berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Lexikalisch-semantische Analyse von Korpora

Lothar Lemnitzer

Workshop *Internetlinguistik und Korpusanalyse*, Hannover, 1. Mai 2015

Inhalt

- Rahmen meiner Darstellung
- Korpora in der lexikographischen Arbeit
- Konzepte und Verfahren der lex.-sem. Korpusanalyse
- Anwendungen (in der Lexikographie)
- Fazit

Der Rahmen – Arbeit am DWDS

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Digitalen Wörterbuchs der deutschen Sprache

- Eigenschaften und Strukturen des Wortschatzes sollen beschrieben werden
- Es wird dabei auf lexikalische Informationen bestehender Wörterbücher aufgebaut, vor allem *Wörterbuch der deutschen Gegenwartssprache, Duden GWB*
- Artikel werden a) neu erarbeitet und b) überarbeitet / aktualisiert
- Dabei wird auf den in Textkorpora dokumentierten Sprachgebrauch Bezug genommen (Methode der „Entdeckung“, Prinzip der Belegung)

Was ein Korpus mir, dem Lexikographen, zeigen könnte…

Bezogen auf eine konventionalisierte Bedeutungsbeschreibung eines sprachlichen Zeichens:

- Anzahl der (neuen) Lesarten des lexikalischen Zeichens
- Beispiele / Belege, die die (typische) Verwendung des Zeichens in einer bestimmten Lesart zeigen
- Paraphrasierung der Bedeutung des sprachlichen Zeichens („Definition“)

Was ein Korpus mir, dem Lexikographen, zeigen könnte…

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Typischerweise in der Umgebung des sprachlichen Zeichens vorkommende Wörter (Wortverbindungen, mehr oder weniger fest…) – syntagmatische Beziehungen
- Auf der Ebene der lexikalisch-semantischen Bedeutung relationierte sprachliche Zeichen (Synonyme, Antonyme, Hyperonyme etc.) – paradigmatische Beziehungen
- Bedeutungswandel als Teil des Sprachwandels (Änderung der Gebrauchsregeln)

Bedeutungsbegriff in der korpusbasierten lex-sem Analyse

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Man kann für eine GROSSE Klasse von Fällen der Benützung des Wortes 'Bedeutung' - wenn auch nicht für ALLE Fälle seiner Benützung - dieses Wort so erklären: Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache. (L. Wittgenstein, PU 43)

Im Gegensatz zu kognitiven Konzeptualisierungen von „Bedeutung“ oder „embodiment“-Theorien des Wissens

Beispiel

etw. riecht stechend, streng, muffig, süßlich, angenehm

etw. schmeckt süß, bitter, sauer, salzig, lecker, vorzüglich

(Daten aus Wortprofil)

AMMONIAK: stechend riechendes Gas

ZUCKER: süß schmeckendes Kohlehydrat

Die sinnliche Erfahrung hinter „stechend riechend“ oder „süß schmeckend“ ist (sprachlich) nicht vermittelbar

Technische Voraussetzung für die lex-sem Analyse

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Schiere Größe der Datenmengen, und, daraus folgend: Verfügbarkeit großer Mengen von Gebrauchsinstanzen, aus denen sich Gebrauchsregeln abstrahieren lassen (Beispiele: Corpora from the Web, IDS, DWDS-Korpora …), (standardisierte) Metadaten
- Anreicherung der Korpora mit Ergebnissen linguistischer Analysen (z.B. Wortart, synt. Struktur), linguistische Abfragesprachen

Technische Voraussetzung für die lex-sem Analyse

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Mathematische / Statistische Verfahren der Datenanalyse und Datenreduktion, deren computationelle Handhabbarkeit
- Computationell verarbeitbare Regelsysteme für die Mustererkennung

Traditionen

Information Retrieval – „gib mir aus der Dokumentenbasis die wichtigsten Dokumente zur Verarbeitung von Druckaufträgen“

Maschinelle Übersetzung: Übersetze „Druck“ als „pressure“, „print“, „compression“ …, je nach der Verwendungsweise im Ausgangstext

Schlüsselwörter korpusbasierter lex-sem Analyse

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

DISTRIBUTIONELLE HYPOTHESE: „Words that occur in similar **contexts** tend to have **similar** meanings“ (Turney and Pantel, 2010, mit Bezug auf Z. Harris, 1957)

Aus der distributionellen Hypothese folgt auch: Lexikalische Zeichen, wenn sie in hinreichend verschiedenen Klassen von Kontexten auftreten, weisen unterschiedliche Verwendungsklassen und damit Lesarten („senses“) auf

Schlüsselwörter korpusbasierter lex-sem Analyse

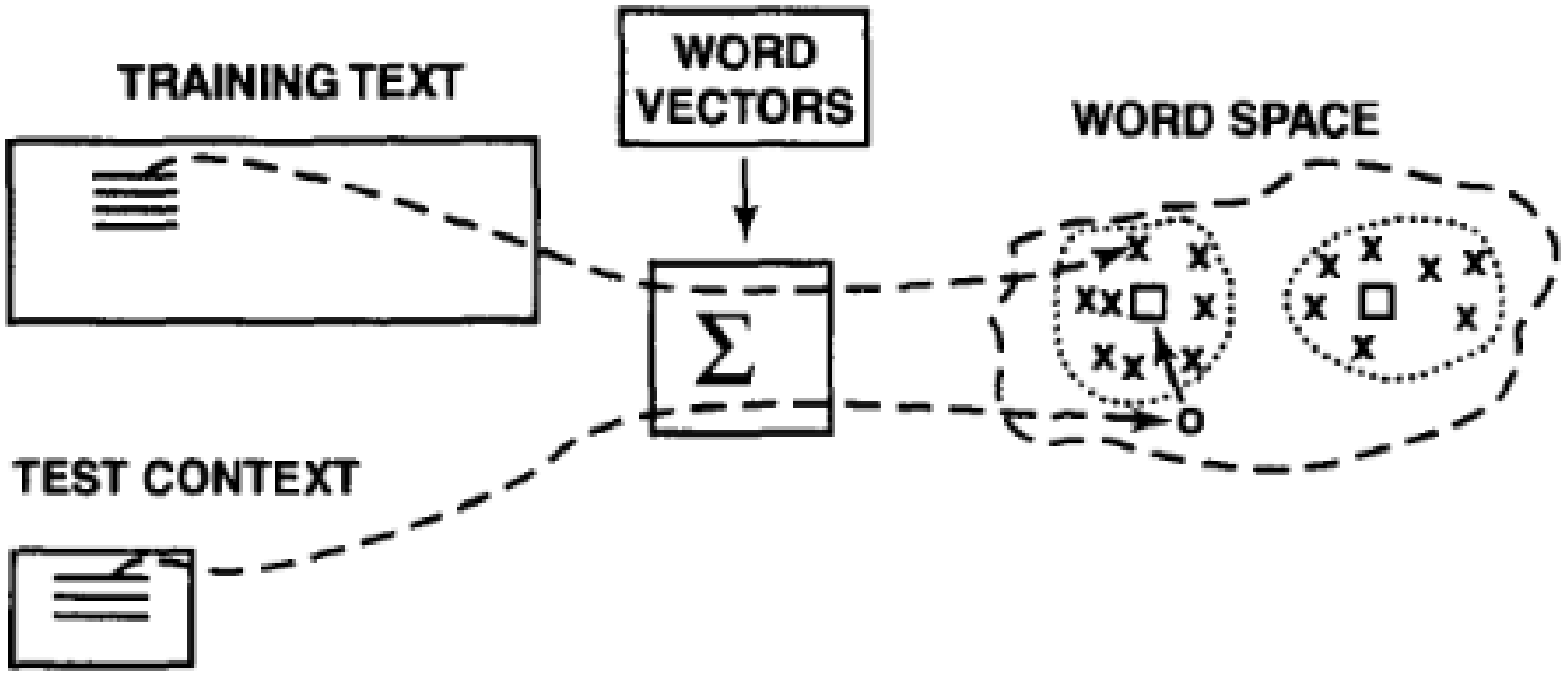
berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

KONTEXT: die Wörter und deren Attribute, die in einer bestimmten Verwendungsinstanz ein bestimmtes lexikalisches Zeichen umgeben → zusammengefasst als Kontextvektoren

ÄHNLICHKEIT / DISTANZ: Ähnlichkeit der Bedeutung zweier Wörter wird als Distanz von Vektoren in einem Vektorraum bzgl. Eeiner Metrik modelliert

CLUSTER: Partitionierung des V.raumes, so dass Häufungen von Vektoren sich ergeben; ZENTROID: zentraler Vektor des C.

Cluster



Schlüsselwörter korpusbasierter lex-sem Analyse

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

MUSTER: Schema der sprachlichen Realisierung eines Äußerungstyps (z.B. Beschreibung der Bedeutung eines Ausdrucks) mit festen und variablen Elementen

Schlüsselwörter korpusbasierter lex-sem Analyse

Von diesen grundlegenden Konzepten abgeleitet:

BEDEUTUNGSÄHNLICHKEIT: Nähe der Kontextvektoren zweier Wörter im Vektorraum

WORD SENSE: Menge von nahe beieinanderliegenden Kontextvektoren für ein lexikalisches Zeichen; Cluster

Anwendungen

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

- Semantisches Information Retrieval: Konzeptuelle Abstraktion von Anfragetermen und Texttermen, Erweiterung der Suchanfrage (welcher Druckauftrag? Welcher Druck?)
- Word Sense Disambiguation: textuelle Disambiguierung auf der Basis eines Verzeichnisses von Lesarten

Anwendungsgebiet Lexikographie

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

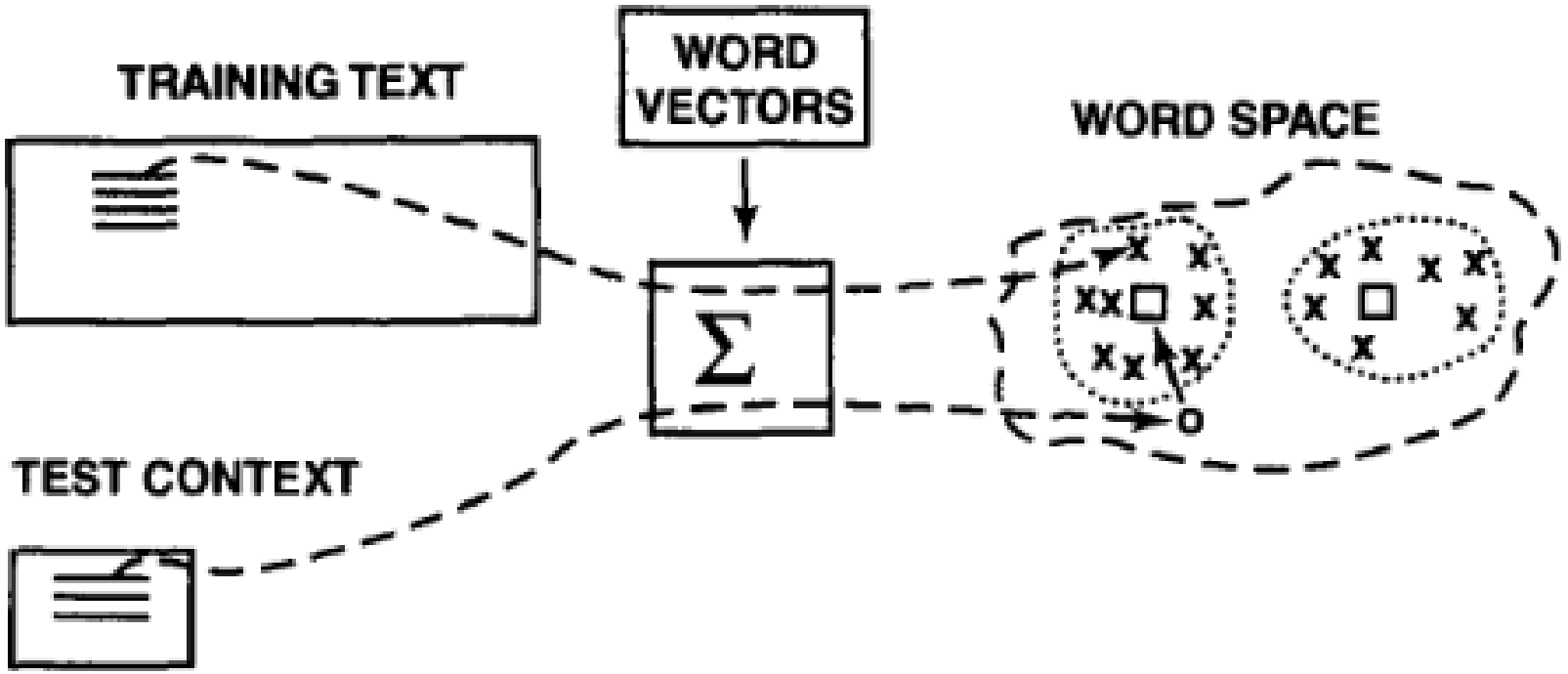
- Anzahl Lesarten ermitteln / Lesarten ergänzen
- Gute Beispiele finden
- Paraphrasierung der Bedeutung des sprachlichen Zeichens („Definition“)
- Kollokationen / Phraseme finden
- Synonyme, Antonyme, Hyperonyme finden
- Bedeutungswandel als Teil des Sprachwandels (Änderung der Gebrauchsregeln) erkennen

Anwendungsgebiet Lexikographie: Lesarten

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

- Festlegung der Lesarten und Unterlesarten ist ein bisher wenig erforschtes Gelände. Beispiel 1, *Flugtag*, im Handout zeigt die Probleme einer automatischen Gruppierung von Verwendungsklassen
- Ergänzung neuer Lesarten – Verfahren der Word Sense Discrimination (Schütze 1998, nächste Folie) – Reduktion um Belege für bekannte Lesarten (s. Beispiel 2, *Menü*, im Handout)

Anwendungsgebiet Lexikographie: Lesarten



Anwendungsgebiet Lexikographie: gute Belege

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Herausforderung: Beispiele / Belege für eine seltene Verwendungsklasse (Lesart) finden.

Ansatz: gibt es zu dieser Lesart ein Synonym, dann ein Cluster hierfür bilden, Wortvektoren für alle Vorkommen des Zielworts „hineinrechnen“ und die dem existierenden Clusterzentrum am nächsten liegenden Belege extrahieren. Bisher noch nicht erprobt (Beispiel: Flugtag, Kategorie Entfernung // Autominute)

Anwendungsgebiet Lexikographie: Definitionen

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Herausforderung: im Korpus Bedeutungsparaphrasen für eine Lesart eines Wortes finden

Ansatz: Textmuster, durch die typischerweise der Äußerungstyp „Bedeutungsexplikation“ realisiert wird (z.B. Cramer 2011)

„Bei NN handelt es sich um …“

Ca. 400 Treffer im DWDS-Korpus

Bei Pulsaren handelt es sich um Neutronensterne , die nach der Explosion einer Supernova als ausgebrannte Reste zurückbleiben . (DWDS, Archiv der Gegenwart)

Anwendungsgebiet Lexikographie: Kollokationen

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Im Allgemeinen werden typische Wortverbindungen auf Grund statistischer Verteilung der Kovorkommen und syntaktischer Muster (Adjektiv-Nomen, Verb-Objekt etc.) ermittelt. Lexikalisch-semantische Eigenschaften von Basis und Kollokator spielen im Allgemeinen keine Rolle.

Lexikographie: bedeutungsverwandte Wörter

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Herausforderung: aus dem Korpus zu einem Stíchwort bedeutungsverwandte Wörter aufzufinden (und zu klassifizieren)

Ansätze: 1. Textmuster finden, in denen typischerweise semantisch verwandte Wörter vorkommen (Jones 2010)

Beispiel: *weder* ADJD *noch* ADJD

In Wirklichkeit haben sich die italienischen Fabrikarbeiter weder als Individuen noch als Gewerkschaften, **weder aktiv noch passiv** den Erneuerungen in den Weg gestellt…

Lexikographie: bedeutungsverwandte Wörter

Ansätze: 2. iterative Kollokationsextraktion. Extrahiert werden zu den Kollokatoren 1…n einer Basis deren Kollokatoren, in der Erwartung, dass die gemeinsame Verwendung vieler Kollokatoren zweier Basen auf eine semantische Ähnlichkeit dieser Basen hindeutet (Biemann et al. 2004)

Als Service verfügbar: http://wortschatz.uni-leipzig.de/

Fazit 1

Korpora und Korpustechnologie sind heute in einem Zustand, dass Aspekte der lexikalisch-semantischen Bedeutung sprachlicher Zeichen aus ihnen erschlossen werden können

Fazit 2

Das Konzept „Bedeutung“ wird dabei so gefasst oder reduziert, dass er operationalisierbar ist, d.h. mit statistischen, geometrischen und algebraischen Methoden oder auch mit Verfahren der robusten Mustererkennung modelliert werden kann

Fazit 3

Die Grenzen der Verfahren liegen

1. In der Erfassung und Beschreibung seltener Phänomene
2. Bei nicht „trennscharfen“ Lesarten
3. In der Erfassung und Beschreibung von Bedeutungsaspekten wie Vagheit, konnotative Bedeutung (‚semantic prosody‘), übertragene Verwendung.

(1) Ist keine prinzipielle Begrenzung, (2) möglicherweise schon

Fazit 4

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Es besteht (wissenschaftlicher wie praktischer) Bedarf an der Erforschung lexikalisch-semantischer Aspekte der Korpuslinguistik und an der Entwicklung von Werkzeugen, die diese Erkenntnisse in der Praxis erproben.

(vorläufiges Ende der Geschichte. Der Vorhang zu und alle Fragen offen)