»Die Greta Garbo der Leichtathletik« – Eine systematische Analyse der Modifier vossianischer Antonomasien mithilfe von Word Embeddings

Schwab, Michel
michel.schwab@hu-berlin.de
Humboldt-Universität zu Berlin

Fischer, Frank
fr.fischer@fu-berlin.de
Freie Universität Berlin

Einführung und Forschungsstand

Die vossianische Antonomasie (VA) ist ein rhetorisches Stilmittel aus der Familie der Antonomasien, eng verwandt mit Metonymie und Metapher. Während bei der klassischen Antonomasie ein Eigenname durch eine typische Eigenschaft ersetzt wird (wenn etwa Michael Schumacher als »der Kerpener« bezeichnet wird), funktioniert die vossianische Antonomasie genau umgekehrt. Hier wird ein typisches Merkmal einer Person durch den Eigennamen einer anderen Person evoziert.

Wenn ein Journalist zum Beispiel Wilson Kipketer, den dänischen Mittelstreckenläufer kenianischer Herkunft, als »Greta Garbo der Leichtathletik« bezeichnet, wird eine typische Eigenschaften der Filmschauspielerin aufgerufen, in diesem Fall ihre distanzierte, zurückhaltende Art, wie dieses Zitat aus der New York Times zeigt: »Kipketer is as guarded [zurückhaltend] as he is fast; some reporters have labeled him the Greta Garbo of track and field.« (NYT, 8. August 1997).

Eine vossianische Antonomasie setzt sich im Normalfall aus drei Teilen zusammen: dem Target (Wilson Kipketer), der Source (Greta Garbo) und dem Modifier (Leichtathletik) (vgl. Bergien 2013). Der Modifier verschiebt eines oder mehrere Merkmale der Source in das Umfeld des Targets. In dieser Arbeit konzentrieren wir uns auf die systematische Analyse des Modifiers.

Die automatisierte Erkennung und Extraktion vossianischer Antonomasien hat sich in den letzten fünf Jahren rasch ausdifferenziert. Während Jäschke et al. 2017 und Fischer et al. 2019 semi-automatisierte Verfahren nutzten, um VA-Ausdrücke in großen Zeitungskorpora ausfindig zu machen, setzten Schwab et al. (2019, 2022) automatisierte Verfahren ein, die meist auf neuronalen Netzen basierten.

Da wir mit einem großen Korpus und Word Embeddings arbeiten, ist unser Forschungsbeitrag der erste, der eine quantitative Untersuchung dieses Phänomens mit einer thematischen Gruppierung der verschiedenen Modifier verbinden kann. Unsere Forschungsergebnisse stellen wir auch über eine interaktive Visualisierung bereit (https://vossanto.weltliteratur.net/dhd2023/modifier.html).

Datensatz

Wir nutzen den VA-Datensatz aus Schwab et al. 2019, welcher mittels eines semi-automatisierten Verfahrens aus dem New York Times -Korpus (Sandhaus 2008) generiert wurde. Das NYT-Korpus besteht aus mehr als 1,8 Mio. Zeitungsartikeln der NYT aus den Jahren 1987–2007. Mit Hilfe des regulären Ausdrucks

\\ b(the|an?)\\s+([\w.,'-]+\\s+){1,5}?(of|for|among)\\b

wurden Kandidatensätze ermittelt, d.h. alle Sätze, die Phrasen enthalten, welche mit »the«, »a« oder »an« anfangen und mit »of«, »for« oder »among« enden, wobei zwischen diesen beiden Polen ein bis fünf Wörter platziert sein können. Die Wörter zwischen Anfangs- und Endsignal stellten die potenzielle Source-Phrase dar und wurden mit einer Wikidata-Liste abgeglichen, die alle Entitätennamen aus Wikidata (inkl. Aliasse) enthielten, die die Eigenschaft ›instanceOf‹ ›human‹ aufweisen. Die Source-Kandidaten wurden also auf Menschen beschränkt, die in Wikidata verzeichnet sind (dabei handelt es sich um eine bewusste Beschränkung bei der Untersuchung des Phänomens – VA können auch mit Orten, Markennamen, Comicfiguren usw. operieren). Anschließend wurden diese Kandidaten mit einer manuell erstellten Sperrliste abgeglichen, um falsche Kandidaten auszuschließen.

Dieser Datensatz wurde in Schwab et al. 2022 verfeinert. Alle VA-Phrasen (Target, Source, Modifier) wurden auf Wortebene innerhalb der Sätze annotiert. Insgesamt enthält der Datensatz 5.995 Sätze, davon enthalten 3.066 einen VA-Ausdruck und 2.929 enthalten keinen, sind aber syntaktisch ähnlich aufgrund des genutzten regulären Ausdrucks.

In Tabelle 1 sind die zehn häufigsten Modifier des Datensatzes aufgelistet. Die häufigsten Ausdrücke sind temporale Ausdrücke (»his day«, »his time«, »the 90s«), geografische Angaben (»Japan«,» China«) und Sportarten (»tennis«, »baseball«, »ballet«).

Tabelle 1: Die zehn häufigsten Modifier im Datensatz inklusive ihrer Häufigkeit.

Modifier	Anzahl
his day	56
his time	35
Japan	32
the 90s	21
China	17
our time	17
tennis	16
his generation	16
baseba ll	16
her time	14

Methode

Wir nutzen kontextabhängige Word Embeddings, um die Modifier-Phrasen in hochdimensionale Vektoren zu transformieren, die die Semantik des Textes wiedergeben sollen.

Mit Hilfe von Word Embeddings wurden in den letzten Jahren viele Benchmarks im Bereich Natural Language Processing erstellt. Insbesondere kontextabhängige Word Embeddings, d.h. die numerische Repräsentation von Wörtern und Tokens in Abhängigkeit ihres Kontexts, haben viel Aufmerksamkeit auf sich gezogen. Der Vorteil dieser Word Embeddings im Gegensatz zu kontextunabhängigen Word Embeddings ist die Möglichkeit, Homonyme korrekt darzustellen. Wir benötigen die numerische Repräsentation der Phrasen, um anschließend ein Clustering-Verfahren durchführen zu können, welches die Modifier in Themenbereiche gruppieren soll.

Wir greifen auf Sentence-Transformers zurück, welches aus Sentence-BERT (Reimers et al. 2019) hervorgegangen ist. Das Modell basiert auf transformerbasierten Sprachmodellen wie BERT (Devlin et al. 2019). Im Gegensatz zu BERT wird S-BERT allerdings mittels einer siamesischen Netzwerkstruktur trainiert, der Output wird durch eine Pooling Operation in einen hochdimensionalen Vektor transformiert. Dadurch kann das trainierte Modell effizient semantische Ähnlichkeiten zwischen Texten errechnen. Wir nutzen das Modell »all-mpnet-base-v2«, welches die besten Resultate in der Anwendung auf verschiedene Datensätze zeigte (siehe https://www.sbert.net/docs/pretrained_models.html).

Dies wenden wir auf die einzelnen Modifier an. Das Netzwerk liefert für jeden Modifier einen 768-dimensionalen Vektor. Diese numerischen Vektoren lassen sich nun durch ein Clustering-Verfahren gruppieren.

Wir entscheiden uns für den k-means-Algorithmus (MacQueen 1967), um die Vektoren in Cluster einzuteilen. Wir nutzen k-means aufgrund verschiedener Annahmen. Einmal gehen wir davon aus, dass es relativ wenige Ausreißer gibt, da die VA-Ausdrücke aus dem Datensatz häufig in ähnlichen Themengebieten in der New York Times vorkommen (vgl. Fischer et al. 2019). Außerdem können wir die Anzahl der Cluster angeben und diese während der Analyse variieren, um zu beobachten, wie sich die Gruppierungen in Abhängigkeit davon verhalten. Dies funktioniert mit dichtebasierten Clustering-Algorithmen nicht so einfach. Da k-means in der Berechnung der Cluster die quadrierte euklidische Distanz nutzt, normalisieren wir die Output-Vektoren, da die normalisierte quadratische euklidische Distanz proportional zur Kosinus-Distanz ist, welche in Reimers et al. 2019 genutzt wird, um die Ähnlichkeit zwischen zwei Vektoren zu berechnen.

Im Anschluss an das Clustering möchten wir den einzelnen Clustern Themen zuordnen, durch ein an das »Topic Modeling« angelehntes Verfahren. Stark vereinfacht basieren klassische Topic-Modeling-Modelle auf der Annahme, dass Wörter, die besonders häufig gemeinsam in Sequenzen vorkommen, ein abstraktes Thema bilden. Meist wird Topic Modeling auf längere Dokumente angewandt, bei denen von signifikanten Wort-Überschneidungen ausgegangen werden kann. 97 Prozent der Modifier-Phrasen bestehen jedoch aus einem bis vier Wörtern und weisen dadurch kaum Überschneidungen auf. Somit sind sie für klassisches Topic Modeling ungeeignet. Selbst beim sogenannten Short Text Topic Modeling wird mit Textsorten wie Tweets oder Rezensionen trainiert, welche immer noch bedeutend länger sind als unsere Phrasen.

Wir nutzen stattdessen den Vorteil, dass viele der Formulierungen Nominalphrasen oder Nomen sind. Dadurch sind sie unter anderem im WordNet (Fellbaum 1998) zu finden, einer lexikalischen Datenbank, die Wortbedeutungen, Synonyme und viele andere Features bereitstellt. Das Projekt WordNet Domains (Bentivogli et al. 2004) hat zusätzlich jedem Wort bzw. jedem Synset (Gruppe ähnlicher Wörter) in WordNet semi-automatisch ein oder mehrere Domains zugeordnet, die für uns als Themengebiete genutzt werden können. Diese Domains sind hierarchisch gegliedert. Dies nutzen wir aus und weisen jeder Modifier-Phrase, soweit vorhanden, ihre Domains zu. Vorher nutzen wir noch das NLTK Toolkit (Bird et al. 2009), um alle Stoppwörter zu entfernen und die Ausdrücke dadurch auf die Nomen zu reduzieren, zum Beispiel »her time« zu »time« oder »the harmonica" zu »harmonica«. Sollte die übrigbleibende Phrase im WordNet nicht vorhanden sein, teilen wir sie in ihre einzelnen Wörter auf und verfahren wie oben beschrieben für jedes Wort der Phrase. Zum Schluss weisen wir die am häufigsten vorkommende Domain der Phrasen je Cluster dem jeweiligen Cluster als Themengebiet zu. In der Web-App kann man sich zusätzlich die zehn hochfrequentesten Domains anschauen.

Anschließend können wir die Cluster visualisieren. Da die Vektoren hochdimensional sind, nutzen wir verschiedene Dimensionsreduktionsalgorithmen, um sie auf zwei Dimensionen zu reduzieren. Wir vergleichen mehrere Algorithmen – PCA (Hauptkomponentenanalyse, Pearson 1901), t-SNE (t-distributed stochastic neighbor embedding Methode, van der Maaten 2008), UMAP (Uniform Manifold Approximation and Projection, McInnes et al. 2018), IVIS (Szubert et al. 2019) –, welche in der Web-App ausgewählt werden können. Nach einigen Durchläufen hat sich die Kombination von PCA und t-SNE als bestes Verfahren herausgestellt, welches wir kurz vorstellen. Wir wenden zuerst PCA an und reduzieren die Vektoren auf eine Länge von 50. Die Hauptkomponentenanalyse vereinfacht Daten, indem die Einträge der Vektoren durch eine geringere Zahl möglichst aussagekräftiger Linearkombinationen (die Hauptkomponenten) genähert werden. Zusätzlich nutzen wir t-SNE, um die 50-dimensionalen Vektoren auf zweidimensionale Vektoren zu reduzieren. Der Vorteil von t-SNE im Gegensatz zu PCA liegt in der Möglichkeit, nichtlineare Abhängigkeiten darzustellen. t-SNE reduziert die Vektoren außerdem so, dass Vektoren, die in der höheren Dimension eine kurze Distanz haben, auch in der reduzierten Dimension eine kurze Distanz zueinander haben. Dadurch wird die lokale wie auch globale Struktur bewahrt.

Erkenntnisse

Die Modifier für vossianische Antonomasien fallen im New York Times -Korpus sehr vielgestaltig aus und sind

nicht auf bestimmte Phrasen oder Wortgruppen limitiert. Abbildung 1 zeigt beispielhaft die Visualisierung mit neun Clustern, wobei die Themen von uns zunächst manuell zugeordnet wurden. Einige der Cluster lassen sich bis auf wenige Ausnahmen eindeutig bestimmten Themen zuordnen, wie Sport, Musik und Tanz, Kunst, Film und Literatur, Geografie, Politik, Wirtschaft oder temporale Ausdrücke.

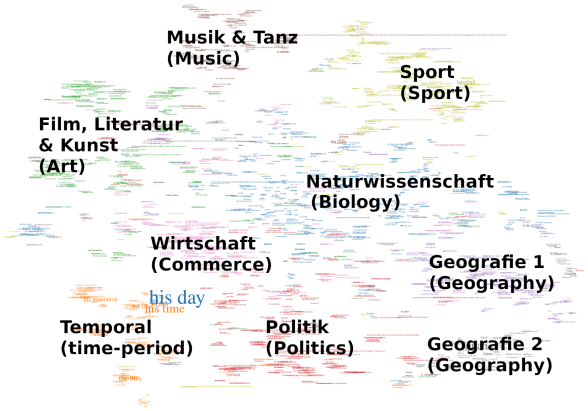

Abbildung 1: Die Abbildung zeigt die Visualisierung des Clusterings nach Dimensionreduktion mit neun Clustern. Den Clustern wurde manuell ein Themengebiet zugeordnet.

Die automatisch gefundenen Themen durch WordNet Domains stimmen mit den von uns manuell zugewiesenen Themen fast vollständig überein, wie Abbildung 1 zeigt. Den von uns manuell annotierten Themen sind in Klammern die automatisch gefundenen Themen beigesellt.

Der blaue Cluster in der Mitte ist allerdings sehr divers und lässt sich nicht genau einer Kategorie zuordnen. Hier tauchen Flora und Fauna auf (»the pumpkins, »Rottweilers«), was zu dem automatisch gefundenen Themengebiet Biologie passen würde. Allerdings kommen auch »space wear«, »Buddhism« sowie »soft drinks« und »the physics world« vor. Wir haben uns für das Rubrum Naturwissenschaft entschieden, welches auf einen Großteil der Phrasen zutrifft.

Einer der Gründe für die Zusammensetzung dieses diversen Clusters ist der Umstand, dass k-means keine Ausreißer zulässt und daher jeder Punkt genau einem Cluster zugeordnet wird. Dadurch finden sich auch Phrasen, die eigentlich nicht in ein bestimmtes Themengebiet gehören oder für die es eigentlich zu wenige ähnliche Phrasen gibt, in einem Cluster wieder.

Ein anderer Grund ist die Diversität der Modifier. Viele Modifier bestehen nicht nur aus einem, sondern aus mehreren Wörtern. Diese Phrasen könnte man verschiedenen Themengebieten oder Subgenres zuordnen, z.B. »ancient Alexandria« (Temporal, Geografie), »Korean radio« (Geografie, Technologie) oder »food writing« (Speisen und Getränke, Literatur). Dies sind auch häufig Phrasen, die durch die Dimensionsreduktion nicht in der Nähe der anderen Phrasen des Clusters liegen, weil zum Beispiel »food writing« in die Nähe von anderen kulinarischen Phrasen verortet wurde, obwohl es ein Subgenre der Literatur ist. An diesem Beispiel sieht man, dass das Clustering-Verfahren das Wort richtig zugeordnet hat (»food writing« gehört zum kulturellen Cluster), aber falsch visualisiert wurde.

Abhängig von der Anzahl der Cluster unterteilt sich zum Beispiel die Kultur nach und nach in Subgenres wie Literatur, Musik, Tanz, Film/TV oder Kunst. Dies kann man in Abbildung 2 gut beobachten. Der linke Teil von Abbildung 2 zeigt einen Ausschnitt der Visualisierung, in der sechs Cluster gebildet wurden. Hier sind die meisten kulturbezogenen Phrasen in einem einzigen Cluster (grün) gruppiert. Im rechten Teil ist der gleiche Ausschnitt zu sehen, allerdings mit zwölf Clustern. Man kann gut erkennen, dass sich der kulturelle Cluster fast vollständig in drei neue Cluster (grau, orange, blau) aufgeteilt hat, nämlich »Kunst«, »Literatur und Film/TV« und »Musik und Tanz«. Auch hier gibt es Grenzfälle wie zum Beispiel »musicals«. Das Clustering hat die Phrase zu »Musik und Tanz« gruppiert, wohingegen in der Visualisierung das Wort in die Nähe des Film-Clusters gerückt wurde, in dem auch theaterbezogene Themen auftauchen.

Auch die Geografie teilt sich mit wachsender Anzahl an Clustern in zwei Hälften, wobei in der einen ein Großteil der US-amerikanischen Geografika angesiedelt sind, allerdings nicht ausschließlich.

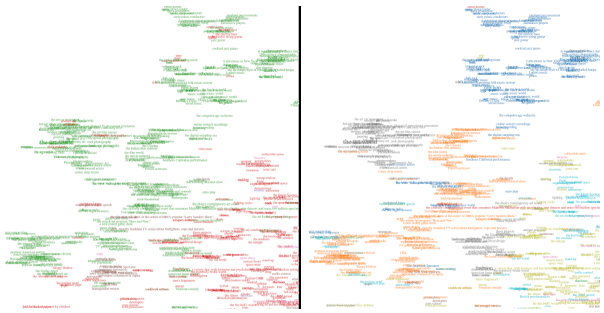

Abbildung 2: Die Abbildung zeigt einen Ausschnitt der Visualisierung mit sechs Clustern auf der linken Seite und zwölf Clustern auf der rechten Seite. Der Ausschnitt zeigt den Großteil der kulturellen Phrasen.

Wie oben bereits angemerkt, stellen wir eine interaktive Visualisierung zur Datenexploration zur Verfügung, in der die oben beschriebenen Fälle nachvollzogen werden können. Die verschiedenen Dimensionsreduktionsverfahren können selbst ausprobiert und die Anzahl der Cluster variiert werden (1–15). Außerdem werden die zehn am häufigsten vorkommenden Domains je Cluster gezeigt, um einen Überblick über die Themen zu bekommen. Die Größe der Labels spiegelt die Anzahl der Vorkommen im Datensatz wider. Zudem kann durch eine Bereichsauswahl gezoomt werden (https://vossanto.weltliteratur.net/dhd2023/modifier.html).

Fazit und Ausblick

Unser Ansatz lenkt den Blick von den Eigennamen in Source und Target einer vossianischen Antonomasie auf den Modifier. Wir konnten zeigen, dass bestimmte Themenfelder besonders häufig sind, also eine besondere Neigung aufweisen, in einer vossianischen Antonomasie Verwendung zu finden. Die Themen wurden in Clustern gruppiert und zweidimensional projiziert. Durch verschie-

dene Verfahren kann das Modell noch verfeinert werden, z.B. durch den Einsatz anderer Cluster- oder Reduktionsverfahren.

Mit Hilfe von Entity Embeddings kann man in Zukunft ähnliche Analysen der Source und des Targets durchführen, um etwa auf Zusammenhänge zwischen den einzelnen Teilen einer vossianischen Antonomasie zu fokussieren. So würde sich zum Beispiel erforschen lassen, in welchen semantischen Abhängigkeiten Source, Target und Modifier eines VA-Ausdrucks zueinander stehen und welche Entitäten signifikant häufig mit welchen Modifier-Gruppen genutzt werden.

Mit Hilfe der Web-App kann man die Daten und Ergebnisse interaktiv explorieren und somit weitere Erkenntnisse erlangen, welche für die automatische Erkennung, aber auch für die automatische Generierung sinnvoller vossianischer Antonomasien eine wichtige Rolle spielen können. Beide Aufgaben verfolgen wir in Zukunft.

Bibliographie

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini und Emanuele Pianta. 2004. "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing." In: COLING 2004 Workshop on "Multilingual Linguistic Resources" , Geneva, Switzerland, S. 101–108.

Bergien, Angelika. 2013. "Names as frames in current-day media discourse." In: Name and Naming. Proceedings of the Second International Conference on Onomastics. Cluj-Napoca: Editura Mega. S. 19–27.

Bird, Steven, Edward Loper und Ewan Klein. 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. https://doi.org/10.48550/arXiv.1810.04805

Fellbaum, Christiane (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fischer, Frank und Robert Jäschke. 2019. "'The Michael Jordan of greatness' – Extracting Vossian antonomasia from two decades of The New York Times, 1987–2007." In: Digital Scholarship in the Humanities 35, no. 1 . S. 34–42. https://doi.org/10.1093/llc/fqy087

Jäschke, Robert, Jannik Strötgen, Elena Krotova und Frank Fischer. 2017. "'Der Helmut Kohl unter den Brotaufstrichen'. Zur Extraktion vossianischer Antonomasien aus groβen Zeitungskorpora." In: Proceedings of DHd 2017 . Universität Bern. https://doi.org/10.5281/zenodo.4646126

MacQueen, J. 1967. "Classification and analysis of multivariate observations." In: 5th Berkeley Symp. Math. Statist. Probability . S. 281–297.

McInnes, Leland, John Healy und James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." arXiv preprint arXiv:1802.03426 .

Pearson, Karl. "LIII. 1901. On lines and planes of closest fit to systems of points in space." In: The London, Edinburgh, and Dublin philosophical magazine and journal of science 2, no. 11. S. 559–572. https://doi.org/10.1080/14786440109462720

Reimers, Nils und Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Hong Kong, China.

Sandhaus, Evan. 2008. "The New York Times Annotated Corpus." LDC2008T19. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/77ba-9x74

Schwab, Michel, Robert Jäschke, Frank Fischer und Jannik Strötgen. 2019. "'A Buster Keaton of Linguistics': First Automated Approaches for the Extraction of Vossian Antonomasia." In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) , Hong Kong, China.

Schwab, Michel, Robert Jäschke und Frank Fischer. 2022. "'The Rodney Dangerfield of Stylistic Devices' – End-to-End Detection and Extraction of Vossian Antonomasia Using Neural Networks." In: Frontiers in Artificial Intelligence 5. https://doi.org/10.3389/frai.2022.868249

Szubert, Benjamin, Jennifer E. Cole, Claudia Monaco und Ignat Drozdov. 2019. "Structure-preserving visualisation of high dimensional single-cell datasets." In: Scientific reports , 9 (1), 1–10. https://doi.org/10.1038/s41598-019-45301-0

van der Maaten, Laurens und Geoffrey Hinton. 2008. "Visualizing Data using t-SNE." In: Journal of Machine Learning Research 9: 2579–2605.