



**Universität  
Zürich**<sup>UZH</sup>

**Institut für Computerlinguistik**

**Dozenten: Dr. Cerstin Mahlow, Dr.-Ing. Michael Piotrowski**

---

# **Sprachvarianten des Deutschen im Zeitraum 1600 bis heute**

Vorlesung: Sprachtechnologie für historische Dokumente:  
Konzepte und Anwendungen (HS 2011)

Simon Hafner, Hernani Marques Madeira, Reto Baumgartner

# Problemstellung

Erkennung der Entstehungszeit  
historischer Texte nach 1600

Lösungsansatz mit Hilfe eines n-Gramm-Sprachmodells

Wo liegen die besten Resultate – bei welcher n-Gramm-  
Ordnung, zeichen- oder Wortbasiert?

# Problemstellung

Erkennung der Entstehungszeit  
historischer Texte nach 1600

Lösungsansatz mit Hilfe eines n-Gramm-Sprachmodells

Wo liegen die besten Resultate – bei welcher n-Gramm-  
Ordnung, zeichen- oder Wortbasiert?

# Trainingskorporus

# Daten für das Trainingskorpus

Digitale Bibliothek

Bereich Literatur

Nach den Regeln der Text Encoding Initiative (TEI)

<http://www.tei-c.org/index.xml>

# TEI-Format

Mehrere Werke in einem XML-File

Einzelne Werke repräsentiert durch Knoten mit dem Tag «TEI»

- ▶ Literaturgenre:
  - ▶ TEI/teiHeader/profileDesc/textClass/keywords/term
  - ▶ prose, drama, verse
- ▶ Erstellungsdatum:
  - ▶ TEI/teiHeader/profileDesc/creation/date
  - ▶ notBefore="1680" notAfter="1983"
- ▶ Text:
  - ▶ TEI/text

# TEI-Format

Mehrere Werke in einem XML-File

Einzelne Werke repräsentiert durch Knoten mit dem Tag «TEI»

- ▶ Literaturgenre:
  - ▶ TEI/teiHeader/profileDesc/textClass/keywords/term
  - ▶ prose, drama, verse
- ▶ Erstellungsdatum:
  - ▶ TEI/teiHeader/profileDesc/creation/date
  - ▶ notBefore="1680" notAfter="1983"
- ▶ Text:
  - ▶ TEI/text

# TEI-Format

Mehrere Werke in einem XML-File

Einzelne Werke repräsentiert durch Knoten mit dem Tag «TEI»

- ▶ Literaturgenre:
  - ▶ TEI/teiHeader/profileDesc/textClass/keywords/term
  - ▶ prose, drama, verse
- ▶ Erstellungsdatum:
  - ▶ TEI/teiHeader/profileDesc/creation/date
  - ▶ notBefore="1680" notAfter="1983"
- ▶ Text:
  - ▶ TEI/text



# Werkzeug zur Erstellung des Trainingskorpus

`korpusbastler.py`

In Python codiert

Einsetzbar ab Python 3.2

Gründe dazu:

- ▶ Eher neue Funktion `xml.etree.cElementTree.itertext()`
- ▶ Einfachere Arbeit mit Encodings ab Python 3.x

# Werkzeug zur Erstellung des Trainingskorpus

Für jedes enthaltene Werk:

- ▶ Literaturgenre:
  - ▶ Weiterverarbeitung der Genres
  - ▶ prose, drama
  - ▶ Keine Textextraktion bei Genres wie verse
- ▶ Erstellungsdatum:
  - ▶ `notBefore`, `notAfter`: Lebensdaten des Autors
  - ▶ Mögliches Erstellungsjahr: Mitte zwischen den Jahren
  - ▶ Einteilung in Korpora nach halben Jahrhunderten mithilfe des Erstellungsjahres
- ▶ Text:
  - ▶ Extraktion des Textes auf allen Tiefen
  - ▶ Mit `xml.etree.cElementTree.itertext()`
  - ▶ Schreiben in entsprechende Korpusdateien

# Werkzeug zur Erstellung des Trainingskorpus

Für jedes enthaltene Werk:

- ▶ Literaturgenre:
  - ▶ Weiterverarbeitung der Genres
  - ▶ prose, drama
  - ▶ Keine Textextraktion bei Genres wie verse
- ▶ Erstellungsdatum:
  - ▶ notBefore, notAfter: Lebensdaten des Autors
  - ▶ Mögliches Erstellungsjahr: Mitte zwischen den Jahren
  - ▶ Einteilung in Korpora nach halben Jahrhunderten mithilfe des Erstellungsjahres
- ▶ Text:
  - ▶ Extraktion des Textes auf allen Tiefen
  - ▶ Mit `xml.etree.cElementTree.itertext()`
  - ▶ Schreiben in entsprechende Korpusdateien

# Werkzeug zur Erstellung des Trainingskorpus

Für jedes enthaltene Werk:

- ▶ Literaturgenre:
  - ▶ Weiterverarbeitung der Genres
  - ▶ prose, drama
  - ▶ Keine Textextraktion bei Genres wie verse
- ▶ Erstellungsdatum:
  - ▶ notBefore, notAfter: Lebensdaten des Autors
  - ▶ Mögliches Erstellungsjahr: Mitte zwischen den Jahren
  - ▶ Einteilung in Korpora nach halben Jahrhunderten mithilfe des Erstellungsjahres
- ▶ Text:
  - ▶ Extraktion des Textes auf allen Tiefen
  - ▶ Mit `xml.etree.cElementTree.itertext()`
  - ▶ Schreiben in entsprechende Korpusdateien

# Trainingskorpora für die Sprachmodelle

Sprachstufe	Anzahl Wörter im Korpus
-------------	-------------------------

1600–1650	
-----------	--

1650–1700	
-----------	--

1700–1750	
-----------	--

1750–1800	
-----------	--

1800–1850	
-----------	--

1850–1900	
-----------	--

1900–	
-------	--



# Testkorpora

- ▶ Aus <http://de.wikisource.org/>
- ▶ 100 Sätze pro Sprachstufe
- ▶ d. h. 20 Sätze pro Jahrzehnt
- ▶ Genres entsprechen dem Trainingskorpus

# Testkorpora

- ▶ Aus <http://de.wikisource.org/>
- ▶ 100 Sätze pro Sprachstufe
- ▶ d. h. 20 Sätze pro Jahrzehnt
- ▶ Genres entsprechen dem Trainingskorpus



# Testkorpora

- ▶ Aus <http://de.wikisource.org/>
- ▶ 100 Sätze pro Sprachstufe
- ▶ d. h. 20 Sätze pro Jahrzehnt
- ▶ Genres entsprechen dem Trainingskorpus

## Testkorpora

- ▶ Aus <http://de.wikisource.org/>
- ▶ 100 Sätze pro Sprachstufe
- ▶ d. h. 20 Sätze pro Jahrzehnt
- ▶ Genres entsprechen dem Trainingskorpus

# Sprachmodelle

