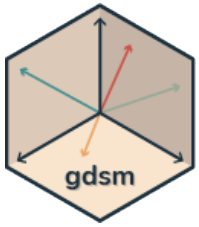


Session 02: Einführung in die Distributionelle Semantik

Viktoria Schneider & Dominic Schmitz

Verein für Diversität in der Linguistik

Was ist Distributionelle Semantik?

Brainstorming

- Die erste Gruppenarbeit ☺
- Nehmt euch 10 Minuten Zeit einmal zu überlegen, was Distributionelle Semantik sein könnte.

Was ist Distributionelle Semantik?

Theoretische Implikationen

- Distributionelle Hypothese:
  - *“Linguistic items with similar distributions have similar meanings.” (e.g., Harris 1954)*
- Unterschied in der Distribution von Wörtern = Unterschied in der Bedeutung von Wörtern
- Wort-Vektoren auf der Basis von computationellen Methoden
  - Kontexte werden benutzt um die Semantik eines Wortes zu bestimmen
- Distanz der Vektoren = Ähnlichkeit/Unähnlichkeit der Wörter aus
  - Hohe Distanz = Unähnlichkeit
  - Niedrige Distanz = Ähnlichkeit
- Für die Distanz können verschiedene Messarten benutzt werden (nächste Session)

Was ist Distributionelle Semantik?

Beispiel *Bank*

- Bank.1: ein Geldinstitut
- Bank.2: eine Sitzgelegenheit

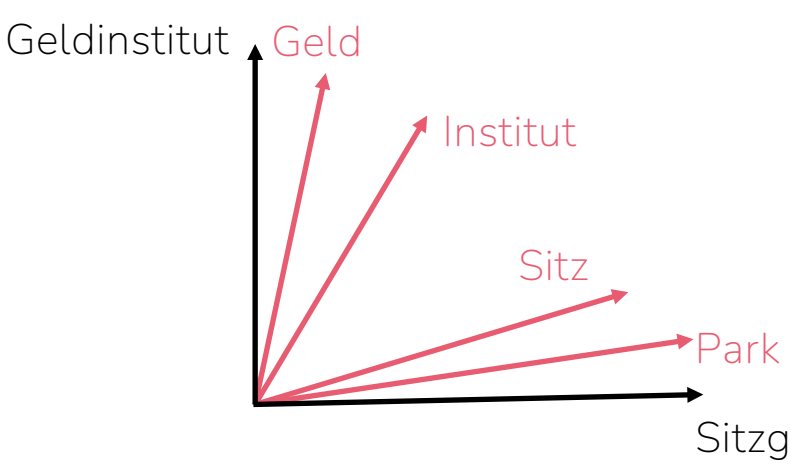
Dimensionen →

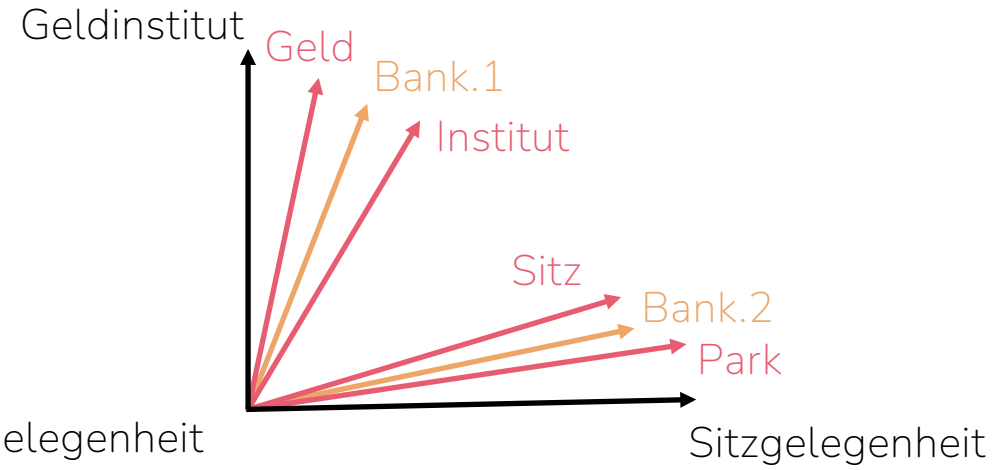
	Geld	Institut	Sitz	Park
Bank.1	35	26	15	0
Bank.2	0	0	37	60

Was ist Distributionelle Semantik?

Beispiel *Bank*

- Bank.1: ein Geldinstitut
- Bank.2: eine Sitzgelegenheit





Was ist Distributionelle Semantik?

Unterschiedliche Methoden

- Generelles Material: Vektorraum
  - Es gibt vorgefertigte Räume (bspw. FastText)
  - Man kann sie selber berechnen (bspw. NDL)
  - Wichtig: alle Zielwörter müssen im Vektorraum
    - Vorhanden sein
    - Oder berechnet werden können
- CBOW → Continuous Bag Of Words
  - Vektoren auf Grundlage von der Summe mehrerer Wörter
- Skip-Gram → *n*-grams
  - Vektoren auf Grundlage von ganzen Wörtern
  - Kann angereichter werden mit *n*-grams

Wir glauben an diese Mathemagie ohne die Formel jemals selber rechnen zu wollen ;-)

Was ist Distributionelle Semantik?

Skip-Gram – ohne *n*-grams

- Vorkommnisse jedes einzelnen Wortes im Vektorraum mit jedem anderen Wort im Vektorraum
- Dimensionen reduziert → je nach Forschungsfrage 100 – unendlich (Wörter im Vektorraum)

	Geld	Institut	Sitz	Park
Bank.1	35	26	15	0
Bank.2	0	0	37	60

Was ist Distributionelle Semantik?

Skip-Gram - mit *n*-grams

- Vorkommnisse jedes einzelnen Wortes im Vektorraum mit jedem anderen Wort im Vektorraum und deren *n*-grams
  - *n*-grams: für Deutsch und Englisch sind 3-6-grams sinnvoll (Bojanowski 2016)
- Dimensionen reduziert → je nach Forschungsfrage 100 – unendlich (Wörter im Vektorraum)

	#ba	ban	ank	nk#
Bank	1	1	1	1
Bar	1	0	0	0

Was ist Distributionelle Semantik?

CBOW

- Vorkommnisse Zielwort im Vektorraum mit der Summe der Vektoren der Wörter in der Umgebung
- Dimensionen reduziert → je nach Forschungsfrage 100 – unendlich (Wörter im Vektorraum)

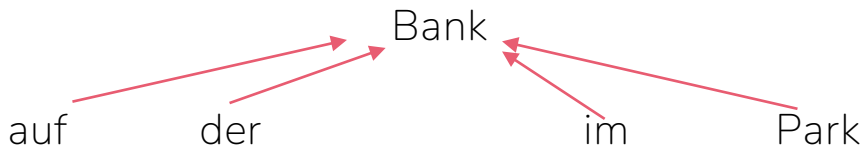
- Ich sitze auf der Bank im Park

Vektor Summe Vektor Summe

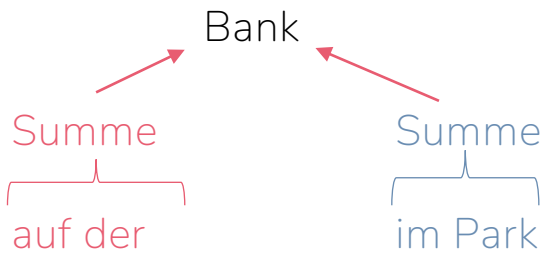
Was ist Distributionelle Semantik?

Beispiel: Fenster 2 Wörter rechts und links vom Zielwort

- Skip-gram: jedes Wort wird mit einbezogen (man kann auch nur content words nehmen)



- CBOW die Bedeutung der Wörter im Kontext zusammen wird benutzt



- Beide versuchen das Zielwort vorher zu sagen, Skip-gram funktioniert wohl besser mit *n*-grams als CBOW

Was ist Distributionelle Semantik?

Fragen über Fragen...

Pause ☺