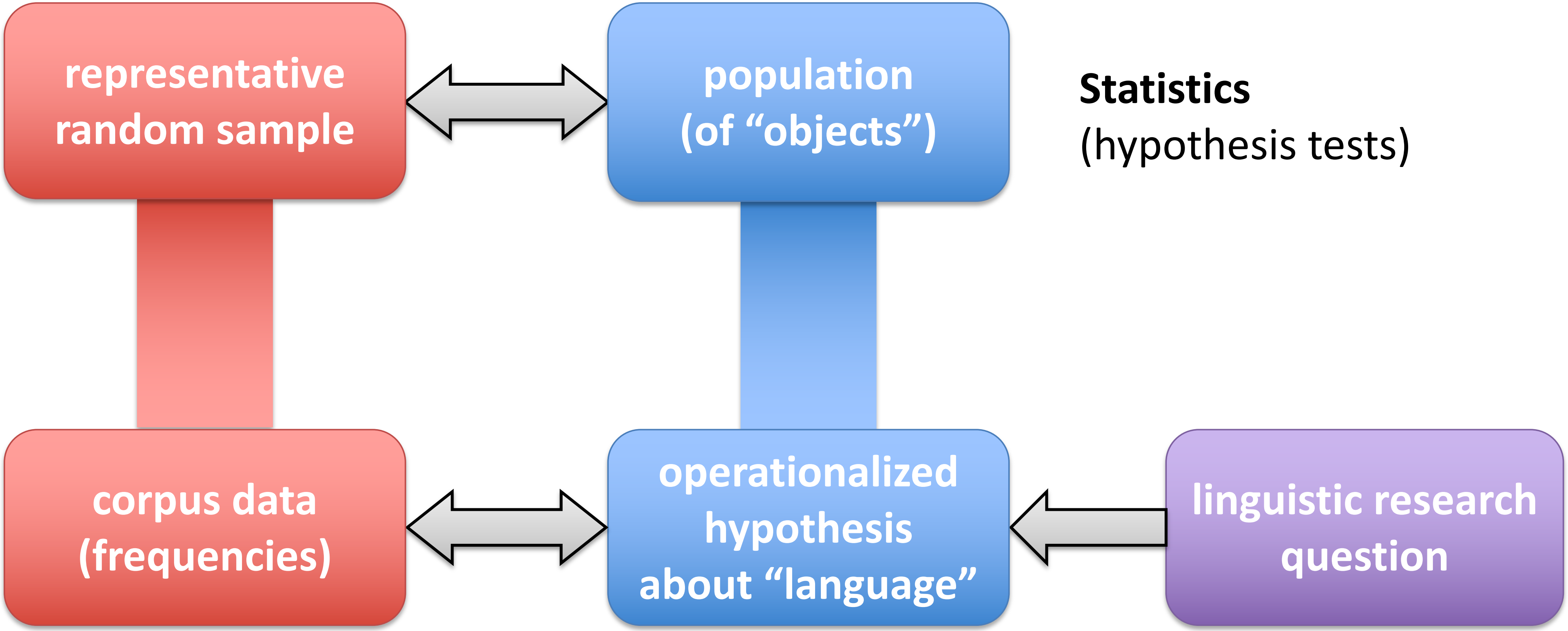
HS Corpus Linguistics / Korpuslinguistik

4. Corpus design & linguistic 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 representativeness?



representative random sample

population (of “objects”)

corpus data (frequencies)

operationalized hypothesis about “language”

linguistic research question

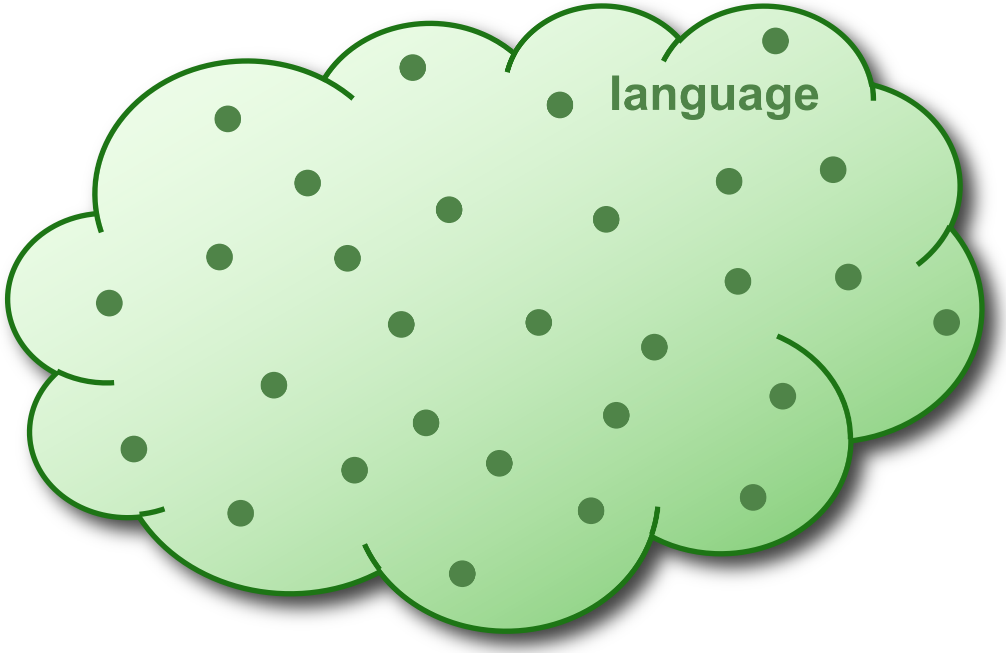
Statistics
(hypothesis tests)

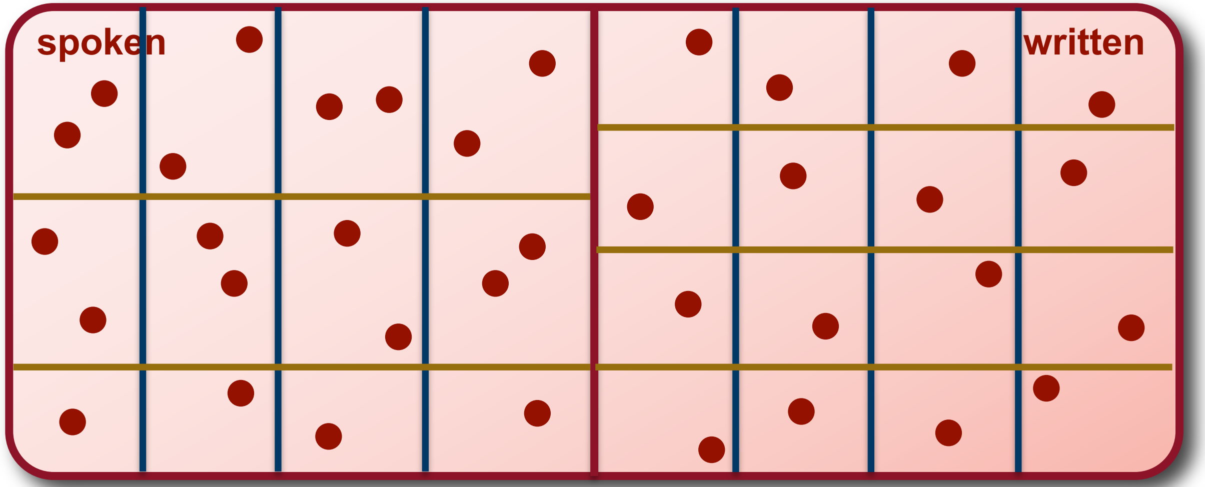
Goals of corpus design

- **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)





sampling frame

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 specifiy **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)

Brainstorming

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

… or another research question?

Assignment of presentation topics

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

Corpus annotation:
raw text + metadata

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

Corpus annotation: tokenization

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 *there*	
	FW	Foreign word	
	IN	Preposition / subordinating conjuction	
*	IN/that	Subordinating conjunction *that*	
	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 (mathemathical/scientific)	
	TO	*to* (any usage)	*fly to Paris, ready to go, …*
	UH	Interjection	
	#	Pound sign	*£*
	$	Dollar sign	*$*

	VB	Verb *be*, base form	
	VBD	Verb *be*, past tense	
	VBG	Verb *be*, gerund/progressive	
	VBN	Verb *be*, past participle	
	VBP	Verb *be*, non-3rd pers. sg. present	
	VBZ	Verb *be*, 3rd pers. sg. present tense	
*	VH	Verb *have*, base form	
*	VHD	Verb *have*, past tense	
*	VHG	Verb *have*, gerund/progressive	
*	VHN	Verb *have*, past participle	
*	VHP	Verb *have*, non-3rd pers. sg. present	
*	VHZ	Verb *have*, 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**	*schon, bald, doch*
APPR	**Präposition / Zirkumposition links**	
APPRART	**Präposition mit Artikel fusioniert**	*zum*
APPO	**Postposition**	*zufolge, wegen*
APZR	**Zirkumposition rechts**	*von … an*
ART	**bestimmter oder unbestimmter Artikel**	
CARD	**Kardinalzahlen** (Ordinalzahl = ADJA)	
FM	**Fremdsprachliches Material**	
ITJ	**Interjektion**	*mhm, ach, tja*
KOUI	**unterordnende Konj. mit *zu* + Inf**	
KOUS	**unterordnende Konjunktion mit Satz**	
KON	**nebenordnende Konjunktion**	*und, oder*
KOKOM	**Vergleichskonjunktion**	*als, wie*
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)**	
PPOSS	**substituierendes Possessivpronomen**	
PPOSAT	**attribuierendes Possessivpronomen**	
PRELS	**substituierendes Relativpronomen**	
PRELAT	**attribuierendes Relativpronomen**	

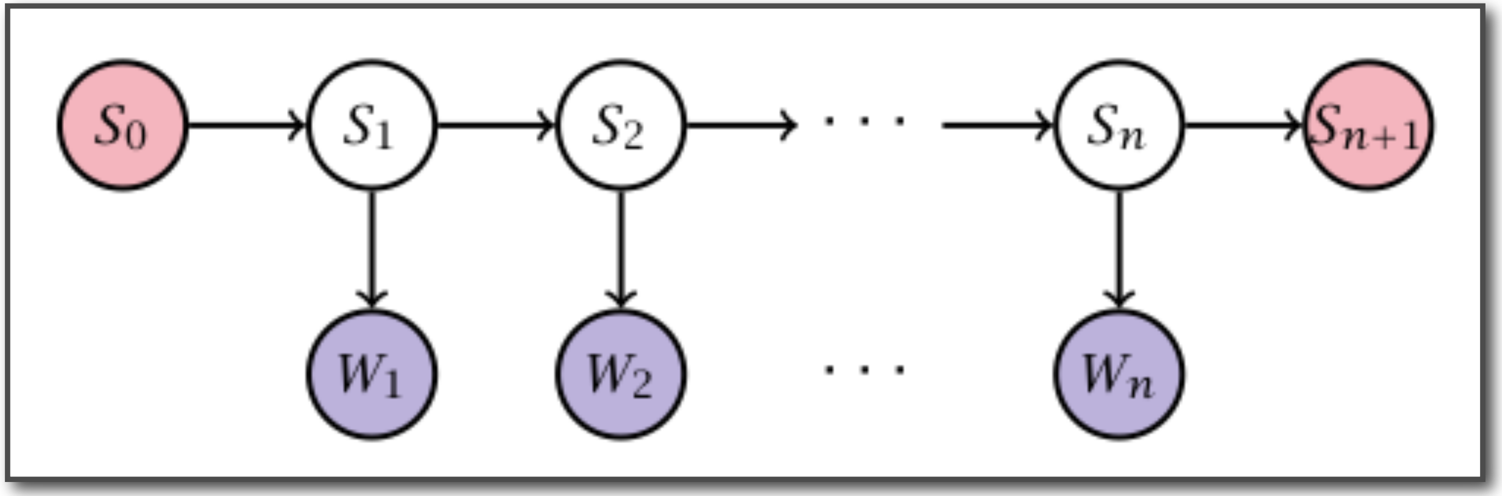
PRF	**reflexives Personalpronomen**	
PWS	**substituierendes Interrogativpron.**	
PWAT	**attribuierendes Interrogativpronomen**	
PWAV	**adverbiales Interrogativ-/Relativpron.**	
PAV	**Pronominaladverb**	*dafür, deswegen*
PTKZU	***zu* vor Infinitiv**	
PTKNEG	**Negationspartikel**	*nicht*
PTKVZ	**abgetrennter Verbzusatz**	*kommt … an*
PTKANT	**Antwortpartikel**	*ja, nein, danke*
PTKA	**Partikel bei Adjektiv/Adverb**	*am, zu*
TRUNC	**Kompositions-Erstglied**	*Unter- und …*
VVFIN	**finites Verb, voll (= lexikalisch)**	
VVIMP	**Imperativ, voll**	
VVINF	**Infinitiv, voll**	
VVIZU	**Infinitiv mit *zu*, 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**	*3:7, H2O*
$,	**Komma**	*,*
$.	**Satzbeendende Interpunktion**	*. ? ! ; :*
$(	**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$_{\text{PP}}$ seemed$_{\text{VBD}}$ a$_{\text{DT}}$ day$_{\text{NN}}$ much$_{\text{RB}}$ as$_{\text{IN}}$ any$_{\text{DT}}$ other$_{\text{JJ}}$ until$_{\text{IN}}$ I$_{\text{PP}}$ happened$_{\text{VBD}}$ to$_{\text{TO}}$ look$_{\text{VB}}$ out$_{\text{RP}}$ of$_{\text{IN}}$ the$_{\text{DT}}$ back$_{\text{JJ}}$ window$_{\text{NN}}$.$_{\text{SENT}}$ </s>

<s> There$_{\text{EX}}$ was$_{\text{VBD}}$ a$_{\text{DT}}$ little$_{\text{JJ}}$ garden$_{\text{NN}}$ behind$_{\text{IN}}$ the$_{\text{DT}}$ house$_{\text{NN}}$;$_{:}$ a$_{\text{DT}}$ well-mown$_{\text{VBN}}$ lawn$_{\text{NN}}$ surrounded$_{\text{VBN}}$ by$_{\text{IN}}$ a$_{\text{DT}}$ neatly$_{\text{RB}}$ cut$_{\text{VBN}}$ hedge$_{\text{NN}}$,$_{,}$ a$_{\text{DT}}$ few$_{\text{JJ}}$ bushes$_{\text{NNS}}$ and$_{\text{CC}}$ colourful$_{\text{JJ}}$ flowers$_{\text{NNS}}$.$_{\text{SENT}}$ </s>

Corpus annotation:
lemmatization

<s> It$_{\text{PP}}^{\text{it}}$ seemed$_{\text{VBD}}^{\text{seem}}$ a$_{\text{DT}}^{\text{a}}$ day$_{\text{NN}}^{\text{day}}$ much$_{\text{RB}}^{\text{much}}$ as$_{\text{IN}}^{\text{as}}$ any$_{\text{DT}}^{\text{any}}$ other$_{\text{JJ}}^{\text{other}}$ until$_{\text{IN}}^{\text{until}}$ I$_{\text{PP}}^{\text{I}}$ happened$_{\text{VBD}}^{\text{happen}}$ to$_{\text{TO}}^{\text{to}}$ look$_{\text{VB}}^{\text{look}}$ out$_{\text{RP}}^{\text{out}}$ of$_{\text{IN}}^{\text{of}}$ the$_{\text{DT}}^{\text{the}}$ back$_{\text{JJ}}^{\text{back}}$ window$_{\text{NN}}^{\text{window}}$.$_{\text{SENT}}^{.}$ </s>

<s> There$_{\text{EX}}^{\text{there}}$ was$_{\text{VBD}}^{\text{be}}$ a$_{\text{DT}}^{\text{a}}$ little$_{\text{JJ}}^{\text{little}}$ garden$_{\text{NN}}^{\text{garden}}$ behind$_{\text{IN}}^{\text{behind}}$ the$_{\text{DT}}^{\text{the}}$ house$_{\text{NN}}^{\text{house}}$;$_{:}^{;}$ a$_{\text{DT}}^{\text{a}}$ well-mown$_{\text{VBN}}^{\text{???}}$ lawn$_{\text{NN}}^{\text{lawn}}$ surrounded$_{\text{VBN}}^{\text{surround}}$ by$_{\text{IN}}^{\text{by}}$ a$_{\text{DT}}^{\text{a}}$ neatly$_{\text{RB}}^{\text{neatly}}$ cut$_{\text{VBN}}^{\text{cut}}$ hedge$_{\text{NN}}^{\text{hedge}}$,$_{,}^{,}$ a$_{\text{DT}}^{\text{a}}$ few$_{\text{JJ}}^{\text{few}}$ bushes$_{\text{NNS}}^{\text{bush}}$ and$_{\text{CC}}^{\text{and}}$ colourful$_{\text{JJ}}^{\text{colorful}}$ flowers$_{\text{NNS}}^{\text{flower}}$.$_{\text{SENT}}^{.}$ </s>

need better
representation format

XML markup of annotation

Standard for data interchange & archiving

root element

element annotated with attributes

start tag of XML element

```
<corpus>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        <token pos="RB"  lemma="much">much</token>
        <token pos="IN"  lemma="as">as</token>
        <token pos="DT"  lemma="any">any</token>
        <token pos="JJ"  lemma="other">other</token>
        <token pos="IN"  lemma="until">until</token>
        <token pos="PP"  lemma="I">I</token>
      </s>
    </p>
  </story>
</corpus>
```

corresponding end tag

XML markup of annotation
Standard for data interchange & archiving

XML declaration

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        <token pos="RB"  lemma="much">much</token>
        <token pos="IN"  lemma="as">as</token>
        <token pos="DT"  lemma="any">any</token>
        <token pos="JJ"  lemma="other">other</token>
        <token pos="IN"  lemma="until">until</token>
        <token pos="PP"  lemma="I">I</token>
        ...
      </s>
    </p>
  </story>
</corpus>
```

XML markup of annotation

Standard for data interchange & archiving

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
 <metadata>
    <author>
      <name>Stefan Evert</name>
      <sex>male</sex>
    </author>
    <publication>
      <title>Very Short Stories</title>
      <type>collection</type>
      <genre>fiction</genre>
    </publication>
  </metadata>
  <story num="6" title="The Garden">
    <p>
      <s>
        <token pos="PP"  lemma="it">It</token>
        <token pos="VBD" lemma="seem">seemed</token>
        <token pos="DT"  lemma="a">a</token>
        <token pos="NN"  lemma="day">day</token>
        ...
```

metadata header

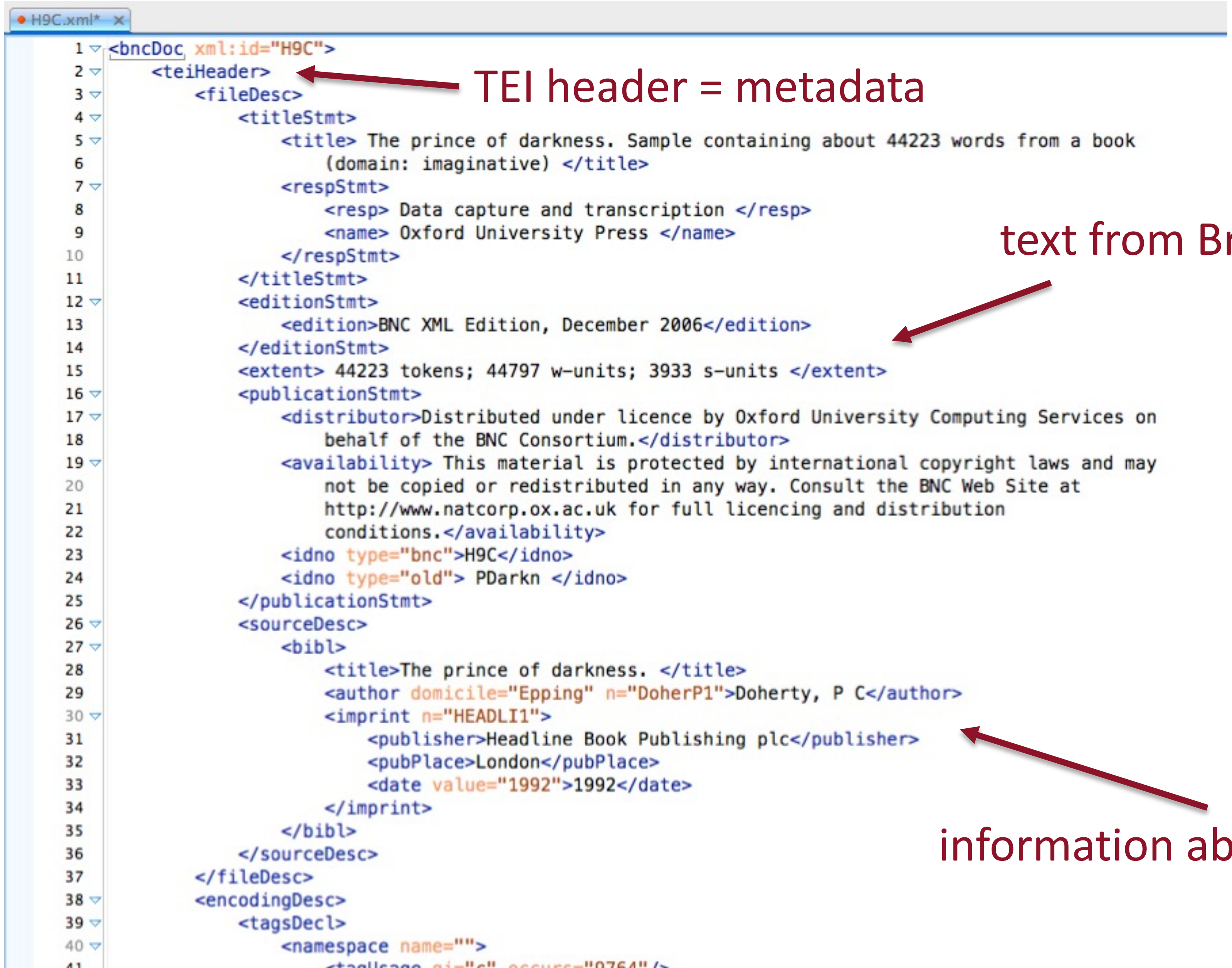
XML standards

- XML (Extensible Markup Language) is a widely-used standard for structured annotation
- A well-formed XML document only specifies the structure of annotation, not its semantics
- DTD (document type declaration) or XML Schema specify valid element & attribute names
  - still doesn't explain semantics without documentation!

- Exchange formats for text corpora:
  TEI (Text Encoding Initiative), XCES (Corpus Encoding Standard),
  ISO 24612: LAF (Linguistic Annotation Framework)
  - but more efficient representation requiredd for corpus search etc.

TEI standard (BNC)

TEI header = metadata

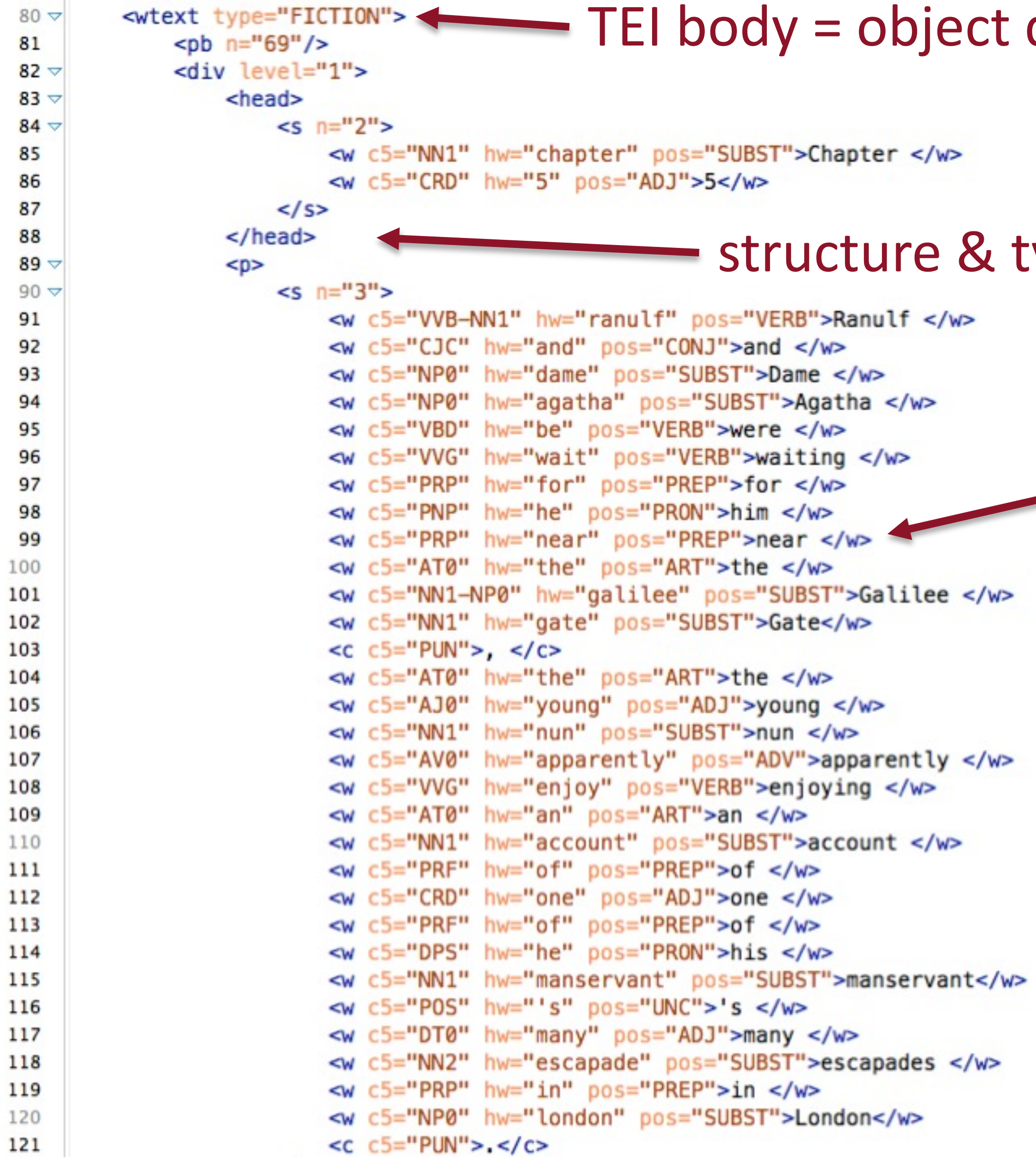


```
<bncDoc xml:id="H9C">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title> The prince of darkness. Sample containing about 44223 words from a book
                    (domain: imaginative) </title>
                <respStmt>
                    <resp> Data capture and transcription </resp>
                    <name> Oxford University Press </name>
                </respStmt>
            </titleStmt>
            <editionStmt>
                <edition>BNC XML Edition, December 2006</edition>
            </editionStmt>
            <extent> 44223 tokens; 44797 w-units; 3933 s-units </extent>
            <publicationStmt>
                <distributor>Distributed under licence by Oxford University Computing Services on
                    behalf of the BNC Consortium.</distributor>
                <availability> This material is protected by international copyright laws and may
                    not be copied or redistributed in any way. Consult the BNC Web Site at
                    http://www.natcorp.ox.ac.uk for full licencing and distribution
                    conditions.</availability>
                <idno type="bnc">H9C</idno>
                <idno type="old"> PDarkn </idno>
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <title>The prince of darkness. </title>
                    <author domicile="Epping" n="DoherP1">Doherty, P C</author>
                    <imprint n="HEADLI1">
                        <publisher>Headline Book Publishing plc</publisher>
                        <pubPlace>London</pubPlace>
                        <date value="1992">1992</date>
                    </imprint>
                </bibl>
            </sourceDesc>
        </fileDesc>
        <encodingDesc>
            <tagsDecl>
                <namespace name="">
                    <tagUsage gi="s" occurs="9764"/>
```

text from British National Corpus

information about this text

TEI standard (BNC)



```
<wtext type="FICTION">
  <pb n="69"/>
  <div level="1">
    <head>
      <s n="2">
        <w c5="NN1" hw="chapter" pos="SUBST">Chapter </w>
        <w c5="CRD" hw="5" pos="ADJ">5</w>
      </s>
    </head>
    <p>
      <s n="3">
        <w c5="VVB-NN1" hw="ranulf" pos="VERB">Ranulf </w>
        <w c5="CJC" hw="and" pos="CONJ">and </w>
        <w c5="NP0" hw="dame" pos="SUBST">Dame </w>
        <w c5="NP0" hw="agatha" pos="SUBST">Agatha </w>
        <w c5="VBD" hw="be" pos="VERB">were </w>
        <w c5="VVG" hw="wait" pos="VERB">waiting </w>
        <w c5="PRP" hw="for" pos="PREP">for </w>
        <w c5="PNP" hw="he" pos="PRON">him </w>
        <w c5="PRP" hw="near" pos="PREP">near </w>
        <w c5="AT0" hw="the" pos="ART">the </w>
        <w c5="NN1-NP0" hw="galilee" pos="SUBST">Galilee </w>
        <w c5="NN1" hw="gate" pos="SUBST">Gate</w>
        <c c5="PUN">, </c>
        <w c5="AT0" hw="the" pos="ART">the </w>
        <w c5="AJ0" hw="young" pos="ADJ">young </w>
        <w c5="NN1" hw="nun" pos="SUBST">nun </w>
        <w c5="AV0" hw="apparently" pos="ADV">apparently </w>
        <w c5="VVG" hw="enjoy" pos="VERB">enjoying </w>
        <w c5="AT0" hw="an" pos="ART">an </w>
        <w c5="NN1" hw="account" pos="SUBST">account </w>
        <w c5="PRF" hw="of" pos="PREP">of </w>
        <w c5="CRD" hw="one" pos="ADJ">one </w>
        <w c5="PRF" hw="of" pos="PREP">of </w>
        <w c5="DPS" hw="he" pos="PRON">his </w>
        <w c5="NN1" hw="manservant" pos="SUBST">manservant</w>
        <w c5="POS" hw="'s" pos="UNC">'s </w>
        <w c5="DT0" hw="many" pos="ADJ">many </w>
        <w c5="NN2" hw="escapade" pos="SUBST">escapades </w>
        <w c5="PRP" hw="in" pos="PREP">in </w>
        <w c5="NP0" hw="london" pos="SUBST">London</w>
        <c c5="PUN">.</c>
```

TEI body = object data + annotation

structure & typographic markup

tokens + token-level annotations

principle:
raw text (= object data) can be reconstructed by deleting all XML tags

Vertical text format (.vrt)

Simpler, more efficient format → used by CWB & NLP tools

```
<corpus>
<story title="The Garden">
<p>
<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
...
</s>
</p>
</story>
</corpus>
```

TAB characters (\t, \x09)

metadata

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

Vertical text format (.vrt)

Text metadata encoded in XML start tags (not in header!)

```
<corpus>
<text title="The Garden" author="Stefan Evert" author_sex="male"
      date="1991-08-05">
<p num="1">
<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
...
</s>
</p>
</text>
</corpus>
```

CQPweb requires <text>,
SketchEngine prefers <doc>

sub-text level metadata

CoNLL format

Vertical text without the metadata

http://universaldependencies.org/docs/format.html

```
# story: "The Garden"
# paragraph #1
1   It      PP   it
2   seemed  VBD  seem
3   a       DT   a
4   fine    JJ   fine
5   day     NN   day
6   .       SENT .

1   There   EX   there
2   was     VBD  be
3   an      DT   a
4   elephantNN   elephant
5   .       SENT .

# this is the end of the file
```

these are just comments

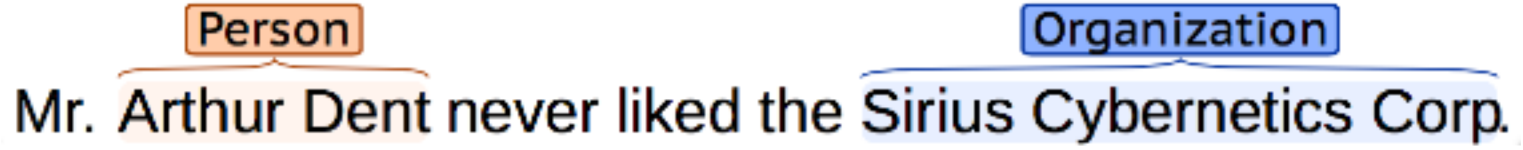
blank lines = sentence boundaries

token numbers (within sentence)

Corpus annotation:
segments and structures

- Automatic recognition and categorization of particular word sequences (segments)

- e.g. named entities (NER = named entity recognition)

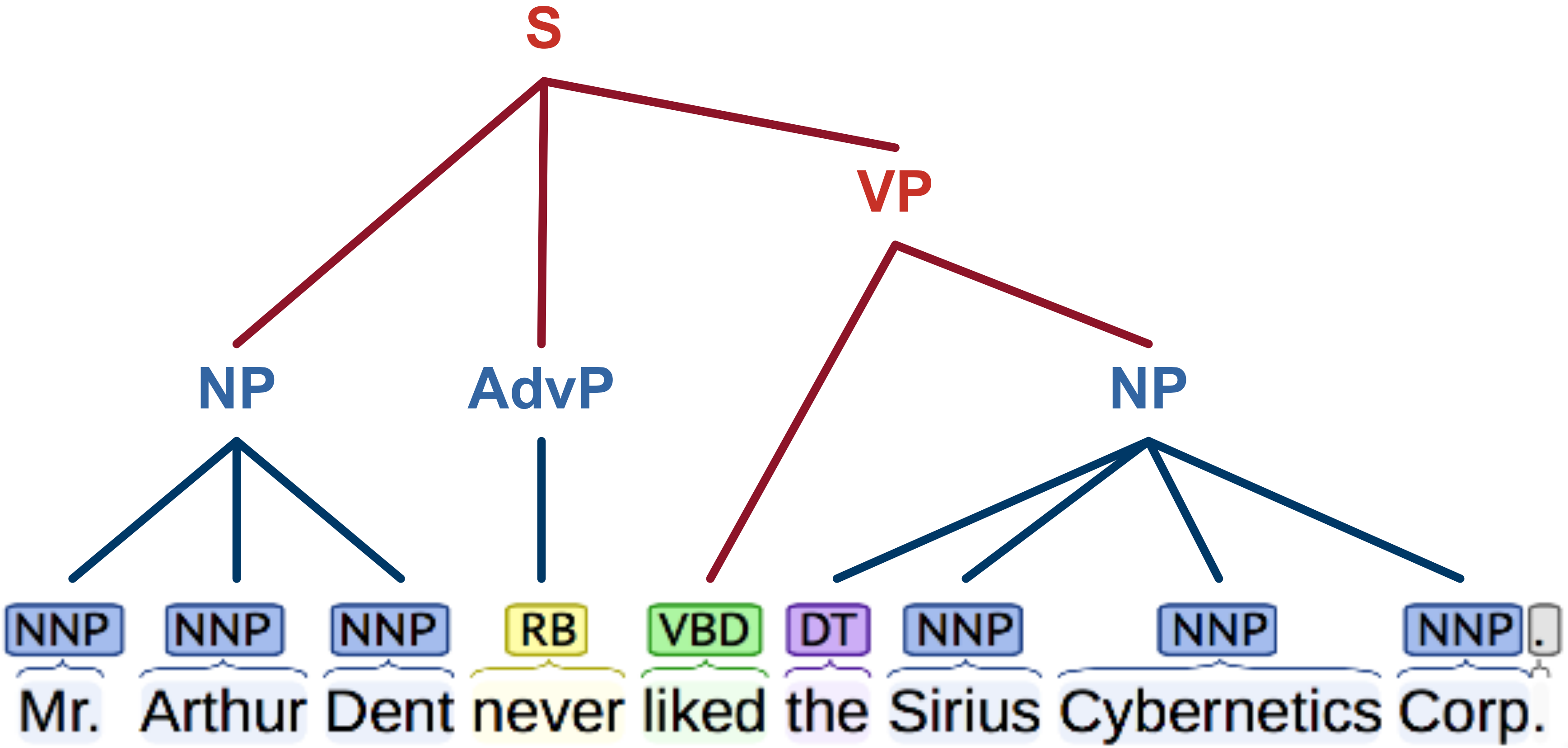


Mr. Arthur Dent never liked the Sirius Cybernetics Corp.

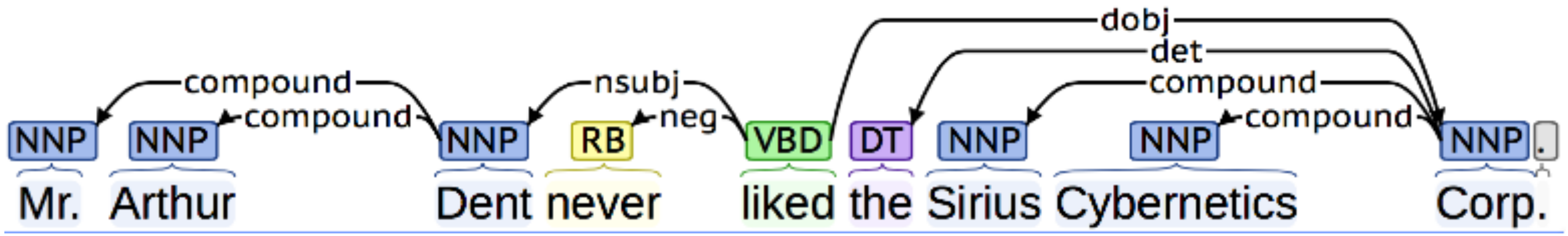
- e.g. time and place expressions: last week, the day after tomorrow, September 15th, in Paris, on the lawn in front of their house, …
- e.g. text spans that need to be masked for anonymization purposes

Corpus annotation:
segments and structures

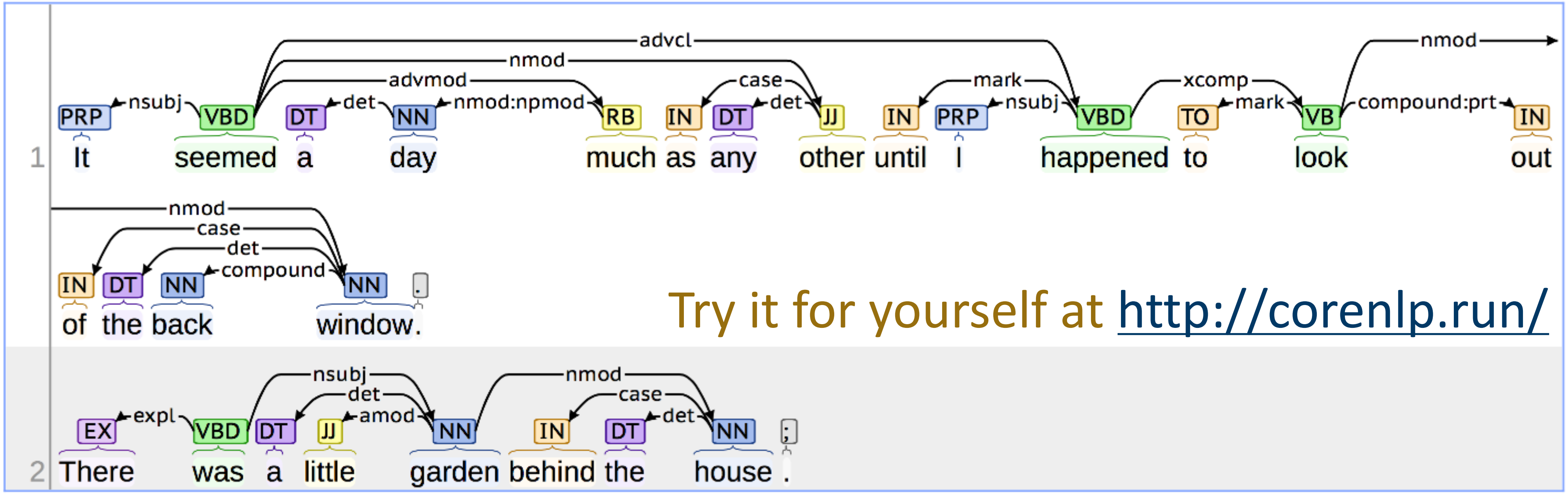
- Syntactic phrase structure analysis
  = parse tree of nested segments corresponding to syntactic units
- „minimal“ phrases as flat segments ➞ chunk parsing



Corpus annotation: syntactic dependency analysis



Mr. Arthur Dent never liked the Sirius Cybernetics Corp.
NNP NNP NNP RB VBD DT NNP NNP NNP .
compound
compound
nsubj
neg
dobj
det
compound
compound



1 It seemed a day much as any other until I happened to look out of the back window.

2 There was a garden behind the house .

Try it for yourself at http://corenlp.run/

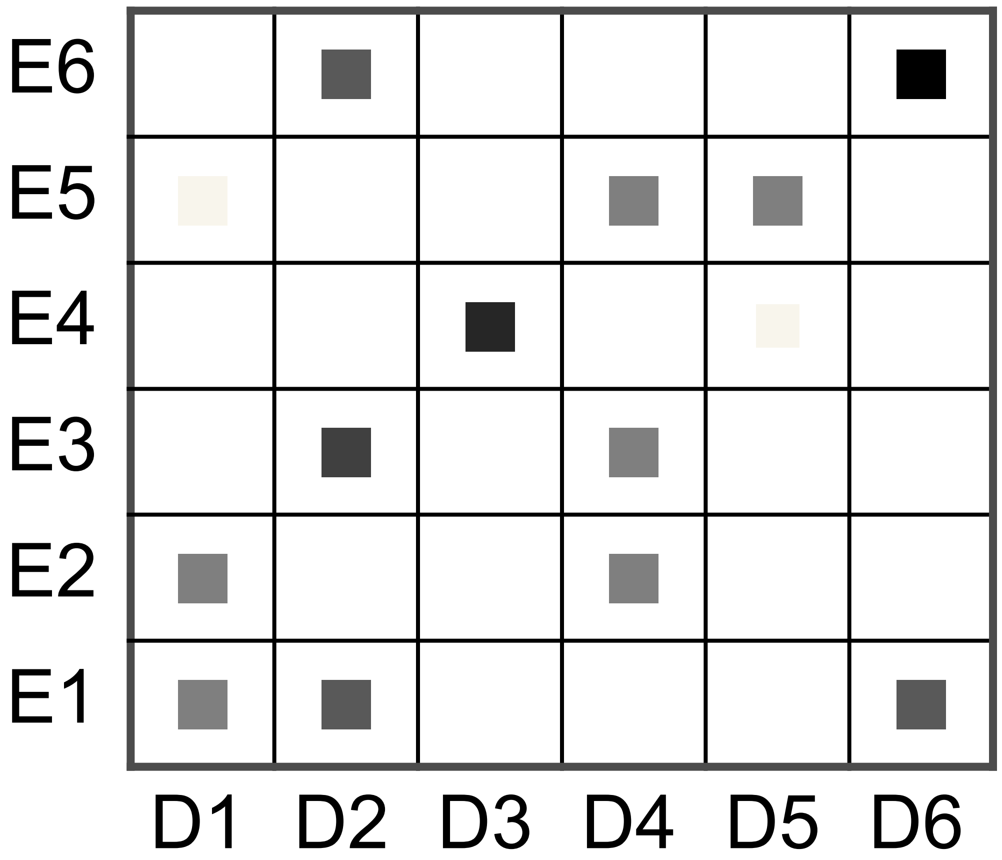
Sentence alignment for parallel corpora

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.	
Darin sind wir uns alle einig.	We all agree on that.

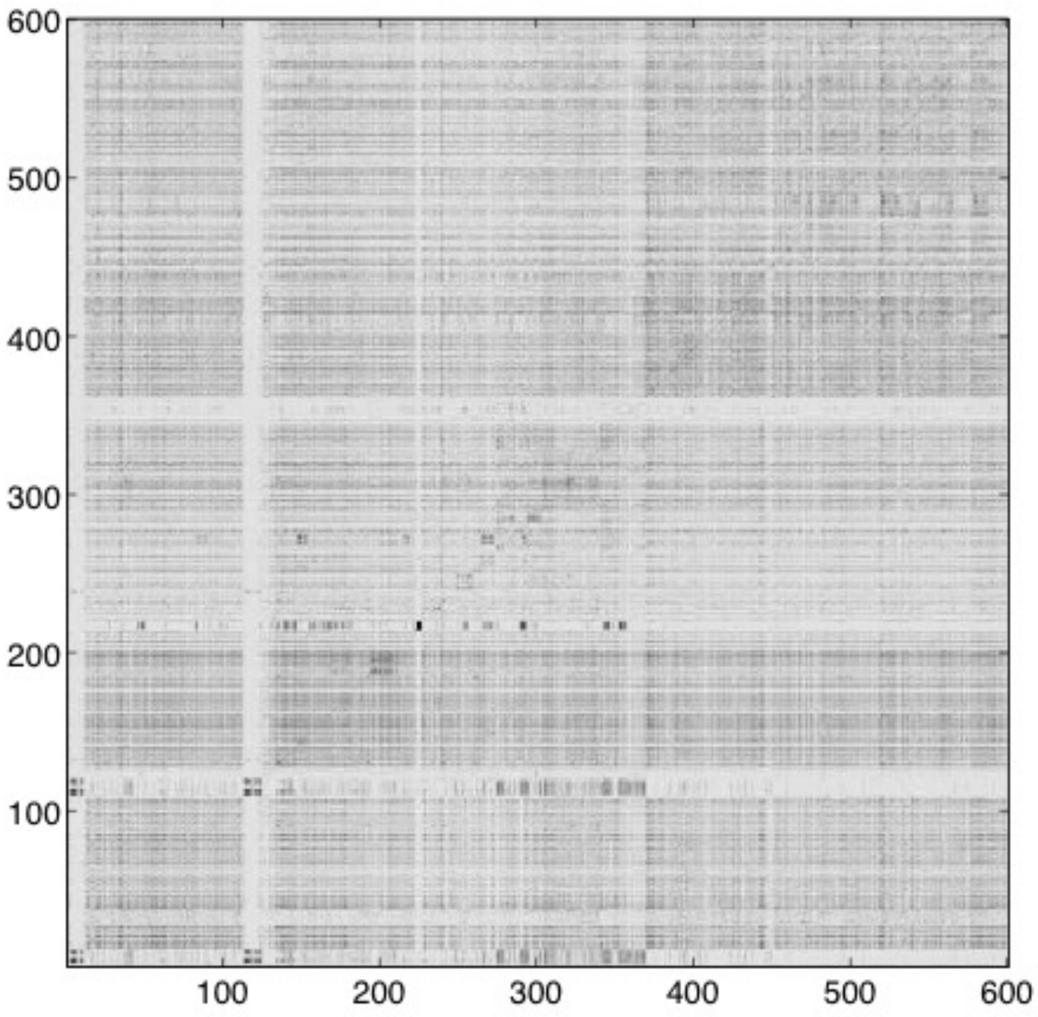
Sentence alignment: bitext map

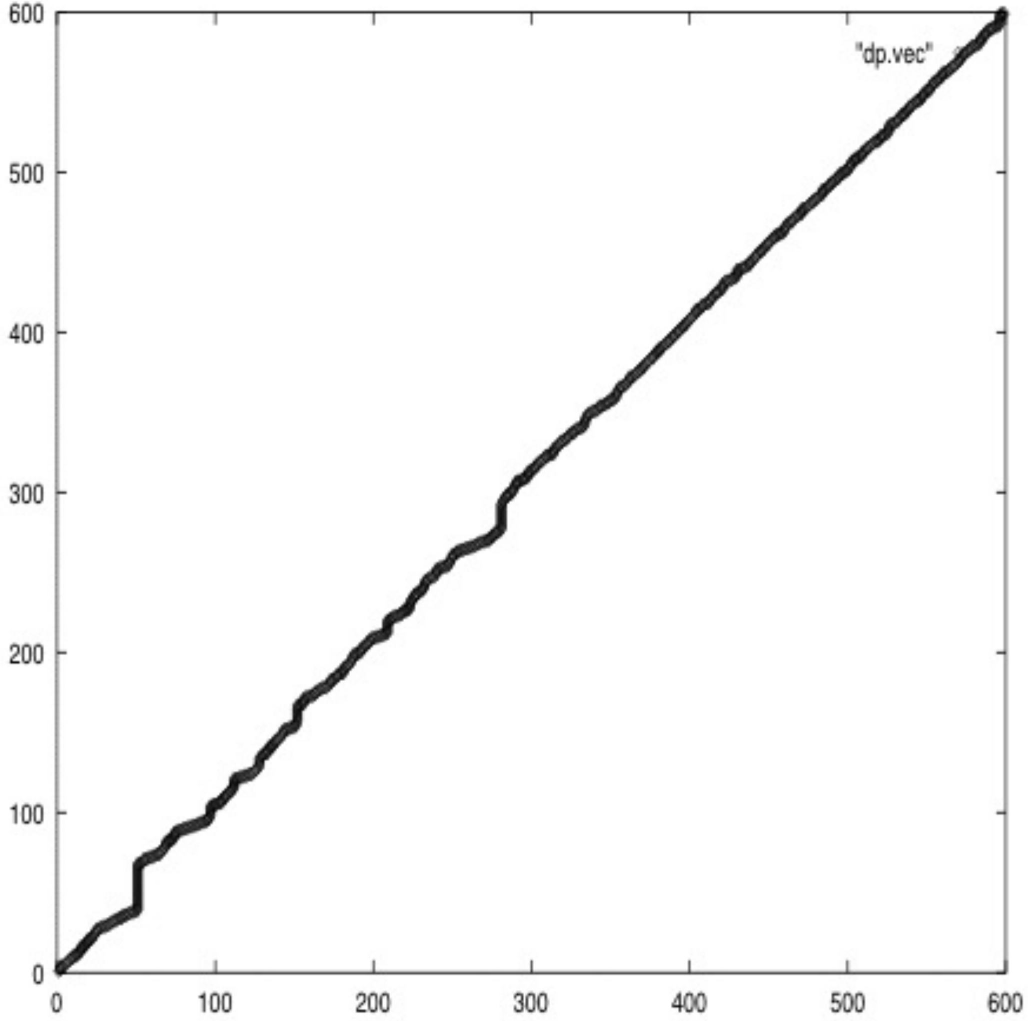
E6						■
E5				■	■	
E4			■			
E3		■				
E2	■					
E1	■					
	D1	D2	D3	D4	D5	D6

Sentence alignment as similarity search



Sentence alignment as similarity search


600
500
400
300
200
100
100
200
300
400
500
600


"dp.vec"