Protokoll zur Sitzung am 10.7.17 (Kolloquium)

Katja Bertholdt: das Zipfsche Gesetz

Am Anfang des Vortrags wurde kurz über George Kingsley Zipf - einen US-amerikanischen Linguisten berichtet, der das vorzustellende Gesetz entwickelte: Zipf studierte an der Harvard University und kam nach seinem Studienabschluss 1925 nach Deutschland. Er hat an der Harvard University promoviert, danach war er als Lehrer für Deutsch und ab 1936 Assistant Professor of German (Germanistik) an der Harvard University tätig.

Das Zipfsche Gesetz beschreibt ein Modell, mit dessen Hilfe es möglich ist, bei bestimmten Größen, die in eine Rangfolge gebracht werden, deren Wert aus ihrem Rang abschätzen zu können. Das Gesetz findet häufige Verwendung in der Linguistik, aber nicht nur da, sondern in vielen Natur- und Sozialwissenschaften. Zipf zeigte beispielsweise auch, dass die Rangverteilung der Einwohnerzahlen amerikanischer Städte dem Zipfschen Gesetz folgt. Außerdem entdeckte der Ökonom Damaso Pareto bereits 1896, dass die Einkommen in einer Volkswirtschaft Zipf-verteilt sind.

Für Linguistik kann man das Gesetz so beschreiben: der Rang i eines Wortes ist indirekt proportional zu seiner relativen Häufigkeit; wenn z.B. Wörter nach der Häufigkeit sortiert werden, ist die Wahrscheinlichkeit ihres Auftretens umgekehrt proportional zur Position innerhalb der Reihenfolge:

$$p(n) \sim \tfrac{1}{n}.$$

Daraus kann man schließen, dass die meisten Wörter selten sind.

Anhand des Prinzips der geringsten Anstrengung wurde das Gesetz nochmals erläutert. Man muss aber konstatieren, dass die Mechanismen, die dem

Zipfschen Gesetz zu Grunde liegen, bis heute immer noch nicht vollständig verstanden sind.

Die Vortragende hat für Demonstration des Gesetzes das Projekt „Deutscher Wortschatz“ (Uni Leipzig) benutzt, das eine Suche in über 26 Millionen deutscher Sätzen (Zeitungstext) möglich macht. Alle Beispiele, die die Vortragende mithilfe dieses Projektes durchgerechnet hat (u.A. „Wie viele Wörter treten mindestens, bzw. genau n mal auf?“, „Wie groß ist das Vokabular?“) entsprachen dem Zipfschen Gesetz, das zusätzlich mithilfe mehrerer Graphiken veranschaulicht wurde. Auffallend war u. A. die Tatsache, dass etwa die Hälfte des Vokabulars wahrscheinlich nur einmal auftritt (Hapax legomena).

Es folgte eine kurze Diskussion über die Bedeutung des Gesetzes für Computerlinguistik, insbesondere über statistische Methoden in der Sprachverarbeitung.