Wissenschaft ohne Geist: Herausforderungen der Digital Humanities am Beispiel der Korpuslinguistik

Bubenhofer, Noah
bubenhofer@cl.uzh.ch
Universität Zürich, Schweiz

Alle, die mit maschinellen Methoden Sprache analysieren, erleben momentan einen tiefgreifenden methodologischen Wandel.[1] Einerseits erfreut man sich vielleicht als Geisteswissenschaftler/in am immer stärkeren Interesse der Ingenieurtechniken und der Informatik für 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 zu machen. Auch in der Politik ist die Analyse von Sprachgebrauch ein wichtiger Faktor, um Wahlkämpfe zu gewinnen.

Andererseits beschert einen dieses Interesse eine Vielzahl von neuen Methoden für die maschinelle Analyse von Text, die auch für geisteswissenschaftliche Fragestellungen interessant sind. Die Digital Humanities sind ein Beispiel für eine Disziplin, die sich den Experimenten mit diesen Methoden verschrieben hat. Auch die Korpuslinguistik profitiert maßgeblich von diesen neuen Methoden.

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. 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.

In der Computerlinguistik wurden bereits 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 (z.B.: Sentiment-Analyse: wie werden Gefühle und Meinungen ausgedrückt; Textklassifikation: wie drückt sich Stil, Autorschaft, Textsorte, Thema etc. aus) besser gelöst wäre.

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 linguistischen Forschungsinteressen eigentlich nicht, da sie keinen Beitrag dazu leisten, sprachliche Phänomene zu verstehen und erklären zu können.

Viel dramatischer ist jedoch, dass die Linguistik offensichtlich nicht in der Lage ist, einen nützlichen Beitrag zur Lösung der Probleme der maschinellen Textanalyse zu leisten. Die Linguistik scheint für die quantitative Analyse von Text weitgehend bedeutungslos zu werden.

Um der Bedeutungslosigkeit zu entgehen, muss die Linguistik ein kritisches Verhältnis zur Forschungslogik in den ingenieurstechnischen Disziplinen pflegen und auf zwei Prinzipien bestehen: 1) Mehr linguistische Theorie. 2) Ergebnisse von quantitativen Analysen müssen gedeutet werden.

Zu 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 führen. Wir benötigen White-Box-, nicht Black-Box-Systeme.

Das Problem der fehlenden Validität zeigt sich z.B. im Feld der sog. „Authorship Attribution“, also der Zuordnung eines Textes X zu einem Autor A, B, C, …. Um dies zu tun, stehen Texte zur Verfügung, von denen die 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 Autoren/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, dass „low-level features like character n-grams are very successful for representing texts for stylistic purposes” (Stamatatos, 2009, S. 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, …? 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 das ausreichend sein, 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.[2]

Die Kritik geht jedoch nicht nur in Richtung des Textminings und der Computerlinguistik, manchmal nicht-valide Kategorien einzusetzen (was zudem oft für die dortigen Zwecke auch sinnvoll ist), sondern auch in die Richtung der Linguistik: Die Computer- und die Korpuslinguistik zeigen beide gleichermaßen, wie wichtig es ist, auch abstrakte Kategorien so zu definieren versuchen, 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 werden, scheitert die quantitativ-maschinelle Lösung unweigerlich.

2) Die Ergebnisse von quantitativen 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 genügend präzise oder nicht. Das Ergebnis ist dann auch im Idealfall 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 ja gerade darin, dass in 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, S. 312).

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 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 (Siegel, 2009; Steinseifer, 2013). Dies sind nun aber Formen, die in der Korpuslinguistik zentral sind: Die Keyword in Context-Liste (zurückgehend etwa auf Zettelkästen im 16. Jahrhundert) etwa kann 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, 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 Gesprächstrankripten, mit der überhaupt erst eine moderne Gesprächslinguistik möglich wurde, zeigt die Kraft von diagrammatischen Umformungen, um Daten neu lesbar zu machen. Hinter diesen diagrammatischen Umformungen stecken diagrammatische Grundfiguren (Bubenhofer im Druck b), die in den Geisteswissenschaften generell wirkmächtig sind.

Ich meine, es lohnt sich, korpuslinguistisches Arbeiten unter diagrammatischer Perspektive zu reflektieren, um die Mechanismen und Möglichkeiten der Gegenstandskonstitution besser zu verstehen. Die Digitalität der Daten und Methoden erlaubt dabei neue Transformationen und macht Daten, egal welcher Modalität, miteinander verrechenbar. Aber die diagrammatischen Grundfiguren führen zu unterschiedlichen Gegenständen: Repräsentiert in einem Vektorraum geben die gleichen Daten einen völlig anderen Gegenstand ab als dargestellt in einer Keyword in Context-Liste. 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. Die algorithmische Repräsentation der Daten folgt dabei ebenfalls den diagrammatischen Transformationen (Beispiel Vektorraum) und kann daher nicht unabhängig davon gedacht werden. Für eine hermeneutische Deutung brauchbare Analysekategorien zu erarbeiten, bedeutet deshalb auch, die damit verbundenen diagrammatischen Operationen zu reflektieren. Dafür nötig sind semiotische und natürlich auch wissenschaftstheoretische Überlegungen, die für alle Disziplinen, die mit maschineller Textanalyse befasst sind, relevant sein müssten.

Fußnoten

1. Dieses „extended Abstract“ ist eine verkürzte und angepasste Fassung des stärker linguistisch ausgerichteten Beitrages von Bubenhofer (im Druck a).
2. Vgl. für eine aktuelle linguistisch motivierte Diskussion von stilometrischen Messmethoden für die Autorschaftsattribution Büttner et al. (2017).

Bibliographie

Bubenhofer, Noah (im Druck a): Wenn „Linguistik“ in „Korpuslinguistik“ bedeutungslos wird. Vier Thesen zur Zukunft der Korpuslinguistik. In: Osnabrücker Beiträge zur Sprachtheorie (OBST).

Bubenhofer, Noah (im Druck b): Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In: Bubenhofer, Noah / Kupietz, Marc (Hg.): Visualisierung sprachlicher Daten. Heidelberg: HeiUP.

Büttner, Andreas / Dimpel, Friedrich Michael / Evert, Stefan / Jannidis, Fotis / Pielström, Steffen / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten (2017): »Delta«#in der stilometrischen Autorschaftsattribution. In: Zeitschrift für digitale Geisteswissenschaften. text/html Format. DOI: 10.17175/2017_006.

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

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

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

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.* Bd. 60, Nr. 3, S. 538–556

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