Frequenzklassenberechnung für die Wortlistenfunktion in der DGD

Thomas Schmidt, Januar 2021

Die Wortlistenfunktion führt neben der absoluten Häufigkeit von Wortformen auch deren Frequenzklasse auf. Die Frequenzklasse ist ein relatives Maß, das die Häufigkeit einer Wortform in Bezug zu den Häufigkeiten anderer Wortformen aus derselben Grundgesamtheit (demselben Korpus) in Bezug setzt. Weil sie von der Korpusgröße unabhängig ist, können über die Frequenzklasse Häufigkeiten zwischen Korpora oder zwischen Teilmengen eines Korpus gut verglichen werden.

Wie bei jeder Quantifizierungsmethode ist auch bei der Berechnung von Frequenzklassen und bei Frequenzklassenvergleichen methodische Sorgfalt und Vorsicht geboten. Die An- oder Abwesenheit von Frequenzklassenunterschieden zwischen zwei Korpora kann vielfältige Gründe haben, die bei einer Interpretation der Zahlen alle in Betracht gezogen werden sollten. Dieses Dokument geht hierauf nicht näher ein, sondern beschränkt sich darauf zu erklären, wie in der DGD Frequenzklassen berechnet werden.

Die Frequenzklasse einer Form ***X*** mit absoluter Häufigkeit ***n*** berechnet sich wie folgt:

Wenn ***N*** die absolute Häufigkeit der häufigsten Form (***Y***) im Korpus ist, dann ist:

$$FC(X) = log_2 \, {}^{N}/_{n} + 0.5$$

Also zum Beispiel für Lemmata in FOLK:

- ***X*** = „Käse“ mit absoluter Häufigkeit ***n*** = 188
- ***Y*** = „d“ (Lemmatisierung von Artikeln) mit absoluter Häufigkeit ***N*** = 103.904
- FC („Käse“) = $\log_2$ (103904/188) + 0.5 = $\log_2$ (552.68) + 0.5 = 9.11 + 0.5 = 9.61
- Also hat das Lemma „Käse“ in FOLK die Frequenzklasse 9

… oder für normalisierte Formen in FOLK:

- ***X*** = „kannst“ mit absoluter Häufigkeit n = 2.249
- ***Y*** = „ja“ mit absoluter Häufigkeit N = 86.875
- Frequenzklasse 5

Voraussetzung für die Berechnung von Frequenzklassen auf einzelnen Korpora ist, dass ***N*** für alle vier Annotationsebenen (transkribiert, normalisiert, lemmatisiert und POS) bekannt ist.

Wie die folgende Tabelle illustriert, sind die häufigsten Formen zwischen Annotationsebenen innerhalb eines Korpus und zwischen Korpora teilweise identisch, teilweise verschieden. Dies ist auch abhängig von der Transkriptionsmethode (hier z.B.: literarisch bei FOLK und GWSS, orthographisch bei DH und ZW) und vom verwendeten Tagger und Tagset.

	Transkribiert	Normalisiert	Lemma	POS
FOLK	ja – 82.823	ja – 86.875	d – 103.904	NN – 281.425
GWSS	äh – 26.784	äh – 45.625	äh – 45.630	NN – 102.577
DH	und – 154.080	und – 154.080	die – 362.746	NN – 1.364.494
ZW	und – 118.824	und – 118.824	d – 338.015	ADV – 639.357

Bei Anfragen mit **mehr als einem Korpus** wird die gemeinsame Wortliste für alle beteiligten Korpora zugrunde gelegt, um ***N*** zu ermitteln (weil Y nicht unbedingt für jedes Korpus gleich ist, können die Werte nicht einfach addiert werden), z.B. für Lemmata in FOLK und GWSS:

- FOLK: ***Y*** = „d“ (Lemmatisierung von Artikeln) mit absoluter Häufigkeit ***N*** = 103.904
- GWSS: ***Y*** = „äh“ (Häsitationspartikel) mit absoluter Häufigkeit ***N*** = 45.630
- FOLK+GWSS: ***Y*** = „d“ (Lemmatisierung von Artikeln) mit absoluter Häufigkeit ***N*** = 143.811

Für Anfragen mit **virtuellen Korpora** wird ***N*** angenähert[1], indem zunächst die absolute Häufigkeit ***N‘*** für alle am virtuellen Korpus beteiligten Korpora ermittelt wird und ***N*** von diesem Wert dann anteilig gemäß der Größe des virtuellen Korpus ***M*** im Verhältnis zum Umfang der Gesamtkorpora ***M‘*** ermittelt wird.

Zum Beispiel für ein virtuelles Korpus mit allen männlichen Sprechern aus FOLK:

- Umfang des virtuellen Korpus: ***M*** = 1.312.158 Tokens
- Umfang des Gesamtkorpus (FOLK): ***M‘*** = 2.719.950 Tokens
- Absolute Häufigkeit der häufigsten Form Y in FOLK ***N‘*** = 103.904
- Angenäherte maximale Frequenz für das virtuelle Korpus: $N = {}^{M}/_{M'} \times N' = {}^{1.312.157}/_{2.719.950} \times 103.904 = 50.125$

… oder für ein virtuelles Korpus mit allen Aufnahmen aus Schleswig-Holstein aus DH und ZW:

- Umfang des virtuellen Korpus: ***M*** =612.111 Tokens
- Umfang der Gesamtkorpora (DH + ZW): ***M‘***= 6.274.343 + 4.063.581 = 10.337.924 Tokens
- Absolute Häufigkeit der häufigsten Form Y in DH+ZW: ***N‘*** = 143.811
- Angenäherte maximale Frequenz für das virtuelle Korpus: $N = {}^{M}/_{M'} \times N' = {}^{612.111}/_{10.337.924} \times 143.811 = 8.515$

[1] Die präzise Berechnung von N wäre mit den aktuellen DGD-Strukturen zu aufwändig. Der angenäherte Wert sollte für viele praktische Zwecke hinreichend genau sein – methodische Vorsicht ist jedoch geboten.