Preprint von: Bubenhofer, Noah (2018): Wenn „Linguistik“ in „Korpuslinguistik“ bedeutungslos wird. Vier Thesen zur Zukunft der Korpuslinguistik. In: Osnabrücker Beiträge zur Sprachtheorie (OBST), 92, S. 17-30.

Wenn „Linguistik“ in „Korpuslinguistik“ bedeutungslos wird

Vier Thesen zur Zukunft der Korpuslinguistik

Noah Bubenhofer, Universität Zürich

Ich möchte versuchen, die von den Herausgeber/innen dieses Bandes aufgeworfenen Fragen zu den theoretischen und methodologischen Problemen der Korpuslinguistik (vgl. Kapitel XX) ganz grundsätzlich anzugehen. Mit einem Blick über die Grenzen unserer Disziplin hinaus, sehe ich nichts weniger als eine Entwicklung, die nicht nur die Korpuslinguistik, sondern auch die Linguistik insgesamt stark prägen wird. Ich glaube, dass linguistische Theorien und Kategorien Gefahr laufen, überall dort unbedeutend zu werden, wo es um die quantitative Analyse von Sprachgebrauch geht. Wenn diese Gefahr abgewendet werden kann, werden damit auch einige der theoretischen und methodologischen Probleme der Korpuslinguistik gelöst. Dies ist der Kern um vier Hypothesen herum, die ich im Folgenden skizzieren möchte – und die sicherlich überspitzt und provokativ formuliert sind.

Im Folgenden werde ich neben der Korpuslinguistik von der Computerlinguistik und dem Textmining sprechen. Mit „Computerlinguistik“ ist die institutionell und wissenschaftlich etablierte Disziplin gemeint, die zwischen Informatik und Linguistik eingebettet ist, sich jedoch in den letzten Jahrzehnten grundlegend wandelt (dazu siehe unten These 2). Mit „Textmining“ sind diverse sehr informatisch-ingenieurstechnische Ansätze gemeint, die linguistisch uninformiert vorgehen. Unter „Korpuslinguistik“ verstehe ich Ansätze der maschinellen Textaufbereitung und Textanalyse, die deutlich von linguistischen Fragestellungen in der ganzen Breite geleitet sind.

These 1: In der Korpuslinguistik sollte es nicht (nur) um Belege gehen

Die sogenannte Keyword-in-Context-Liste ist nach wie vor emblematisch für korpuslinguistisches Arbeiten. Sie ist in der Tat ein entscheidender Schritt, der einen korpuslinguistischen Denkstil beförderte, wie ich in These 4 noch genauer diskutieren werde. Trotzdem ist es erstaunlich, dass die Ausgabe der Belege nach wie vor zu den unvermeidlichen Standardausgaben korpuslinguistischer Abfragesysteme gehört. Dabei ist schon länger klar, dass Korpora mehr sind als elektronische Zettelkästen (Perkuhn und Belica 2006) und dass für die meisten korpuslinguistischen Fragestellungen nicht diese Belege von Interesse sind, sondern die Zusammenfassung dieser Belege oder die Abstraktion davon. Diese Zusammenfassung oder Abstraktion wird auch meist gemacht, etwa indem Korrelationen der Belege mit textinternen oder textexternen Eigenschaften beobachtet oder Kollokationen berechnet werden. Und natürlich dadurch, dass die Bedeutung der Belegmengen für eine bestimmte Fragestellung hermeneutisch gedeutet wird.

Wenn in der Computerlinguistik oder im Textmining mit Korpora gearbeitet wird (was sehr oft der Fall ist), finden sich in wissenschaftlichen Publikationen normalerweise keine Belege. Dafür wird eine breite Palette von automatischen Verfahren eingesetzt, um natürliche Sprache der algorithmischen Auswertung zuzuführen. Damit werden dann abstraktere Ziele verfolgt als das schnelle Auffinden von Belegstellen, z.B. die automatische Klassifikation von Texten, die maschinelle Übersetzung oder die Analyse der Texte hinsichtlich bestimmter Kriterien wie Bewertungen, Akteur/innen etc.

Einblick in die Praxis der Korpuslinguistik erhält man über eine Analyse der Zeitschriften „International Journal of Corpus Linguistics“ und „Corpus Linguistics and Linguistic Theory“.[1] Eine Auswertung aller 75 Beiträge seit 2015 zeigt: Bei vielen Beiträgen ist eine deutliche linguistische Fragestellung vorhanden, die sorgfältig operationalisiert und mit unterschiedlichen Methoden korpuslinguistisch analysiert wird. Dabei kommen manuelle, semi-automatische oder automatische Kategorisierungen zur Anwendung (bei über 50 % der Fälle). Über solche (kategorisierten) Belegsammlungen gehen einige Studien korpuslinguistisch nicht hinaus. Immerhin in 61 % aller Artikel werden diese Belegmengen statistisch deskriptiv aber auch analytisch (Korrelationsanalysen) und multifaktoriell (Regressionsanalysen, ANOVA etc.) ausgewertet. Allerdings greift nur eine Minderheit von 17 % auf sogenannte maschinelle Lernverfahren zurück – davon ausgenommen solche, die im Rahmen der Korpusaufbereitung durch POS-Tagging, Parsing etc. sowieso oft Werkzeuge nutzen, die auf statistischen Modellen beruhen.

Es gibt also in den angrenzenden Disziplinen wie der Computerlinguistik und dem Textmining noch ein großes methodisches Potenzial, das für linguistische Fragestellungen fruchtbar gemacht werden könnte (vgl. dazu Bubenhofer und Scharloth 2015). Deshalb sind die dortigen methodologischen und theoretischen Entwicklungen aus korpuslinguistischer Sicht von großem Interesse.

These 2: Bei der quantitativen Analyse von Sprache droht die Linguistik bedeutungslos zu werden

So groß das Interesse der Korpuslinguistik an der maschinellen Textanalyse mancherorts ist – und zumindest sein müsste, so groß ist das Interesse der Ingenieurtechniken und der Informatik an Sprache. Dies kann durchaus als Erfolg der Linguistik betrachtet werden, zurückgehend auf den Linguistic Turn, der viele andere Disziplinen schon seit Jahrzehnten beeinflusst. Unternehmen interessieren sich für ihre Reputation im massenmedialen Diskurs oder sind der Überzeugung, ihr in unzähligen Dokumenten versprachlichtes Wissen besser verwalten zu können, wenn sie es nach sprachlichen Kriterien neu ordnen. Das Geschäftsmodell von Internetunternehmen basiert ganz erheblich darauf, sprachliche Kommunikation maschinell zu verarbeiten, um daraus Wissen aufzubauen und Vorhersagen über das Handeln von Kunden/innen zu machen. Auch in der Politik ist die Analyse von Sprachgebrauch ein wichtiger Faktor, um Wahlkämpfe zu gewinnen. Und automatische Spracherkennung und Übersetzung gehören zu den herausragenden Features technischer Gadgets.

Dies führt dazu, dass Algorithmen entwickelt werden, die mit sogenannt „unstrukturierten Daten“, also Text, umgehen können. In den Augen der Informatik ist Sprache unstrukturiert: Anders als eine Tabelle oder Datenbank ist für sie ein Text nicht eine Sammlung von Variablen und Ausprägungen, sondern eine ungeordnete Aneinanderkettung von meist ambigen Informationen, die erst in ein „ordentliches“, also strukturiertes Format überführt werden müssen.

Die Computerlinguistik widmete sich dieser Aufgabe bereits früh und verband dafür das Know-how zweier Disziplinen: der Linguistik und der Informatik. Es war offensichtlich, dass linguistisches Wissen über die Struktur von Sätzen hilft, Text einem automatischen Verstehen zugänglich zu machen. Ziel war der Bau von maschinenlesbaren Grammatiken, also Sammlungen

[1] Die beiden Zeitschriften sind zwar wichtig für die korpuslinguistische Forschung, lassen aber gleichzeitig viele wichtige Bereiche außen vor, vor allem eher sozial- und kulturwissenschaftlich interessierte korpusanalytische Arbeiten.

von Regeln, die es ermöglichen, die syntaktische Struktur eines Satzes zu erkennen und dann beispielsweise die Lexeme in einem Wörterbuch nachzuschlagen. Diese sogenannten regelbasierten Ansätze, einer strukturalistischen Linguistik entstammend, bestimmten wesentlich den ingenieurstechnischen Zugriff auf Sprache. Linguistische Theorien waren damit zentral, um diese unstrukturierten Sprachdaten in den Griff zu bekommen.

Mit der statistischen Wende in der Computerlinguistik[2] änderte sich dies jedoch fundamental. Während beispielsweise ein regelbasiertes maschinelles Übersetzungssystem versuchte, erst die syntaktische Struktur des Satzes in der Ausgangssprache zu „verstehen" und diese dann mittels Transformationsregeln zu übersetzen, gehen statistische Verfahren einen anderen Weg: Sie „lernen" auf der Basis großer Korpora von übersetzten Texten, wie Textfragmente normalerweise übersetzt werden. Eine syntaktische oder gar semantische Analyse ist völlig unnötig, es genügt das statistische Modell, das bei ausreichend großem Trainingsmaterial ziemlich zuverlässige Aussagen darüber treffen kann, welche Übersetzung wahrscheinlich ist. „Google Translate" ist ein bekanntes Beispiel für ein solches Verfahren, das noch nicht einmal domänenspezifisch angepasst ist (es würde noch besser funktionieren, wenn es für eine bestimmte Domäne trainiert wäre). Es liefert zwar keine perfekten Übersetzungen, ist aber im Vergleich zu einem regelbasierten System unheimlich „robust".

Aktuell erfahren in der Computerlinguistik und generell im Data Mining neuronale Netze großen Zuspruch, die den Prozess des maschinellen Lernens nach dem Modell des menschlichen Gehirns gestalten.[3] Solche Systeme, „Deep Learning-Systeme" genannt, sind in der Lage, Muster in den Daten zu erkennen, ohne dass vorher explizit die Eigenschaften festgelegt werden, die getestet werden sollen. Zudem findet das Lernen auf mehreren verborgenen Ebenen statt, so dass das Lernen nicht beobachtet und damit auch die Frage, welche Eigenschaften nun welchen Einfluss auf das gelernte Modell haben, kaum beantwortet werden kann.

Solche Verfahren können auf beliebige Datensätze angewandt werden, seien es Sprachdaten (geschrieben oder gesprochen), Zahlenwerte, Bilder etc., solange diese Daten ein digitales Format aufweisen. Um wiederum ein Beispiel aus dem Bereich der maschinellen Übersetzung zu nennen: Gegenwärtig wird das Startup-Unternehmen „DeepL" (vgl. www.deepl.com) rege diskutiert, das ein Übersetzungssystem auf der Basis eines neuronalen Netzes anbietet. Die öffentlichkeitswirksame Bezeichnung für solche Systeme ist „künstliche Intelligenz". Auf der Website des Unternehmens kann man lesen: „Wir suchen Mitarbeiter: Mathematiker, Informatiker." Und weiter steht:

> Bist du Mathematiker, Informatiker oder Physiker mit großem Interesse an neuronalen Netzwerken, dann melde dich bei uns. Wir freuen uns über außergewöhnliche Talente und helle Köpfe mit unterschiedlichem Background und mit Verständnis für neuronale Netzwerktechnologie und der Mathematik, die dahinter steckt. In unserem Team wirst du zum hochqualifizierten Experten für neuronale Netzwerke und du gestaltest die Zukunft der KI-Entwicklung mit. (https://www.deepl.com/jobs.html, 19. Sep. 2017)

---

[2] Vgl. für eine Übersicht zu den Forschungsparadigmen in der Computerlinguistik Uszkoreit (2009).

[3] Die ACL-Anthology-Datenbank, die alle wichtigen Computerlinguistik-Arbeiten aufführt, zeigt den Trend eindrucksvoll: Bis und mit 2012 gibt es jeweils unter zehn Arbeiten pro Jahr, die „Deep Learning" enthalten. 2013 sind es 27, 2014: 71, 2015: 99, 2016: 225 und bis September 2017 bereits 51 (vgl. http://aclasb.dfki.de/).

Eines der attraktivsten Forschungsthemen der Computerlinguistik (Carstensen u. a. 2001: 520) kommt also vollständig ohne Linguistik aus – und ohne Computerlinguistik.

In der Korpuslinguistik wollen wir keine maschinellen Übersetzungssysteme bauen. Das Forschungsinteresse liegt nicht primär darin, Werkzeuge zu entwickeln, die natürliche Sprache maschinell verarbeiten können, sondern darin, bestimmte Analysen vorzunehmen. Wie in These 1 oben dargelegt, reicht es dafür allerdings auch nicht aus, sich nur Belegmengen anzusehen und zu deuten, sondern es sind Verfahren der Datenaggregation sinnvoll. Dafür, aber auch schon bei der Aufbereitung von Korpusdaten, kommen Tools der maschinellen Sprachverarbeitung zum Einsatz.

Ein Wortarten-Tagger, wie der häufig benutzte TreeTagger (Schmid 1994), ist ein Beispiel für ein solches Tool, das bei vielen Korpora zur maschinellen Annotation von Wortartklassen verwendet wird. Er basiert auf einem gelernten Modell typischer Abfolgen von Wortarten. Die linguistische Motivation für die Modellierung dieser Funktionsweise liegt wahrscheinlich in den grundlegenden Prämissen von Kontextualismus und Distributionalismus. Die Funktionsweise eines solchen Taggers ist deshalb aus linguistischer Sicht theoretisch mehr oder weniger begründbar.

Einen ähnlichen theoretischen Hintergrund haben sogenannte Word-Embedding-Ansätze. Sie überführen Lexeme in einen Vektorraum, indem ihr Kollokationsverhalten in einem großen Korpus als Vektor repräsentiert wird, der ausdrückt, mit welchen anderen Lexemen es überzufällig auftritt. Der Vektor repräsentiert also ein Kollokationsprofil, eine Größe, die in der Linguistik auf eine lange Tradition der theoretischen Begründung und Erprobung blickt (vgl. für eine Übersicht des Kollokationsbegriffs z.B. Evert 2009). Während in der Linguistik solche Kollokationsprofile zwar eine wichtige Rolle spielen, etwa in der Lexikographie oder auch in der Diskurslinguistik, um die Gebrauchsbedeutung eines Lexems zu ergründen, ist es die Computerlinguistik, die mit der Repräsentation des Profils in Form eines Vektors den Schritt ermöglicht, mit Kollokationsprofilen zu rechnen. Denn als Vektor im Vektorraum sind nun geometrische Operationen möglich, etwa ähnliche (Synonyme) oder entgegengesetzte (Antonyme) Vektoren zu berechnen.

Die beiden Beispiele sind jedoch transparente Verfahren, sogenannte „White Box"-Modelle, deren Funktionsweise einsehbar ist. Zudem sind sie mehr oder weniger theoretisch begründet. Bei vielen anderen Anwendungen des maschinellen Lernens und insbesondere beim Deep Learning handelt es sich jedoch um Black Boxes, die nicht eingesehen werden können. D.h., der Erfolg des gelernten Modells kann natürlich anhand eines manuell erstellten „Goldstandards" sorgfältig evaluiert werden, warum das Modell jedoch erfolgreich ist (oder nicht), ist weitgehend unklar. Und unklar ist auch, wie das Modell genau funktioniert, also welche Eigenschaften welchen Einfluss haben.

In der Computerlinguistik wurden aber für viele Probleme Deep-Learning-Algorithmen eingesetzt, meist mit Erfolg. Erfolg bedeutet, dass die statistischen Modelle besser den Goldstandard voraussagen können, aber nicht, dass das grundlegende linguistische Problem besser gelöst wäre (z.B.: Sentiment-Analyse: Wie werden Gefühle und Meinungen ausgedrückt; Textklassifikation: Wie drückt sich Stil, Autorschaft, Textsorte, Thema etc. aus; maschinelles Übersetzen: Wie sieht die Praxis aus, sprachliche Äquivalente zu einer Ausgangssprache in der Zielsprache zu finden).[4]

[4] Beim SemEval-Shared-Task 2016 (Task 4: Sentiment Analysis in Twitter) waren durchgängig diejenigen Ansätze am erfolgreichsten, die Deep Learning nutzten: „A general trend that emerges from SemEval-2016 Task 4 is that most teams who were ranked at the top in the various subtasks

Allerdings gibt es auch Versuche, die Deep-Learning-Algorithmen nicht ungezielt auf alle möglichen Eigenschaften anzusetzen, sondern spezifische linguistische Eigenschaften, die für das Problem relevant sind, zu integrieren. So wurde beispielsweise ein Ansatz in der maschinellen Übersetzung vorgeschlagen, der Lemmaformen, Wortartklassen, syntaktische Dependenz- und morphologische Informationen hinzufügt, um klassische Ambiguitäts- oder Stellungsprobleme zu lösen (Sennrich und Haddow 2016).

Trotzdem: Überall wo Sprachgebrauch quantitativ und maschinell analysiert wird, gibt es einen starken Trend, möglichst ohne linguistischen Kategorien und Theorien auszukommen und Black-Box-Systeme zu verwenden. Das ist nachvollziehbar, da es in den meisten Fällen darum geht, ein System zu bauen, das eine klar definierte Aufgabe sehr gut lösen kann. Obwohl diese Ansätze natürlich auch für die Korpuslinguistik interessant sind, genügen sie so, wie sie beispielsweise in der Computerlinguistik und dem Textmining eingesetzt werden, unseren Forschungsinteressen nicht, da sie keinen Beitrag dazu leisten, sprachliche Phänomene auch nur verstehen, geschweige denn erklären zu können.

Viel dramatischer ist jedoch, dass die Linguistik offensichtlich nicht in der Lage ist, einen nützlichen Beitrag zu den Problemen der maschinellen Textanalyse zu leisten. Die Linguistik scheint für die quantitative Analyse von Text weitgehend bedeutungslos zu werden.[5]

These 3: Um der Bedeutungslosigkeit zu entgehen, sind zwei Dinge wichtig: Linguistische Theorie und die hermeneutische Deutung der Ergebnisse quantitativer Analysen

Die Entwicklungen im Kontext von Computerlinguistik und Textmining sind also eigentlich von großem Interesse in einer linguistisch interessierten Korpuslinguistik, um über manuelles oder halbautomatisches Kategorisieren und Auswerten von Belegstellen hinaus zu kommen (vgl. These 1). Zudem gibt es deutliche Überschneidungen zwischen den Fragestellungen, die mit computerlinguistischen Tools und Textmining beantwortet werden sollen, und den Fragestellungen in der Linguistik. Auch im Textmining ist man daran interessiert, Emotionalität, Bewertungen oder stilistische Eigenheiten in Texten maschinell zu identifizieren, um Beispiele zu nennen. Die Linguistik muss dabei aber ein kritisches Verhältnis zur Forschungslogik in den ingenieurstechnischen Disziplinen pflegen und auf zwei Prinzipien bestehen, um bei der quantitativen Textanalyse nicht bedeutungslos zu werden:

---

used deep learning, including convolutional NNs, recurrent NNs, and (general-purpose or task-specific) word embeddings. In many cases, the use of these techniques allowed the teams using them to obtain good scores even without tuning their system to the specifics of the subtask at hand [...]. Conversely, several teams that have indeed tuned their system to the specifics of the subtask at hand, but have not used deep learning techniques, have performed less satisfactorily. This is a further confirmation of the power of deep learning techniques for tweet sentiment analysis" (Nakov u. a. 2016).

[5] Fragen der Bedeutung der Linguistik innerhalb der Computerlinguistik wurden am „EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics" durchaus kontrovers diskutiert, etwa von Johnson (2009), der nicht davon ausgeht, dass zukünftig die theoretische Linguistik noch eine große Rolle spielen wird, oder von Moore (2009), der findet, dass bestimmte Probleme ohne linguistisches Wissen nicht erfolgreich gelöst werden können.

1) Nicht nur für die Linguistik, sondern für alle geistes- und sozialwissenschaftlichen Disziplinen gilt: Eine theoretische Fundierung der Analysekategorien ist essentiell. Dafür werden valide Analysekategorien benötigt, die deutbar sind. Dieses Prinzip richtet sich jedoch keinesfalls gegen datengeleitete Verfahren, im Gegenteil: Sie sind es, die die theoretischen Modelle herausfordern und schärfen können. Aber das Ziel aller Analysen muss darin liegen, ein Puzzleteil zu einem besseren Verständnis sprachlicher Strukturen, von Sprachgebrauch oder gesellschaftlichen und kulturellen Bedeutungen von Sprache zu liefern. Wir benötigen White-Box-, nicht Black-Box-Systeme (vgl. dazu auch Bubenhofer und Scharloth 2015).

Fehlende Validität ist ein ernsthaftes Problem, was z.B. das Feld der sogenannten „Authorship Attribution", also der Zuordnung eines Textes X zu einem Autor/einer Autorin A, B, C, …, zeigt. Um entsprechende Zuordnungen zu treffen, stehen normalerweise Texte zur Verfügung, deren Autorschaft bekannt ist. Die Frage ist dann also, ob über die sprachlichen Merkmale des Textes X automatisch bestimmt werden kann, wer der Autor/die Autorin (aus der Menge der möglichen Autor/innen) von Text X ist. Genauer lautet die Frage aber, ob und wie sich persönlicher Schreibstil sprachlich niederschlägt.

Besonders erfolgreich für diese Aufgabe sind Methoden maschinellen Lernens, die das Problem als Klassifikationsaufgabe auffassen und anhand von Trainingskorpora typische sprachliche Merkmale der Texte der jeweiligen Autor/innen lernen. Dabei zeigt sich: „low-level features like character n-grams are very successful for representing texts for stylistic purposes" (Stamatatos 2009: 24). Das bedeutet, solche Modelle, die auf der Distribution von Buchstaben-N-Grammen beruhen, sind, gemessen an einem Goldstandard, am erfolgreichsten. Allein: Solche Modelle lassen sich nicht linguistisch deuten, da völlig unklar ist, was sie eigentlich messen. Ist es Stil, Thema, Textsorte, sprachliche Handlungsqualität…? Es handelt sich also weder um eine valide, noch um eine deutbare Kategorie (insbesondere, wenn das statistische Modell nicht einsehbar ist). Für spezifische Aufgaben der Autorschaftsattribution mag eine solche Modellierung genügen, aber bereits für forensische Anwendungen, beispielsweise vor Gericht, ist eine solche Modellierung fragwürdig und gefährlich. Und für eine linguistische Deutung des Phänomens Autorschaftsstil ist sie gänzlich unbrauchbar.[6]

Die Kritik geht jedoch nicht nur in Richtung des Textminings und der Computerlinguistik, manchmal nicht-valide Kategorien einzusetzen (was für die dortigen Zwecke auch oft sinnvoll ist), sondern auch in die Richtung der Linguistik: Die Computer- und die Korpuslinguistik zeigen beide gleichermaßen, wie wichtig es ist, auch abstrakte Kategorien möglichst so zu definieren, dass überhaupt eine Chance besteht, sie für eine quantitative Analyse operationalisierbar zu machen. Wenn eine linguistische Kategorie so vage ist, dass sich selbst (geschulte) Menschen uneinig darüber sind, wenn sie an authentischem Sprachegebrauch angewendet wird, scheitert die quantitativ-maschinelle Lösung unweigerlich. Diesen Aspekt muss ich aber vor einem korpuslinguistischen Publikum nicht weiter vertiefen.

2) Die Ergebnisse quantitativer Analysen sind nicht Antworten auf Fragestellungen, sondern neue Daten, die vor einem geistes- und sozialwissenschaftlichen Hintergrund genauso hermeneutisch gedeutet werden müssen wie einzelne Texte. Das ist vielleicht das größte Missverständnis, wenn Textminer und Computerlinguistinnen mit Korpuslinguistinnen zusammenarbeiten: Erstere wollen, dass ein Werkzeug ein Ergebnis hervorbringt, das an einem Goldstandard evaluiert werden kann. Das Ergebnis ist dann im Einzelfall richtig oder falsch und in der Gesamtheit

[6] Vgl. für eine kritische Diskussion verschiedener Methoden zur Autorschaftsattribution aus korpuslinguistischer Sicht die Blogbeiträge von Joachim Scharloth: http://www.security-informatics.de/blog/?tag=authorship-detection (24. September 2017).

genügend präzise oder nicht. Und das Ergebnis ist im Idealfall dann auch die Lösung der Forschungsfrage. Bei den meisten geistes- und sozialwissenschaftlichen Fragestellungen beginnt auf der Grundlage dieser Ergebnisse jedoch ein Interpretationsprozess, um (meist in Kombination mit weiteren Analysen) eine plausible Deutung zu ermöglichen – eine vorläufige Deutung. Die Stärke der Geistes- und Sozialwissenschaften liegt dabei gerade darin, dass ihrer Methodologie ein Zweifeln inhärent ist, mit dem die „gegenwärtig besiegelten Bedeutungen jeweils eingeklammert oder angezweifelt [werden], um zu prüfen, inwiefern sich nach rationalem Ermessen nicht bessere Lösungen, überlegenere Interpretationen oder zustimmungsfähigere Regelungen finden lassen“ (Honneth 2016: 312).

These 4: Korpuslinguistisches Arbeiten ist diagrammatisches Operieren

Neben der Suche nach validen Analysekategorien und dem Hochhalten geisteswissenschaftlicher Prinzipien der Deutung sehe ich einen weiteren Aspekt, der helfen sollte, der Korpuslinguistik eine deutliche linguistische Prägung zu verleihen. Es ist der Versuch, korpuslinguistisches Arbeiten als „diagrammatisches Operieren“ aufzufassen. Mit dem Diagramm-Begriff[7] folge ich Krämer (2016), die deutlich macht, dass Diagramme als Formen der Visualisierung von Daten „Denkzeuge“ sind, mit denen operiert wird: Ich kann Daten in einem Diagramm darstellen (auf einer Karte, in einem Netzwerkgraph, einem Punkteplot, …) und danach damit operieren, um neue Erkenntnisse daraus zu ziehen. Wenn man einem breiten Diagramm-Begriff folgt, wird deutlich, dass auch Listen, Tabellen und dergleichen diagrammatischen Charakter haben, da sie Informationen räumlich anordnen und damit diese grafische Anordnung bedeutsam wird und neue Lesarten ermöglicht (Siegel 2009; Steinseifer 2013). Listen und Tabellen sind nun aber Formen, die in der Korpuslinguistik zentral sind: Die Keyword-in-Context-Liste (zurückgehend etwa auf Zettelkästen im 16. Jahrhundert) kann etwa als Keimzelle eines völlig neuen Textverständnisses angesehen werden, mit dem die Einheit des Textes zerstört wird, um eine neue Sicht auf Textdaten zu gewinnen. Viele weitere Formen der Anordnung von Textdaten spielen ebenfalls wichtige Rollen, entscheidend etwa die Überführung von Textdaten in den Vektorraum (vgl. These 2), in dem operiert werden kann (z.B. in Form geometrischer Operationen – Lagen von Vektoren und ihren Winkeln zueinander). Aber auch die Erfindung der Partiturdarstellung bei Gesprächstranskripten (Sacks u. a. 1974; Ehlich und Rehbein 1976; Selting u. a. 1998), mit der überhaupt erst eine moderne Gesprächslinguistik möglich wurde, zeigt die Kraft diagrammatischer Umformungen, um Daten neu lesbar zu machen (vgl. für weiterführende Überlegungen und Beispiele zu Diagrammen in der Linguistik Bubenhofer im Druck a; Bubenhofer u. a. im Druck; Bubenhofer im Druck b).

Ich meine, es lohnt sich, korpuslinguistisches Arbeiten unter diagrammatischer Perspektive zu reflektieren, um die Mechanismen und Möglichkeiten der Gegenstandskonstitution besser zu verstehen. Denn repräsentiert in einem Vektorraum, geben die gleichen Daten einen völlig anderen Gegenstand ab als dargestellt in einer Keyword in Context-Liste. Die Wahl einer diagrammatischen Form ist gegenstandskonstitutiv. Und es müsste vordringliches Ziel sein, noch ganz andere Formen der diagrammatischen Darstellung von Text zu finden, um damit andere Gegenstandskonstitutionen und Fragestellungen zu ermöglichen. Dafür sind semiotische und natürlich auch wissenschaftstheoretische Überlegungen nötig, die für alle Disziplinen, die mit maschineller Textanalyse befasst sind, relevant sein müssten.

---

[7] Der Diagramm-Begriff geht auf Charles Sanders Peirce zurück und erfährt gegenwärtig mit verschiedenen theoretischen Arbeiten zu Diagrammatik eine Renaissance (Bauer und Ernst 2010; Bender und Marrinan 2014; Krämer 2011; Stetter 2005; Stjernfelt 2007).

Danksagung

Dieser Text und die dazugehörige Forschung wurden ermöglicht durch ein Ambizione-Stipendium des Schweizer Nationalfonds.

Bibliographie

Bauer, Matthias; Ernst, Christoph (2010): *Diagrammatik / Einführung in ein kultur- und medienwissenschaftliches Forschungsfeld*. Bielefeld: transcript.

Bender, John B.; Marrinan, Michael (2014): *Kultur des Diagramms*. Berlin: Akademie Verlag (Actus et imago).

Bubenhofer, Noah (im Druck a): *Visual Linguistics: Plädoyer für ein neues Forschungsfeld*. In: Bubenhofer, Noah; Kupietz, Marc (Hrsg.) *Visual Linguistics*. Heidelberg: Heidelberg University Publishing.

Bubenhofer, Noah (im Druck b): *Visualisierungen in der Korpuslinguistik. Diagrammatische Operationen zur Gegenstandkonstitution, -Analyse und Ergebnispräsentation*. In: Kupietz, Marc; Schmidt, Thomas (Hrsg.) *Korpuslinguistik*. Berlin / New York: De Gruyter.

Bubenhofer, Noah; Rothenhäusler, Klaus; Affolter, Katrin; u. a. (im Druck): *The Linguistic Construction of World – an Example of Visual Analysis and Methodological Challenges*. In: Scholz, Ronny (Hrsg.) *Quantifying Approaches to Discourse for Social Scientists*. Basingstoke: Palgrave Macmillan.

Bubenhofer, Noah; Scharloth, Joachim (2015): *Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate*. In: *Zeitschrift für Germanistische Linguistik*. 43 (1), S. 1–26.

Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; u. a. (Hrsg.) (2001): *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Heidelberg/Berlin: Spektrum.

Ehlich, Konrad; Rehbein, Jochen (1976): *Halbinterpretative Arbeitstranskriptionen (HIAT)*. In: *Linguistische Berichte*. 45 , S. 21–41.

Evert, Stefan (2009): *58. Corpora and collocations*. In: Lüdeling, Anke; Kytö, Merja (Hrsg.) *Corpus Linguistics*. Berlin, New York: Mouton de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft), S. 1212–1248.

Honneth, Axel (2016): *Denaturierung der Lebenswelt. Vom dreifachen Nutzen der Geisteswissenschaften*. In: Panteos, Athena; Rojek, Tim (Hrsg.) *Texte zur Theorie der Geisteswissenschaften*. Stuttgart: Reclam (Reclams Universal-Bibliothek), S. 283–315.

Johnson, Mark (2009): *How the Statistical Revolution Changes (Computational) Linguistics*. In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Athens, Greece: Association for Computational Linguistics S. 3–11.

Krämer, Sybille (2011): *Diagrammatische Inskriptionen: Über ein Handwerk des Geistes*. In: *Sehen und Handeln*. Berlin: Akademie.

Krämer, Sybille (2016): *Figuration, Anschauung, Erkenntnis: Grundlinien einer Diagrammatologie*. Berlin: Suhrkamp Verlag.

Moore, Robert C. (2009): *What Do Computational Linguists Need to Know about Linguistics?* In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Athens, Greece: Association for Computational

Linguistics S. 41–42.

Nakov, Preslav; Ritter, Alan; Rosenthal, Sara; u. a. (2016): *SemEval-2016 Task 4: Sentiment Analysis in Twitter*. In: *Proceedings of the 10th International Workshop on Semantic Evaluation*. San Diego, California: Association for Computational Linguistics (SemEval '16).

Perkuhn, Rainer; Belica, Cyril (2006): *Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik*. In: *Sprachreport*. 22 (1), S. 2–8.

Sacks, Harvey; Schegloff, Emanuel A.; Jefferson, Gail (1974): *A Simplest Systematics for the Organization of Turn-Taking for Conversation*. In: *Language*. 50 (4), S. 696–735, doi: 10.2307/412243.

Schmid, Helmut (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Selting, Margret; Auer, Peter; Barden, Birgit; u. a. (1998): *Gesprächsanalytisches Transkriptionssystem (GAT)*. In: *Linguistische Berichte*. 173 , S. 91–122.

Sennrich, Rico; Haddow, Barry (2016): *Linguistic Input Features Improve Neural Machine Translation*. In: *arXiv:1606.02892 [cs]*.

Siegel, Steffen (2009): *Tabula: Figuren der Ordnung um 1600*. Berlin / Boston: Akademie-Verlag.

Stamatatos, Efstathios (2009): *A Survey of Modern Authorship Attribution Methods*. In: *J. Am. Soc. Inf. Sci. Technol.* 60 (3), S. 538–556, doi: 10.1002/asi.v60:3.

Steinseifer, Martin (2013): *Texte sehen – Diagrammatologische Impulse für die Textlinguistik*. In: *Zeitschrift für germanistische Linguistik*. 41 (1), S. 8–39.

Stetter, Christian (2005): *Bild, Diagramm, Schrift*. In: *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*. München: Wilhelm Fink Verlag (Kulturtechnik).

Stjernfelt, Frederik (2007): *Diagrammatology: an investigation on the borderlines of phenomenology, ontology, and semiotics*. Dordrecht; London: Springer.

Uszkoreit, Hans (2009): *Linguistics in Computational Linguistics: Observations and Predictions*. In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Athens, Greece: Association for Computational Linguistics S. 22–25.