Statistik in der Computerlinguistik, 2023

UNI ZAGREB

UNI WIEN

ADRIANA ĐUGUM

2023

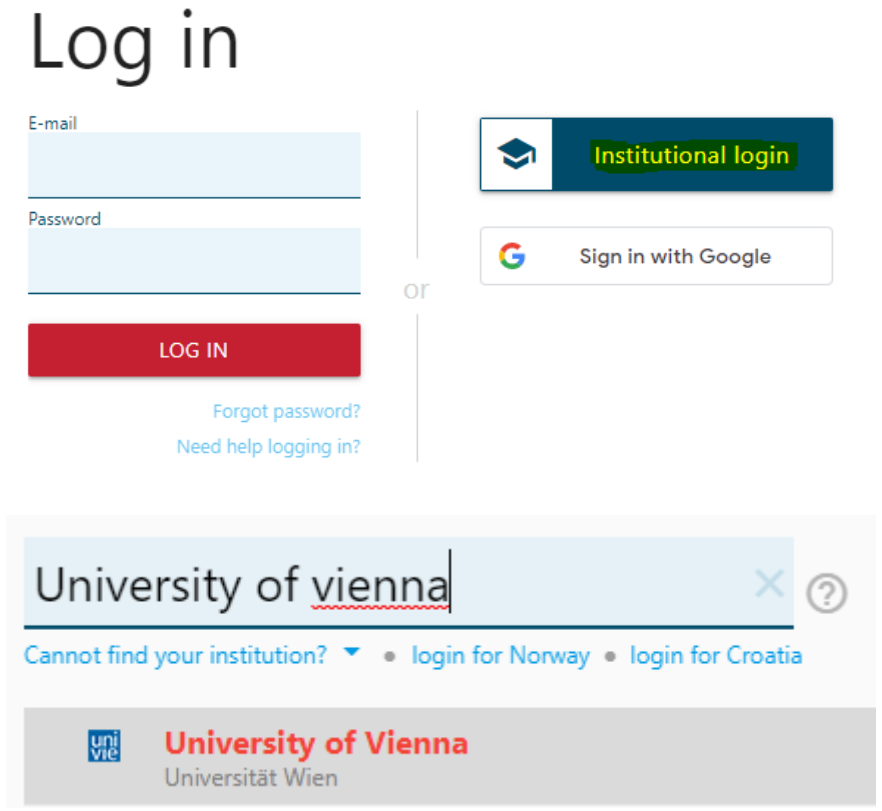
Statistik und Stichprobe

- **Stichprobe**
  - Wichtig: wie wir die eine Stichprobe generieren
- Computer Linguistik – Stichprobe
  - Bsp. - The British National Corpus – 100.000.000 token (100 M)
    - ***Type*** *- the ratio of the number of different words*
    - ***Token*** *– the total number of words*
  - Bsp. Aufgabe: Untersuchung des Tokens „in" im BNC
  - Wir entnehmen 10 Stichproben des Textes mit 1000 Tokens und wir messen die Frequenz des Tokens „in"

1	the	5578746
2	of	2728247
3	and	2313881
4	to	2289460
5	a	1926141
6	in	1687481
7	that	923054
8	is	886650
9	it	834407
10	was	801621
11	for	760425
12	's [possessive]	631302
13	on	630343
14	be	601369

Korpus - Corpus

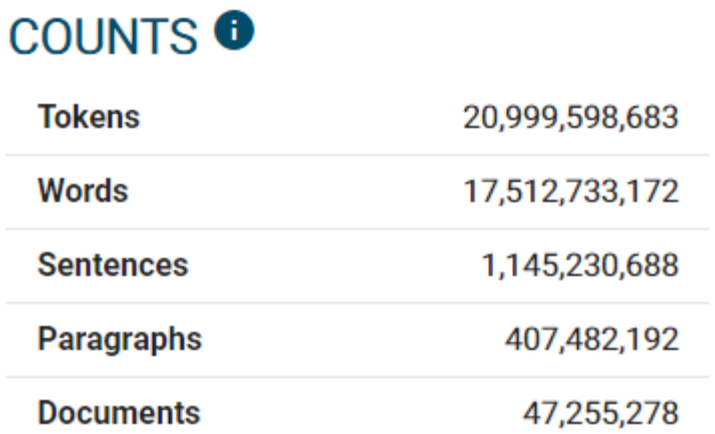
- Sketch Engine - www.sketchengine.eu
- Uni Wien log in (gleich wie u:space)
- YouTube (how to):
  - [How to start with Sketch Engine – YouTube]
- **Korpus, das**: *[als Datenbank angelegte] Sammlung einer begrenzten Anzahl von Texten, Äußerungen o. Ä. als Grundlage für sprachwissenschaftliche Untersuchungen*



Korpus in Sketch Engine

- Einen Korpus auswählen – German Web 2020:
  - deTenTen – German corpus from the web | Sketch Engine

- Einen Korpus erstellen mit Sketch Engine

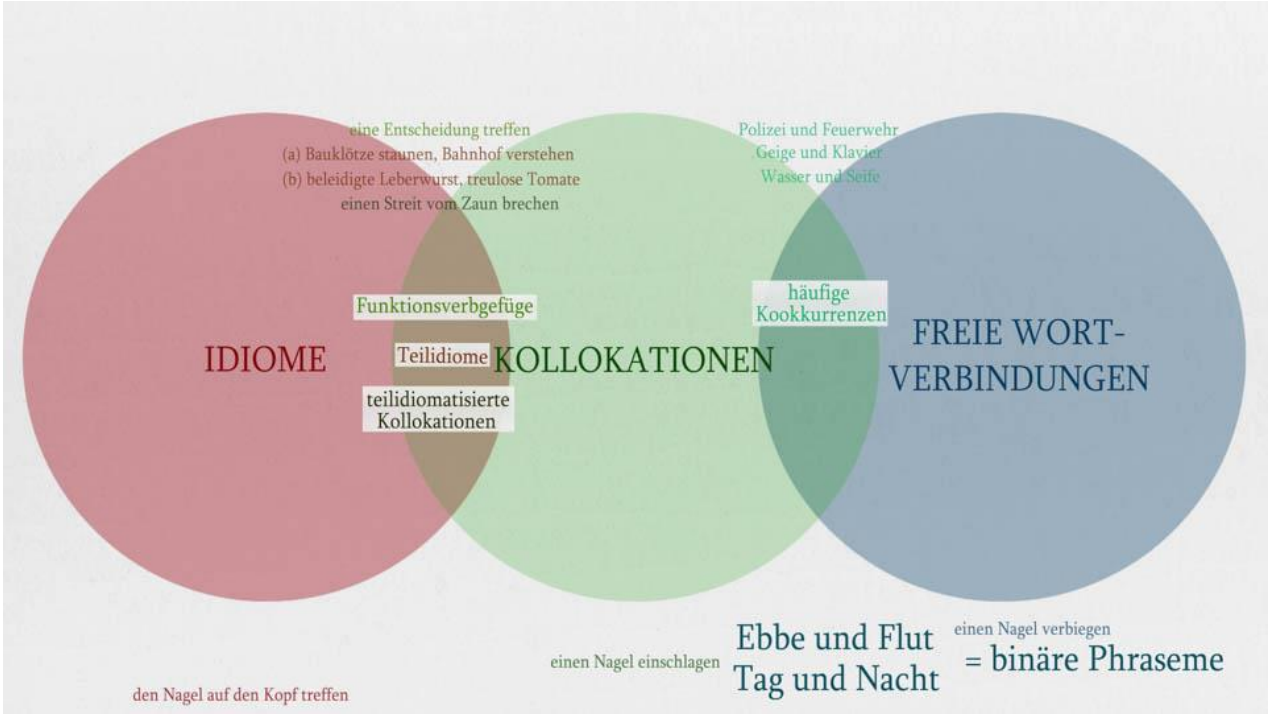


COUNTS

Tokens	20,999,598,683
Words	17,512,733,172
Sentences	1,145,230,688
Paragraphs	407,482,192
Documents	47,255,278

Kollokationen

- John Rupert **Firth** 1957. eingeführt - *You shall know a word by the company it keeps.*
- Manche Wörter haben die Tendenz sich nebeneinader zu befinden
- Bsp. *dick* + *Buch*, aber nicht *dick* + *Haus*
- 3 Eigenschaften der Kollokationen:
  - *Unteilbar (non-compositionality)*
  - *Nicht Substituierbar (non-substitutability)*
  - *Nicht Modifizierbar (non-modifiability)*



Warum sind Kollokationen wichtig?

- Die Statistik gibt uns eine gute Einschätzung ob etwas eine Kollokation ist oder nicht, aber warum und wie kann uns das helfen?
- Sprache als System → alle Tokens sind nicht komplett freiwillig distribuiert

- Maschinelle Translation
- Maschinelle Bearbeitung der Sprache
- Lexikographie
- Sprachen lernen

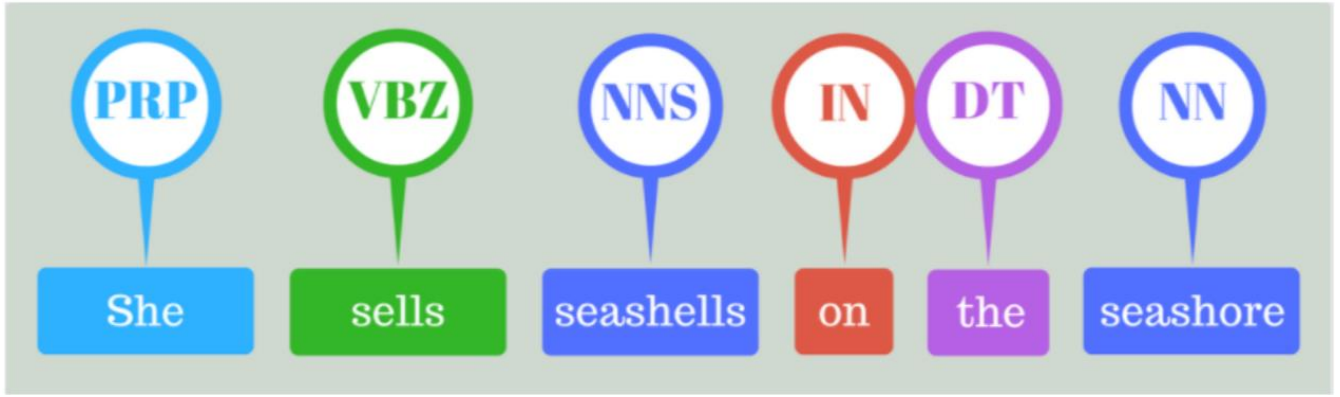
Maschine Translation

- Kann zu Problemen führen wenn wir eine Wort-für-Wort Übersetzung machen
  - Bsp. Ugasiti (požar) svjetlo - *das Licht löschen
    - Das Licht ausmachen/ausschalten
    - Fehler bei Google Translate

Statistik & Sprache 2.0

- POS tagging – Part of Speach
- Grammatik & Statistik

- D D N D D N D D N D D N D D N D X



- Mit der Statistik kann man auch noch nie gesehene Texte „taggen“
  - Wenn man eine gute Anzahl von Texten manuell taggt dann kann die Maschine alle weiteren Texte taggen mit einer sehr hohen Richtigkeit
- Google Search, Text Prediction, Voice Recognition, Spell Check, Grammar check…