

Frequency of Low-frequency Words in Text Corpora

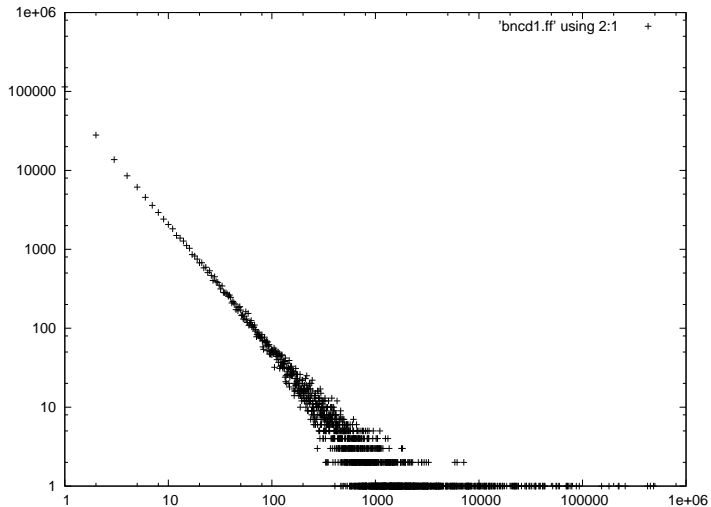
Pavel Rychlý

pary@fi.muni.cz

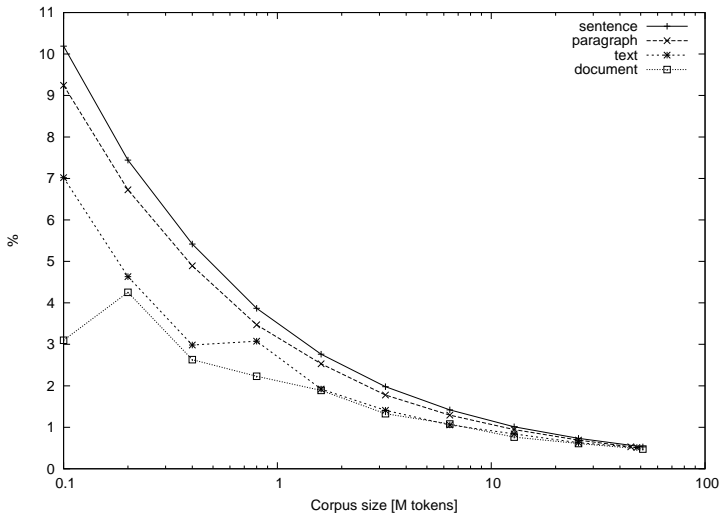
Contents

- 1 Freq-Freq
- 2 Stability of Frequences
- 3 Sampling units
- 4 Stability in documents

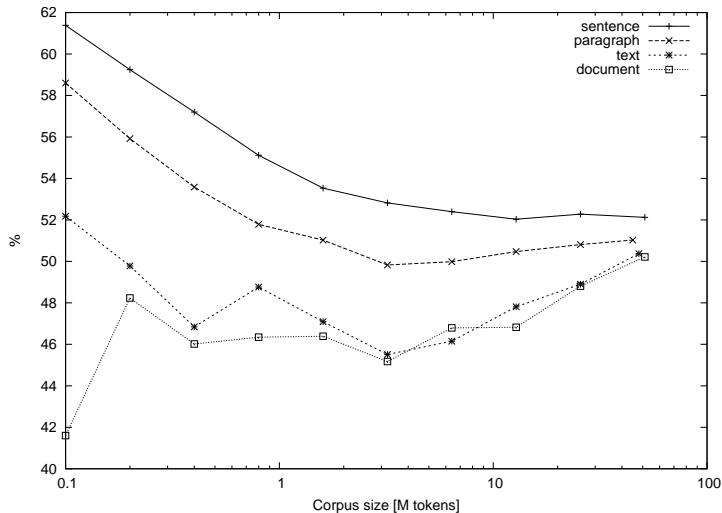
Freq-Freq – Frequencies of Frequencies



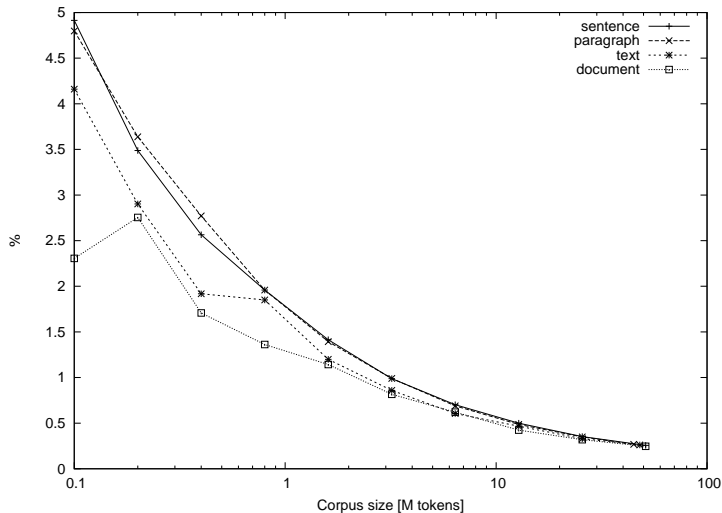
Percentage of text covered by hapax legomena



Percentage of word types of hapax legomena



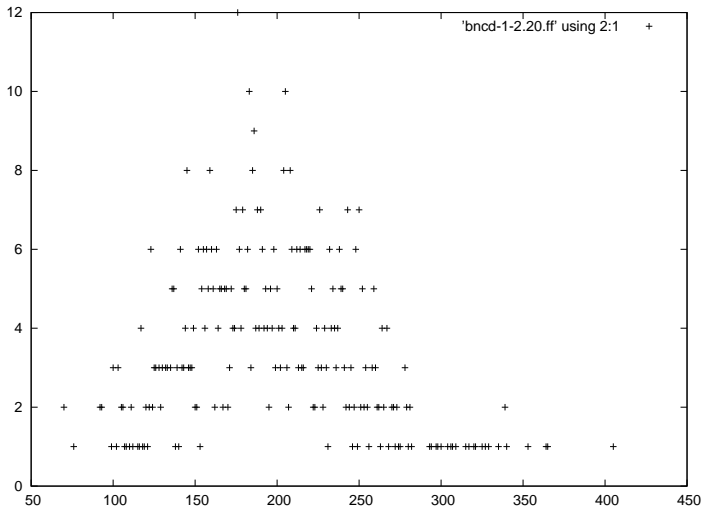
Percentage of word types of dis legomena



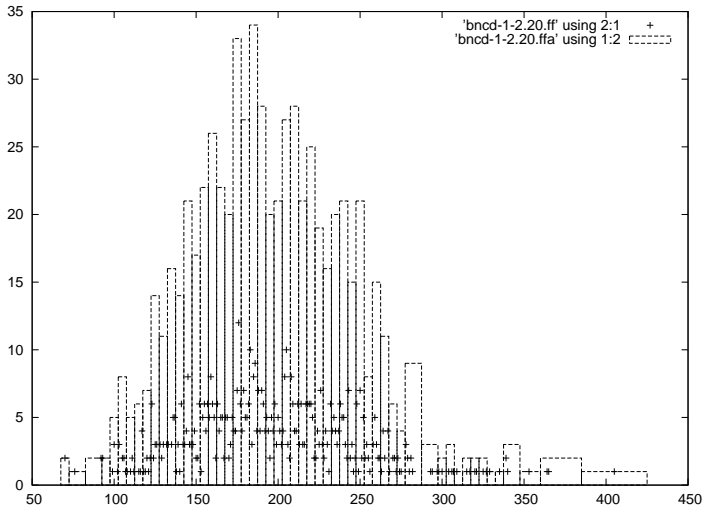
Contents

- 1 Freq-Freq
- 2 Stability of Frequences**
- 3 Sampling units
- 4 Stability in documents

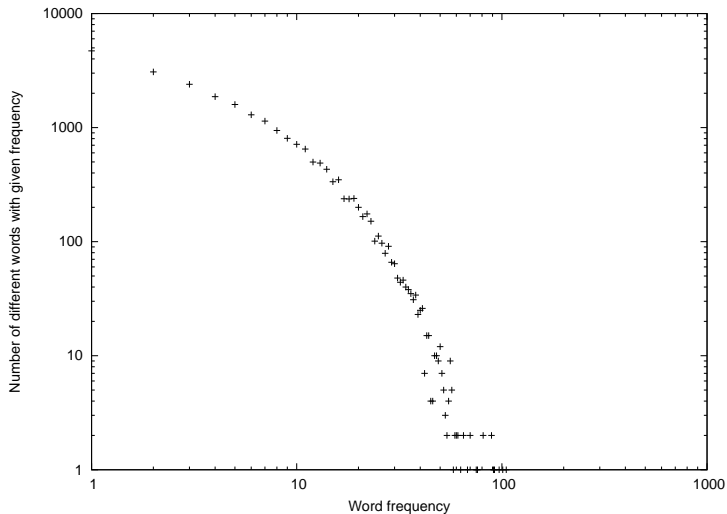
Stability 10M-100M, freq=20



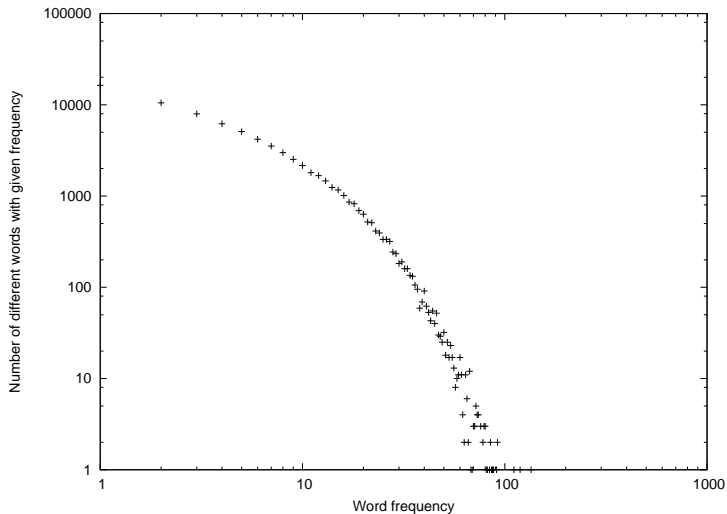
Stability 10M-100M, freq=20, aggregated(5)



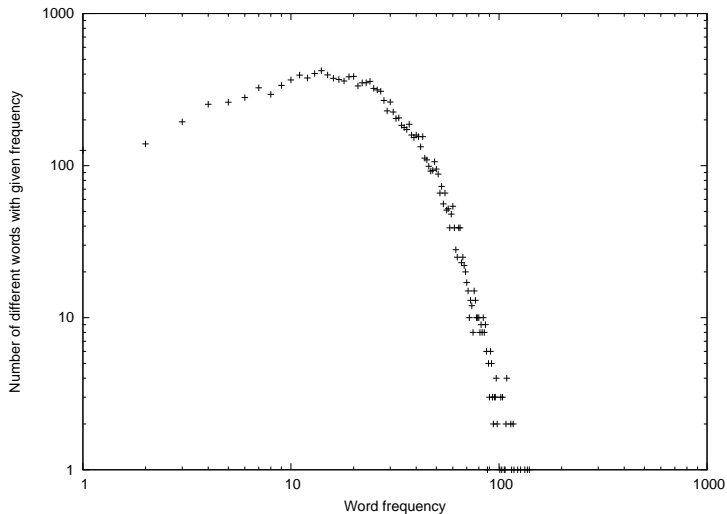
Hapax Legomena, 1M-10M



Hapax Legomena, 10M-100M



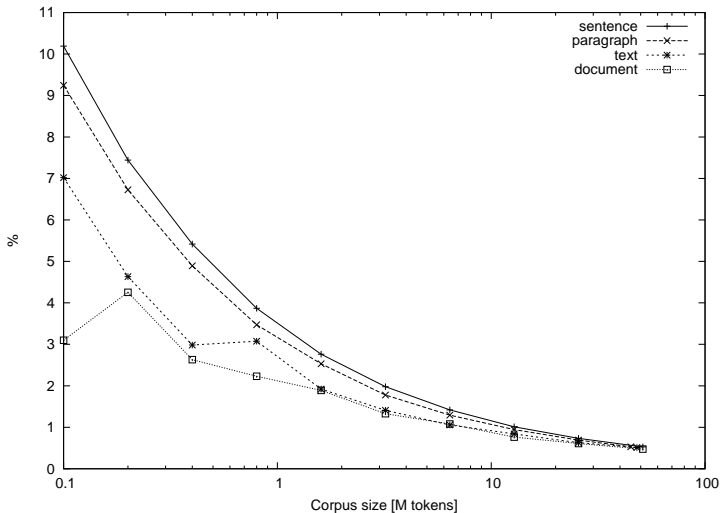
Tris Legomena, 10M-100M



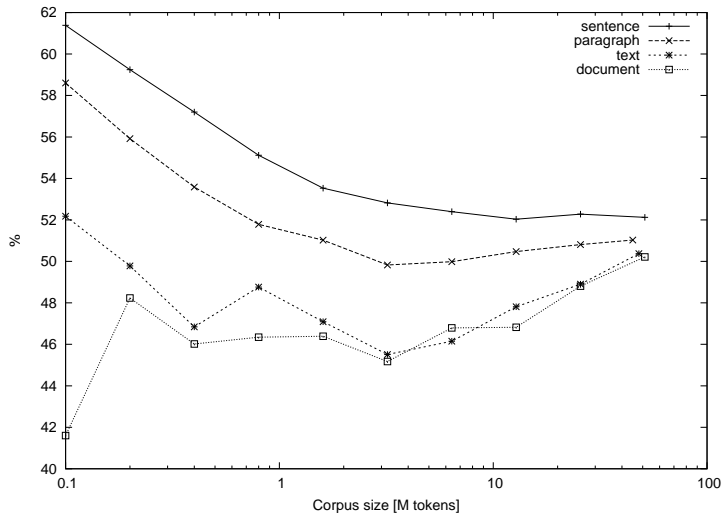
Contents

- 1 Freq-Freq
- 2 Stability of Frequences
- 3 Sampling units**
- 4 Stability in documents

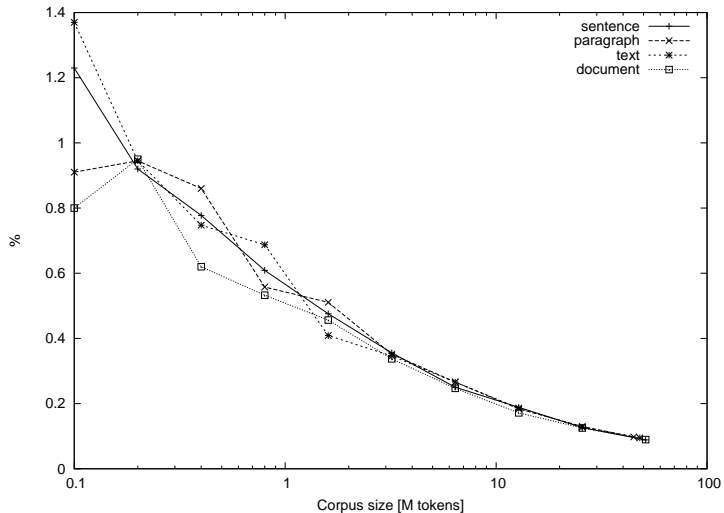
Percentage of text covered by hapax legomena



Percentage of word types (size of the corpus lexicon) of hapax legomena



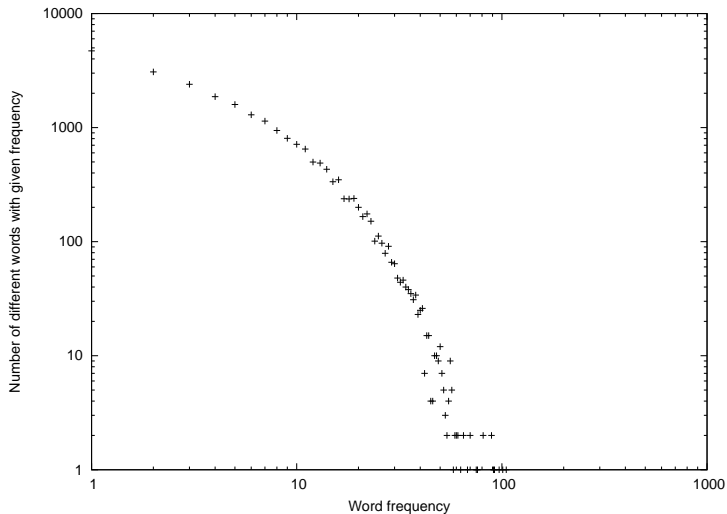
Percentage of text covered by words occurring exactly 10 times in the corpus



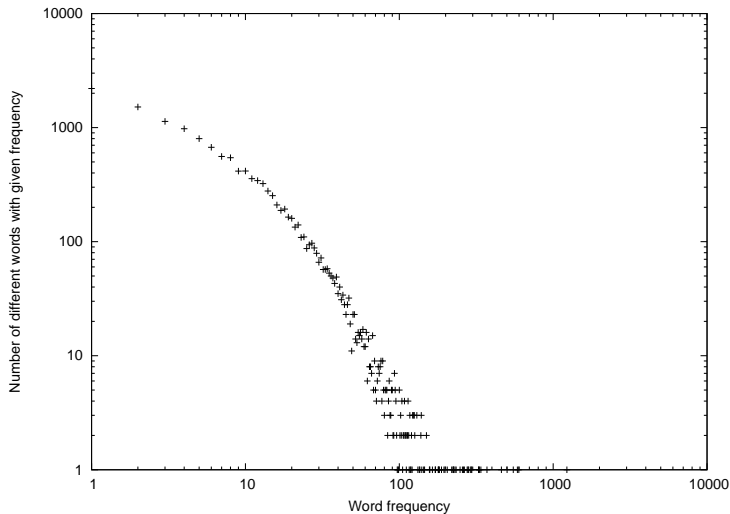
Contents

- 1 Freq-Freq
- 2 Stability of Frequences
- 3 Sampling units
- 4 Stability in documents**

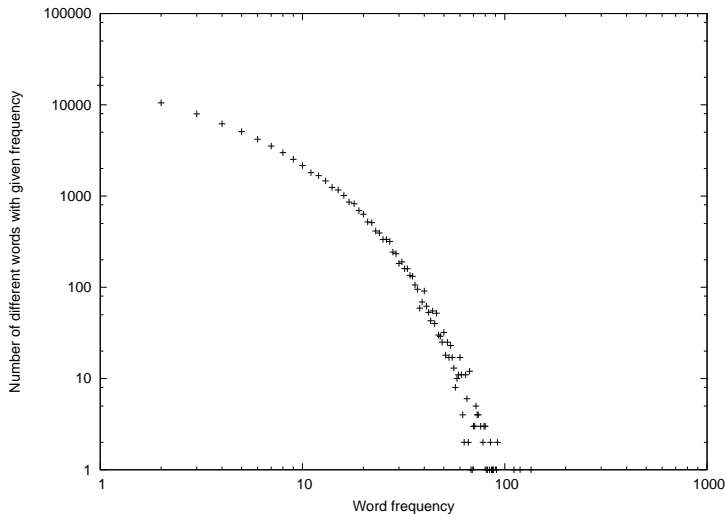
Hapax Legomena, 1M-10M, sentences



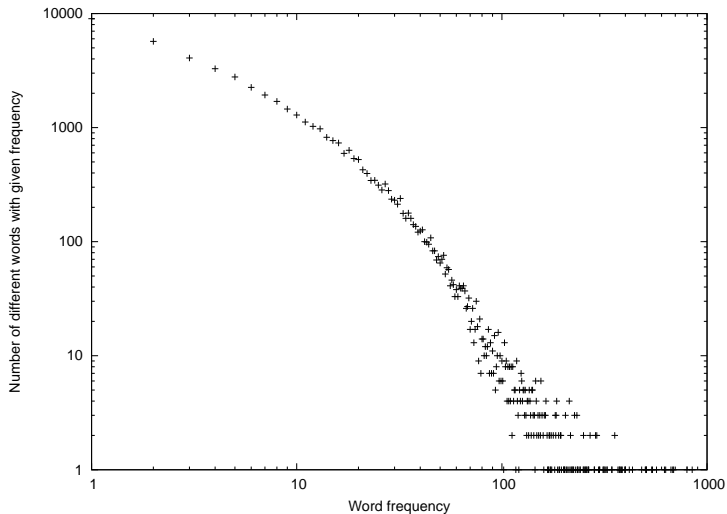
Hapax Legomena, 1M-10M, documents



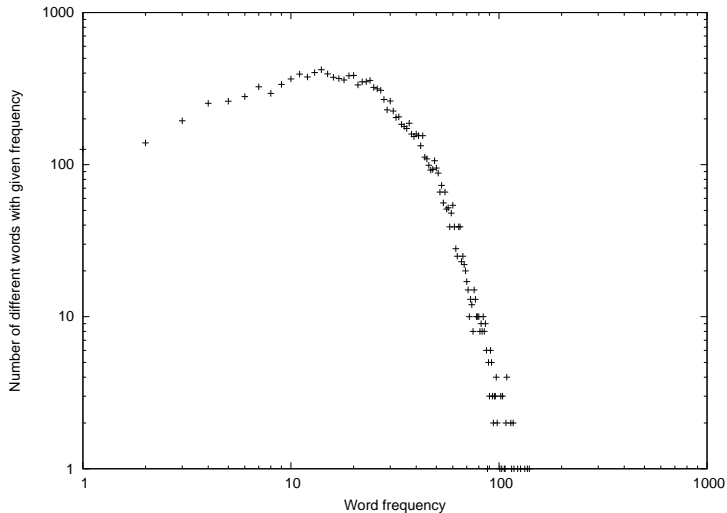
Hapax Legomena, 10M-100M, sentences



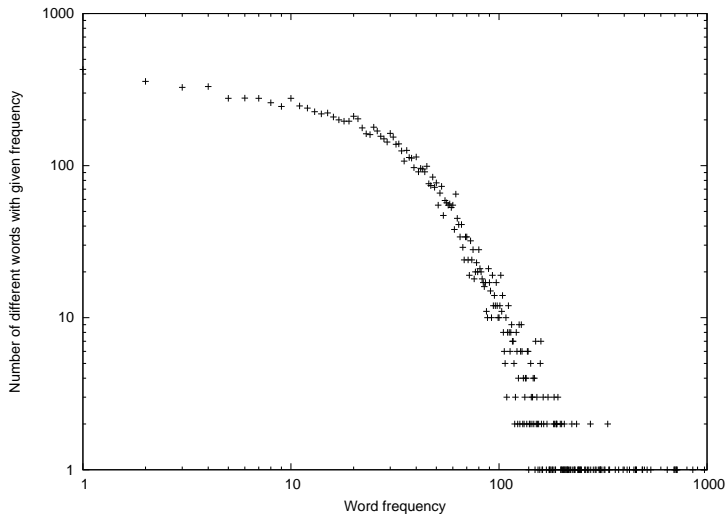
Hapax Legomena, 10M-100M, documents



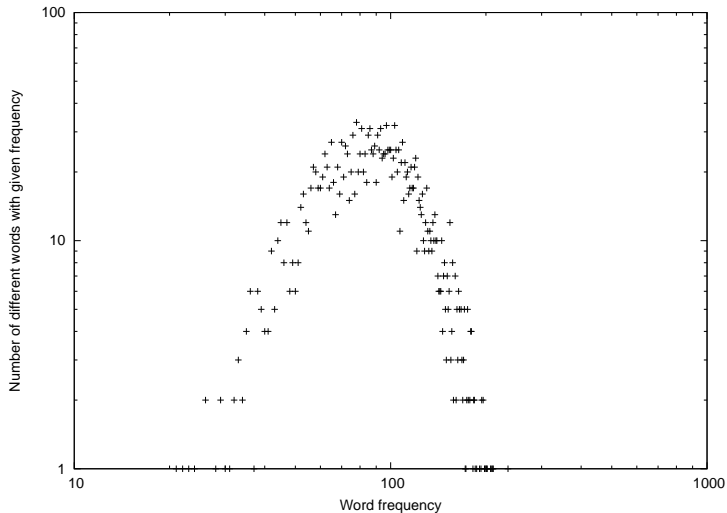
Tris Legomena, 10M-100M, sentences



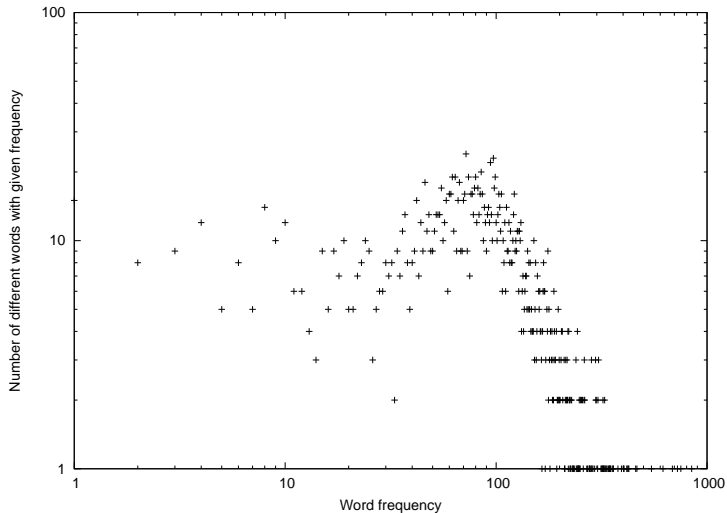
Tris Legomena, 10M-100M, documents



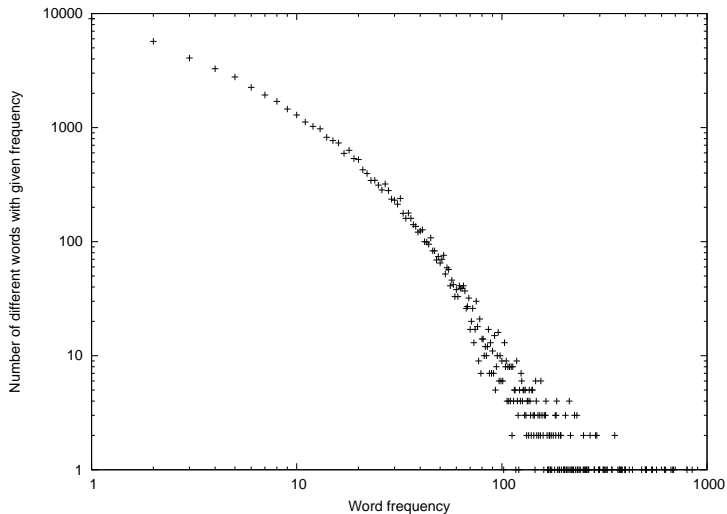
freq=10, 10M-100M, sentences



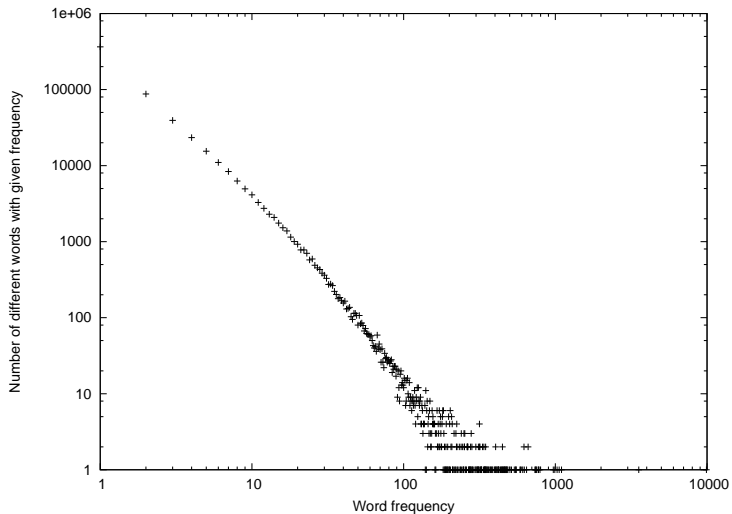
freq=10, 10M-100M, documents



Hapax Legomena, 10M-100M, documents



freq=0, 10M-100M, documents



Conclusion

- low-frequency words has much lower frequencies
- sampling unit have big impact on frequency distribution