Zur Relevanz der Kombinierbarkeit von Metadaten bei der korpus-linguistischen Untersuchung von Neologismen

Sascha Wolfer, Leibniz-Institut für Deutsche Sprache Mannheim

Abstract

Das IDS-Neologismenwörterbuch bietet eine inzwischen fast dreißig Jahre umspannende Datenbasis von fast 2000 redaktionell bearbeiteten Artikeln zu neuen Wörtern in der deutschen Sprache. Eine solche Datenbasis ermöglicht u.a. auch korpuslinguistische Untersuchungen, die von der Ebene des Einzellemmas abstrahieren. In diesem Vortrag stelle ich eine Beispieluntersuchung anhand von 239 Lemmata vor, in der wir uns mit der folgenden Forschungsfrage beschäftigen: Werden aus anderen Sprachen entlehnte Neologismen zögerlicher in der deutschen Sprache „akzeptiert" als Neologismen, die Wortbildungsprodukte des Deutschen darstellen? Diese Frage wurde in Form einer Behauptung im Call for Papers einer Konferenz formuliert, was den Ausgangspunkt der Untersuchung darstellte.

Hierzu haben wir Abfragen in allen Archiven von Cosmas 2, dem Abfragesystem für das Deutsche Referenzkorpus am IDS, formuliert, was ca. 5,5 Millionen Treffer ergab. Da die oben genannte Forschungsfrage eine zeitliche Dimension der Untersuchung nahelegt, wurde zunächst der Frequenzverlauf der verschiedenen Lemmata über die Zeit hinweg analysiert. In einem nächsten Schritt nutzten wir die KWIC-Ergebnisse, um linguistisch markierte Vorkommen zu identifizieren. Die häufigste Art der linguistischen Markierung (in der Forschungsliteratur auch als „flagging" bezeichnet) sind Anführungszeichen um die jeweilige Wortform. Es sind auch andere Formen der Markierung möglich, bspw. das Hinzufügen von „sogenannt" vor der Wortform, nachgestellte Erklärungen, die mit „das heißt" oder ähnlichen Formen eingeleitet werden und einige mehr. Aufgrund der Ergebnismenge lassen sich diese Markierungen nur anhand automatisierter Heuristiken identifizieren.

Im Vortrag werde ich insbesondere darauf eingehen, wie relevant Metadaten gerade bei dieser Art von automatisierter Analyse von Massendaten sind. Ich werde zeigen, dass jede Untersuchung an eine unüberwindbare Grenze stößt, wenn das jeweilige Korpus-Abfragesystem nicht zu *jedem einzelnen Treffer* eine beliebige Anzahl von Metadaten assoziiert. Ich argumentiere weiter, dass die Kreuzkombination von Metadaten *kein* methodisches „Nischenbedürfnis" ist, das womöglich nur für die Untersuchung von Neologismen relevant ist, sondern für Analysen auf allen linguistischen Ebenen unabdingbar ist.