Digitale Korpora zur Sprachgeschichte jenseits von Morphologie und Syntax: Einige konzeptionelle Überlegungen mit Beispielen

Wolf Peter Klein (Universität Würzburg)

Aus sprachhistorischer Sicht gibt es üblicherweise einige zentrale Wünsche, die man mit der Erstellung eines digitalen Korpus verbindet: Es soll möglichst umfangreich und repräsentativ sein, daneben gut gehandhabt und durchsucht werden können, alles reichhaltig morphosyntaktisch annotiert sowie mit einschlägigen Metadaten und Schreib- und Schriftrepräsentationen versehen. Diese Wünsche sind natürlich sehr nachvollziehbar und werden die Erstellung digitaler Korpora zur Sprachgeschichte mit Recht auch in Zukunft bestimmen. In Ergänzung dazu soll im Vortrag darauf hingewiesen werden, dass angesichts der immer zahlreicheren Digitalisierungen, die von verschiedenen Anbietern überall auf der Welt zur Verfügung gestellt werden, auch alternative sprachhistorische Möglichkeiten der Erstellung und Nutzung digitaler (Quasi-?) Korpora im Raum stehen. Ausgangspunkt ist dabei der Umstand, dass digitale Objekte in Zukunft sicherlich stärker und häufiger in stabilen, frei verfügbaren Internet-Umgebungen verankert sein werden als zu Beginn des digitalen Zeitalters. Technisch sind sie über feste URL-Adressen („Perma-Links“ o.ä.) greifbar. Diese digitalen Objekte (z.B. Handschriften-, Buch-, Zeitschriften-, Plakat-, Postkarten-, Ton-, Filmdigitalisate) lassen sich in übergeordneten Projekten mit Blick auf bestimmte sprachhistorisch relevante Fragestellungen nutzen und vernetzen, ohne dass im Projekt selber Digitalisierungen sprachhistorischer Quellen erstellt werden müssten. Der Vorteil solcher Projekte liegt also darin, dass sie auf die arbeits- und zeitaufwändige Erstellung eigener Digitalisierungen verzichten können; zudem ist die Zahl möglicher Korpus-Einheiten potenziell sehr groß. Der Nachteil solcher Projekte liegt darin, dass nur diejenigen Dinge für die linguistische Analyse genutzt werden können, die anderenorts bereits in die Digitalisierung gesteckt wurden. Auf morphosyntaktische (oder ähnliche, unmittelbar linguistisch motivierte) Annotationen lässt sich folglich in der Regel nicht zurückgreifen. Stattdessen können aber Analysekategorien entwickelt und genutzt werden, die einem bestimmten digitalen Objekt als Ganzes (z.B. einem einzelnen Handschriften-, Buch-, Zeitschriften-, Plakat-, Postkarten-, Ton-, Filmdigitalisat) oder relevanten Teilen daraus (z.B. einer einzelnen Handschriften-, Buch- oder Zeitschriftenseite) zugeordnet werden können. Die Projekt-Zuschnitte und die Analyse-Möglichkeiten, die in solchen Korpus-Perspektiven denk- und machbar sind, sollen im Vortrag auch mit zwei Datenbanken veranschaulicht werden, die an der Universität Würzburg mit der Semantic MediaWiki – Software auf den Weg gebracht wurden: einerseits ZweiDat, eine Datenbank, die sprachliche Zweifelsfälle aufgreift und zur Erforschung des neuhochdeutschen Sprachkodex herangezogen werden kann, andererseits FTDB, eine Datenbank, die einen systematischen Zugang zu Quellentexten anbietet, die für die Analyse der frühen Geschichte der deutschen Wissenschaftssprache einschlägig sind.