

Entwurf und Entwicklung eines (semi-)automatischen semantischen Annotationssystems für deutschsprachige Korpora

59. Jahrestagung

des Leibniz-Instituts für Deutsche Sprache

Dieses Forschungsprojekt wird durch das Förderprogramm Formación de Profesorado Universitario vom spanischen Bundesministerium für Universitäten unter dem Förderkennzeichen FPU21/00188 gefördert





Erste Beobachtung:

Textkorpora verfügen in der Regel über keine semantische relationale und ontologische Annotation. Daher lassen sich sprachliche Muster nur schwer disambiguieren.

Zweite Beobachtung:

Mithilfe von Textkorpora lassen sich formale und morphosyntaktische Daten quantitativ erheben.

[lemma="Debatte"][tag="ART\.(Def|Indef)|PRO.(Dem|Poss).Attr]).Gen.*"] CQL-Abfrage: [tag="N.*"]

Ergebnisse: Debatte der Gegenwart | Debatte der Historiker | Debatte der Hasssprache

Forschungsfrage

vorgestellt) — erstellt werden?

Hypothese:

Automatisch getaggte Daten mit semantischer — kategorieller und relationaler (Engel, 2004) — Information tragen zur Operationalisierung und Beschreibung von Argumentstrukturmustern bei.

Forschungsfragen:



Ist die Entwicklung eines (semi)automatischen Annotationssystems und seine





Theoretische Grundlagen

Semantische Valenz in der nominalen Argumentstruktur

Substantive verfügen über die Fähigkeit, als Prädikate zu fungieren.

Eine Argumentstruktur ist "die semantische Repräsentation einer Proposition und besteht […] aus genau einem Prädikat und der ihr zugeordneten Liste von Argumenten [...]." (Engelberg, 2019)

Die semantische Valenz "reflektiert den Umstand, dass Wörter (als Valenzträger) bestimmte Kontextpartner mit bestimmten Bedeutungsmerkmalen (als Valenzpartner) fordern, andere Kontextpartner mit anderen Bedeutungsmerkmalen aber ausschließen. Die semantische Valenz regelt somit die Besetzung von Leerstellen mit Klassen von Partnern, die semantisch durch bestimmte Bedeutungsmerkmale festgelegt sind." (Helbig 1992, S. 8)

Theory of Norms and Exploitations (TNE) | Corpus Pattern Analysis (CPA)

Um die Bedeutungsnuancen lexikalischer Einheiten festlegen zu können, schlägt Hanks (2013) einen Ansatz (TNE) vor, auf dem die Herangehensweise CPA beruht. Die Grundidee besteht darin, dass durch Abstraktionen der Bedeutungsmerkmale und die daran anlehnende Erstellung lexikalischer Sets das semantische — auch kombinatorische — Verhalten von Wörtern leichter zu fassen ist.

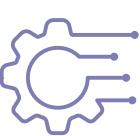
PORTLEX, MultiGenera und MultiComb

PORTLEX



Mehrsprachiges Wörterbuch zur Valenz des Nomens mit Beschreibung der syntaktischsemantischer Schnittstelle.

MultiGenera und MultiComb



Automatische Generatoren von Argumentstrukturen des Nomens mittels einer Filterung von Bedeutungsmerkmalen.

MultiTools



Erstellung von semantisch annotierten Vokallisten mit Rückgriff auf Wordnet.

Erste Ebene Zweite Ebene Dritte Ebene Methode menschliches Wesen menschlicher Umstand Erziehung Dozentin Zahnärztin menschliches Wesen menschlicher Umstand Medizin Wahre Positive Entwicklung eines Skripts ad Bestimmung formaler CQLhoc. Abgleich zwischen den aus Auswahl einer semantisch Auswertung der Ergebnisse Wahre Negative Optimierung des Skripts und Abfragen zur Erhebung nnotierten Vokabelliste, wie Korpora stammenden Daten mittels einer weitere Tests. linguistischer Daten. und den lexikalischen z.B. {menschlich, Beruf}. Falsche Positive Konfusionsmatrix. Einheiten in der Vokabelliste. Falsche Negative

Vorläufige Ergebnisse

Tatsächlich Präzision: 9.9% Spezifizität: 99,66% Sensitivität: 45,92% 11.6% 78.3% F1-Maß:

Auf einen zweckgemäß funktionierenden Prototyp hinweisende Faktoren

Lexikalische Einheiten, die die Funktion anderer Ergänzungen erfüllen, werden nicht annotiert. Das Annotationssystem annotiert Genitivrealisierungen, die als Subjektivergänzungen fungieren (z.B. Debatte der Republikaner). Andere Ergänzungen (wie z.B. die Objektivergänzung, Debatte der Frage) werden nicht getaggt. Dieser Standpunkt entspricht dem Zweck unserer Untersuchung. Auch lexikalische Einheiten, die der Rolle einer Angabe entsprechen, werden nicht markiert (Debatte der Woche), was auch das Ziel ist.

Zu verbessernde Aspekte

Lexikalische Einheiten, die in der ausgewählten Vokabelliste nicht vorhanden sind, werden nicht getaggt. Das ist der Fall der meisten Komposita bei Nominalphrasen wie Debatte der Parteivorsitzenden.

Lösungsvorschlag: Anwendung regulärer Ausdrücke wie .*vorsitzende

Eigennamen, Akronyme und Abkürzungen werden nicht annotiert.

Partizipien I und II, die als Nomina fungieren, werden nicht anerkannt und nicht getaggt.

Lösungsvorschlag: Einbeziehung von Partizipien in die Vokabelliste oder Umsetzung regulärer Ausdrücke wie .*nden

Lexikalische Einheiten, die reguläre Polysemie aufweisen, werden nicht annotiert. Das ist der Fall von Strukturen, die mindestens zwei Lesartı: Der Bericht ist in der Zeitung zu finden. Lesarten aufweisen: z.B. Bericht der Zeitung. Lesart2: Es wird über die Zeitung (als Institution) berichtet.

Ausgewählte Literatur: Arias Arias, Iván (2022): Anotação semântica (semi)automática de corpora: a frase nominal em alemão (Masterarbeit). Braga: RepositoriUM. https://hdl.handle.net/1822/79649 (Stand: 22.12.2022).

Domínguez Vázquez, María José (2011): Kontrastive Grammatik und Lexikographie: spanisch-deutsches Wörterbuch zur Valenz des Nomens. München: Niemeyer.

Domínguez Vázquez, María José (2022): Contribución de la semántica combinatoria al desarrollo de herramientas digitales multilingües. In: Círculo de Lingüística Aplicada a la Comunicación, 90, S. 171-188. DOI: 10.5209/clac.73849. Engel, Ulrich (2004): Deutsche Grammatik. Neubearbeitung. München: Iudicum. Engelberg, Stefan (2019): Argumentstrukturmuster. Ein elektronisches Handbuch zu verbalen Argumentstrukturen im Deutschen. In: Czicza, Dániel /Dekalo, Volodymyr/ Diewald, Gabriele (Hg.): Konstruktionsgrammatik IV.

Varianz in der konstruktionalen Schematizität. Tübingen: Staffenburg, S. 13-38. Hanks, Patrick (2013): Lexical Analysis. Norms and Exploitations. Cambridge: MIT Press.

Teubert, Wolfgang (1979): Valenz des Substantivs. Attributive Ergänzungen und Angaben. Düsseldorf: Schwann.

Ontologie = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D. (2021). Ontología léxica. http://portlex.usc.gal/ontologia/. [Stand: 02/01/2023]. PORTLEX = Domínguez V., M. J., Valcárcel R., C., Mirazo B., M., Sanmarco B., M. T., Simões, A. & Vale, M. J. (2018). Portlex. Diccionario multilingue de la valencia del nombre. Universidade de Santiago de Compostela. http://portlex.usc.gal/portlex/ [Stand: 20/01/2023].

Sketch Engine = Lexical Computing CZ S.R.O. (n. d.). Sketch Engine. https://www.sketchengine.eu [Stand: 02/02/2023].

español y francés. Universidade de Santiago de Compostela. http://portlex.usc.gal/combinatoria/usuario [Stand: 02/02/2023].

Simulation: Genitivrealisierungen in der Funktion einer Subjektivergänzung.

Xera = Domínguez V., M. J., Valcárcel R., C., Bardanca O., D., Calañas C., J. A., Catalá T., N., López I., N., Martín G., R., Mirazo B., M., Sanmarco, B., M. T. & Pino S., L. (2020). Xera. Prototipo online para la generación monoargumental de la frase nominal en alemán,

Helbig, G. (1992). Probleme der Valenz- und Kasustheorie. Berlin, Boston: Max Niemeyer Verlag.

Online Ressourcen: