SuW GRADUIERTENPLATTFORM
Forschungsnetzwerk Sprache und Wissen

DIGITALE TAGUNG
01. BIS 02. OKTOBER 2020

Sprache und Wissen hin und zurück – iterative Annotation als linguistische Forschungsmethode

PROGRAMM

Donnerstag, 01.10.2020

09:00	Begrüßung und inhaltliche Einführung, organisatorische Hinweise
09:30-10:30	SARAH SCHRÖDER (UNIVERSITÄT PADERBORN): Direktiva in Aufrufen: Erkenntnisse einer manuellen, digitalen Annotation
10:30-11:30	ULRIKE LOHNER (UNIVERSITÄT HEIDELBERG): „In abgehobenen akademischen Sphären“: iterative Annotation in einer Stiluntersuchung rechter Drohbriefe und linker Bekennerschreiben
11:30	Pause
12:00-12:00	JÖRAN LANDSCHOFF (UNIVERSITÄT HEIDELBERG): Sprache im Netz – Wege zu einer soziolinguistischen Korpusannotation
13:00	Ende des ersten Tages

Freitag, 02.10.2020

09:00-10:00	THOMAS JURCZYK (UNIVERSITÄT BOCHUM): Projektvorstellung „ReligionML – Annotation religiöser Texte für Machine Learning“
10:00-11:00	JÖRN STEGMEIER (TECHNISCHE UNIVERSITÄT DARMSTADT): Maschinelle Konformitätsprüfung: Entwicklung eines Annotationsschemas zur Identifizierung semantischer Unsicherheit in Normtexten
11:00	Pause
11:30-12:30	ANA SCHENK (TECHNISCHE UNIVERSITÄT DARMSTADT): Vorurteilsfrei annotieren? Die Rolle des Vorwissens im qualitativ-hermeneutischen Annotationsprozess
12:30-13:00/13:30	Abschlussdiskussion
13:00/13:30	Ende der Tagung

ORGANISATORISCHES

- Die Tagung findet digital via Zoom statt; die Vorträge werden live gehalten
- Wer an der Tagung teilnehme möchte, den bitten wir um eine Anmeldung per Mail bis zum 28.09.2020 an Michael Bender (michael.bender@tu-darmstadt.de) und Katharina Jacob (katharina.jacob@gs.uni-heidelberg.de); nach der Anmeldung erhalten Sie den Zoom-Link.
- Die Vorträge dauern 30 Minuten, weitere 30 Minuten sind für die Diskussion vorgesehen.

TAGUNGSTHEMA

Schon die analoge Kulturtechnik des Annotierens mit ihren antiken Wurzeln (Glossen und Scholien z.B.) stellt eine sprachsensible Praktik dar, durch die das Zusammenwirken von verschriftlichter Sprache und kontextualisierend-ergänzendem Wissen explizit und sichtbar gemacht wird. Annotationen dienen einerseits der Anreicherung von Dokumenten mit mehr oder weniger kontextspezifisch relevantem Wissen. Andererseits zielen Annotationen aber auch auf die Erschließung von Texten, das Erkennen von Zusammenhängen und Mustern, die Analyse und Interpretation – also das Gewinnen von Erkenntnissen aus sprachlichen Daten.

Digitale Annotation stellt als wissenschaftliche Methode keine rein mimetisch ins Digitale überführte Form einer analog geprägten Praktik dar, sondern hat sich unter den spezifischen Bedingungen der Digitalität (weiter)entwickelt und ausdifferenziert. Insbesondere die Aspekte der Automatisierbarkeit und der algorithmischen Weiterverarbeitung bzw. Analyse eröffnen neue Perspektiven – gerade für die Korpuslinguistik und die digitale Diskurslinguistik. Doch auch in anderen Bereichen der Linguistik (bspw. der Pragmatik oder Soziolinguistik) wird Annotation immer mehr angewendet. Derzeit verändert sich das methodische Konzept des Annotierens in diesen Forschungsrichtungen dahingehend, dass Annotation als Erkenntnisprozess, der selbst Forschung ist, stärker im Vordergrund steht. Das durch die Computerlinguistik schon länger eingeführte Verständnis von Annotation hingegen ist vor allem auf die Auszeichnung meist systemlinguistischer Kategorien nach einem zuvor festgelegten Gold-Standard, die Messung des Inter-Annotator-Agreements und die Automatisierung bzw. automatisierte Auswertbarkeit ausgerichtet. Es hat sich tendenziell dahingehend entwickelt, dass unter Annotation oft ‚nur' das routinemäßige Tagging als mehr oder weniger standardisierter Erschließungsschritt angesehen wird, und zwar als ein Arbeitsgang vor der eigentlichen Analyse und Interpretation der Daten.

In linguistischen Disziplinen, in denen hermeneutische Interpretation eine zentrale Rolle spielt und etwa auch implizite sprachliche Phänomene analysiert werden, wie zum Beispiel in der Pragmatik (vgl. Archer/Culpeper/Davies 2008 und Weisser 2015, 2018), ist eine andere Konzeption des Annotierens notwendig. Hier besteht der methodische Kern aus dem Zusammenspiel von deduktiver und induktiver Kategorienbildung vor dem Hintergrund linguistischer Theorien sowie der iterativen Ausdifferenzierung von Tagsets und Guidelines in einem teils kollaborativen und diskursiven Verfahren. Die Operationalisierung von Forschungsfragen und linguistischen Phänomenen in Kategorien-Definitionen und -Systemen sowie in den entsprechenden Annotationsrichtlinien ist nicht mehr nur eine Voraussetzung für die Analyse, sondern selbst eine wichtige Ergebnisdimension (vgl. Bender 2020, Bender/Müller 2020).

Schon 2007 wurde im vielzitierten Sammelband von Hermanns/Holly „Linguistische Hermeneutik" dazu angeregt (hier Haß 2007), über eine „Korpus-Hermeneutik" nachzudenken. Wir wollen diesen Gedanken aufgreifen und dafür plädieren, dass sie den aktuellen Forschungsstand zu Verfahren des Annotierens einbezieht, sich interdisziplinär und auch im Bezug zu den Digital Humanities verortet, aber linguistisch präzise positioniert. Ziel dabei sollte sein, das linguistische Wissen, welches wir an die Daten herantragen, transparent, plausibel und praktikabel aufzubereiten, also zu explizieren und

zu operationalisieren, und in einen iterativ-dynamischen Forschungsprozess zu integrieren. So werden Forschungsansätze möglich, die über klassische korpuslinguistische Analysen der Frequenz bestimmter sprachlicher Kategorien u.ä. hinausgehen. So können etwa soziolinguistische und pragmatische Parameter inkludiert, auf verschiedene Ebenen angewendet und in Relation gesetzt werden.

Eine weitere Perspektive ist die iterative Verzahnung qualitativ-hermeneutischer Annotationsverfahren mit quantifizierend-algorithmischen Ansätzen. Dabei stellt sich die Frage, welche Kriterien die hermeneutisch-interpretative, qualitative Annotation erfüllen muss, damit sie mit automatisierten Verfahren, etwa des maschinellen Lernens, interagieren kann (vgl. Ide 2017).

Die Tagung adressiert auch Early-Career- und Student-Researcher, Interessierte können auch gerne als Zuhörer*in und Diskutant*in teilnehmen.

Wir freuen uns auf eine spannende Tagung und verbleiben mit den besten Grüßen,

Dr. Michael Bender (TU Darmstadt) und Dr. Katharina Jacob (Universität Heidelberg)

Literatur

Archer, Dawn; Culpeper, Jonathan; Davies, Matthew (2008): Pragmatic Annotation. In: Lüdeling, Anke; Kytö, Merja (Hg.): Corpus Linguistics. An International Handbook. Berlin: de Gruyter, S. 613-641.

Bender, Michael (2020): Annotation als Methode der digitalen Diskurslinguistik. In: Diskurse digital. Theorien – Methoden – Fallstudien. Band 2, Heft 1/2020: 1-35. DOI: https://doi.org/10.25521/diskurse-digital.2020.140

Bender, Michael; Müller, Marcus (2020): Heuristische Textpraktiken in den Wissenschaften. Eine kollaborative Annotationsstudie zum akademischen Diskurs. In: Zeitschrift für Germanistische Linguistik 48 (1)/2020: 1-46.

Haß, Ulrike (2007): Korpus-Hermeneutik. Zur hermeneutischen Methodik in der lexikalischen Semantik. In: Hermanns, Fritz; Holly, Werner (Hg.): Linguistische Hermeneutik. Theorie und Praxis des Verstehens und Interpretierens. Tübingen: Niemeyer, S. 241-261.

Hermanns, Fritz; Holly, Werner (Hg.) (2007): Linguistische Hermeneutik. Theorie und Praxis des Verstehens und Interpretierens. Tübingen: Niemeyer.

Ide, Nancy (2017): Introduction: The Handbook of Linguistic Annotation. In: Ide, Nancy; Pustejovsky, James (Hg.): Handbook of Linguistic Annotation. Vol. I, Dordrecht: Springer, S. 1-18.

Weisser, Martin (2015): Speech Act Annotation. In: Ajmer, Katrin; Rühlemann, Christoph (Hg.): Corpus Pragmatics. A Handbook. Cambridge: Cambridge University Press, S. 84-116.

Weisser, Martin (2018): How to Do Corpus Pragmatics on Pragmatically Annotated Data. Amsterdam, Philadelphia: John Benjamins.

Sarah Schröder (Universität Paderborn)

Direktiva in Aufrufen: Erkenntnisse einer manuellen, digitalen Annotation

Durch etwa das Zusammenspiel der hermeneutischen oder qualitativen Interpretation und einer quantitativen Analyse stellt die Annotation mehr dar als einen reinen Routineschritt. Schon anhand des Annotierens selbst bzw. der Festlegung der Annotationsrichtlinien etc. werden neue Erkenntnisse gewonnen. Annotationen können so eine Schnittstelle zwischen „Gegenstand und Methode“ bilden (Bender 2020: 8).

Dies wird auch anhand der Untersuchung direktiver Sprechakte in Texten der Sozialisten bzw. Sozialdemokraten und Kommunisten deutlich. Dazu wurden im Rahmen der Masterarbeit historische und bisher kaum untersuchte Aufrufe und Flugblätter der beiden Gruppen aus der Zeit der Weimarer Republik und des Widerstands gegen den Nationalsozialismus mithilfe des Analysetools CATMA analysiert. Aufrufe der untersuchten Gruppen richteten sich dabei teils nur an die eigene Gruppe, teils aber auch an die allgemeine Bevölkerung, was eine nähere Betrachtung der Direktiva besonders interessant macht.

Es besteht generell eine Fülle an Forschungsliteratur zu direktiven Sprechakten, jedoch wurde dieses Thema bisher noch nicht an der gewählten Fragestellung und nicht mithilfe der Annotation untersucht. Ziel der Untersuchung ist nicht nur die reine Erfassung der Sprechakte. Zunächst aus der Literatur (Hindelang 1978, Ickes 2008, Stelzel 2003) abgeleitete Kategorien werden stattdessen im Rahmen der Annotation und Entwicklung des Tagsets induktiv erweitert, die Annotation führt so zur Gewinnung einer Typologie der Direktiva im Kontext der untersuchten Schriften.

Literatur

Bender, Michael (2020): Annotation als Methode der digitalen Diskurslinguistik. In: Diskurse – digital. Theorien – Methoden – Fallstudien. Band 2, Heft 1/2020: 1-35. doi: 10.25521/diskurse-digital.2020.140.

Hindelang, Götz (1978): Auffordern: Die Untertypen des Aufforderns und ihre sprachlichen Realisierungsformen.

Göppingen: Kümmerle (= Göppinger Arbeiten zur Germanistik, Bd. 247).

Ickes, Andreas (2008): Parteiprogramme: Sprachliche Gestalt und Textgebrauch. Darmstadt: Büchner-Verlag.

Stelzel, Ulla (2003): Aufforderungen in den Schriften Herzogin Elisabeths von Braunschweig-Lüneburg. Eine Untersuchung zum wirkungsorientierten Einsatz der direktiven Sprachhandlung im Frühneuhochdeutschen. Hildesheim [u.a.]: Olms (= Documenta linguistica: Studienreihe, Bd. 5).

Ulrike Lohner (Universität Heidelberg)

„In abgehobenen akademischen Sphären“: iterative Annotation in einer Stiluntersuchung rechter Drohbriefe und linker Bekennerschreiben

Schon seit geraumer Zeit diskutieren Linguisten darüber, welche Analysemethoden für die Forensische Linguistik, speziell für die Autorenanalyse, am angemessensten seien: quantitative und leicht (teil)automatisierbare Verfahren einerseits (vgl. z.B. Chaski 2001, Nini 2017) oder rein qualitative und manuelle Analysen (etwa Coulthard 2004, 2005) andererseits. Während die Forschung von quantitativen Untersuchungen verlässlichere Statistiken erwartet, richtet sich die Methodik in der forensischen Praxis tatsächlich nicht nur nach der Fragestellung (vgl. Fobbe 2017: 272), sondern auch nach den Erfordernissen des vorliegenden Textmaterials, so dass „very few cases require exactly the same selection from the linguist’s toolkit“ (Coulthard/Johnson 2007: 6). Eine iterative Annotation spielt hier also eine zentrale Rolle, weil die Relevanz einzelner Merkmale von Fall zu Fall variiert und eine gleichbleibende Annotationsgrundlage letztlich nicht zum gleichen Kategorienkatalog führten muss.

Im vorgestellten Dissertationsprojekt wird versucht, übliche Kategorien aus der Autorenanalyse als Grundlage für eine stilistische Korpusuntersuchung nutzbar zu machen. Ziel ist es, in einem Korpus aus insgesamt 165 authentischen Texten (115 Droh- und Schmähbriefe mit rechtsextremen, 50 Bekennerschreiben mit linksextremen Inhalten) aufgrund autorenspezifischer Sprachmerkmale distinktive Stilausprägungen sichtbar zu machen. Die Annotation der Texte wurde dabei manuell mithilfe des Programms MAXQDA durchgeführt. Als Ausgangspunkt dienten zunächst Kategorien, die bereits in anderen Untersuchungen zum Autorenstil als relevant eingestuft wurden, etwa in Braun (1989), Dern (2003), Kniffka (2000) oder Krieg-Holz/Hahn (2016). Anhand konkreter Beispiele wird im Vortrag gezeigt, wo diese bestehenden Kategorien im untersuchten Korpus an Grenzen stießen und auf welche Weise sie modifiziert wurden, um dem speziellen Charakter der untersuchten Texte gerecht zu werden. Ebenso wird auf einige Kategorien eingegangen, die erst durch die Arbeit mit den Texten entstanden sind, und erläutert, wie sie in den dynamischen Annotationskatalog integriert wurden. Schließlich werden Kategorien vorgestellt, deren Anwendung vielversprechend schien, die sich in der praktischen Umsetzung im Rahmen des Projekts jedoch noch unhandlich erweisen. Mögliche Anpassungen dieser Kategorien werden zur Diskussion gestellt.

Literatur

Braun, Angelika (1989): "Linguistische Analysen im forensischen Bereich. zu den Möglichkeiten einer Texturheberschaftsuntersuchung". In: Bundeskriminalamt (ed.): *Symposium: Forensischer Linguistischer Textvergleich*. *Referate und Zusammenfassungen der Diskussionsbeiträge*. Wiesbaden: Bundeskriminalamt Wiesbaden: 143–166.

Chaski, Carole E. (2001): "Empirical evaluations of language-based author identification techniques". *Forensic Linguistics* 1/8: 1–65.

Coulthard, Malcolm (2004): "Author Identification, Idiolect, and Linguistic Uniqueness". *Applied Linguistics* 4/25: 431–447.

Coulthard, Malcolm (2005): "Some forensic applications of descriptive linguistics". *Veredas* 1/9: 9–28.

Coulthard, Malcolm/Johnson, Alison (2007): *An introduction to forensic linguistics*. *Language in evidence*. London: Routledge.

Dern, Christa (2003): "Sprachwissenschaft und Kriminalistik. Zur Praxis der Autorenerkennung". *Zeitschrift für germanistische Linguistik* 31: 44–77.

Fobbe, Eilika (2017): "Forensische Linguistik". In: Felder, Ekkehard/Vogel, Friedemann (eds.): *Handbuch Sprache im Recht*. Berlin, Boston: de Gruyter: 271–290.

Kniffka, Hannes (2000): "Anonymous Authorship Analysis Without Comparison Data? A Case Study With Methodological Implications". *Linguistische Berichte* 182: 179–198.

Krieg-Holz, Ulrike/Hahn, Udo (2016): "CodE Alltag. Ein deutsches E-Mail-Korpus für die Forensische Linguistik". In: Bülow, Lars et al. (eds.): *Performativität in Sprache und Recht*. Berlin: de Gruyter: 245–264.

Nini, Andrea (2017): "Register variation in malicious forensic texts". *International Journal of Speech Language and the Law* 01/24: 99–126.

Jöran Landschoff (Universität Heidelberg)

Abstract: Sprache im Netz – Wege zu einer soziolinguistischen Korpusannotation

Korpuslinguistisch arbeitende Diskursanalysen befassen sich heute überwiegend mit rekurrent auftretenden sprachlichen Formationen, sei es auf semantischer, grammatischer oder lexikalischer Ebene. Solche Musterhaftigkeiten oder Sprachgebrauchsmuster (Bubenhofer 2009) werden in großen Textmengen (Korpora) mit Hilfe automatisierter Tagger oder auch (semi-)manuell arbeitender Annotationsprogramme ausfindig gemacht. Der Standard ist bis heute das part-of-speech- oder postagging, wobei auch syntaktische Annotationen immer besser werden und semantische in Einzelprojekten erfolgreich Anwendung finden (Gries/Behrens 2017, S. 382ff.). Ein großes Problem korpuslinguistischer Diskursanalysen ist allerdings die höchst einseitige Befassung mit Medientexten, die eine sehr besondere Textsorte darstellen (ebd., S. 381). Besonders für die Frage nach den Diskursakteuren bieten diese Texte wenig Aufschluss, da die journalistische Praxis keineswegs nur O-Töne produziert und dennoch mehrere Diskursbeiträge bündelt.

Die Akteursebene ist für diskurslinguistische Fragen zentral, weil sie eine Gelenkposition zwischen Ereignis und Struktur einnimmt, an der individuelles zu kollektivem Wissen werden kann (Spitzmüller/Warnke 2011, S. 173). Wollen wir als Diskursforschende am methodischen Individualismus (Keller 2009, S. 13-14) festhalten, müssen handlungstheoretische Überlegungen in die Erforschung kollektiver Strukturen einbezogen werden. Constanze Spieß plädiert daher für die stärkere Berücksichtigung der Akteure und ihrer jeweiligen „Kontextuniversen“, die als Strukturressource zum Aufbau kommunikativer Kontexte Handlungsvoraussetzung sind (2018, S. 347-348). Diskurse wären so als Orte von Aushandlungen anzusehen, an die linguistische Fragestellungen gerichtet werden können, wodurch die Bereiche der Kultur und des Sozialen betreten sind.

Ich möchte meine Überlegungen hinsichtlich der Analyse sprachlicher Daten darstellen, die darauf abzielt, Kollektivbildungen in gesellschaftlich offenen Kommunikationsarenen mittels linguistischer Untersuchungen nachzuzeichnen. Ausgangspunkt sind einerseits obige diskurslinguistische Annahmen und soziologische Methoden, die aus der Relationalen Soziologie und Netzwerkforschung hervorgegangen sind. Dabei werden kulturelle Artefakte als Produkte sozialer Aushandlungen zwischen Individuen verstanden, die über Netzwerkstrukturen distribuiert, sanktioniert und integriert werden (Stegbauer 2016, S. 33). Andererseits bediene ich mich soziolinguistischer Theorien, die die Funktion von Sprache für Gruppenidentität, Selbstpositionierung und Abgrenzung hervorheben (Löffler 2016, S. 113).

Linguistisch geht es also um die Korrelation von Sprachwandel- und Gruppenbildungsprozessen. Um den Faktor „Individuum“ als Auslöser, nicht aber zielgerichteten Lenker diskursiver Dynamiken (Albert 2018, S. 417) einzubeziehen, werden Daten benötigt, die sich auf Akteure als Produzenten zurückführen lassen. Ich befasse mich daher mit diskursiven Beiträgen, die vom Mikrobloggingdienst *Twitter* gesammelt wurden. Ziel ist es, die linguistische Annotation der Korporatexte (=*Tweets*) durch eine soziolinguistischen Annotation zu erweitern, indem die Verlinkungsstrukturen, die sich aus den Kommunikationsfunktionen von *Twitter* ergeben (Wer schreibt wann in Antwort worauf an wen?), im Sinne einer Netzwerkanalyse als Textmetadaten annotiert werden. Auf diese Weise sollten semantische, lexikalische und grammatische Phänomene auf ihre Distribuierung, Verfestigung und Ablehnung hin untersucht werden können.

Literatur

Albert, Georg (2018): Diskurslinguistik und sprachliche Innovation. In: Warnke, Ingo H. (Hg.): Handbuch Diskurs. Berlin/Boston: de Gruyter, S. 405-425.

Bubenhofer, Noah (2009): Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin/New York: de Gruyter.

Gries, Stephan Th./Berez, Andrea L. (2017): Linguistic Annotation in/for Corpus Linguistics. In: Ide, Nancy/Pustejovsky, James (Hg.): Handbook of Linguistic Annotation, Dodrecht: Springer, S. 379-409.

Keller, Rudi (2009): Konventionen, Regeln, Normen. Zum ontologischen Status natürlicher Sprachen. In: Konopka, Marek/Strecker, Bruno (Hg.): Deutsche Grammatik. Regeln, Normen, Sprachgebrauch (Jahrbuch des Instituts für Deutsche Sprache), Berlin: de Gruyter, S. 9-22.

Löffler, Heinrich (2016): Germanistische Soziolinguistik. Berlin: Erich-Schmidt-Verlag.

Spieß, Constanze (2018): Diskurs und Handlung. In: In: Warnke, Ingo H. (Hg.): Handbuch Diskurs. Berlin/Boston: de Gruyter, S. 339-362.

Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin/Boston: de Gruyter.

Stegbauer, Christian (2016): Grundlagen der Netzwerkforschung. Wiesbaden: Springer Fachmedien Wiesbaden.

Thomas Jurczyk (Ruhr-Universität Bochum)

Projektvorstellung „ReligionML – Annotation religiöser Texte für Machine Learning“

Das in meinem Vortrag vorzustellende Projekt *ReligionML* wird derzeit als interne Arbeitsgruppe am Centrum für Religionswissenschaftliche Studien (CERES, Ruhr-Universität Bochum) durchgeführt. Die Arbeitsgruppe bestehend aus Religionswissenschaftler:innen verschiedener Schwerpunktbereiche hat es sich zum Ziel gesetzt, religionswissenschaftlich relevante Texte gemeinsam zu annotieren, um so mit der Zeit ein verlässlich annotiertes religionswissenschaftliches Corpus zu schaffen, das sowohl für die automatisierte als auch manuelle Bearbeitung religionswissenschaftlicher Fragen herangezogen werden kann.

Obwohl das finale Corpus für unterschiedliche Forschungsfragen nutzbar sein soll, steht im theoretischen Zentrum0F[1] der Gruppe derzeit die Frage, wie religiöse Semantik1F[2] in unterschiedlichen gesellschaftlichen Kontexten (Politik, Kunst, Wirtschaft, Religion etc.) verwendet wird. Das Corpus soll es dabei ermöglichen, diese Frage nicht nur punktuell, sondern möglichst umfangreich und repräsentativ bearbeiten zu können. Außerdem sollen Machine Learning Modelle, die auf Basis der annotierten Daten des Corpus trainiert wurden, dabei helfen, unbekannte Daten vorzufiltern und beispielsweise einzuordnen, ob es sich bei einem Text, der religiöse Semantik beinhaltet, um religiöse oder nicht-religiöse Kommunikation handelt2F[3] und aus welchem gesellschaftlichen Bereich diese stammt. Die Erstellung solcher automatisierter Klassifizierungsmodelle würde nicht nur die Filterung großer Datenmengen ermöglichen, um spezifischere Fragen zu bearbeiten,3F[4] sondern auch Rückschlüsse auf die Besonderheiten der religiösen bzw. nicht-religiösen Verwendung religiöser Semantik ermöglichen, die in den Klassifizierern zugrunde liegenden Entscheidungsparametern erkennbar sind.

Das Projekt *ReligionML* befindet sich noch in der Anfangsphase. Es basiert technisch auf einer von mir erstellten Webapplikation und konzentriert sich inhaltlich derzeit auf die Annotation von englischen Tweets, die Wörter wie „holy“ oder „religion“ enthalten, wobei das Corpus stetig erweitert werden soll und bereits wird. Es wurden bisher zwei Annotationsschritte implementiert: Zum einen werden die Tweets als Ganzes von den Annotatoren:innen klassifiziert bzw. annotiert.4F[5] Zum anderen haben die Annotatoren:innen die Möglichkeit, einzelne Wörter aus den Tweets separat zu annotieren. Wir arbeiten dabei bewusst nicht mit einem Goldstandard, sondern die Annotationskategorien werden während regelmäßiger Treffen weiterentwickelt. Anfangs sind wir dabei von einer binären religiös/nicht-religiös Klassifizierung ausgegangen, haben dann allerdings schnell gemerkt, dass dies nicht ausreichend ist, und unsere Annotationen immer weiter differenziert bzw. ausgeweitet.

[1] Frei nach dem ersten Schritt des MATTER Modells in (Pustejovsky, Bunt, and Zaenen 2017).
[2] Beispielsweise Heiligkeitssemantiken.
[3] Erste Tests mit simplen Machine Learning Modellen wie KNN und Logistic Regression wurden dabei bereits durchgeführt.
[4] Zum Beispiel, wenn die Frage im Zentrum steht, wie religiöse Semantiken in politischer Kommunikation verwendet werden.
[5] Dabei sieht das derzeitige Annotationsschema der Tweets auf der Makroebene wie folgt aus: **a) Inner-religion** means that the text has a religious meaning and was stated from within religion. A typical example would be a Christian or Muslim saying something about his or her belief. **b) Religion-transcendence** means that although the text does not belong to an inner-religious sphere, the overall context is still referring to religion as a social system dealing with the immanence/transcendence distinction. A typical example is two non-religious persons talking about religion as a worldview and religious truth claims compared to, for instance, philosophical ones. **c) Religion-immanence** means that the text still uses religious semantics but more in the sense of a general distinction marker. A good example is "religion" as an ethnic/political category. **d) Metaphorical-use** means that the terminology is no longer directly referring to "religion" but uses the "religion" domain in another context. A typical example are sentences such as "Electronic music is my religion." **e) No-religion** means that the text has nothing to do with religion at all.

Während meines Vortrages möchte ich den bisherigen Stand unserer Diskussion sowie insbesondere unser iteratives Vorgehen vorstellen und diskutieren. In diesem Zusammenhang möchte ich als besonderes Merkmal unseres Vorhabens hervorheben, dass wir bewusst mit der Ambiguität der Texte bzw. der Annotationen umzugehen versuchen. So sehen wir beispielsweise divergierende Kategorisierungen durch die einzelnen Annotatoren:innen nicht als Problem an, sondern vielmehr als Teil der Phänomenbeschreibung, die es uns erlaubt, Einordnungswahrscheinlichkeiten von Texten prozentual wiederzugeben, anstatt eine eindeutige Zuordnung vorzunehmen, die so oftmals auch in den Texten schlicht nicht gegeben ist.

Besonders interessiert bin ich darüber hinaus an existierenden Annotationsschemata aus dem Bereich der Annotation religiöser Texte, die mir bisher nicht bekannt sind, sowie an einem allgemeinen Austausch und der Knüpfung von Kontakten, da wir uns noch in einer sehr frühen Projektphase befinden.

Literatur

Pustejovsky, James, Harry Bunt, and Annie Zaenen. 2017. “Designing Annotation Schemes: From Theory to Model.” In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, 73–113. Dordrecht: Springer.

Jörn Stegmeier (Universität Darmstadt)

Maschinelle Konformitätsprüfung: Entwicklung eines Annotationsschemas zur Identifizierung semantischer Unsicherheit in Normtexten

Das Pilotprojekt ist ein Teilprojekt im Rahmen des SFB 805 ("Beherrschung von Unsicherheit in lasttragenden Systemen des Maschinenbaus") an der TU-Darmstadt und hat zum Ziel, ein Annotationsschema zu entwickeln, mit dem verschiedene Arten semantischer Unsicherheit in einschlägigen DIN-Normen sicht- und quantifizierbar gemacht werden können.

Semantische Unsicherheit wird im normalen Sprachgebrauch über hermeneutisch-kognitive Prozesse desjenigen Kommunikationspartners aufgelöst, der in einem Produzent-Rezipient-Gefüge als Rezipient agiert. Ist der Rezipient nicht in der Lage, die in einem Text befindlichen Unsicherheiten aufzulösen, bedarf es weiterer kommunikativer Akte, die die Unsicherheit beseitigen. Die Verwendung semantisch unsicherer Formulierungen in Normtexten ist dabei kein Versehen oder gar böse Absicht. Sie entspringt vielmehr der Notwendigkeit, fallspezifische Lösungen zulassen zu müssen, da nicht jede Anwendungssituation antizipiert werden kann.

Semantische Unsicherheiten, die nicht auf Mangel an Fachwissen ruhen, liegen z. B. dann vor, wenn mehrere Optionen als gleichermaßen gültig dargestellt werden. Ein prototypisches Beispiel hierfür ist die folgende Formulierung: "Der Bügel sollte eine Isolierung tragen, um die Übertragung von Körperwärme zu verringern." Die Optionen "Isolierung tragen" und (implizit) "Keine Isolierung tragen" sind aufgrund der Formulierung mit "sollen" beide gleichermaßen gültig, wodurch Unsicherheit hinsichtlich der Normkonformität erzeugt wird. Weitere Formen semantischer Unsicherheit gehen z. B. auf den Gebrauch von Ausdrücken zurück, die unterspezifizierte Konzepte evozieren. Beispiel: "Bei erdverlegten Leitungen mit nicht zugfesten Rohrverbindungen sind an Bögen und Abzweigen **ausreichend bemessene** Widerlager anzuordnen." Der Normanwender kann in diesem Fall nur dann normkonform arbeiten, wenn ihm Zusatzinformationen darüber vorliegen, was ein Widerlager als "ausreichend bemessen" konstituiert.

Im Vortrag wird ein Zwischenstand des Projekts vorgestellt, der alle Schritte von der Datenbeschaffung bis zur Ergebnisaufbereitung umfasst.

Ana Schenk (TU Darmstadt)

Vorurteilsfrei annotieren? Die Rolle des Vorwissens im qualitativ-hermeneutischen Annotationsprozess

Dieser Vortrag widmet sich der Rolle und Legitimität von Vorwissen im hermeneutisch-qualitativen Prozess des Annotierens. Insbesondere wird die Frage im Vordergrund stehen, zu welchem Zeitpunkt der Einbezug von linguistischem und extralinguistischem Wissen in den Forschungsprozess sinnvoll sein kann und wie sich die Spezifikation der Forschungsfrage auf die Relevanz der produktiven „Vorurteile“ im Sinne Gadamers (1975) auswirken kann.

Zur Veranschaulichung und Perspektivierung sollen Erkenntnisse und Hürden aus dem eigenen induktiven Kategoriefindungsprozess im Rahmen der Analyse unterschiedlicher Textsorten aus der Autismusforschung angeführt werden. Im Zuge dieser diskurslinguistischen Untersuchung unterschiedlicher Praktiken der Begriffskonstruktion- und Bewertung wurden bisher unterschiedliche linguistische Modelle herangezogen, kombiniert, verallgemeinert oder spezifiziert, um Annotationskategorien ableiten oder zumindest begründen zu können.

Welche Arten des Vorwissens produktiv oder hinderlich waren und inwiefern Vor-Urteile über den Diskurs, Textsorten oder das Korpus sich auf die Funktionalität und Anwendbarkeit der Kategorien auf die Daten ausgewirkt haben, soll in diesem Beitrag gegenstandsnah umrissen werden.
Im Kontext der Reflexion über unterschiedliche Szenarien der Nutzung von Vorwissen zur Erschließung des Untersuchungsgegenstands wird auch auf scheinbar vorurteilsfreie Vorgehensweisen wie die Grounded-Theory-Methodologie (Glaser & Strauss 1967; Corbin & Strauss 2015) eingegangen und diesbezügliche Implikationen, Grenzen und Variationen diskutiert.

Literatur

Bender, Michael (2020): Annotation als Methode der digitalen Diskurslinguistik. In: Diskurse digital. Theorien – Methoden – Fallstudien. Band 2, Heft 1/2020: 1-35. DOI: https://doi.org/10.25521/diskurse-digital.2020.140.

Corbin, Juliet & Strauss, Anselm (2015 [1990]): Basics of Qualitative Research. Techniques and Procedures for Developing Grounded Theory. London: Sage Publications, Inc.

Gadamer, Georg (1975): Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik. 4. Auflage. Tübingen: J.C.B. Mohr.

Glaser, Barney & Strauss, Anselm (1976): The Discovery of Grounded Theory. Strategies for Qualitative Research. Chicago IL: Aldine.

Hermanns, Fritz; Holly, Werner (Hg.) (2007): Linguistische Hermeneutik. Theorie und Praxis des Verstehens und Interpretierens. Tübingen: Niemeyer.

Mey, Günter & Mruck, Katja (Hg.) (2011): Grounded Theory Reader. 2., aktualisierte und erweiterte Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften.

Rapp, Andrea (2017): Manuelle und automatische Annotation. In: Fotis Jannidis, Hubertus Kohle und Malte Rehbein (Hrsg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, 253-267.