

Computerlinguistische Aspekte der Phraseologie

Heid, Ulrich (2007): Computational Linguistic Aspects of Phraseology
Rothkegel, Annely (2007): Computerlinguistische Aspekte der Phraseologie
(beides in Burger et al. (2007), pp. 1027–1044)

1 Begriffe der Computerlinguistik

NLP (*Natural Language Processing*) – Oberbegriff für Technologien wie Automatische Übersetzung (*Machine Translation MT*), Dialogsysteme, Information Retrieval, Information Extraction, Spracherkennung, Sprachsynthese, elektronische Wörterbücher und Benutzerunterstützungssysteme.

Wort – Zeichenkette zwischen zwei Leerzeichen

Mehrwort-Einheit (*multiword unit*) – zwei oder mehr Worte, die signifikant häufig zusammen vorkommen. Darunter zählen:

- Komposita (*FR pomme de terre*)
- Kollokationen (*Notiz nehmen*)
- Idiome (*jmdm. einen Bären aufbinden*)
Die ersten drei Typen werden unter einem erweiterten Begriff *Phrasem* subsummiert.
- Mehrwort-Adverben, -Präpositionen, -Konjunktionen, -Adjektive
- Routineformeln und konventionalisierte Ausdrücke

CL vs. klassische Phraseologieforschung

1. Erkenntnisse der klass. Phraseologie werden für algorithmische Verarbeitung von Sprache genutzt
2. CL ermöglicht quantitative Korpusanalysen und multimediale / hypertextuelle Repräsentationen

2 Phraseologie: Fragestellungen der CL

- **Extraktion:** Phraseme automatisch aus Textkorpora gewinnen und quantitativ untersuchen
- **Identifikation:** Phraseme automatisch als solche erkennen
- **Semantische Repräsentation:** die spezifische Bedeutungskonstruktion erfassen
- **Standardisierung:** standardisierte Formate für wiederverwendbare elektronische Phrasem-Lexika entwickeln

2.1 Korpuslinguistik und Datengewinnung

Manuell Phraseme zu extrahieren ist teuer und aufwändig, also braucht man Automatisierungsmethoden. Die Werkzeuge müssen „Mehrwort-Einheiten“ *identifizieren, morphosyntaktische Eigenschaften ermitteln* und im Idealfall auch eine *Klassifizierung* vornehmen.

- Statistische Methoden: **Kookkurrenzhäufigkeiten**
 - hohe Frequenz: Kollokationen und Komposita
 - niedrige Frequenz: Idiome
- Probleme dabei:
 - besonders seltene Phänomene (e.g. 5mal im Korpus) könnten Zufallsprodukt sein
 - verschiedene Berechnungsformeln erzeugen stark unterschiedliche Resultate
 - statistisch signifikante Wortkombinationen bedeuten nicht immer Phrasemstatus (cf. *Kaffee trinken*)

Lösungsansätze: neben Kookkurrenz auch syntaktische Muster als Extraktionskriterium nutzen.

2.2 Phrasemidentifikation und NLP

Im Unterschied zu Korpusanalysen soll hier *entschieden* werden, ob eine Kette von Lexemen ein Phrasem darstellt. Dafür braucht man *zusätzlichen Input* (reicht von einfachen Wortlisten bis zu komplexen Repräsentationen in Lexikon und Grammatik).

Verschiedene Arten von Phrasemen erfordern verschiedene Strategien:

1. **fixer Ausdruck** (*Schritt für Schritt*) – Abgleich mit Phrasemlisten
2. auftretende **potentielle Phrasemkonstituenten** erlauben die Entscheidung, ob es sich um ein Phrasem handelt (*{Katze, Sack, lassen}* vs. *{Katze, Haus, lassen}*)
– Miteinander-Vorkommen, aber nicht Nebeneinander-Vorkommen
3. Phraseme, die aufgrund von **syntakt. Beschränkungen** identifiziert werden können (*Blinder Passagier* vs. *Der Passagier ist blind*) – erfordert syntaktische Analyse (Parsing)

Phrasemidentifikation ist entscheidend für MT-Systeme. Sie wird Teil des Syntax-Parsers (s.o. Punkt 3) und damit werden Phraseme mit dem übrigen Teil des Systems gleich behandelt.

2.3 Semantische Repräsentation im Lexikon

- Für **Information Retrieval, Information Extraction**, einfache **Dialogsysteme** und **Sprachein- / -Ausgabe** werden Phraseme i.d.R. *monolexikalisiert*, also wie einzelne Lexeme behandelt.
- fortgeschrittenere Repräsentationsformen braucht man für **Wörterbücher** (e.g. Abbilden von Synonymie zwischen Mehrwort-Phrasem und einfachem Lexem)

Gängige Methoden zur Repräsentation von Phrasemsemantik sind:

- Gleichsetzung mit einem Einwort-Synonym
- Paraphrasieren – erlaubt, die innere syntaktische Struktur beizubehalten, bestimmte Konstituenten zu erweitern und die Möglichkeit von Attribuierungen abzuleiten.

Bsp: *jmdm. einen [schönen] Bären aufbinden*

2.4 Standardisierung

Bisher gibt es nur Versuche in Richtung eines standardisierten Beschreibungsschemas, um wiederverwendbare phraseologische Wörterbücher zu ermöglichen. Die Abkürzungen ISLE und XMELLT sind Hyperlinks und führen zu den entsprechenden Projektseiten im Netz.

- ISLE – International Standards for Language Engineering
- XMELLT – XML-basiertes Format zur Codierung morphosyntaktischer und semantischer Eigenschaften von Phrasemen
- Odijk (2004) – Vorschlag einer Beschreibung von Idiomen mit Hilfe syntaktischer Muster (*patterns*) und Listen von Lexemen, die in die Muster gefüllt werden. Er konnte mit 481 patterns 80% der über 14.000 Idiome einer Datenbank erfassen