Perspektiven auf ein Korpus. Kombinationen quantitativ-qualitativer Analysemethoden zur Ermittlung von Textgliederungsprinzipien

Haaf, Susanne
haaf@bbaw.de
Berlin-Brandenburgische Akademie der Wisesnschaften, Deutschland

Einführung

Im Bereich digital basierter Untersuchungen wird zunehmend eine Verzahnung quantitativen und qualitativen Arbeitens gefordert. In der konkreten Arbeit der Korpusanalyse wird aus dieser scheinbaren Dichotomie jedoch schnell eine Methodenvielfalt, denn gerade durch Kombinationen verschiedener Perspektiven auf die Daten werden unterschiedliche Phänomene greifbar und entfaltet sich das volle Potential quantitativ-qualitativen Arbeitens. Das hier präsentierte Poster soll dies an einem konkreten Beispiel veranschaulichen.

Fragestellung

Inhaltlicher Ausgangspunkt ist die Frage nach Textgliederungsprinzipien, welche für bestimmte erbauliche Textsorten kennzeichnend sind. Mittel der Textgliederung können als Mittel der Markierung von Teiltexten innerhalb eines Gesamttextes beschrieben werden (Hausendorf/Kesselheim 2008: 41) und können wiederum für Textsorten charakteristisch sein. Sie finden sich auf verschiedenen Ebenen des Textes, u.a. im Bereich der Typographie (Stein 2003: 422).

Gerade diese typographischen Gliederungsmerkmale stellen einen guten Ausgangspunkt für eine quantitative Analyse von Textgliederungsmerkmalen dar, da sie in TEI-annotierten Korpora z.B. durch die XML-Strukturierung automatisch greifbar werden. Anders als bei dem Verfahren, textbezogene Phänomene reduziert auf bestimmte TEI-Strukturen zu untersuchen (z.B. Schöch 2016: 351ff., Haaf 2016), gelangen hier die im Korpusvergleich möglicherweise signifikanten Häufigkeiten der TEI-Strukturierungen selbst in den Blick.

Die inhaltliche Frage nach Textgliederungsprinzipien erbaulicher Textsorten wird ausführlich behandelt in Haaf (in Vorber.). Im vorliegenden Beitrag stehen – der thematischen Ausrichtung der Konferenz entgegenkommend – Überlegungen zur adäquaten Methodik einer solchen Untersuchung im Vordergrund.

Korpus- und Analysegrundlage

Der hier präsentierten Studie liegen drei Teilkorpora des 17. Jahrhunderts aus dem Deutschen Textarchiv (2017) zugrunde:

- Prosaische Erbauungsliteratur: 25 Bände (10 Autoren, 10.501 Seiten)
- Funeralschriften: 334 Schriften (14.316 Seiten)
- Referenzkorpus: 187 Bände verschiedener Textsorten (60.798 Seiten)

Die Texte des DTA-Korpus wurden nach einheitlichen Richtlinien und mittels eines TEI-Subsets, das Ambiguitäten der Auszeichnung möglichst reduziert, ausgezeichnet (Haaf et al. 2014/15).

Für die vorliegende Untersuchung wurden einzelne TEI-Strukturen hinsichtlich der Häufigkeit ihres Auftretens (relativ zur Token-Anzahl) und ihrer Verteilung im jeweiligen Korpus verglichen, um speziell die Unterschiede in der Textgliederung zwischen den untersuchten Textsorten herauszuarbeiten. Dabei wurden solche Tags einbezogen, die voraussichtlich Textgliederungsmerkmale repräsentieren. So kann z.B. tei:div die Kapitelstruktur eines Textes anzeigen, durch tei:l wird der Wechsel zwischen Prosatext und Lyrik greifbar, tei:note zeigt Metatexte in Form von Anmerkungen, z.B. Marginalien, an, tei:hi repräsentiert Hervorhebungen von Textpassagen gegenüber dem Grundtext. Die Ergebnisse wurden einer qualitativen Beurteilung unterzogen.

Ergebnisse

Zur Evaluation eines Merkmals war hier nicht allein der Blick auf seine relativen Häufigkeiten in und deren signifikante Unterschiede zwischen den untersuchten Korpora relevant. Die signifikant erhöhte Häufigkeit eines Merkmals kann vielmehr unterschiedliche Gründe haben. So kann sie einerseits zwar durchaus (1) auf die höhere Relevanz des Merkmals im Korpus hindeuten, wie sich am Merkmal der Marginalie zeigt, das signifikant häufig und regelmäßig verteilt im Korpus der Funeralschriften auftritt (Abb. 1). Sie kann andererseits aber auch (2) aufgrund der unausgewogenen Verteilung des Merkmals im Korpus gar nicht aussagekräftig sein, entweder weil (2a) sich das Korpus selbst als in sich unausgewogen und nicht repräsentativ für den zu beschreibenden Gegenstand erweist oder weil (2b) das Merkmal im gegebenen Kontext nicht relevant ist, wie etwa die horizontale Trennlinie zwischen Textteilen, die in allen drei Vergleichskorpora unregelmäßig verteilt war (Abb. 2). Andererseits kann es (3) auch vorkommen, dass die bestehende Ausgewogenheit

der Verteilung eines Merkmals in einem Korpus letztlich nicht aussagekräftig für dessen Relevanz ist.

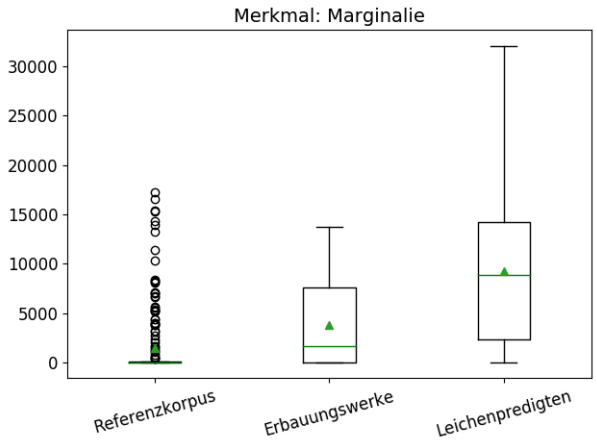

Abb. 1: Relative Häufigkeiten je 1 Mio. Token und deren Verteilung in den drei untersuchten Korpora für das Merkmal „Marginalie“

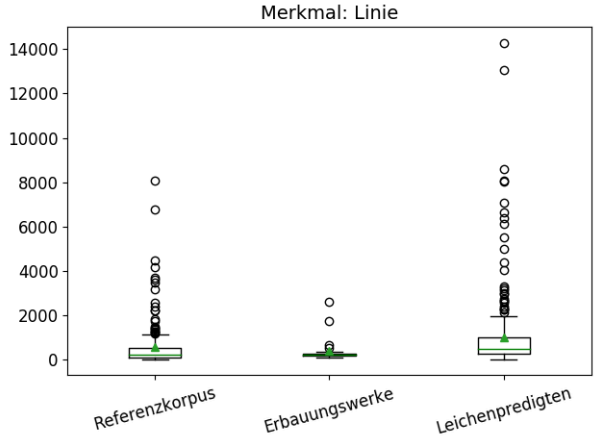

Abb. 2: Relative Häufigkeiten je 1 Mio. Token und deren Verteilung in den drei untersuchten Korpora für das Merkmal „Horizontale Trennlinie“

Weiterhin konnten im gegebenen Kontext (4) auch Merkmale mit geringeren Häufigkeiten relevant sein, und schließlich ist nicht zuletzt (5) auch der Ort, an dem ein Merkmal im Dokument auftritt, zu berücksichtigen. Beide Aspekte (4 und 5) zeigen sich z.B. am Merkmal der Liste, die in den erbaulichen Prosawerken erwartungsgemäß selten, aber relativ regelmäßig auftritt, und zwar in Form von Registern am Buchbeginn oder Buchende (in tei:front oder tei:back).

Methodisch zeigte sich also, dass für eine adäquate Beurteilung der untersuchten Merkmale verschiedene Blickwinkel notwendig sind. Das Poster veranschaulicht anhand der erwähnten und weiterer Beispiele diese genannten methodischen Aspekte.

Inhaltlich führte die Untersuchung zutage, dass z.B. Merkmale, die den Zugang zum Text erleichtern und Hilfe zur Orientierung im Text geben, für die erbaulichen Textsorten relevant sind (Näheres vgl. Haaf (in Vorber.)).

Bibliographie

Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin 2017. http://www.deutschestextarchiv.de [letzter Zugriff: 24.09.2017]

Haaf, Susanne (i. Vorb.): „Art und Funktion von typographischen Mitteln zur Textgliederung in erbaulichen Textsorten des 17. Jahrhunderts. Automatische Analyse im Korpusvergleich und qualitative Einordnung“, in: Simmler, Franz / Baeva, Galina (Hrsg.): *Textgliederungsprinzipien. Ihre Kennzeichnungsformen und Funktionen vom 8. bis 18. Jahrhundert. Akten zum Internationalen Kongress vom 22. bis 24. Juni 2017 an der Universität St. Petersburg*. Berlin: Weidler [2018].

Haaf, Susanne (2016): “Corpus Analysis based on Structural Phenomena in Texts. Exploiting TEI Encoding for Linguistic Research”, in: Nicoletta Calzolari et al.: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 23.–28. Mai 2016, Portorož (Slovenia)*. Paris.

Haaf, Susanne / Geyken, Alexander / Wiegand, Frank (2014/15): “The DTA ‘Base Format’. A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources”, in: *Journal of the Text Encoding Initiative* 8.

Hausendorf, Heiko / Kesselheim, Wolfgang (2008): *Textlinguistik fürs Examen*. Göttingen: Vandenhoeck & Ruprecht.

Schöch, Christoph (2016): „Ein digitales Textformat für die Literaturwissenschaften. Die Richtlinien der Text Encoding Initiative und ihr Nutzen für Textedition und Textanalyse“, in: *Romanische Studien* 4.

Stein, Stephan (2003): *Textgliederung. Einheitenbildung im geschriebenen und gesprochenen Deutsch. Theorie und Empirie*, Berlin.