Data-Mining und Visualisierung zur Erfassung der semantischen Entwicklung von Konstruktionen

Catherine Diederich und Florent Perek, Universität Basel
Kontakt: catherine.diederich@unibas.ch, florent.perek@unibas.ch

Ausgangslage

"Digital Humanities" in der Sprachwissenschaft:
- Korpusanalysen: Gebrauch von grossen digitalen Datenbanken um grosse Mengen von Daten auf grammatikalische Eigenschaften und Muster zu untersuchen.
  - Problem: Ein Korpus bietet nur Zugang zu Wortformen → Bedeutung ist nicht direkt beobachtbar

Fragestellungen: Inwiefern eignen sich Data-Mining Techniken um den Gebrauch von syntaktischen Konstruktionen diachronisch zu untersuchen? Wie können wir die semantische Entwicklung dieser Konstruktionen aufzeigen?

Möglicher Ansatz:
- Verteilungsanalysen (‚distributional analyses')
  - Berechnungsmethoden zur automatischen Annotation
  - Vektor-Raum Modelle basieren auf der Verteilungs-Hypothese, d.h. **Wörter, die in ähnlichen Kontexten auftreten, zeigen ähnliche Bedeutungen auf**

Vektor-Raum Modelle

Ausgangspunkt: Aufbau einer Kookkurrenz-Matrix basierend auf Korpus-Daten.
- Die Matrix enthält die Menge an Kookkurrenzen für jedes Vorkommnis des Schlagworts in einem festgelegten Kontextfenster
- Funktionswörter werden normalerweise nicht berücksichtigt
→ Beispiel: das lexeme *kiss* (Kontextfenster: +/− fünf Wörter)

	frizzy	hair,	felt	him	**kiss**	me	on	the	cheek	with
give	her	my	flowers,	to	**kiss**	her	hand,	but	did	not
Diana,	and	bent	down	to	**kiss**	him	on	the	cheek.	She
towards	him	and	tried	to	**kiss**	her	on	the	mouth.	Over
and	I	lean	over	and	**kiss**	her	on	the	mouth.	I

Kookkurrenzen

frizzy	1
hair	1
feel	1
cheek	2
give	1
flower	1
hand	1
Diana	1
bend	1
try	1
mouth	2
lean	1

Resultat: $N_{Wörter} \times N_{wörter}$
- Einschränkung der Dimensionalität auf die informativsten Aspekte der Wortverteilung
- Jedes Wort wird einer Zeile der Matrix zugeordnet
- Die Ähnlichkeit der Reihen reflektiert die semantische Ähnlichkeit

Anwendung von Vektor-Raum Modellen

Vorteile der Vektor-Raum Modelle
- Das informelle Konzept der semantischen Repräsentation wird in ein empirisch-prüfbares Modell gewandelt
- Erlauben die Quantifizierung der semantischen Ähnlichkeit

Repräsentation von psychologischer Realität
- Resultate korrelieren mit menschlichen Leistungen in verschiedenen Aufgaben: Synonyme beurteilen, Wortassoziationen, semantisches Priming... (Lund et al. 1995, Landauer et al. 1998).

Vektor-Raum Modelle werden seit kurzem in der Linguistik verwendet
- Um semantische Klassen zu bestimmen
  - z.B. Gries und Stefanowitsch (2010): Clustering von Verben und ihrer Vorkommnis in Konstruktionen anhand der häufigen Kollokationen

Frage: Können Vektor-Raum Modelle auch angewendet werden, um syntaktische Produktivität zu untersuchen?
- d.h. um zu analysieren wie Sprachbenutzer auf kreative Weise Wörter kombinieren
- in einer diachronischen Perspektive entspricht Produktivität der lexikalischen Erweiterung einer Konstruktion

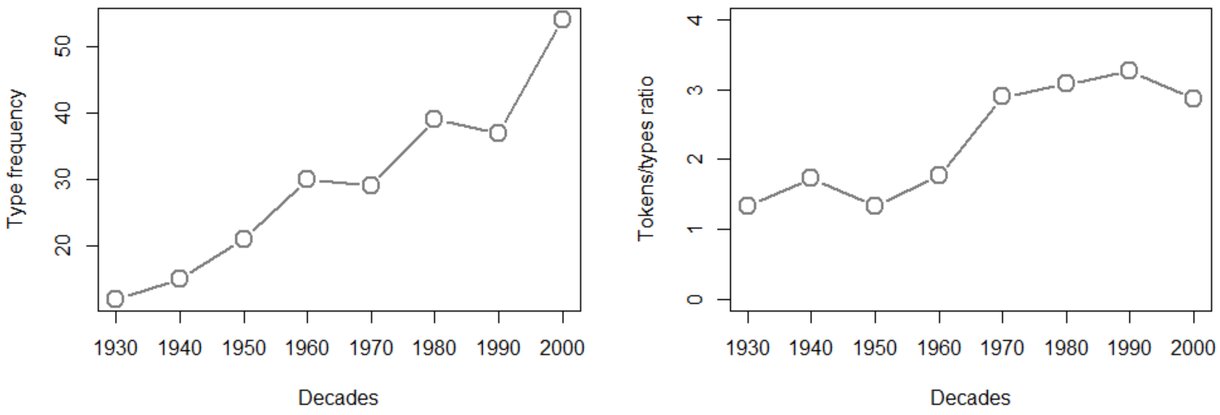
Korpusanalyse: Transitiv-Konstruktion mit *hell*

Konstruktion: **V *the hell out of* NP** → z.B. *You scared the hell out of me!*
- Vermittelt eine semantische Intensivierung
- Einige vulgäre Varianten, z.B. *crap*, *shit*, oder *fuck* anstelle von *hell*
- Intransitive Verben kommen auch vor, z.B. *I've been listening to the hell out of your tape* (COCA)
- Eignet sich für eine konstruktionsgrammatische Analyse:
  - Die intensivierende Funktion wird in der ganzen Konstruktion vermittelt
  - **Wie hat sich die Konstruktion im Verlauf der Zeit verändert?**

Daten: Corpus of Historical American English (COHA, Davies 2010)
- 1810 – 2009, ca. 20 Mio. Wörter pro Jahrzehnt , Amerikanisches English
- Geschriebene Sprache aus verschiedenen Gattungen: Zeitungen, Magazine, Belletristik, Sachbücher

- Neue Konstruktion: Erste Befunde in den 1930er
- Die Konstruktion zentriert sich auf zwei Verben: *scare* und *beat* (30% und 25% in 2000)
- Andere Beispiele: *Then I [...] avoided the hell out of his presence*
  But you drove the hell out of it!
  The Russians understood the hell out of that.

Resultate/ Visualisierungen

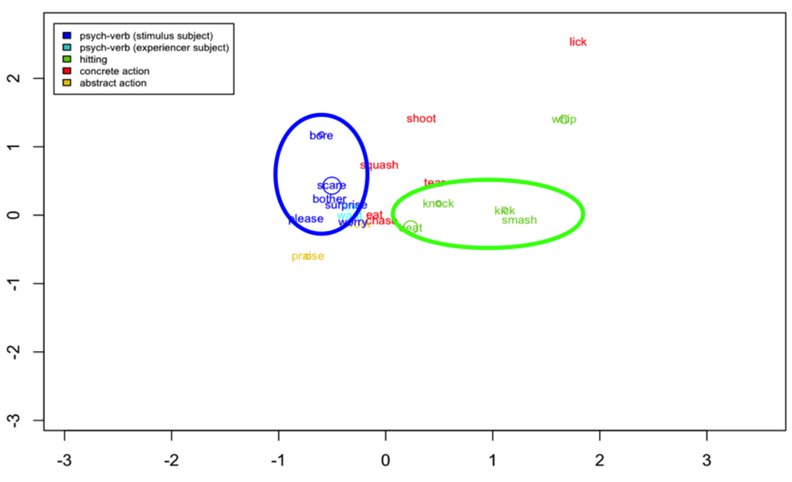


- Ein **Vektor-Raum Model** um die Verben in der *hell*-Konstruktion semantisch zu evaluieren
- **Visualisierung** des Vektor-Raumes anhand **‚multidimensional scaling'** (MDS; zweidimensional)

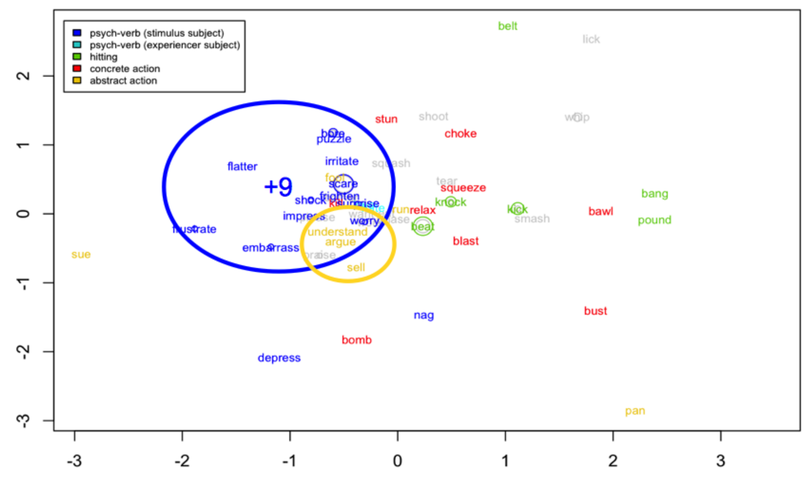
- Im Verlauf der Zeit kommen immer mehr verschiedene Verben in der Konstruktion vor
- Daraus resultierende Fragen:
  - Welche Art von Verben sind dazugekommen und wann?
  - Werden bestimmte semantische Domänen von der Konstruktion bevorzugt und hat sich dies im Verlaufe der Zeit verändert?

Diachronische Analyse → vier Zeiträume von je 20 Jahren:

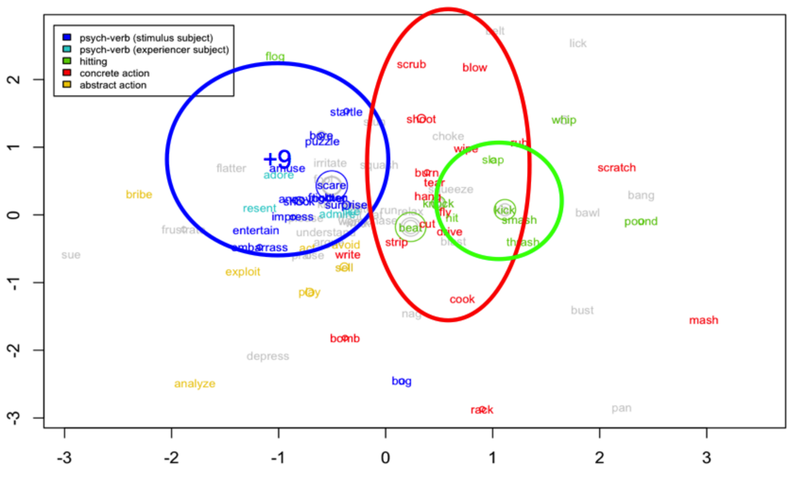
The *hell*-construction in diachrony: 1930s-1940s



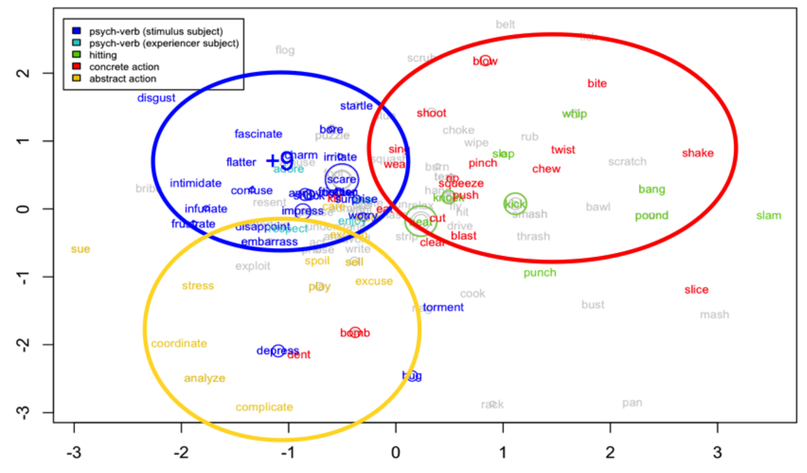
The *hell*-construction in diachrony: 1950s-1960s



The *hell*-construction in diachrony: 1970s-1980s



The *hell*-construction in diachrony: 1990s-2000s



Diskussion

Es werdet vorwiegend **psychologische Verben** bevorzugt (v.a. mit einem Stimulus-Subjekt)
- Sie bilden von Anfang an (1930er–) ein dichtes Cluster
- Meist besiedelte „Region" in allen Jahrzehnten
- Diese Gruppe zieht regelmässig neue Mitglieder an

Was ist mit den anderen semantischen Domänen?
- Weniger dicht besiedelt, d.h. niedrige Type-Frequenz, hohe semantische Variabilität
- Weniger vertreten von 1930 – 1940
- Anfangs ziehen diese Domänen wenig neue Mitglieder an
- Die lexikalische Produktivität nimmt mit der Zeit zu, z.B.:
  - konkrete Handlungen in den 1970er – 1980er
  - abstrakte Handlungen in den 1990er – 2000er

Die Ergebnisse entsprechen den aktuellen usage-based Modellen der Produktivität
- d.h. Produktivität ist ein Produkt der Type-Frequenz und der semantischen Variabilität
- Dicht besiedelte Regionen ziehen am ehesten neue Mitglieder an
- In weniger dicht besiedelten Regionen ist eine kritische Masse an Verben notwendig um Produktivität anzukurbeln
- Token-Frequenz ist nicht besonders entscheidend für Produktivität:
  - z.B. das häufig vorkommende Verb *kick* bleibt weitgehend isoliert
  - Gelegentlich erscheinen neue Verben in der Nähe von *kick*: Vermutlich analoge Ausweitungen eines prominenten Modells (Barðdal 2008; Bybee and Eddington 2006)

Referenzen

- Barðdal, J. (2008). *Productivity: Evidence from Case and Argument Structure in Icelandic*. Amsterdam: John Benjamins.
- Bybee, J. and D. Eddington (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language* 82 (2), 323–355.
- Davies, M. (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009*. Available online at http://corpus.byu.edu/coha/.
- Gries, S. and A. Stefanowitsch (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 73-90. Stanford, CA: CSLI.