

Corpus-Based Work

vorbereitet von Olga Walker

Neue Begriffe

- **Tokenisierung** - Segmentierung eines Textes in Einheiten
- **Lexem** - kleinste semantische Einheit, Träger der lexikalischen Bedeutung
- **Lemmatisierung** – Rückführung einer Wortform auf ihr Lemma oder Lexem

2

Probleme der Textaufbereitung

- Filtern vor der Aufbereitung
- Groß- Kleinbuchstaben
- Tokenisierung
- Worttrennung in versch. Sprachen
- Apostroph: Auslassungen vs. Genetiv
- Bindestrich: Silbentrennung vs. zusammengesetzte Worte
- Punkt: Abkürzungen vs. Satzende
- Verschiedene Schreibweisen

3

Morphologie

- Definition: Lehre von Flexion und Wortbildung
- Probleme:
 - Wortformen gruppieren?
 - großes, redundantes Lexikon
 - Wörter liegen selten in ihrer Grundform im Text vor

4



Sätze

- Probleme:
 - Satzgrenzerkennung:
Bedeutung des Punktes, geschachtelte Sätze (z. B. mit indirekten Rede)
- Algorithmus zur Satzgrenzerkennung

5



Textauszeichnung

- Metasprachen für Definition von Auszeichnungssprachen
 - **SGML**
 - **XML**
- Tag Sets

6