Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken

Zusammenfassung

Die vorliegende Arbeit setzt sich mit verschiedenen Aspekten der Nutzung korpusbasierter Statistiken in quantitativen typologischen Untersuchungen auseinander. Die einzelnen Abschnitte der Arbeit können als Teile einer automatischen und sprachunabhängigen Prozesskette angesehen werden, die somit umfassende Untersuchungen zu den verschiedenen Sprachen der Welt erlaubt. Es werden dabei die Schritte von der Erstellung der grundlegenden Ressourcen über die mathematisch fundierten Methoden bis hin zum fertigen Resultat der verschiedenen typologischen Analysen betrachtet. Außer durch die Nutzung webbasierter Korpora hebt sich die Arbeit dabei durch strikte Automatisierung von anderen Arbeiten des Feldes ab.

Hauptaugenmerk der Untersuchungen liegt zunächst auf den Textkorpora, die der Analyse zugrundeliegen. Dabei werden auf der einen Seite die Beschaffung aus dem World Wide Web und auf der anderen Seite die Verarbeitung unter technischen Gesichtspunkten betrachtet. Unter Nutzung der vorgestellten Verfahren werden Textkorpora in vielen hundert Sprachen erstellt.

Es schließen sich Abhandlungen zur Nutzung der Korpora im Gebiet des lexikalischen Sprachvergleich an, wobei eine Quantifizierung sprachlicher Beziehungen mit empirischen Mitteln erreicht wird. Als Grundlage dienen dabei Orthographie oder Verteilung der im Text vorkommenden Wortformen.

Darüber hinaus werden die Korpora als Basis für automatisierte Messungen sprachlicher Parameter verwendet. Zum einen werden derartige messbare Eigenschaften vorgestellt, zum anderen werden sie hinsichtlich ihrer Nutzbarkeit für sprachtypologische Untersuchungen systematisch betrachtet. Abschließend werden Beziehungen dieser Messungen untereinander und zu sprachtypologischen Parametern mit quantitativen Methoden untersucht. Aufgrund des strikten Einsatzes sprachunabhängiger und automatischer Verfahren, können dabei umfassende Textressourcen in mehreren hundert Sprachen untersucht werden, was zu Ergebnissen hoher statistischer Signifikanz führt.