Aufbau eines historisch-literarischen Metaphernkorpus für das Deutsche

Pernes, Stefan
stefan.pernes@uni-wuerzburg.de
Universität Würzburg, Deutschland

Keller, Lennart
jan.keller@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

Peterek, Christoph
christoph.peterek@stud-mail.uni-wuerzburg.de
Universität Würzburg, Deutschland

1. Überblick

Metaphorischer Sprachgebrauch umfasst komplexe gedankliche Würfe genauso wie alltägliche Begrifflichkeiten. Die Metapher gilt als Untersuchungsgegenstand nicht nur in den Literaturwissenschaften, Sprachwissenschaften und der Anthropologie, sondern hat auch Relevanz für so disparate Forschungsprogramme wie das der Künstlichen Intelligenz und der Kritischen Diskursanalyse. Darüber hinaus stellt die Erkennung und Auflösung von Metaphern ein wichtiges Desiderat in sprachtechnologischen Anwendungen dar, deren Gegenstand die Disambiguierung von Wortbedeutungen umfasst. Korpusuntersuchungen zeigen, dass metaphorischer Sprachgebrauch in gängigen Textsorten durchschnittlich in jedem dritten Satz zu finden ist (vgl. Steen et al. 2010; Shutova und Teufel 2010) – ein Beleg für die Ubiquität der Metapher, die in erster Linie darin begründet liegt, dass die idealtypische Karriere eines metaphorischen Ausdrucks als kühne Betonung beginnt und als konventionelle Form endet. Eine Sprachressource zum Metapherngebrauch kann also eine wichtige Ergänzung bei der automatischen inhaltlichen Erschließung von Textbeständen darstellen. Dabei stellt das hier entwickelte Korpus annotierter Sätze, dessen Grundlage eine Sammlung deutschsprachigen Romane aus dem 19. Jahrhunderts bildet, einen spezifischen Beitrag zur Erschließung von historischen Textbeständen dar.

Die große Mehrheit der heute verfügbaren Metaphernkorpora basiert auf dem Prinzip, einige wenige Zielbegriffe sowie unter Umständen ausgewählte konzeptuelle Domänen zu definieren und alle passenden sprachlichen Realisierungen aus einem großen Textbestand zu extrahieren. Diese Herangehensweise lässt vermuten, dass die damit modellierten Eigenschaften sich nicht auf arbiträren Text in realen Anwendungsszenarien übertragen lassen, denn jedes vordefinierte lexikalische oder konzeptuelle Inventar wird dabei zu kurz greifen (vgl. Shutova 2015). Im Gegensatz dazu enthält das hier entwickelte Korpus keine Einschränkungen hinsichtlich der konzeptuellen Domänen oder der erfassten sprachlichen Konstruktionen, bis auf die Tatsache, dass es sich aus literarischen Prosatexten zusammensetzt. Es sollte noch darauf hingewiesen werden, dass mit der Hamburg Metaphor Database eine weitere deutschsprachige Ressource zur Metapher existiert, diese jedoch nach wesentlich anderen Gesichtspunkten erstellt wurde und lediglich eine kleine Zahl ausgewählter Beispielsätze enthält.

1. Korpuserstellung

Grundlage für die Erstellung des Korpus bildet die Romansammlung der Digitalen Bibliothek des Projektes TextGrid. Die Sammlung umfasst insgesamt 454 Werke vom frühen 16. bis zum frühen 20. Jahrhundert, wobei der Bedarf nach orthographisch normalisiertem Text die Datengrundlage auf 383 Romane aus den Jahren 1830 bis 1940 eingeschränkt hat. Zur Ziehung der zu annotierenden Sätze wird eine balancierte Sampling-Strategie hinsichtlich zeitlicher Streuung und Gender der AutorInnen eingesetzt. Es handelt sich dabei um eine Quotenstichprobe, die aus jedem 10-Jahres-Abschnitt und zu gleichen Teilen männlicher und weiblicher Autorinnen Sätze auswählt. Darüber hinaus wird im Rahmen des Samplings eine automatische Vorauswahl getroffen, sodass die Hälfte der Sätze Metaphern enthält. Möglich wird dies durch einen Classifier, der anhand von TF-IDF Scores – auf Grundlage einer lemmatisieren Version des gesamten Romankorpus – feststellen kann wie “ungewöhnlich” ein zu klassifizierender Satz ist. Anhand eines empirisch festgestellten, von der Größe des TF-IDF Korpus abhängigen, Schwellenwerts ist es anschließend möglich, eine Vorauswahl zu treffen, die indirekt Metaphorizität erfasst. Es handelt sich dabei um eine vereinfachte Form des von Schulder & Hovy (2014) entwickelten Klassifikationsansatzes. Ziel des hier entwickelten Korpus ist es, einen Gesamtumfang von bis zu 2000 annotierten Sätzen zu erreichen. Als Grundlage dafür wurden insgesamt 3000 Sätze ausgewählt.

1. Annotation

Wir orientieren uns an der Metaphor Identification Procedure (MIP) der Pragglejaz Group (Pragglejaz Group 2007; Steen et al. 2010) und sehen zunächst jedes Wort im Text als potentielle Metapher. Gegenstand der Metaphernannotation ist es somit, jedes Wort als metaphorisch beziehungsweise nicht metaphorisch zu klassifizieren. Die Aufgabe ist dabei auf metaphorische Äußerungen auf der Wortebene beschränkt, das heißt Satzmetapher und Textmetapher sowie Phänomene grammatischer Metapher sind ausgenommen. Aufgrund der Neigung des Deutschen

zur Kompositabildung wird jedoch eine automatische Kompositazerlegung durchgeführt. Da der Umfang der zu annotierenden Sätze eine Herausforderung für eine solche detallierte Herangehensweise wie das MIP darstellt, wird eine automatische Vorselektion potentieller Metaphernkandidaten durchgeführt.

Auf Grundlage von Part-of-Speech-Informationen und Dependency-Bäumen werden aus den Sätzen folgende Konstruktionstypen als Kandidaten für eine metaphorische Verwendung extrahiert (zu den Typen vgl. Skirl und Schwarz-Friesel 2007): Substantivmetapher – dazu gehören Komposita, Kopulakonstruktionen ("X ist ein Y"), Simile ("X ist wie ein Y"), Genitivmetapher – sowie Verb-, Adjektiv-, Präpositions- und 'als'-Metapher. Wir folgen mit diesem Ansatz der automatischen Vorauswahl Gandy et al. (2013), die dadurch bei der Metaphernauszeichnung eine Übereinstimmung der Annotatoren von Kappa = 0.80 erreichen. Die extrahierten Konstruktionen werden anschließend zusammen mit den Sätzen für die Annotation in ein geeignetes Format exportiert und in die Annotationsumgebung WebAnno (Yimam et al. 2014) geladen.

Für den manuellen und bei weitem umfassendsten Teil der Arbeit wurde ein Annotationsleitfaden verfasst, der die Identifikationsstrategie MIP reproduziert, Hinweise zum Umgang mit lexikalisierten metaphorischen Ausdrücken und der Abgrenzung zur Metonymie enthält. Darüber hinaus ist dargestellt, welche Konstruktionstypen vormarkiert werden und welche Ausnahmen dabei zu erwarten sind (Eigennamen und Hilfsverben sind von der Markierung ausgenommen, vgl. Shutova und Teufel 2010). Schließlich wird festgelegt wie mit Ausnahmen und fehlerhaften Sätzen zu verfahren ist. Satzfragmente, starke dialektale Formen sowie Sätze, die ohne Kontext nicht interpretierbar sind, werden als Ausschuss markiert, fehlerhafte Vormarkierungen werden gekennzeichnet und im Rahmen der Kuration der annotierten Sätze verbessert.

1. Ergebnis

Es kann von den folgenden vorläufigen Ergebnissen der automatischen Vorauswahl und der manuellen Annotation berichtet werden:

Die automatische Extraktion der möglicher Metaphernkandidaten hat es ermöglicht, ein korpusgestütztes Bild darüber zu erlangen wie die Konstruktionen in einer relativ offenen Domäne – einem Romankorpus, das diverse Gattungen umfasst – verteilt sind. Des Weiteren ist ausgehend von der Annotationspraxis festzustellen, dass die erhobenen Konstruktionen – zumindest im Rahmen der hier zugrundeliegenden Texte – theoretisch alle Vorkommen von metaphorischen Äußerungen auf der Wortebene abdecken. In der Praxis kommt es jedoch aufgrund komplexer Hypotaxen und fehlender automatischer Koreferenz-Auflösung zu fehlerhaften Vormarkierungen.

Die vorbereitende, automatische Klassifikation der Sätze in Metapher beziehungsweise Nicht-Metapher führt zu einem Anteil von 48% an Sätzen, die lebendige Metaphern enthalten. Werden lexikalisierte metaphorische Ausdrücke mit eingerechnet, steigt der Anteil der Sätze, die Metaphern enthalten, auf 61%. Ein erheblicher Vorteil, der sich aus der Klassifikation der Sätze ergibt, ist die Fülle des Materials, die sich dadurch generieren lässt. Ohne Vorauswahl liegt die durchschnittliche Anzahl von metaphorischen Ausdrücken pro Satz – je nach Textsorte – zwischen 0.12 und 0.54 (vgl. Shutova & Teufel 2010), während mit dem hier vorgestellten Ansatz ein Wert von 1.91 Metaphern pro Satz erreicht wird. Eine genaue Auswertung der Präzision des Classifiers steht noch aus, in Bezug auf die Struktur der so ausgewählten Sätze kann jedoch festgestellt werden, dass die Klassifizierung keine Auswirkung auf die Verteilung der erhobenen Konstruktionstypen hat.

Für die Übereinstimmung zwischen zwei Annotatoren beim Aufbau des hier vorgestellten Metaphernkorpus kann ein Wert von 0.87 (Cohen’s Kappa) berichtet werden. Werden lediglich die vormarkierten Konstruktionen als Grundlage der Berechnung herangezogen, schwankt Kappa je nach Einbezug lexikalisierter Äußerungen zwischen 0.77 bis 0.80.

Bibliographie

Gandy, Lisa / Allan, Nadji / Atallah, Mark / Frieder, Ophir / Howard, Newton / Kanareykin, Sergey / Koppel, Moshe / Last, Mark / Neuman, Yair / Argamon, Shlomo (2013): „Automatic identification of conceptual metaphors with limited knowledge“, in: *Proceedings of AAAI 2013*.

Schulder, Marc / Hovy, Eduard (2014): „Metaphor detection through term relevance“, in: *Proceedings of the Second Workshop on Metaphor in NLP*.

Shutova, Ekaterina / Teufel, Simone (2010): „Metaphor corpus annotated for source - target domain mappings“, in: *Proceedings of LREC 2010* 3255–3261.

Shutova, Ekaterina (2015): „Design and Evaluation of Metaphor Processing Systems“, in: *Computational Linguistics* 41 (4): 579–623.

Skirl, Helge / Schwarz-Friesel, Monika (2007): *Metapher*. Universitätsverlag Winter.

Steen, Gerard J. / Dorst, Aletta G. / Herrmann, J. Berenike / Kaal, Anna A. / Krennmayr, Tina / Pasman, Trijntje (2010): *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam / Philadelphia: John Benjamins.

Yimam, Seid Muhie / Eckart de Castilho, Richard / Gurevych, Iryna / Biemann Chris (2014): „Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno“, in: *Proceedings of ACL-2014*, demo session, Baltimore, MD, USA.