Klyueva Olga
CIS, LMU München
10.07.1917

Katja Bertholdt: Zipfisches Gesetz.

In dem Vortrag geht es um das Zipfische Gesetz, das von dem amerikanischen Linguisten George Kingsley Zipf erfunden wurde. Er versuchte die Linguistik als Naturwissenschaft zu betrachten.
Das Zipfische Gesetz enthält das Prinzip der geringsten Anstrengung. In den natürlichen Sprachen bedeutet es, dass bestimmte Worte häufiger vorkommen als andere. Die Worte, die selten vorkommen, sind lange Worte. Die Worte, die oft vorkommen, sind kurz und sind meistens Funktionsworte, die keine große Bedeutung haben.
Als nächstes wurde eine Zipfische Formel erwähnt. Für die Formel braucht man die Häufigkeit des Wortes in einem Text und einen Rang für jedes Wort. Das häufigste Wort hat einen Rang eins, das nächste Rang zwei und so weiter.
Eine Zipfische Formel sagt, dass das Produkt aus Rang und Häufigkeit über einen Text etwa konstant ist. Mit der entstehenden Zipf-Verteilung lässt sich beschreiben, dass beispielweise das Wort mit dem Rang zwei durchschnittlich nur halb so oft im Text vorkommt wie das Wort mit dem Rang eins. Das Wort mit dem Rang drei tritt im Text drei Mal seltener auf als das Wort mit dem Rang eins und so weiter.
Für die Anwendung der Formel wurde das Projekt Deutscher Wortschatz ausgewählt. Das Projekt besteht aus online-verfügbaren Archiven von Zeitungstexten und Fachtexten. Die Anzahl der Types enthält fünf Millionen, die Anzahl der Tokens über 20 Millionen.
Um die Formel zu überprüfen, wurde eine Häufigkeitsliste erstellt und der jeweilige Rang zu jedem Wort zugewiesen. Man bildet das Produkt aus der Häufigkeit und dem jeweiligen Rang (10, 100, 500, 1000, 5000). Die Konstante für den betrachteten Korpus beträgt 18 Millionen.
Es wurde gezeigt, was man mit der Formel noch berechnen kann. Mit der Formel wurde die Anzahl der Wortformen berechnet, die mindestens 100 Mal in dem Projekt Deutscher Wortschatz vorkommen. Die Anzahl der Wortformen lässt sich auf 178 031 schätzen.
Man kann mit der Formel schätzen, wie groß das gesamte Vokabular eines Textes ist. Dazu wird die Annahme getroffen, dass das seltenste Wort nur einmal vorkommt. Die Größe des Vokabulars lässt sich auf circa 70 Millionen schätzen.
Es wurde berechnet wieviele Worte 50 Mal vorkommen. Dazu berechnet man die Formel für den untersten Rang von 50 und für den untersten Rang von 51 und zieht die Resultate voneinander ab. Das Resultat beträgt über 80 Millionen Wortformen, die genau 50 Mal im Projekt Deutscher Wortschatz vorkommen.
Die Konsequenzen der Zipfischen Verteilung bestehen darin, dass es schwer ist, das Verhalten des Wortes in unterschiedlichen Kontexten und die Eigenschaften des Wortes festzustellen.
Zweite Konsequenz der Zipfischen Verteilung ist, dass die ersten vier Worte 10 Prozent des gesamten Korpus im Projekt Deutscher Wortschatz abdecken.
Am Schluss wurden die Bereiche, bei denen die Zipfische Verteilung benutzt wird, präsentiert:

- Grössenverteilung von Städten
- Verteilung von Macht und sozialer Status
- Musik (Noten in einem Stück)
- Malerei (Bestimmung der Anteilsverteilung von Farbflächen)
- Internet (Häufigkeit des Aufrufens von Internetseiten)
- Kommunikationsverhalten zwischen Tieren