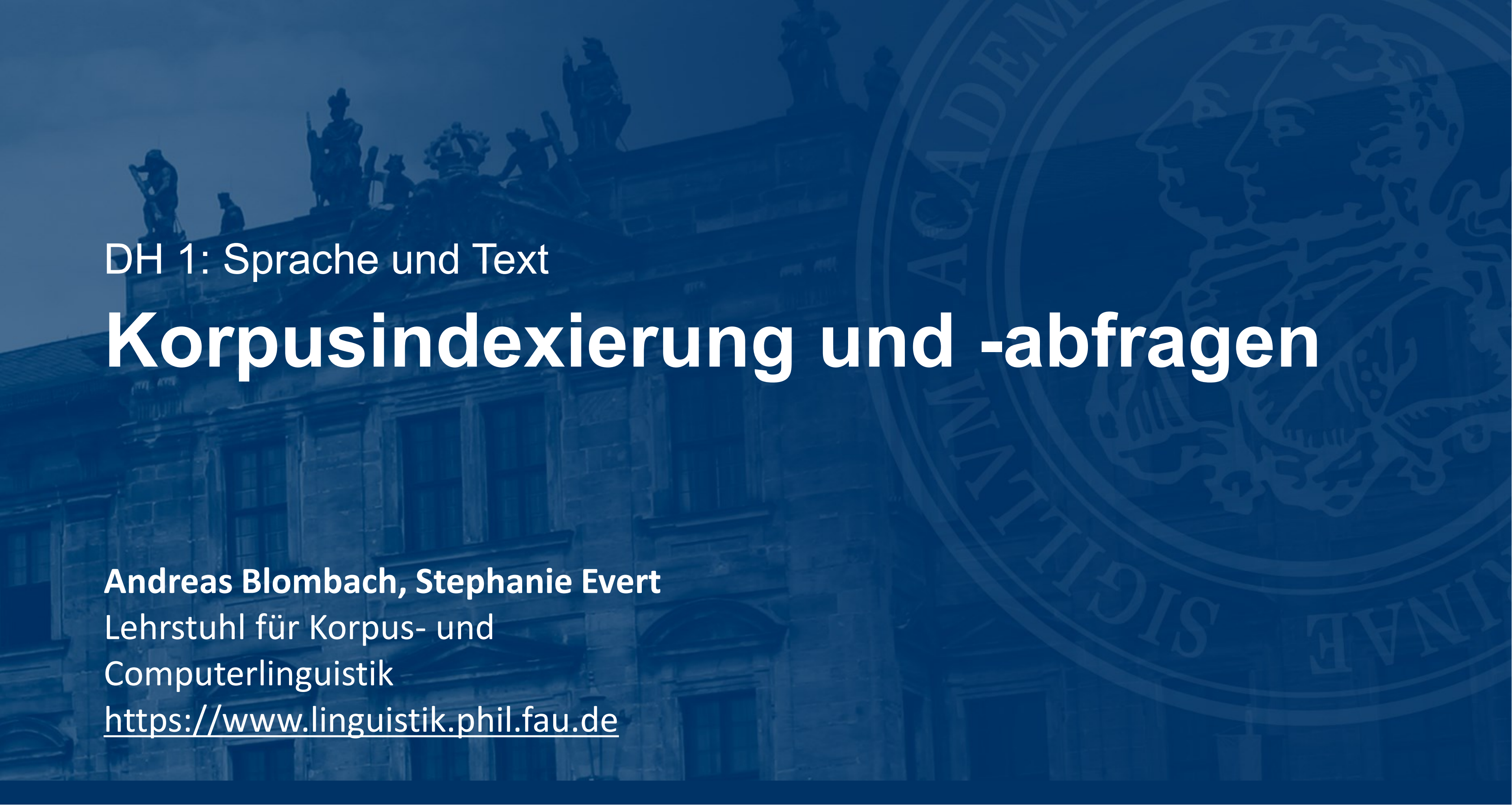


DH 1: Sprache und Text

# Korpusindexierung und -abfragen

**Andreas Blombach, Stephanie Evert**
Lehrstuhl für Korpus- und
Computerlinguistik
https://www.linguistik.phil.fau.de

Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

# Repräsentationsformate & Indexierung

Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

# Korpusannotation: Rohtext + Metadaten

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.

### Metadaten

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

# Korpusannotation: Tokenisierung

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 .

# Korpusannotation: Satzsegmentierung

<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>

# Korpusannotation: POS-Tagging (Wortartenannotation)

<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>

# Korpusannotation: Lemmatisierung

<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$_{\text{NN}}^{\text{window}}$ .$_{\text{SENT}}^{\text{.}}$ </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$_{\text{NN}}^{\text{house}}$ ;$_{\text{:}}^{\text{;}}$ a$_{\text{DT}}^{\text{a}}$ well-mown$_{\text{VBN}}^{\text{well-mown}}$ 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$_{\text{NN}}^{\text{hedge}}$ ,$_{\text{,}}^{\text{,}}$ 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$_{\text{NNS}}^{\text{flower}}$ .$_{\text{SENT}}^{\text{.}}$ </s>

# XML-Markup der Annotation

## Standard für Datenaustausch und -archivierung

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <story 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-Deklaration

# XML-Markup der Annotation

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <story 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>
```

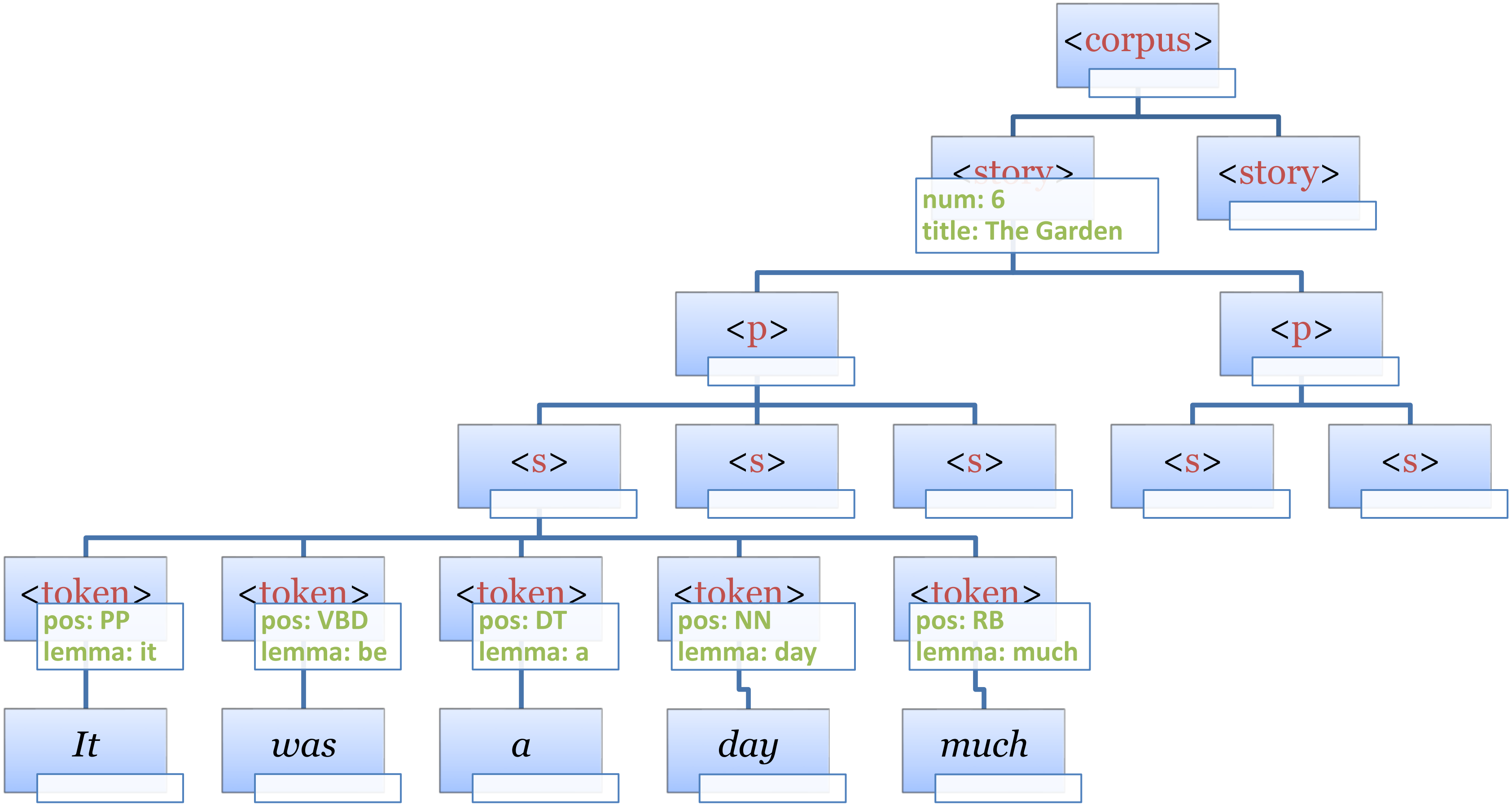
Wurzelelement / root element

Start-Tag des XML-Elements

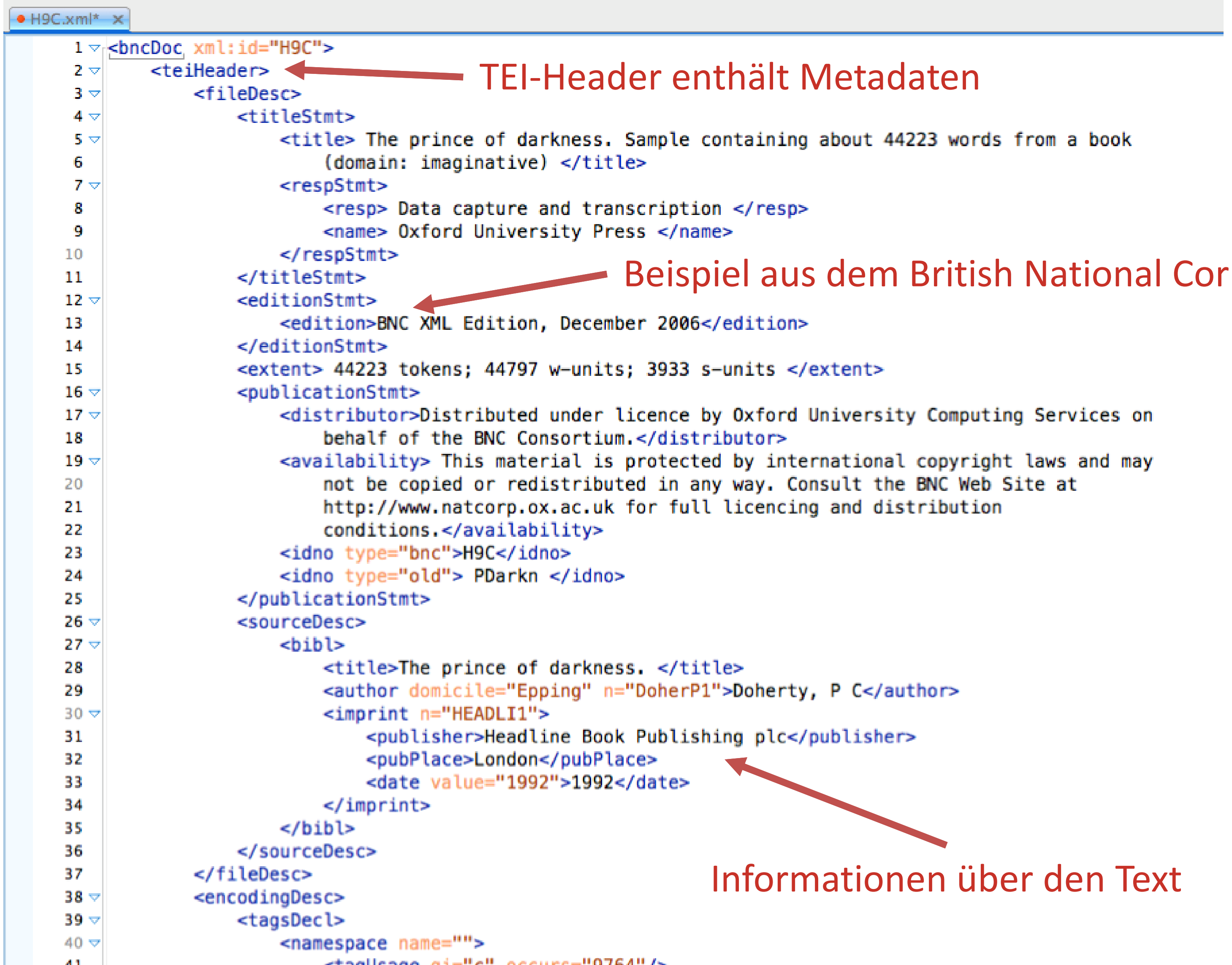
Element mit Attributen

korrespondierendes End-Tag

# XML-Dokument = geordneter Baum



# XML als Repräsentationsformat



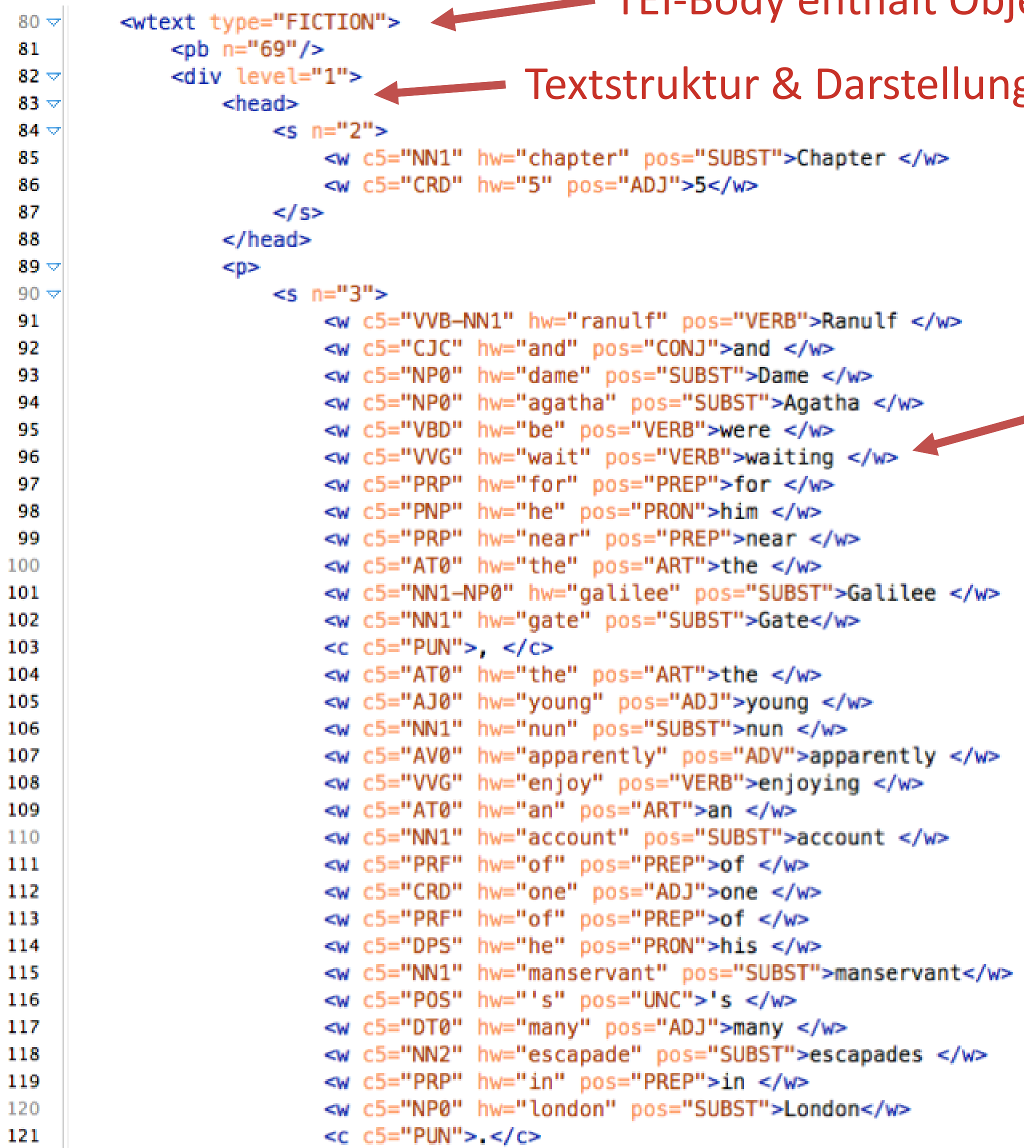
```
<bncDoc xml:id="H9C">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title> The prince of darkness. Sample containing about 44223 words from a book
                    (domain: imaginative) </title>
                <respStmt>
                    <resp> Data capture and 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="">
```

TEI-Header enthält Metadaten

Beispiel aus dem British National Corpus

Informationen über den Text

# XML als Repräsentationsformat



```
<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 enthält Objektdaten

Textstruktur & Darstellung

Token mit Annotationen

**XML-Prinzip:**
Durch Entfernen aller XML-Tags kann der ursprüngliche Objekttext wiederhergestellt werden

# XML-Standards

- XML (*Extensible Markup Language*) ist ein weitverbreiteter Standard für hierarchisch strukturierte Annotationen
- ein wohlgeformtes XML-Document legt nur die Struktur der Annotation/Auszeichnung fest, nicht die Semantik (also was was bedeutet)
- eine DTD (*document type declaration*) oder ein XML-Schema legt gültige Element- und Attributnamen fest
  - … erklärt aber ohne Dokumentation noch immer nicht die Semantik!
- Austauschformate für Textkorpora:
  TEI (*Text Encoding Initiative*), XCES (*Corpus Encoding Standard*),
  ISO 24612: LAF (Linguistic Annotation Framework)
  - ideal für Archivierung und Interoperabilität
  - für Korpusabfragen u.a. wird aber effizientere Implementierung benötigt

# CWB

- IMS Open Corpus Workbench
  - ursprünglich 1993–1996 entwickelt (IMS Stuttgart)
  - Anwendungen: statistische Sprachverarbeitung, Lexikographie, Korpuslinguistik
  - Open-Source-Veröffentlichung 2005 (GPL)
  - Aktuelle Version: CWB 3.5 (UTF-8, Korpora bis zu 2 Milliarden Wörter)
- Standardoberfläche: CQPweb
  - diverse simplere und/oder spezialisiertere Web-Oberflächen verfügbar
  - Kommandozeilen-CQP für erfahrene Benutzer (Uli Heid)
- SketchEngine: gleiches Datenmodell, gleiche Abfragesyntax
  - aber unterschiedliche Implementierung

http://cwb.sf.net/

CWB
The IMS Open Corpus Workbench

# Vertical text format (.vrt)

## Einfacheres, effizienteres Format → wird von CWB & NLP-Tools verwendet

```
<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>
```

Tabulatorzeichen (\t, \x09)

**metadata**

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

# Vertical text format (.vrt)

## Textmetadaten in den XML-Start-Tags (nicht im 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(e)

## Vertical Text ohne Metadaten (weit verbreitet in NLP)

```
# 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    elephant NN   elephant
5    .        SENT .

# this is the end of the file
```

Metainformation in Kommentaren

Leerzeile = Satzgrenze

Token durchnummeriert (je Satz)

# CWB: Datenmodell

| # | word | pos | lemma |
|---|---|---|---|
| 0 | A | DET | a |
| 1 | fine | ADJ | fine |
| 2 | example | NN | example |
| 3 | . | PUN | . |
| | | | |
| 4 | Very | ADV | very |
| 5 | fine | ADJ | fine |
| 6 | examples | NN | example |
| 7 | ! | PUN | ! |

implizite Nummerierung der Token:
Korpusposition (**cpos**)

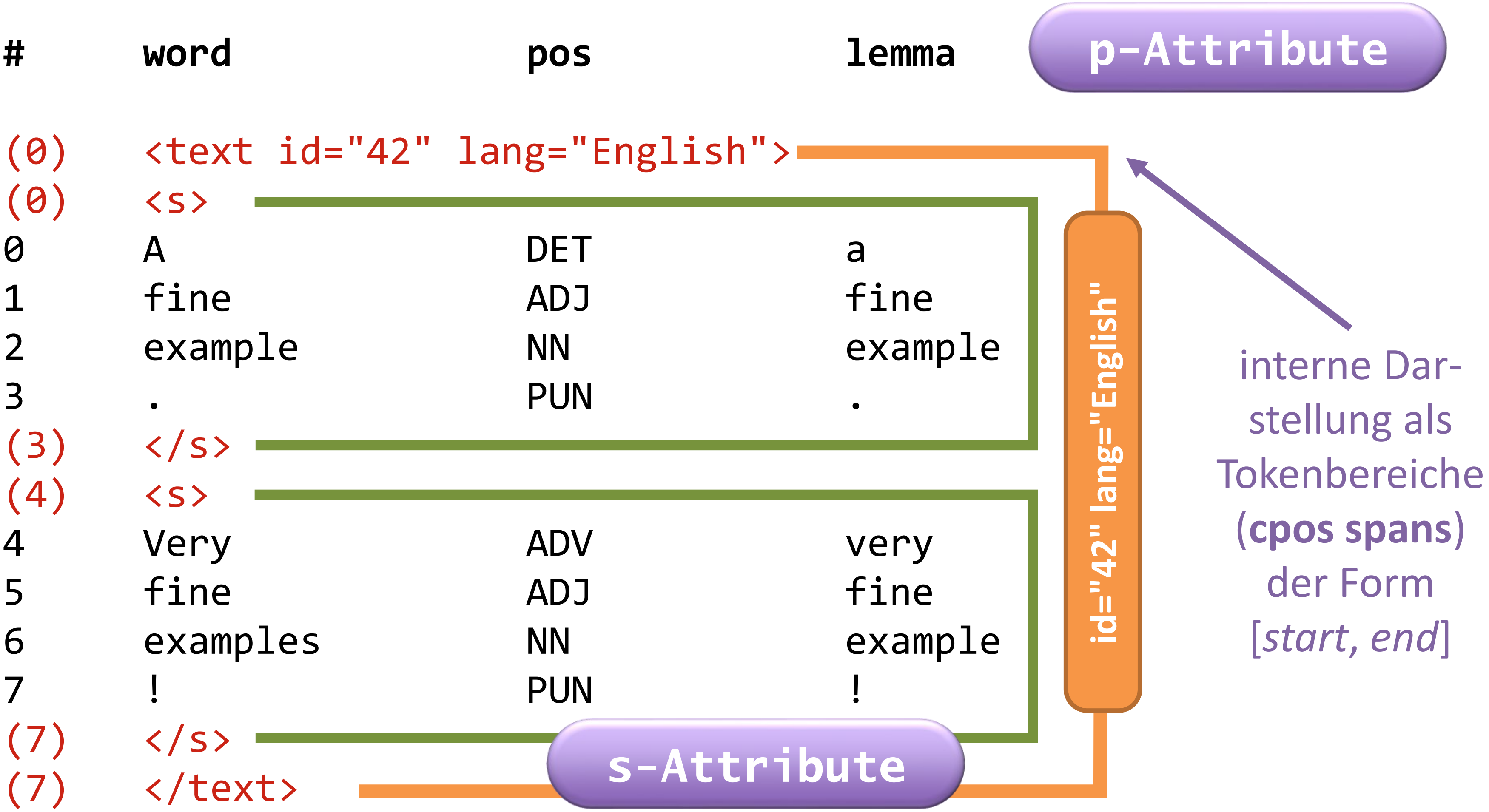
# CWB: Datenmodell

```
#      word                pos            lemma

(0)    <text id="42" lang="English">
(0)    <s>
0      A                   DET            a
1      fine                ADJ            fine
2      example             NN             example
3      .                   PUN            .
(3)    </s>
(4)    <s>
4      Very                ADV            very
5      fine                ADJ            fine
6      examples            NN             example
7      !                   PUN            !
(7)    </s>
(7)    </text>
```

XML-Tags als „unsichtbare“ Token

# CWB: Datenmodell



```
#      word               pos           lemma

(0)    <text id="42" lang="English">
(0)    <s>
0      A                  DET           a
1      fine               ADJ           fine
2      example            NN            example
3      .                  PUN           .
(3)    </s>
(4)    <s>
4      Very               ADV           very
5      fine               ADJ           fine
6      examples           NN            example
7      !                  PUN           !
(7)    </s>
(7)    </text>
```

interne Darstellung als Tokenbereiche (**cpos spans**) der Form [*start*, *end*]

# Korpusabfragen mit CQP

Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

# Zum Ausprobieren: CQPweb & BNCweb

- **CQPweb** (flexible Web-Oberfläche für CWB)
  https://corpora.linguistik.uni-erlangen.de/cqpweb/
  - Login: `studentX` (1 … 15)
  - Passwort: `erlangen`
  - Dokumentation: Tutorial-Videos auf YouTube
    https://www.youtube.com/user/CorpusWorkbench

- **BNCweb** (speziell für British National Corpus)
  https://corpora.linguistik.uni-erlangen.de/bncweb/
  - gleicher Gäste-Login wie für CQPweb

# Simple query syntax

- viele Web-Oberflächen unterstützen eine „einfache“ Abfragesyntax
  - Wort oder Wortfolge einfach direkt eingeben
  - Platzhalter (*wildcards*) für variable/optionale Elemente (eingeschränkt)
- hier: CEQL-Syntax (*Common Elementary Query Language*)
  - relativ mächtige einfache Abfragesprache
  - unterstützt von BNCweb, CQPweb und einigen anderen Web-Oberflächen
- Tutorial & Dokumentation
  - **Kap. 6** aus Hoffmann, Sebastian *et al.* (2008): *Corpus Linguistics with BNCweb – a Practical Guide*. Frankfurt a.M. [etc.]: Peter Lang.
  - CQPweb simple query manual
    https://cqpweb.lancs.ac.uk/doc/cqpweb-simple-syntax-help.pdf

## CEQL in a nutshell

| | |
|---|---|
| `speak` | exakt diese Wortform |
| `{speak}` | alle flektierten Formen (Lemma) |
| `at the end of the day` | exakt diese Wortfolge |
| `is n't it \?` | Tokenisierungsregeln & Escapes |
| `*able` | Suffix *-able* |
| `+able` | ohne das Wort *able* selbst |
| `light_JJ` | *light* als Adjektiv |
| `Mr _N*` | männl. Person (*Mr* + Substantiv) |
| `[Mr,Mrs,Ms] _N*` | Person (männlich oder weiblich) |
| `Mr _N* {be} _J*` | Aussage über die Person<br>(wird später noch verbessert) |

# CEQL in a nutshell

- `Smith:C` Groß- und Kleinschr. beachten
- `deja:d vu:d` Diakritika ignorieren
- `\D` Zahl (mind. eine Ziffer)
- `\u\u\u\u:C` Akronym (4 Großbuchstaben)
- `\u\L:C` nur erster Buchstabe groß
- `take * off` optionales Token
- `take ++*** off` zwei bis fünf optionale Token
- `in ( _JJ* )? time` optionales Adjektiv
- `Mr (_N*)+ {be} (_RB)? _J*` + = 1 oder mehr
- `his ( _JJS | most _JJ )* _N*` Alternativen
- `<s> but` Satzanfang
- `<ne_type=PERSON> (+)+ </ne_type>` XML-Element

# CQP Query Syntax

- formale Abfragenotation
  - basiert auf regulären Ausdrücken auf mehreren Ebenen
  - ermöglicht es, das Suchmuster sehr präzise festzulegen
  - deutlich flexibler und mächtiger als die CEQL-Syntax
- von allen Web-Oberflächen unterstützt, die auf CWB basieren!
- Tutorial & Dokumentation
  - **Kap. 12** aus Hoffmann, Sebastian *et al.* (2008): *Corpus Linguistics with BNCweb – a Practical Guide*. Frankfurt a.M. [etc.]: Peter Lang. (= English Corpus Linguistics 6)
  - CQP Query Language Tutorial (online version)
    http://cwb.sourceforge.net/files/CQP_Tutorial.pdf http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/

# CQP-Abfragen: einzelne Tokens

- reg. Ausdruck in Anführungszeichen „matcht“ Oberflächenform von Tokens
  - `"(over|under)\w+"` / `'(over|under)\w+'`
  - um Anführungszeichen zu finden: `""""` / `'"'`
  - immer regulärer Ausdruck ➞ Escapes bei Metazeichen nicht vergessen

- Suchoptionen:

| | | |
|---|---|---|
| `"deja"%c` | … | *case-insensitive* (Groß- und Kleinschreibung ignorieren) |
| `"deja"%d` | … | *diacritic-insensitive* (Diakritika ignorieren) |
| `"deja"%cd` | … | beides |
| `"?"%l` | … | exakte Zeichenkette (keine Metazeichen) |

- Beispiele hier in BNCweb mit dem CLAWS C5-Tagset

# CQP-Abfragen: einzelne Tokens

- Abfage von Token-Annotation mit einem Attribut-Wert-Paar aus p-Attribut und regulärem Ausdruck für den Wert des Attributs:
  - `[lemma = "(over|under)\w+_ADJ"]` (im BNC)
  - `[pos = "AJS"]` … Superlative (BNC)
  - `"deja"%cd` ist Kurzform für `[word = "deja"%cd]`

- Bedingungen mit Booleschen Operatoren kombinieren:
  - Operatoren: `&` (und), `|` (oder), `!` (nicht), `!=` (ungleich)

    token description
  - `[(word="can"%c) & (pos!="VM.*")]`
  - äquivalent zu `[(word="can"%c) & !(pos="VM.*")]`

# CQP-Abfragen: Token-Folgen

- CQP-Abfragen sind reguläre Ausdrücke über *token descriptions* (`[…]`)
  - `"in" [pos="AJ.*"]? [hw="year"]` … optional
  - `"in" [pos="AJ.*"]+ [hw="year"]` … mindestens eins
  - `"in" [pos="AJ.*"]{2} [hw="year"]` … genau zwei
  - `([pos="AJS"] | "most"%c [pos="AJ0"])` … entweder oder

- Abstände
  - `[]` … matchall (beliebiges Token)
  - `"dog" []{0,4} "cat"` … maximal vier Token dazwischen
  - `"dog" []{0,4} "cat" within s` … im selben Satz (s-Attribut)

# CQP-Abfragen: s-Attribute

- XML-Tags „matchen“ Start- und End-Tag von s-Attribut-Bereichen (*regions*)
  - `<head> "UK"` … als erstes Wort einer Überschrift
  - `"UK" </head>` … als letztes Wort einer Überschrift
  - `<head>` … findet nichts (weil Länge von 0 Token)
  - `<mw> []* </mw>` … passende Start-/End-Tags matchen ganzen Bereich

- innerhalb eines Bereichs suchen:
  - `"Twain" within quote;`
  - `[pos="NN.*"] :: match.mw_pos = "PRP";`
    … Abfrage von s-Attribut-Annotationen im „global constraint“
    … so können auch Bedingungen an Metadaten in der Query gestellt werden
  - vordefinierte Anker: `match`, `matchend`, `target` (`@`)

# CQP-Abfragen: Token-Folgen

- Quantoren und Alternativen können verschachtelt werden, um komplexe lexikalisch-grammatische Muster zu finden:

```
([pos="AJS"] | "most"%c [pos="AJ0"])
(
  "(and|\,)"%c
  ([pos="AJS"] | "most"%c [pos="AJ0"])
)+
[pos = "NN.*"]
```

Was tut diese Abfrage?

- Voreinstellung: möglichst wenig Wiederholungen (*non-greedy*)
  - `"ho"%c ("," "ho"%c)+` … findet immer *ho, ho*
  - `(?longest) "ho"%c ("," "ho"%c)+`
    … Strategie auf *greedy matching* umstellen (in CQPweb auch über UI möglich)

# Überlegungen vor der Recherche (1)

Vor einer ernsthaften Korpusrecherche gilt es, einige wichtige Überlegungen anzustellen:

- Was ist die sprachliche Grundgesamtheit, die einen interessiert? Standarddeutsch? (In Deutschland? In der Schweiz? Geschrieben und/oder gesprochen?) Eine bestimmte Varietät? Eine bestimmte Textsorte? Texte mit einem bestimmten Themenbezug? Texte aus den letzten paar Jahren oder aus vergangenen Jahrhunderten?
- Gibt es ein Korpus oder Teilkorpus, das dafür geeignet ist?
- Ist das gewählte Korpus groß genug? Ist daher für das zu untersuchende Phänomen mit genügend Treffern zu rechnen, um eine differenzierte Auswertung zu ermöglichen?

# Überlegungen vor der Recherche (2)

- Wie ist das Korpus genau aufgebaut? Welche Texte welcher Autoren, Textsorten, Entstehungszeiten usw. enthält es?
- Welche Zweifel an der Repräsentativität des Korpus gibt es? Wie gravierend sind diese? Worauf ist bei der Auswertung zu achten?

Wenn sich kein passendes Korpus finden lässt oder erste Testabfragen keine oder nicht genügend Treffer liefern, muss man u.U. darüber nachdenken, ein eigenes Korpus zusammenzustellen (oder die Fragestellung ganz anders anzugehen – z.B. mit einer Befragung).

# Übersicht: Web-Interfaces für Korpusabfragen

# CQPweb

- https://corpora.linguistik.uni-erlangen.de/cqpweb/
  - Login: `studentX` (`1` … `15`)
  - Passwort: `erlangen`

CEQL*

CWB
The IMS Open Corpus Workbench

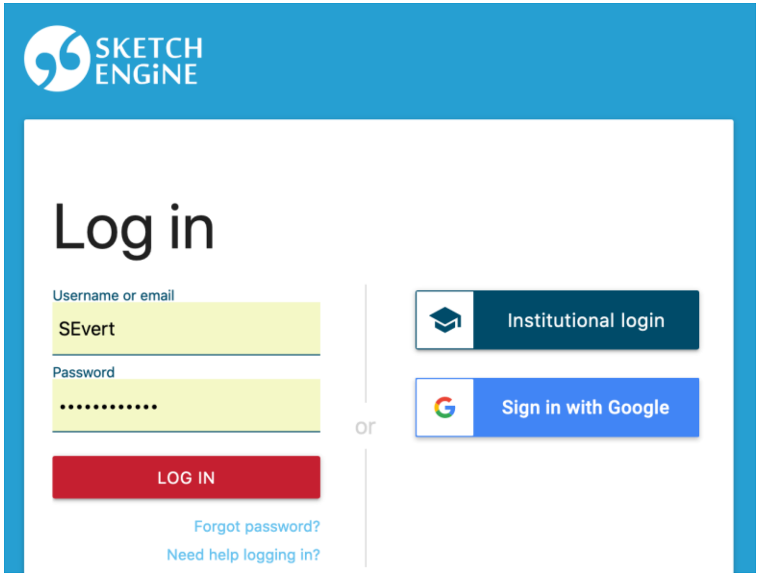
- Hintergrund:
  - Hardie (2012); Evert & Hardie (2011)
  - http://cwb.sourceforge.net/

- Dokumentation: Tutorial-Videos auf YouTube
  https://www.youtube.com/user/CorpusWorkbench

*http://cwb.sourceforge.net/ceql.php

# Sketch Engine

- https://app.sketchengine.eu/
  - sehr große Auswahl von Korpora in vielen verschiedenen Sprachen verfügbar
  - kommerzieller Anbieter: €90 / Jahr und Person (für Forschungszwecke)
  - Erstellen/Hochladen eigener Korpora bis zu einer Größe von 1 Mio. Wörter möglich (Upgrades möglich: z.B. auf 50 Mio. Wörter für €300 / Jahr)
  - FAU hat leider keine Uni-Lizenz (€3000 / Jahr)



- Bedienungsanleitung: https://www.sketchengine.eu/user-guide/

# Andere Web-Oberflächen @ FAU

- BNCweb (British National Corpus)
  https://corpora.linguistik.uni-erlangen.de/bncweb/ CEQL*
  - Login: studentX
  - Passwort: erlangen
  - v.a. auch für Beispiele und Übungsaufgaben aus *Corpus Linguistics with BNCweb – a Practical Guide* (Hoffmann et al. 2008)
- Debatten aus dem Europäischen Parlament
  http://corpora.linguistik.uni-erlangen.de/demos/CQP/Europarl/ CEQL*
- deutsche Zeitungen aus 1990er Jahren (HGC: Huge German Corpus)
  http://corpora.linguistik.uni-erlangen.de/demos/auth/HGC/ CEQL*
  - Login: demo
  - Passwort: demo
  - morphologisch annotiert

*http://cwb.sourceforge.net/ceql.php

# Weitere Web-Oberflächen (CWB-basiert)

- OPUS: Sammlung von Parallelkorpora
  http://opus.nlpl.eu/
- Leeds IntelliText (diverse Sprachen, Web-Korpora)
  http://corpus.leeds.ac.uk/itweb/htdocs/Query.html
- BFSU CQPweb (chinesische & englische Korpora @ BFSU)
  http://111.200.194.212/cqp/ http://www.bfsu-corpus.org/channels/corpus
- Linguateca AC/DC (Portugiesisch)
  http://www.linguateca.pt/ACDC/
- Ungarisches Nationalkorpus
  http://corpus.nytud.hu/mnsz/index_eng.html
- Corpus del Español Actual (Spanisch)
  http://spanishfn.org/tools/cea/english
- Varitext (Französisch)
  http://syrah.uni-koeln.de/varitext
- Spraakbanken (Schwedisch)
  https://spraakbanken.gu.se/korp/
- KorpusDK (Dänisch)
  http://ordnet.dk/korpusdk/

# Weitere Web-Oberflächen (CWB-basiert)

- TSCorpus (Türkisch)
  http://tscorpus.com/
- CORIS/CODIS (Italienisch)
  http://corpora.ficlit.unibo.it/
- SSLMIT La Repubblica (italienische Zeitungen)
  http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica
- BwanaNet (Katalanisch, Spanisch, Englisch)
  http://bwananet.iula.upf.edu/
- Georgetown University CQPweb (diverse Korpora)
  https://corpling.uis.georgetown.edu/cqp/
- Perugia Corpus (Italienisch)
  https://www.unistrapg.it/cqpweb/
- CorpusEye (diverse Sprachen)
  http://corp.hum.sdu.dk/
- TEI:TOK (mehrere Sprachen)
  http://www.teitok.org/index.php?action=projects

# Andere beliebte Web-Oberflächen

- BYU Corpora (von Mark Davies)
  http://www.english-corpora.org/
  - COCA, COHA, Seifenopern, GloWbE, TIME, Spanisch, Portugiesisch, …
- Google Web 1T 5-Grams (N-gramm-Datenbank)
  http://corpora.linguistik.uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5_freq.perl
  - N-Gramme durchsuchen, vorberechnete (Quasi-)Kollokationen
  - NetSpeak: hübschere Web-Oberfläche
    http://www.netspeak.org/
- Google Books Ngram Viewer (Info)
  https://books.google.com/ngrams/ https://books.google.com/ngrams/info
  - Visualisierung von Frequenzänderungen im Laufe der Zeit (Wörter, Phrasen)
- Linguee: Englisch, Deutsch, Französisch
  http://www.linguee.com/ http://www.linguee.de/ http://www.linguee.fr/
  - Web-crawled parallel corpora for many language pairs
  - nützlich, um mögliche Übersetzungen zu finden (*caveat emptor* …)
- Treebank.info (automatisch syntaktisch annotierte Korpora)
  http://treebank.info/

# Andere beliebte Web-Oberflächen

- DWDS-Korpora *
  https://www.dwds.de/r | http://kaskade.dwds.de/dstar/
  - Korpusabfragen, (diachrone) Kollokationsanalyse, Wortverlaufskurven
- COSMAS II *
  https://cosmas2.ids-mannheim.de/cosmas2-web/
  - Korpora des Instituts für deutsche Sprache (IDS), inkl. Kollokationsanalyse
- DGD: Datenbank für gesprochenes Deutsch des IDS *
  https://dgd.ids-mannheim.de
- KorAP: neue Web-Oberfläche für IDS-Korpora (CQL-Abfragen möglich)
  https://korap.ids-mannheim.de/
- ANNIS-Interface der HU Berlin: diverse, v.a. kleinere Korpora
  https://korpling.german.hu-berlin.de/annis3/

* Kurzanleitung für DWDS, COSMAS II und DGD:
http://sprachwissenschaft.fau.de/personen/daten/blombach/korpora.pdf