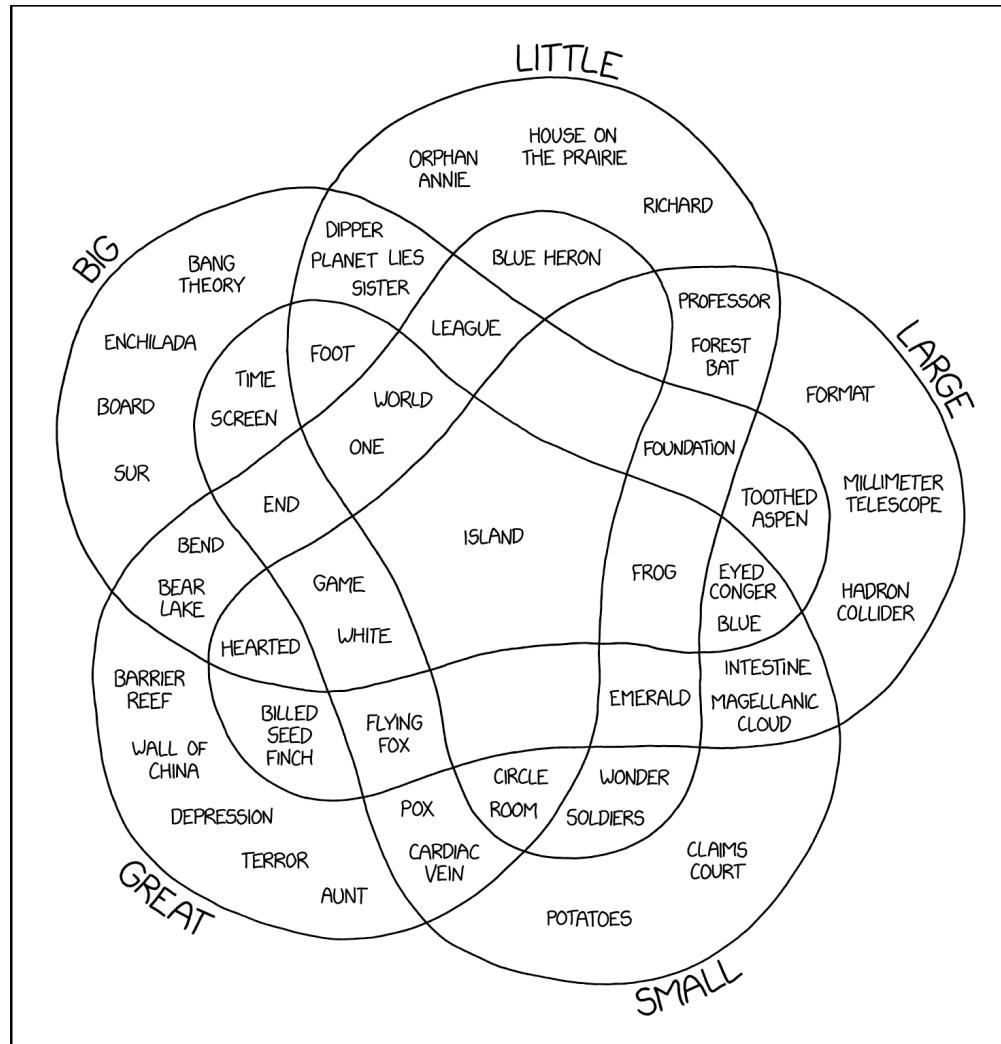
ORIGIN late Middle English (denoting a human or animal body): from Latin, literally '**body**'. **SENSE 1** dates from the early 18th cent.

- Corpus as **electronic text collection** (archive / database)
 - wide sense: common in computational linguistics & digital humanities
- Corpus as **representative sample** of language
 - narrow sense: dominant in corpus linguistics (in narrow sense)

What is corpus linguistics?

In a nutshell ...

The empirical
scientific study
of language
based on
authentic samples
of language use



<https://xkcd.com/2122/>

What is corpus linguistics?

- Corpus linguistics as a **methodology**
 - (one form of) empirical evidence for linguistic research
 - corpus design & practical aspects
 - development of software tools
 - quantitative analysis techniques, often in terms of frequency comparison
- Corpus linguistics as **discipline**
 - concordance, collocations, keywords
 - language in use, not language as system
 - language as a medium → applied CL
 - typical research topics
- Corpus-based research
 - verification of hypotheses and theoretical predictions
 - compilation of specialised corpora
 - exploration of specific constructions
 - statistical significance of differences, data analysis as discovery procedure
- Corpus-driven research
 - “trust the text” (Sinclair 2004)
 - purely empirical, inductive, contextualised
 - socio-political discourses & rhetoric
 - sociolinguistic & functional aspects
 - language variation
 - computational lexicography
 - language pedagogy (→ CALL)

...

Corpus linguistics vs. Theoretical linguistics

Corpus linguistics

- long tradition in esp. British linguistics since 1950s (Quirk, Firth, Sinclair, Leech, ...)
- based on corpus data,
i.e. authentic language use
- empiricist / inductive
- patterns & preferences
- frequency
- “ordinary” language in use
- empirical validation of theory,
induction of new hypotheses
- goals: language use, language as medium

Generative linguistics

- mainstream of theoretical linguistics since early 1960s (Chomsky 1957)
- based on introspection and grammaticality judgements
- rationalist / deductive
- rules
- yes / no (grammaticality)
- language system, border cases
- theory developed on basis of individual (invented) examples
- goal: cognitive account of language faculty

Pre-history of corpus linguistics

- First corpus-based quantitative studies in late 19th century
- Orthography and frequency lists
 - Käding (1897): German frequency dictionary based on corpus of ca. 11 million words (manual work!)
- Language acquisition
 - first longitudinal studies ca. 1876–1926 (parent diaries)
 - large cross-sectional studies ca. 1927–1957
- Lexicography
 - Murray: several million index cards for OED (1879–1928)
- Foreign language teaching
 - basic vocabulary, vocabulary levels, collocations (e.g. Palmer 1933)
- Structuralist language documentation
 - Boas (1940), J.R. Firth (1930–1955), ...
- Comparative philology
 - Eaton (1940): semantic frequency lists for English, French, German, Spanish

Chomsky (1957)

- Rationalism (introspection) vs. empiricism (data-oriented)
 - corpus linguistics: empirical description of language patterns
 - Chomsky: explanatory theory, must be cognitively plausible
- Competence vs. performance
 - Chomsky: corpus reflects speaker performance (with mistakes), empirical frequency data irrelevant for language competence
 - counter-argument: *armchair linguistics* based on invented examples
- Representativity
 - human speakers can produce infinitely many well-formed utterances; even a large corpus only contains a small subset → not representative
 - Chomsky: “Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, ...”
- Learnability: the poverty of stimulus argument
 - corpus data insufficient for language acquisition: no negative examples

Recent history of corpus linguistics

- Since 1950: Humanities Computing (→ Digital Humanities)
- 1950–1960: Mechanolinguistics (Juillard: contrastive corpora)
→ quantitative / mathematical linguistics (Harris 1968)
- 1960–1980: Corpus linguistics as European counter-movement against mainstream of generative linguistics (→ Chomsky)
- Corpus-based grammars (e.g. Quirk/Greenbaum)
 - Survey of English Usage (SEU) since 1960
 - Brown Corpus 1961–1963
- British contextualism (Firth & Sinclair)
 - Firth (1957) building on Malinowski & Jones, Sinclair (1991), computational lexicography (COBUILD)
 - principles of collocation & colligation, “trust the text”
- Since 1990: Corpus linguistics established as subdiscipline of linguistics, begins to accept methodological innovations from other subdisciplines
- Since 2010: Corpus evidence has become an essential part of linguistic research

Applications of corpus linguistics (I)

- Dialectology & contrastive linguistics
- Historical linguistics & language change
- Language description (e.g. endangered languages)
- Language teaching, language acquisition, CALL
- Language variation, register studies (→ Biber's MDA)
- Lexical semantics (e.g. semantic prosodies)
- Lexicography & lexicology (→ COBUILD, collocation dictionaries)
- Morphology (→ quantitative productivity)
- Phonology (esp. studies of phonological variation)

See McEnery, Xiao & Tono (2006, Unit A10)
for examples and references

Applications of corpus linguistics (II)

- Pragmatics & discourse analysis (→ rhetoric, ideology, politics)
- Psycholinguistics & psychology (e.g. frequency & association norms)
- Sociolinguistics (e.g. gender studies, language ideology, power relations)
- Stylometry & literary studies (e.g. authorship attribution, literary stylistics)
- Syntax & grammar (→ Quirk/Greenbaum, LGSWE)
- Theoretical linguistics (→ validation of theoretical predictions)
- Translation studies (→ “translationese”, CAT)
- Usage-based linguistics (→ cognitive linguistics, construction grammar)

See McEnery, Xiao & Tono (2006, Unit A10)
for examples and references

The stages of a typical corpus study

1. Operationalization

- research question → quantitative hypothesis, definition of population (*sampling frame*)

2. Corpus compilation

- selection of texts (from sampling frame), digitization / format conversion
- collection of metadata, legal & ethical issues
- shortcut: reuse existing corpus (often by selecting a suitable subcorpus)



today

3. Linguistic annotation

- manual annotation, GUI, annotator agreement
- automatic annotation with NLP tools for larger corpora



Thursday

4. Representation format

- standards: Unicode, XML, TEI, XCES, ... important for archiving and data exchange
- efficient binary index formats (e.g. CWB) for corpus search & quantitative analysis

The stages of a typical corpus study

5. Indexing & search

- search for keyword, phrase, linguistic pattern, ...
- view results as concordance (“kwic” = keyword in context)
- analysis = grouping & structuring of concordance in order to identify recurrent patterns
- efficient search based on binary index form

Your query “[word=“what”%c & !bound(s)] “a”%c [pos=“JJ.*”]+ “night”%c” returned 18 matches in 15 different texts (in 98,511,777 words [9,802 texts]; frequency: 0.18 instances per million words)
[4.616 seconds]



Solution 1 to 18 Page 1 / 1		
My dearJarmila ,	what a great night	for you .
oes crazy on a hot night , and maybe that 's	what a hot night	is for .
Wow ,	what a miserable night	.
tee that you 'll have a ball Come one and all	What a great night	you 've got in store You 'll wanna keep comi
e Hey , everyone , let 's go on with the show	What a great night	you 've got in store I 'll bet you 'll wanna kee
I am glad on ` t.	What a fearful night	is this !
Oh , dear ,	what a terrible night	.
noes on and on Oh , Robin ,	what a beautiful night	.
morning skyline Whoa oh	What a weird night	, huh ?
It 's a wonder	what a good night	's sleep will do for you .
ink it 's sunset and see	what a nice night	it is , they 'll muster in the lobby .
God ,	what a beautiful night	, Jack .
God ,	what a beautiful night	, huh ?

6. Quantitative analysis

- many insights based on systematic analysis of frequency data (esp. for large corpora)
- frequency comparison, keywords, co-occurrence
- statistical hypothesis tests, data analysis, visualisation

Friday

7. Interpretation

Types of corpora

- written **vs.** spoken **vs.** multimodal/multi-media
- reference corpus **vs.** specialized corpus
- synchronic **vs.** diachronic (discrete, continuous)
- closed corpus **vs.** monitor corpus
- monolingual **vs.** multilingual (parallel, comparable)
- unannotated (raw text) **vs.** annotated
 - metadata = information about texts & speakers/authors
 - linguistic annotation = systematically coded interpretation
- corpus size: small & clean **vs.** large & messy
 - measured in M = million (or G = billion) running words

Some corpora everybody should know

- Brown Corpus (Francis & Kucera 1964)
 - American English, written (edited), texts published in 1961
 - 500 samples @ 2000 words from 15 text genres (*categories*)
- Brown Family
 - Brown (AmE, 1961), LOB (BrE, 1961) – Frown (AmE, 1991), FLOB (BrE, 1991)
– BLOB (BrE, 1931), BE2006 (BrE, 2006)
- Penn Treebank (Marcus, Santorini & Marcinkiewicz, 1993)
 - ca. 3 million words of AmE with syntactic analyses (*parse trees*)
- British National Corpus (Aston & Burnard 1998)
 - British English, 90% written / 10% spoken, collected ca. 1991
 - approx. 100 million words in 4048 files (= texts / collections)
- Web as Corpus: WaCky (Baroni et al. 2009)
 - ca. 2 billion words of text from automatically crawled Web pages for each of DE, EN, FR, IT
 - many other Web as Corpus projects: larger corpora, additional languages (Arachnea, COW, SkE 10^{10})

Recommended textbooks

- McEnery, Tony and Wilson, Andrew (2001). *Corpus Linguistics*. Edinburgh University Press, 2nd ed.
- McEnery, Tony; Xiao, Richard; Tono, Yukio (2006). *Corpus-Based Language Studies: An advanced resource book*. Routledge, London/New York. <https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/CBLS.htm>
- McEnery, Tony and Hardie, Andrew (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.
- Lüdeling, Anke and Kytö, Merja (eds.) (2008). *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.
- Kennedy, Graeme D. (1998). *An Introduction to Corpus Linguistics*. Longman (Pearson Education Ltd), London and New York.
- Hoffmann, Sebastian *et al.* (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, vol. 6 of English Corpus Linguistics. Peter Lang, Frankfurt.
- Lemnitzer, Lothar and Zinsmeister, Heike (2015). *Korpuslinguistik: Eine Einführung*. Narr, Tübingen, 3rd edition.

Research community

- Important conferences
 - **Corpus Linguistics** (Lancaster / Birmingham / UK)
 - ICAME = International Computer Archive of Modern and Medieval English
 - ACL = American Association for Corpus Linguistics
 - CILC = International Conference on Corpus Linguistics
- Scientific journals
 - *International Journal of Corpus Linguistics* (IJCL)
 - *Corpora*
 - *ICAME Journal*
 - *Corpus Linguistics and Linguistic Theory* (CLLT)
- Web resources
 - **David Lee's bookmarks** (maintained by Martin Weisser): <http://tiny.cc/corpora> (navigation: CBL Links)
 - Linguistics Web (Sabine Bartsch): <http://www.linguisticsweb.org/>



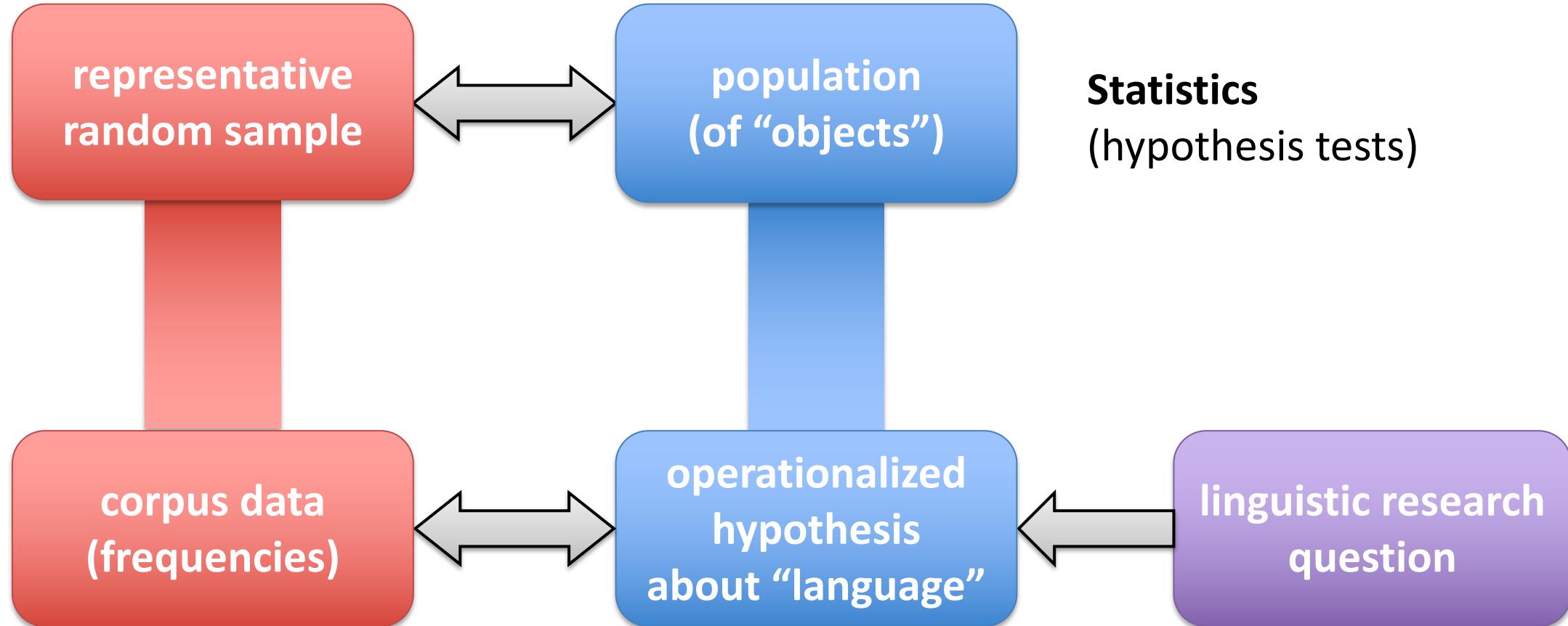
Corpus design



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

What is representativeness?



Goals of corpus design

● Representativeness

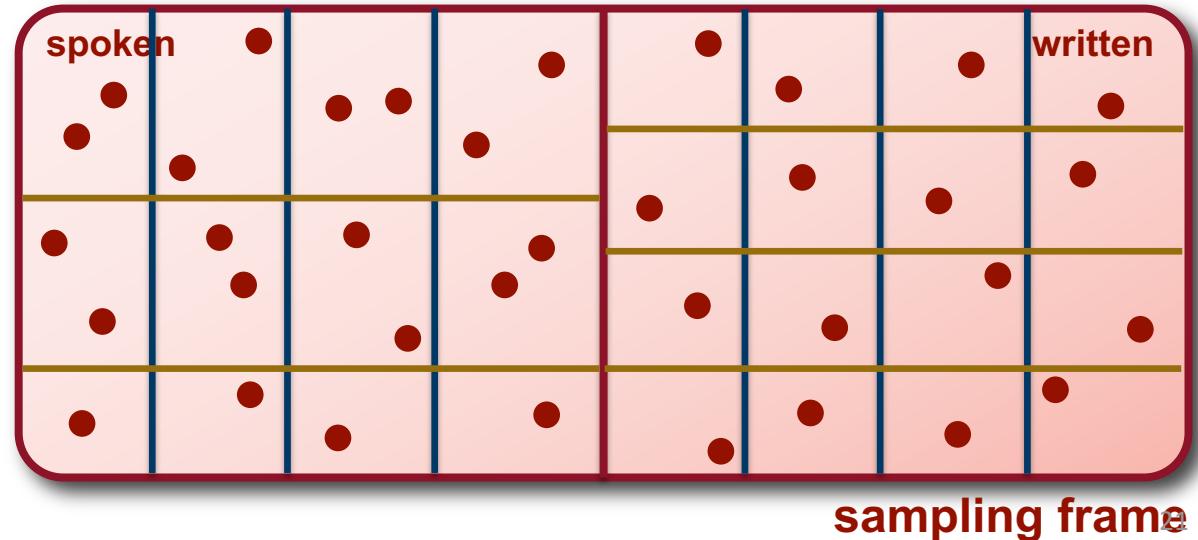
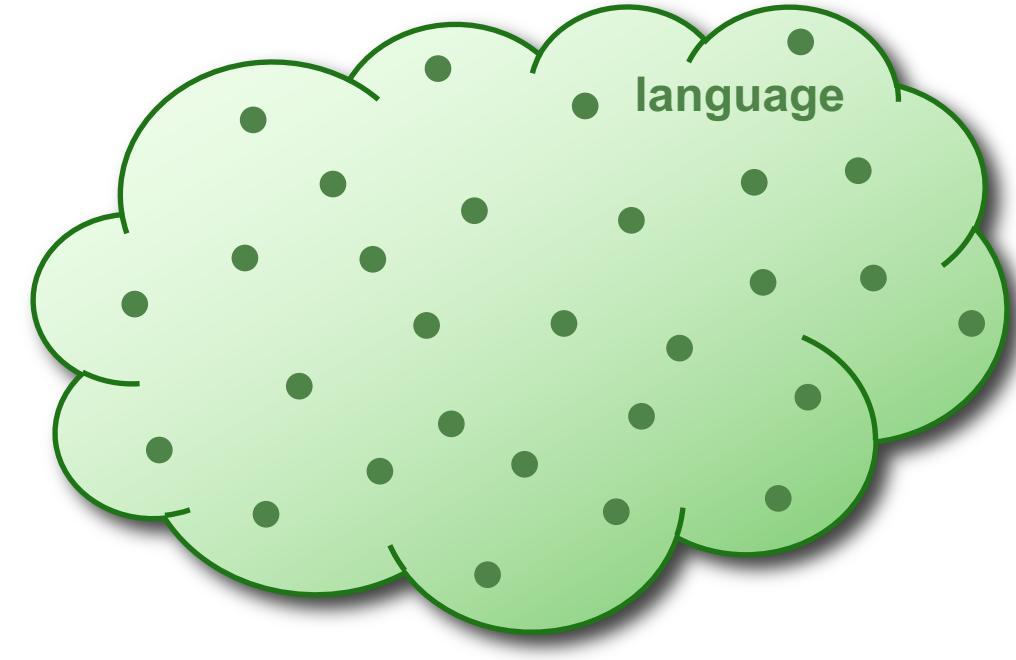
- a corpus should be representative of the (sub-)language to be studied
- statistics: random sample
- full representativeness difficult to achieve
- must at least be **balanced** (= good coverage of different registers, speakers, ...)
- and avoid **bias** or **skew** towards any particular group of speakers, text type, topic, ...

● Comparability

- corpus linguistic analysis often builds on frequency comparison between different corpora or sub-corpora
- prerequisite: comparable corpora

Representativeness & sampling

- Statistics: completely **random sample**
 - extensional population of interest, i.e. a (possibly infinite) collection of objects
 - randomly select n objects from population
- But what about language?
- Design criteria → **sampling frame**
 - dices up and defines linguistic population
→ make relevant texts **identifiable**
 - “A sampling frame is an operational definition of the population, an itemized listing of population members from which a representative corpus can be chosen.”
(Biber 1993, 244)
 - pick specified number of items from each cell (related to stratified sampling)



Representativeness & sampling

- Definition of a sampling frame
 - fundamental distinctions: mode (spoken/written/written-to-be-spoken), medium
 - text characteristics: (publication) date, author (single/multi/anon), region, target audience, ...
 - function of text: genre / text type (factuality, purpose, situation, ...), topic domain, ...
 - properties of author/speaker: sex, age, dialect, social class, ...
 - see Atkins et al. (1992) for a comprehensive system of categories
- Balance
 - include texts from all (combinations of) categories in the sampling frame = grid cells
 - avoids bias/skew → balanced coverage of the “language” population
- Representativeness
 - sampling frame makes population identifiable (for each combination of categories)
→ random selection of texts for each cell
 - must specify **proportion of texts** to be sampled from each category = prevalence in language

Further reading

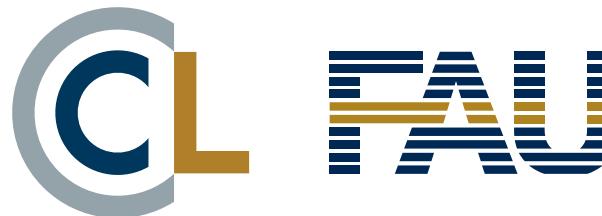
- Atkins, Sue; Clear, Jeremy; Ostler, Nicholas (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1–16.
- Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243– 257.
- HSK 29.1 *Corpus Linguistics*, Art. 9
- HSK 5.4 *Dictionaries: Computational Lexicography*, Art. 96 (Ch. XVIII)

How would you design a corpus for a study
of evaluative language in music reviews?

... or another research question?



Annotation



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

A corpus consists of ...

- **Object data** = texts
 - primary data, main object of analysis
- **Metadata** = information about the texts
 - title, author, publication date, text type, medium, ...
 - age, sex, education, region, dialect, ... of authors
 - always include all variables used to define sampling frame
- **Typographic markup & text structure**
 - paragraphs, headings, bold/italics, typeface, itemized lists, footnotes, ...
- **Annotation** = linguistic interpretation
 - simple (token level) vs. structured (e.g. syntax tree)
 - essential for querying and analyzing large corpora

It seemed a day much as any other until I happened to look out of the back window. There was a little garden behind the house; a well-mown lawn surrounded by a neatly cut hedge, a few bushes and colourful flowers.

metadata

title: The Garden
author: Stefan Evert
author sex: male
date: 05.08.1991

It seemed a day much as any other until I happened to look out of the back **window** . There was a little garden behind the **house** ; a well-mown lawn surrounded by a neatly cut **hedge** , a few bushes and colourful **flowers** .

Corpus annotation: sentence segmentation

< s > It seemed a day much as any other until I happened to look out of the back window . < /s >

< s > There was a little garden behind the house ; a well-mown lawn surrounded by a neatly cut hedge , a few bushes and colourful flowers . < /s >

Corpus annotation: part-of-speech (POS) tagging

< s > It_{PP} seemed_{VBD} a_{DT} day_{NN} much_{RB} as_{IN} any_{DT} other_{JJ} until_{IN} I_{PP}
happened_{VBD} to_{TO} look_{VB} out_{RP} of_{IN} the_{DT} back_{JJ} window_{NN} ·SENT </s>

< s > There_{EX} was_{VBD} a_{DT} little_{JJ} garden_{NN} behind_{IN} the_{DT} house_{NN} ;
a_{DT} well-mown_{VBN} lawn_{NN} surrounded_{VBN} by_{IN} a_{DT} neatly_{RB} cut_{VBN}
hedge_{NN} , a_{DT} few_{JJ} bushes_{NNS} and_{CC} colourful_{JJ} flowers_{NNS} ·SENT
</s>

English: Penn tagset

* with TreeTagger-internal modifications

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential <i>there</i>
FW	Foreign word
IN	Preposition / subordinating conjunction
* IN/that	Subordinating conjunction <i>that</i>
JJ	Adjective (positive)
JJR	Adjective (comparative)
JJS	Adjective (superlative)
LS	List item marker
MD	Modal verb
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending ('s)
PP	Personal pronoun
PP\$	Possessive pronoun
RB	Adverb
RP	Particle
SYM	Symbol (mathematical/scientific)
TO	<i>to</i> (any usage) <i>fly to Paris, ready to go, ...</i>
UH	Interjection
#	Pound sign £
\$	Dollar sign \$

VB	Verb <i>be</i> , base form
VBD	Verb <i>be</i> , past tense
VBG	Verb <i>be</i> , gerund/progressive
VBN	Verb <i>be</i> , past participle
VBP	Verb <i>be</i> , non-3rd pers. sg. present
VBZ	Verb <i>be</i> , 3rd pers. sg. present tense
* VH	Verb <i>have</i> , base form
* VHD	Verb <i>have</i> , past tense
* VHG	Verb <i>have</i> , gerund/progressive
* VHN	Verb <i>have</i> , past participle
* VHP	Verb <i>have</i> , non-3rd pers. sg. present
* VHZ	Verb <i>have</i> , 3rd pers. sg. present tense
* VV	Lexical verb, base form
* VVD	Lexical verb, past tense
* VVG	Lexical verb, gerund/progressive
* VVN	Lexical verb, past participle
* VVP	Lexical verb, non-3rd pers. sg. present
* VVZ	Lexical verb, 3rd pers. sg. present tense
WDT	Wh-determiner
WP	Wh-pronoun
WP\$	Possessive wh-pronoun
WRB	Wh-adverb
SENT	Sentence-final punctuation . ! ?
,	Comma ,
:	Colon, semi-colon : ;
()	Comma ([])
‘ ’ “ ” ‘ ’ “ ”	Comma “ ” ‘ ’ “ ”

German: STTS tagset

ADJA	attributives Adjektiv
ADJD	adverbiales / prädikatives Adjektiv
ADV	Adverb <i>schon, bald, doch</i>
APPR	Präposition / Zirkumposition links
APPART	Präposition mit Artikel fusioniert <i>zum</i>
APPO	Postposition <i>zufolge, wegen</i>
APZR	Zirkumposition rechts <i>von ... an</i>
ART	bestimmter oder unbestimmter Artikel
CARD	Kardinalzahlen (Ordinalzahl = ADJA)
FM	Fremdsprachliches Material
ITJ	Interjektion <i>mhm, ach, tja</i>
KOUI	unterordnende Konj. mit <i>zu + Inf</i>
KOUS	unterordnende Konjunktion mit Satz
KON	nebenordnende Konjunktion <i>und, oder</i>
KOKOM	Vergleichskonjunktion <i>als, wie</i>
NN	normales Nomen
NE	Eigenname
PDS	substituierendes Demonstrativpron.
PDAT	attribuierendes Demonstrativpron.
PIS	substituierendes Indefinitpron.
PIAT	attrib. Indefinitpron. ohne Determiner
PIDAT	attrib. Indefinitpron. mit Determiner
PPER	Personalpronomen (nicht reflexiv)
PPOS	substituierendes Possessivpronomen
PPOSAT	attribuierendes Possessivpronomen
PRELS	substituierendes Relativpronomen
PRELAT	attribuierendes Relativpronomen

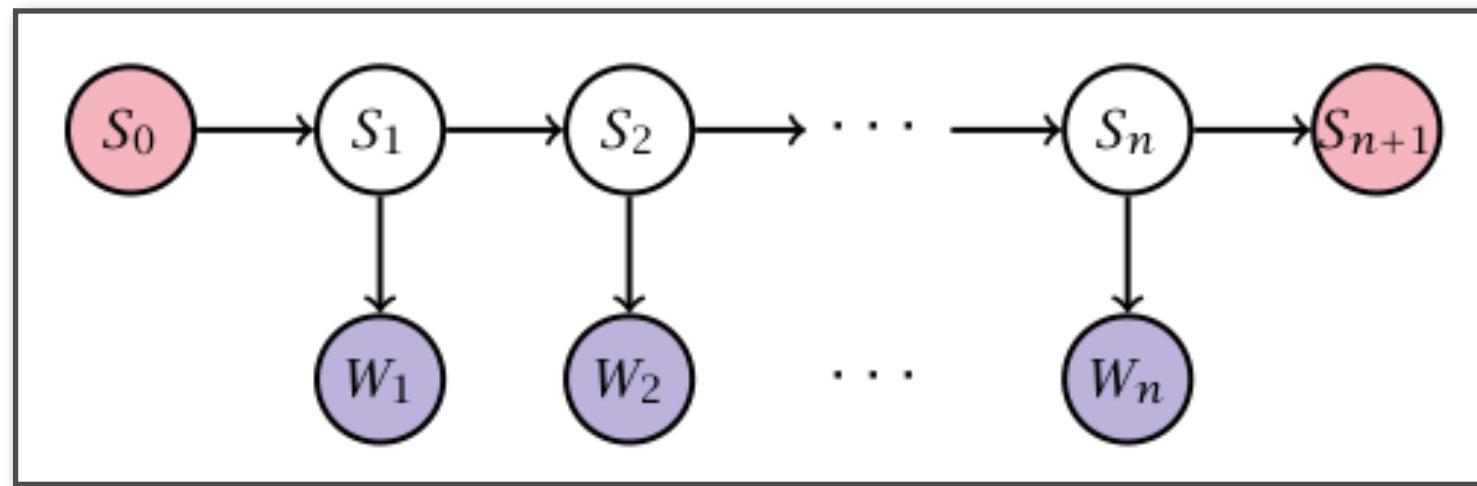
PRF	reflexives Personalpronomen
PWS	substituierendes Interrogativpron.
PWAT	attribuierendes Interrogativpronomen
PWAV	adverbiales Interrogativ-/Relativpron.
PAV	Pronominaladverb <i>dafür, deswegen</i>
PTKZU	zu vor Infinitiv
PTKNEG	Negationspartikel <i>nicht</i>
PTKVZ	abgetrennter Verbzusatz <i>kommt ... an</i>
PTKANT	Antwortpartikel <i>ja, nein, danke</i>
PTKA	Partikel bei Adjektiv/Adverb <i>am, zu</i>
TRUNC	Kompositions-Erstglied <i>Unter- und ...</i>
VVFIN	finites Verb, voll (= lexikalisch)
VVIMP	Imperativ, voll
VVINF	Infinitiv, voll
VVIZU	Infinitiv mit <i>zu</i> , voll
VVPP	Partizip Perfekt, voll
VAFIN	finites Hilfsverb
VAIMP	Imperativ, Hilfsverb
VAINF	Infinitiv, Hilfsverb
VAPP	Partizip Perfekt, Hilfsverb
VMFIN	Finites Modalverb
VMINF	Infinitiv, Modalverb
VMPP	Partizip Perfekt, Modalverb
XY	Nichtwort mit Sonderzeichen <i>3:7, H2O</i>
\$,	Komma ,
\$.	Satzbeendende Interpunktions . ? ! ; :
\$()	sonstige Satzzeichen (intern) - [] ()

Manual annotation

- Manual annotation for small, high-quality corpora
 - e.g. digital edition, political speeches, poetry/song texts, ...
- Annotation schema and categories
- Guidelines = detailed instructions for annotators
 - plus collection of examples for unclear / difficult cases
- Annotation tools (usually Web-based)
 - e.g. INCEpTION (<https://inception-project.github.io>), Prodigy (<https://prodi.gy>)
- Inter-Annotator Agreement (IAA)
 - reliability and validity of the annotation
 - annotator mistakes vs. systematic differences

Automatic annotation

- Most successful approach: machine learning
- Need to cast annotation as **classification task**
- **Gold standard** = corpus with manual annotation
 - annotation must be consistent, errors seem unproblematic
 - separate into training, development and test data
- Example: tagging with **Hidden Markov Model (HMM)**
 - see e.g. Brants (2000), Schmid (1995)



(Evert et al. 2009)

Corpus annotation: lemmatization

< s > It_{PP} seemed_{VBD} a_{DT} day_{NN} much_{RB} as_{IN} any_{DT} other_{JJ} until_{IN} I_{PP}
happened_{VBD} to_{TO} look_{VB} out_{RP} of_{IN} the_{DT} back_{JJ} window_{NN} ·SENT </s>

< s > There_{EX} was_{VBD} a_{DT} little_{JJ} garden_{NN} behind_{IN} the_{DT} house_{NN} ;
a_{DT} well-mown_{VBN} lawn_{NN} surrounded_{VBN} by_{IN} a_{DT} neatly_{RB} cut_{VBN}
hedge_{NN} , a_{DT} few_{JJ} bushes_{NNS} and_{CC} colourful_{JJ} flowers_{NNS} ·SENT
</s>

Corpus annotation: lemmatization

< s > It_{PP} it seemed_{VBD} seem a_{DT} a day_{NN} day much_{RB} much as_{IN} as any_{DT} any other_{JJ} other until_{IN} until I_{PP} I happened_{VBD} happen to_{TO} to look_{VB} look out_{RP} out of_{IN} of the_{DT} the back_{JJ} back window_{NN} window .SENT. </ s >

< s > There_{EX} there was_{VBD} be a_{DT} a little_{JJ} little garden_{NN} garden behind_{IN} behind the_{DT} the house_{NN} house ; ; a_{DT} a well-mown_{VBN} ??? lawn_{NN} lawn surrounded_{VBN} surround by_{IN} by a_{DT} a neatly_{RB} neatly cut_{VBN} cut hedge_{NN} hedge , , a_{DT} a few_{JJ} few bushes_{NNS} bush and_{CC} and colourful_{JJ} colorful flowers_{NNS} flower .SENT. </ s >



Corpus annotation: segments and structures

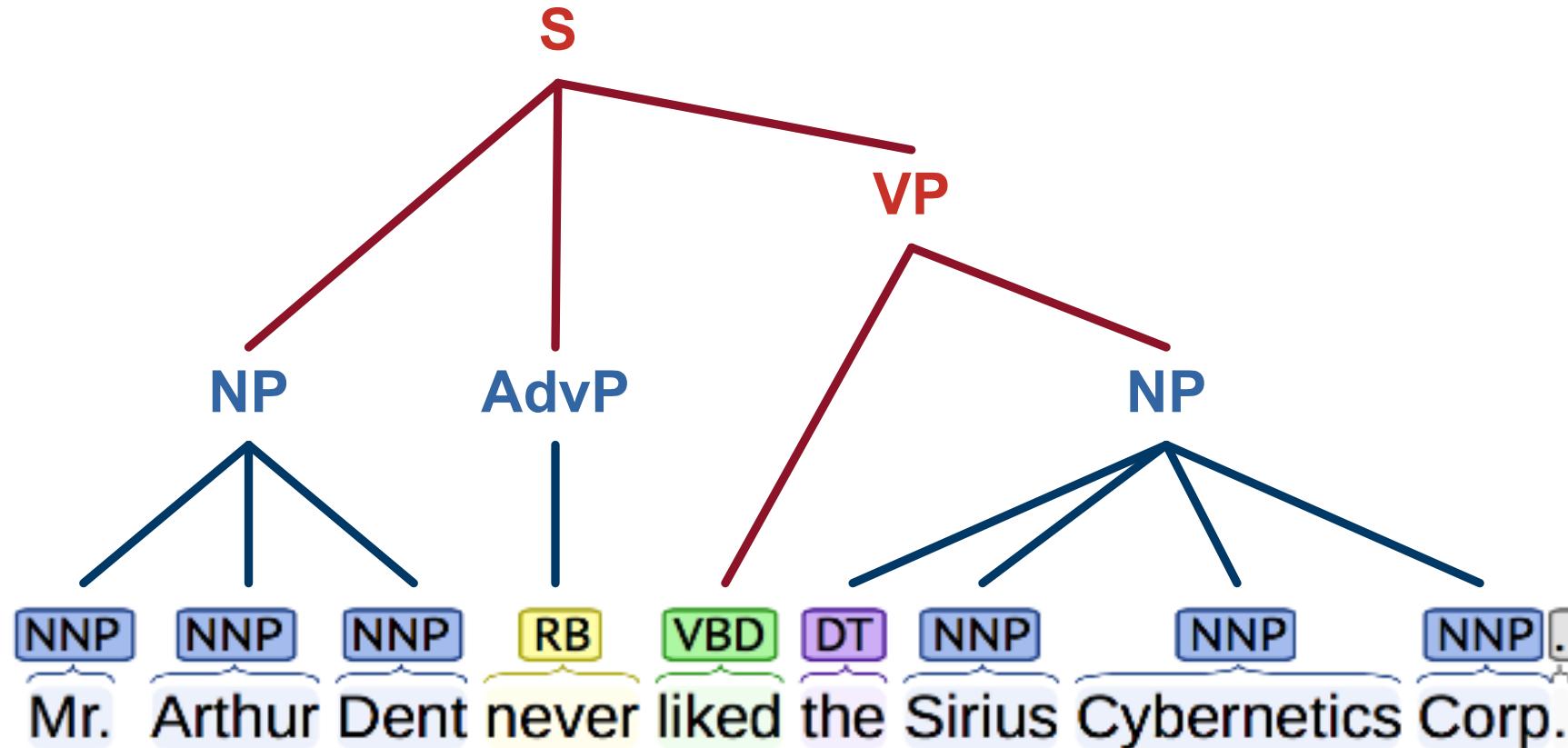
- Automatic recognition and categorization of particular word sequences (segments)
- e.g. named entities (NER = named entity recognition)

The diagram shows the sentence "Mr. Arthur Dent never liked the Sirius Cybernetics Corp." with specific words highlighted and categorized. The word "Mr." is underlined and enclosed in an orange box labeled "Person". The words "Sirius Cybernetics Corp." are underlined and enclosed in a blue box labeled "Organization".

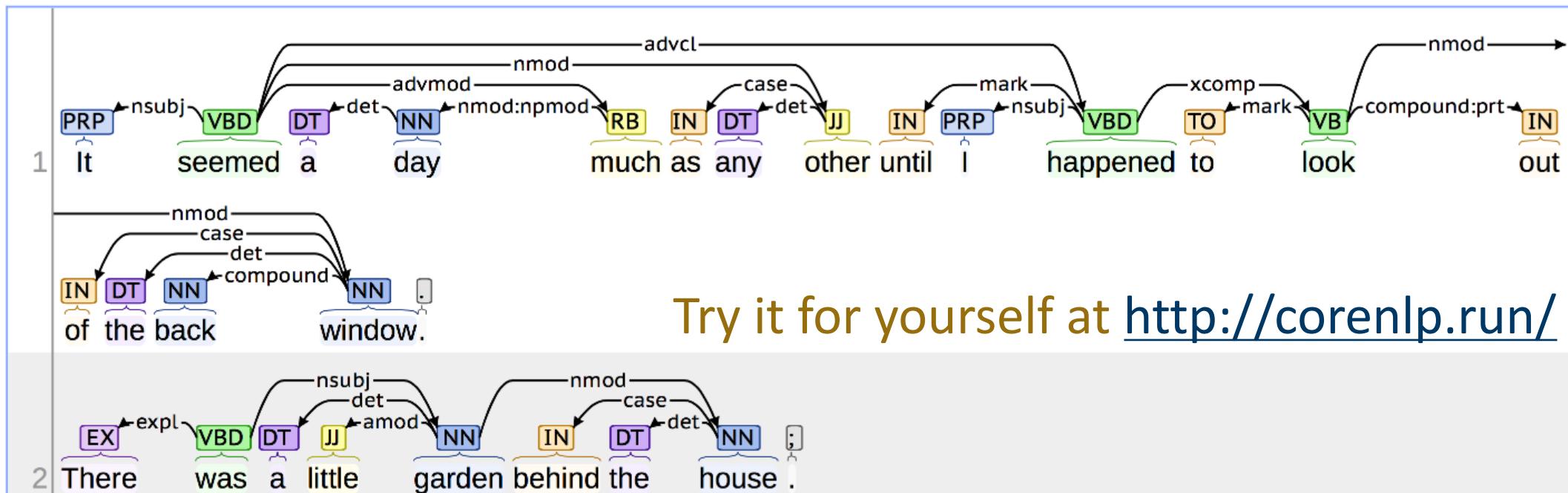
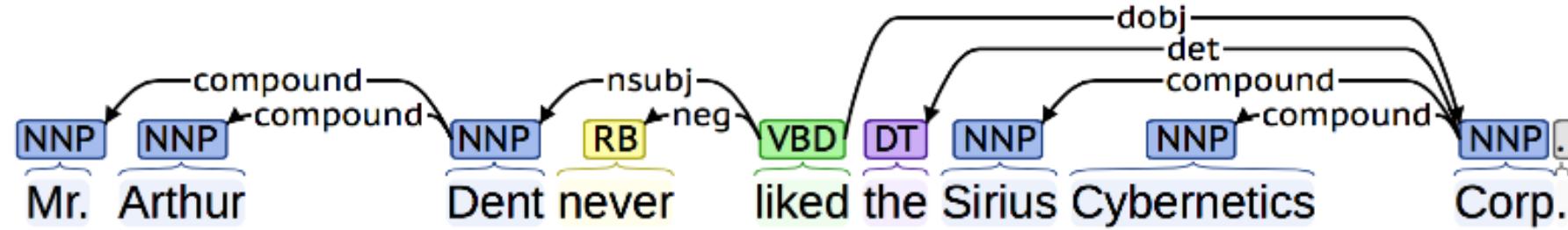
- e.g. time and place expressions: last week, the day after tomorrow, September 15th, in Paris, on the lawn in front of their house, ...
- e.g. text spans that need to be masked for anonymization purposes

Corpus annotation: segments and structures

- Syntactic phrase structure analysis
= parse tree of nested segments corresponding to syntactic units
- „minimal“ phrases as flat segments → chunk parsing



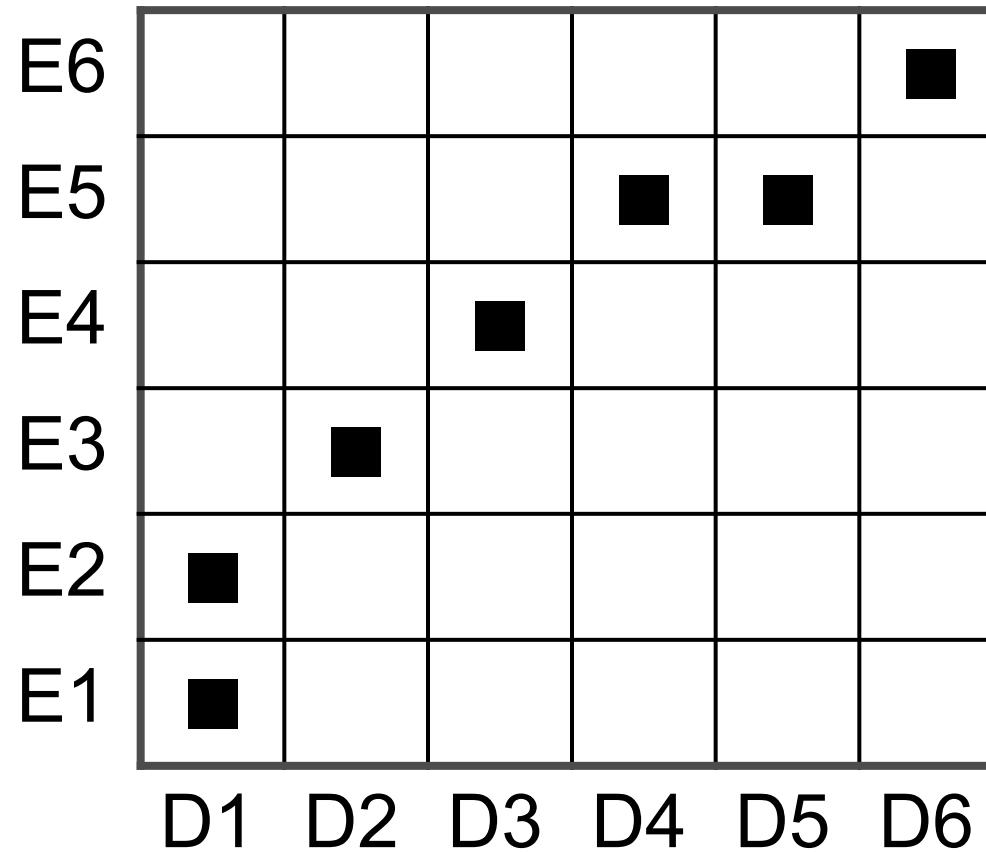
Corpus annotation: syntactic dependency analysis



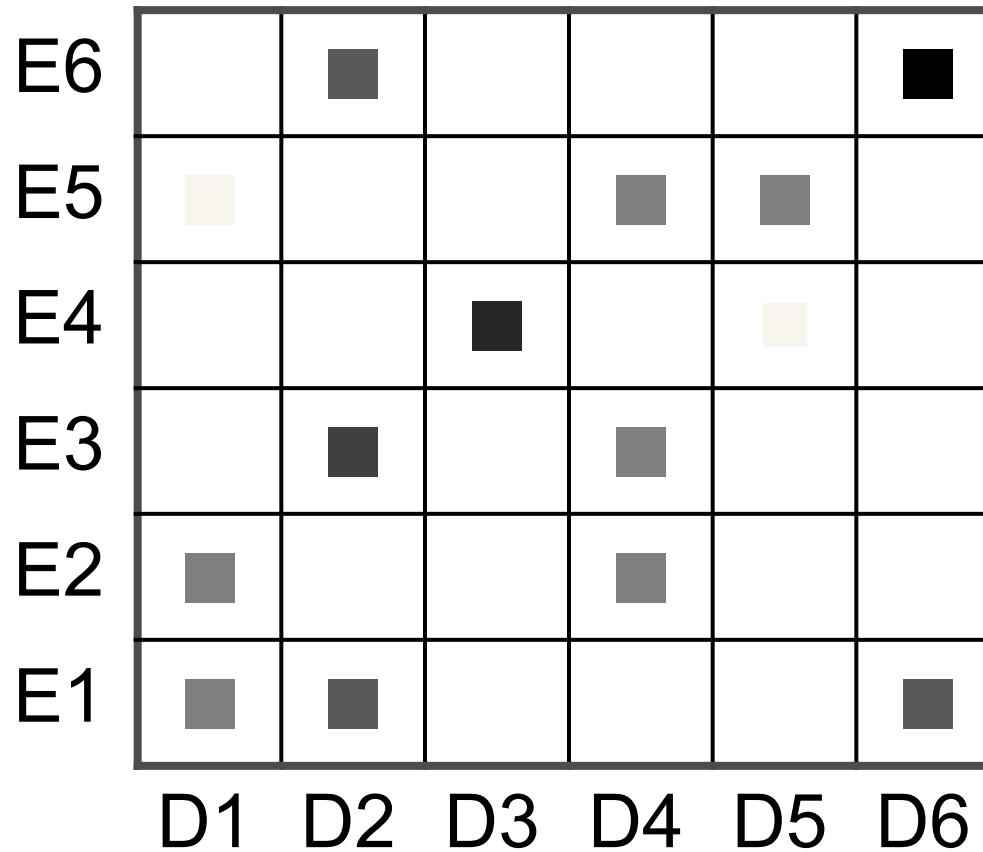
Sentence alignment for parallel corpora

Das stört mich keineswegs, ich halte das für eine gute Initiative, aber wiederum ist Europa nicht zur Stelle.	That is no problem for me. I think it is a good initiative, but again Europe is absent.
Es darf nicht wieder geschehen!	It should not happen again, Mr President.
Meine Fraktion verlangt, daß die italienische Präsidentschaft hier vor uns erklärt, welche Rolle sie spielt.	My Group wants the Italian presidency to come here and explain what its role is.
Herr Präsident, liebe Kolleginnen und Kollegen!	Mr President, ladies and gentlemen, I think it is important that we should discuss the situation in the Middle East this week.
Ich halte es für wichtig, daß wir diese Woche über die Situation im Nahen Osten reden.	We all agree on that.
Darin sind wir uns alle einig.	

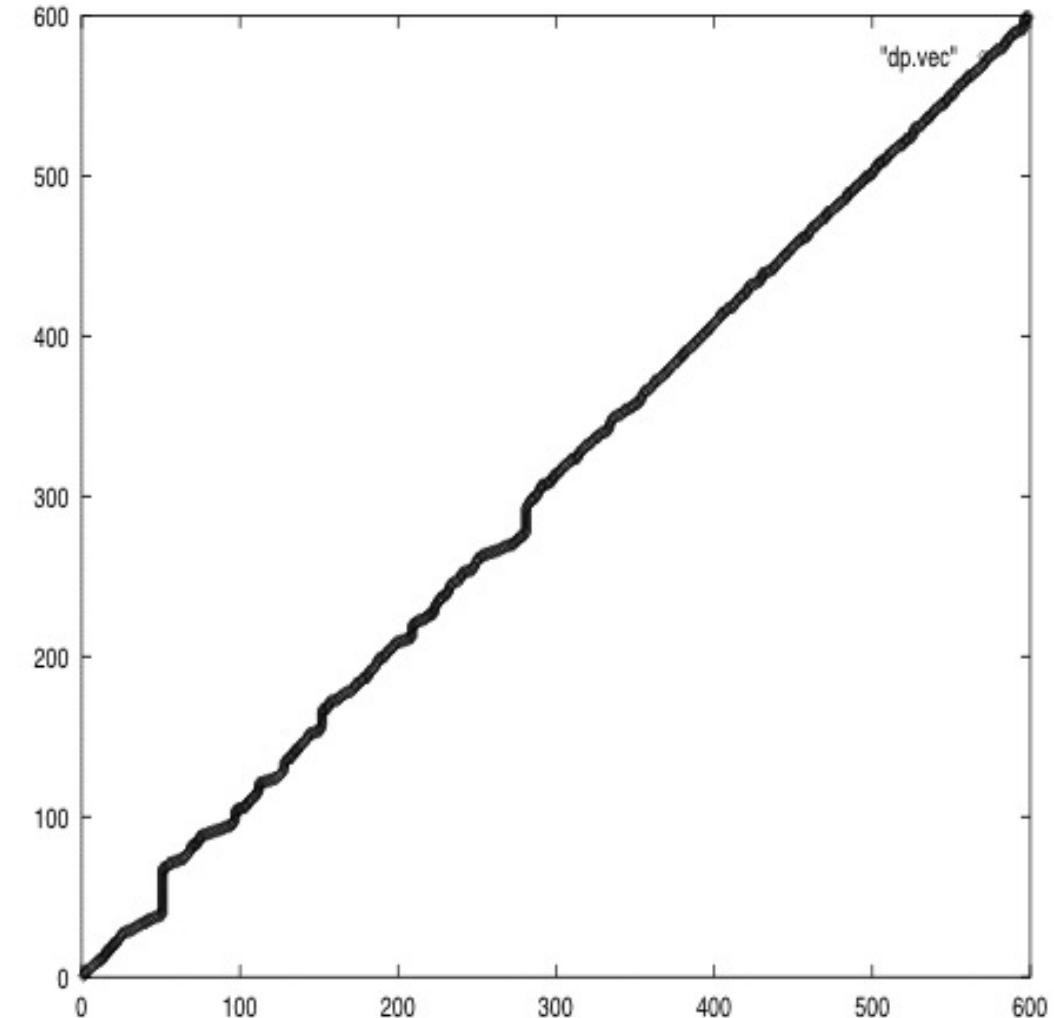
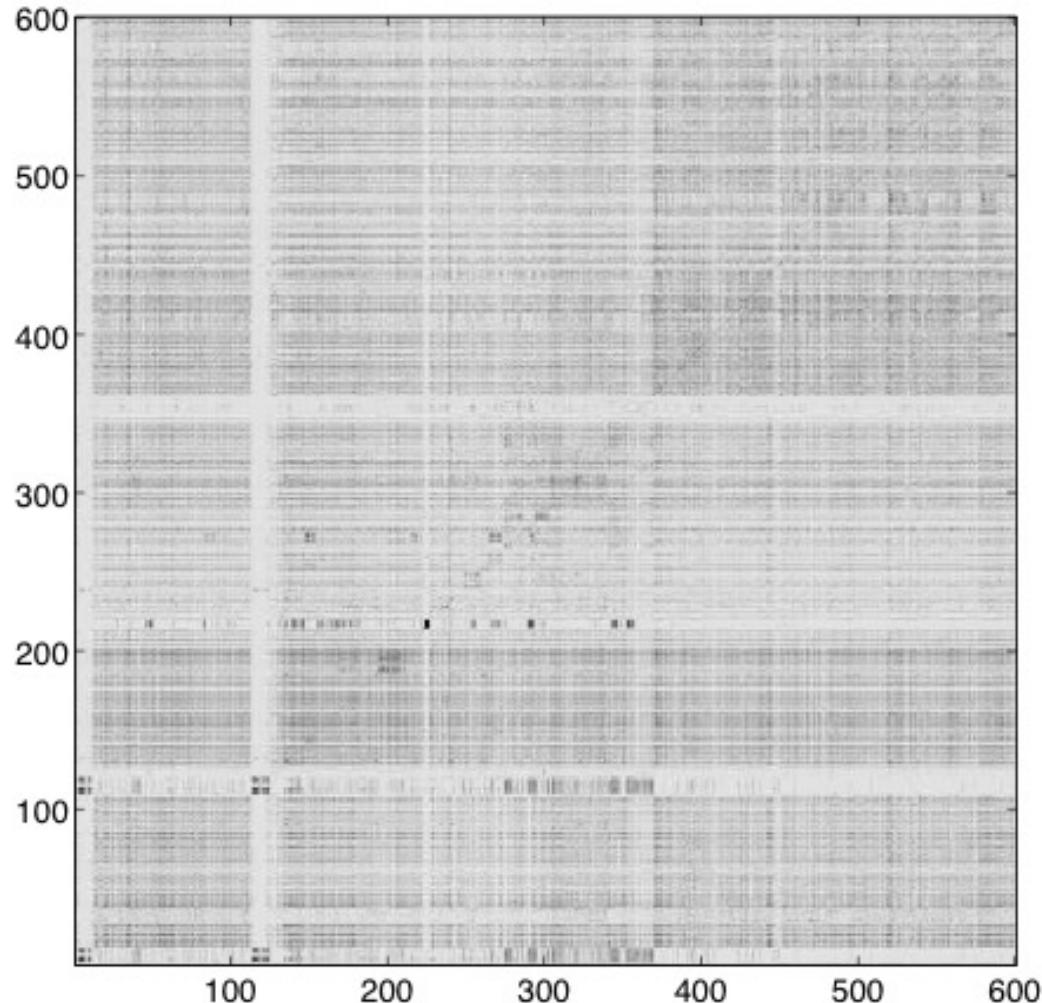
Sentence alignment: bitext map



Sentence alignment as similarity search

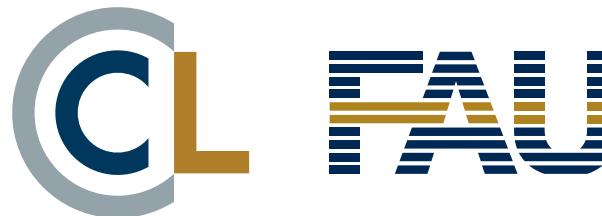


Sentence alignment as similarity search





Thank you for listening!



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE