1. Zipf’sches Gesetz - Verständnisfragen

Bewerten Sie die folgende Aussage als richtig oder falsch und begründen Sie Ihre Entscheidung kurz!

(a) Stellt man die Ränge und Häufigkeiten von Wörtern aus einem Korpus in einem Koordinatensystem dar, erhält man nach dem Zipfschen Gesetz eine Gerade.

(b) Aus dem Zipfschen Gesetz folgt, dass die Anzahl verschiedener Wortformen (types) in einem Text proportional zu seiner Länge ist.

(c) Trägt man die Kurven zum Zipfschen Gesetz für mehrere Korpora einer Sprache in einem doppelt logarithmischen Koordinatensystem ein, erhält man näherungsweise parallele Kurven.

(d) Die sprachspezifische Konstante c ist für verschiedene Sprachen unterschiedlich.

2. Formulierung und Anwendung des Zipfschen Gesetzes

Gegeben seien u. a. aus einer Gesamtzahl von 71.370 Wortformen (tokens) die folgenden Daten fürs Englische:

Nr.	**Wort**	**Absolute Häufigkeit**	**Rang**
1	he	877	10
2	but	410	20
3	comes	16	500
4	applausive	1	8000

(a) Leiten Sie aus den Daten das sog. Zipf'sche Gesetz in zwei Formulierungen ab, und zwar zum einen bezogen auf einen **Text** und zum anderen bezogen auf eine **Sprache**. Welchen Wert hat die Zipf'sche Konstante fürs **Englische** und fürs **Deutsche**?

(b) Wie hoch wird die Anzahl unterschiedlicher Wortformen **(types)** sein? Wie viele davon werden nur ein Mal im Text vorkommen? Bezogen auf die Gesamtzahl der **Tokens**: Wie hoch wird der Anteil an Wortformen sein, die nur ein Mal im Text vorkommen?

3. Weitere Anwendungen des Zipf’schen Gesetzes

Gegeben seien ein Textkorpus des Deutschen mit 1 Million Sätzen und 16 015 429 tokens.

(a) Wie viele Wörter kommen nach Abschätzung des Zipfschen Gesetzes 100 mal oder öfter in diesem Text vor?

(b) Wie groß ist nach Abschätzung des Zipfschen Gesetzes das Vokabular?

(c) Wie viele Wörter kommen nach Abschätzung des Zipfschen Gesetzes genau 100 mal in diesem Text vor?

(d) Wie viele Wörter kommen nur 1 mal in diesem Text vor?