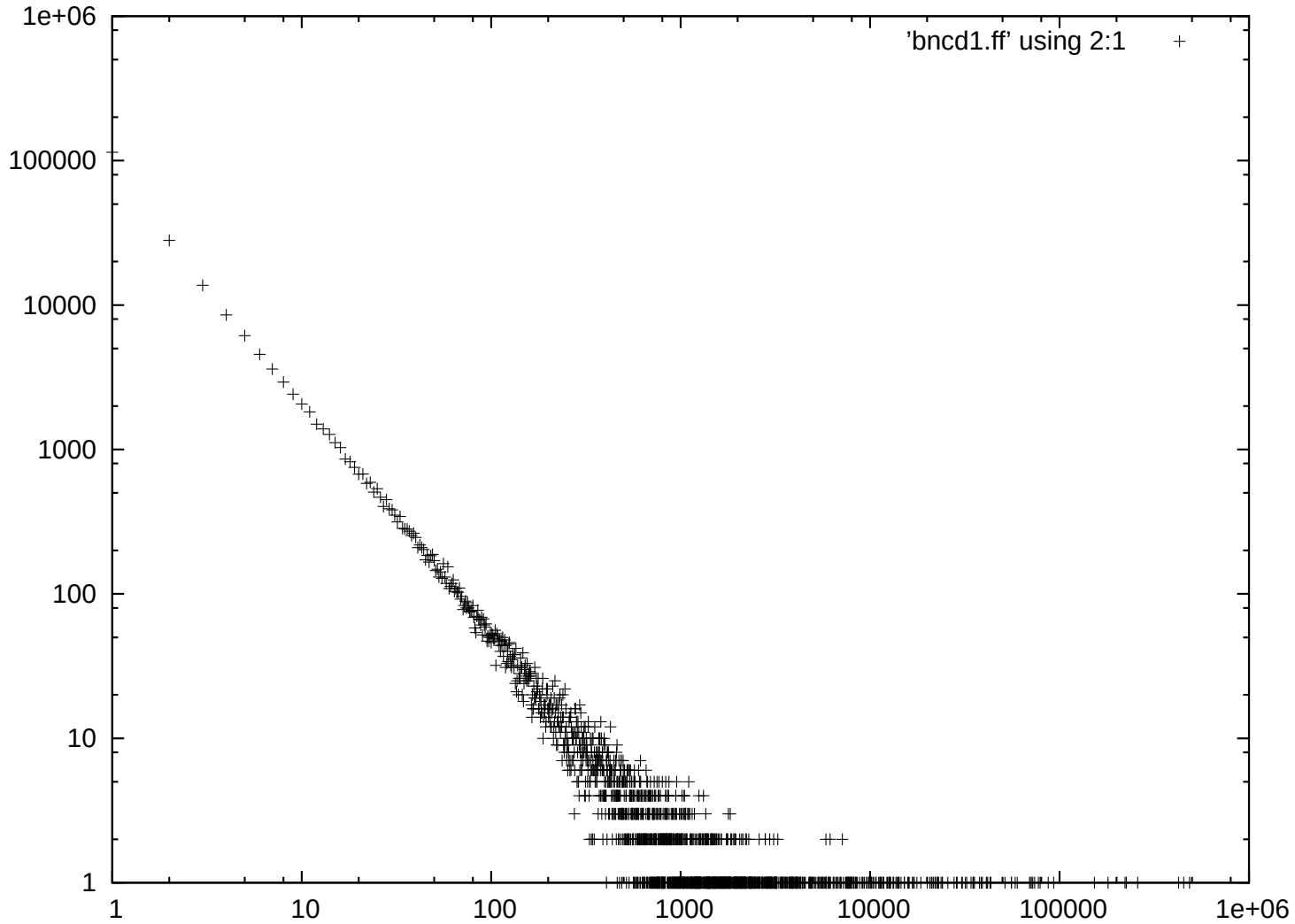
Frequency of Low-frequency Words in Text Corpora

Pavel Rychlý

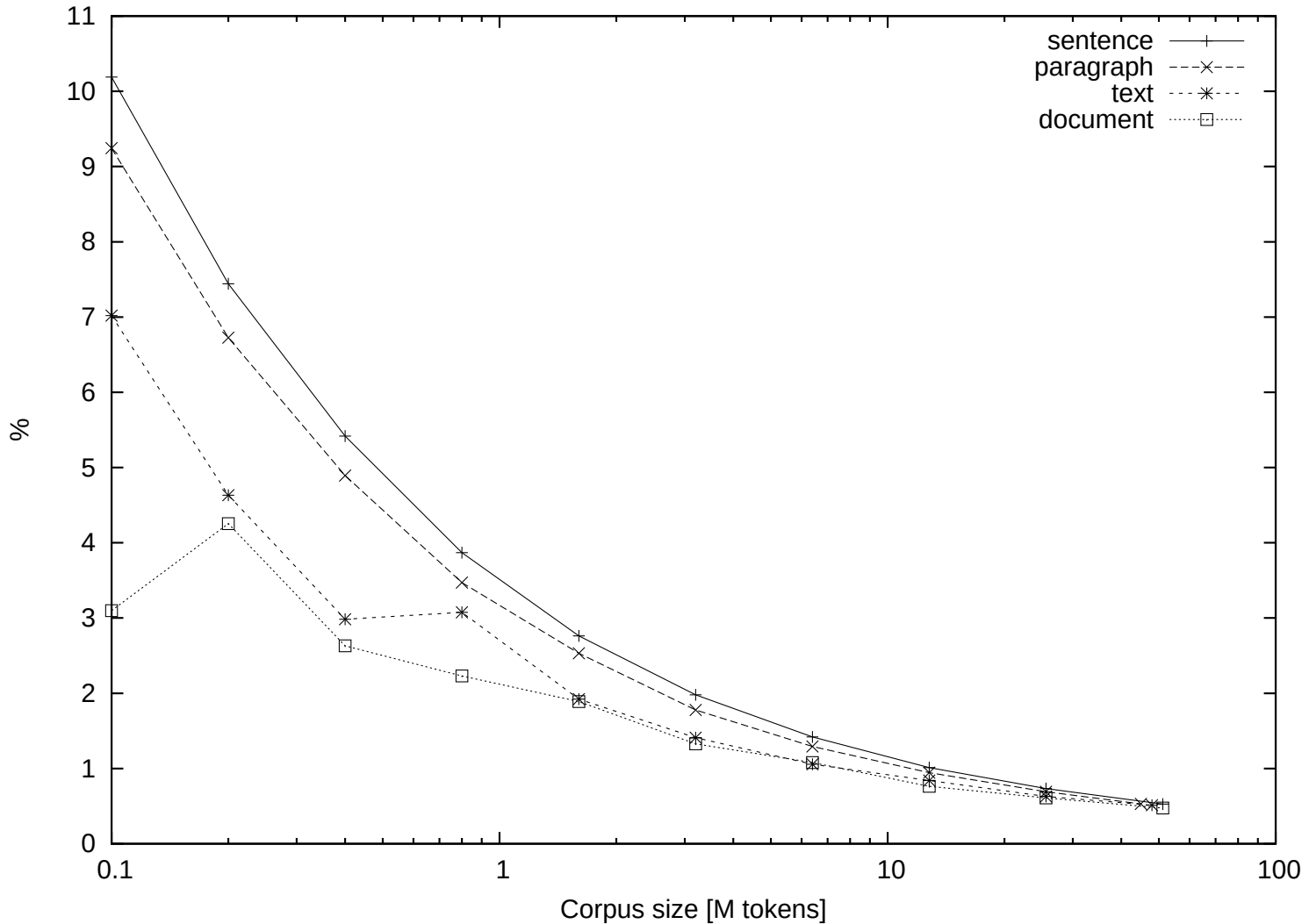
pary@fi.muni.cz

Contents

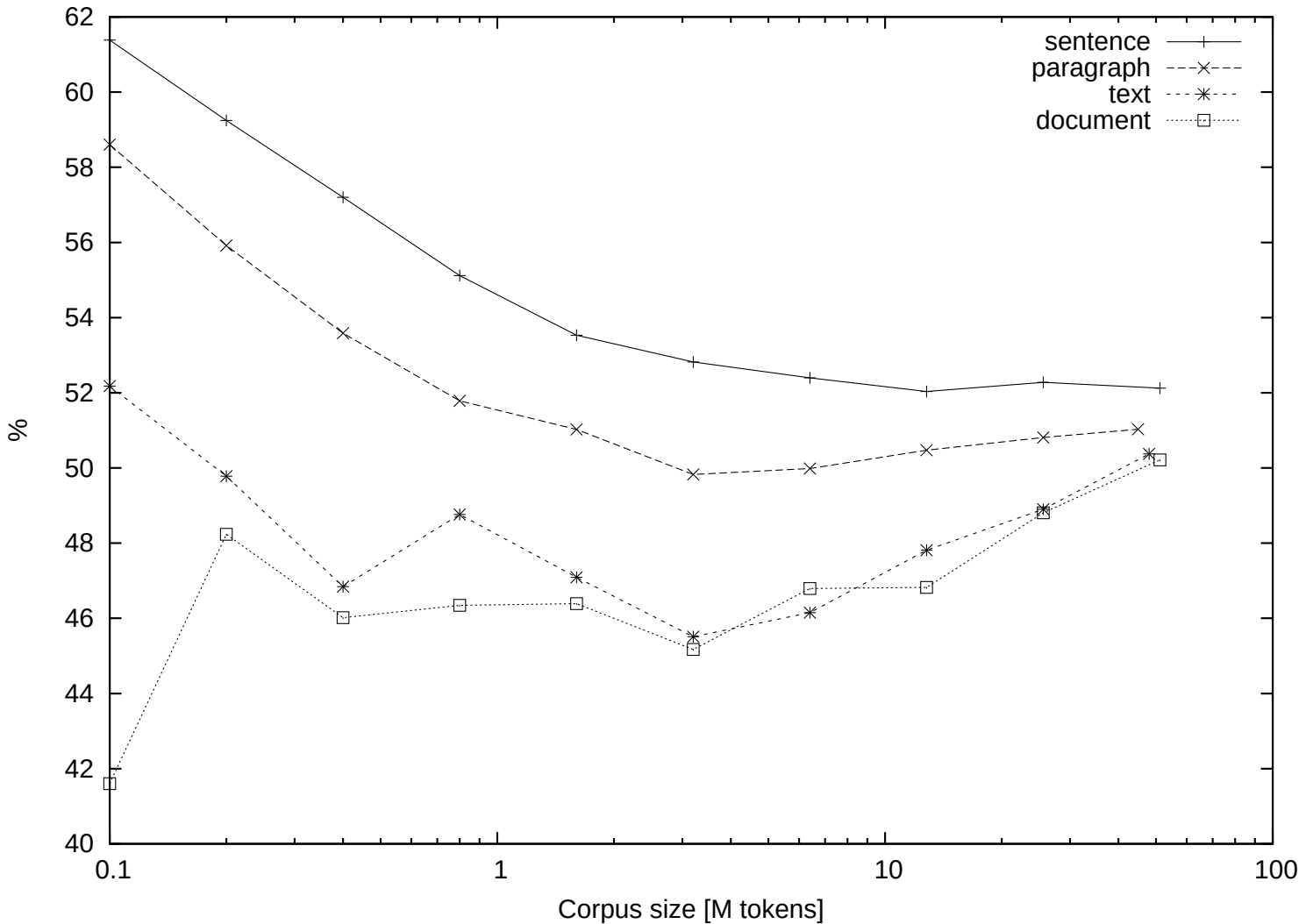
Freq-Freq – Frequences of Frequences



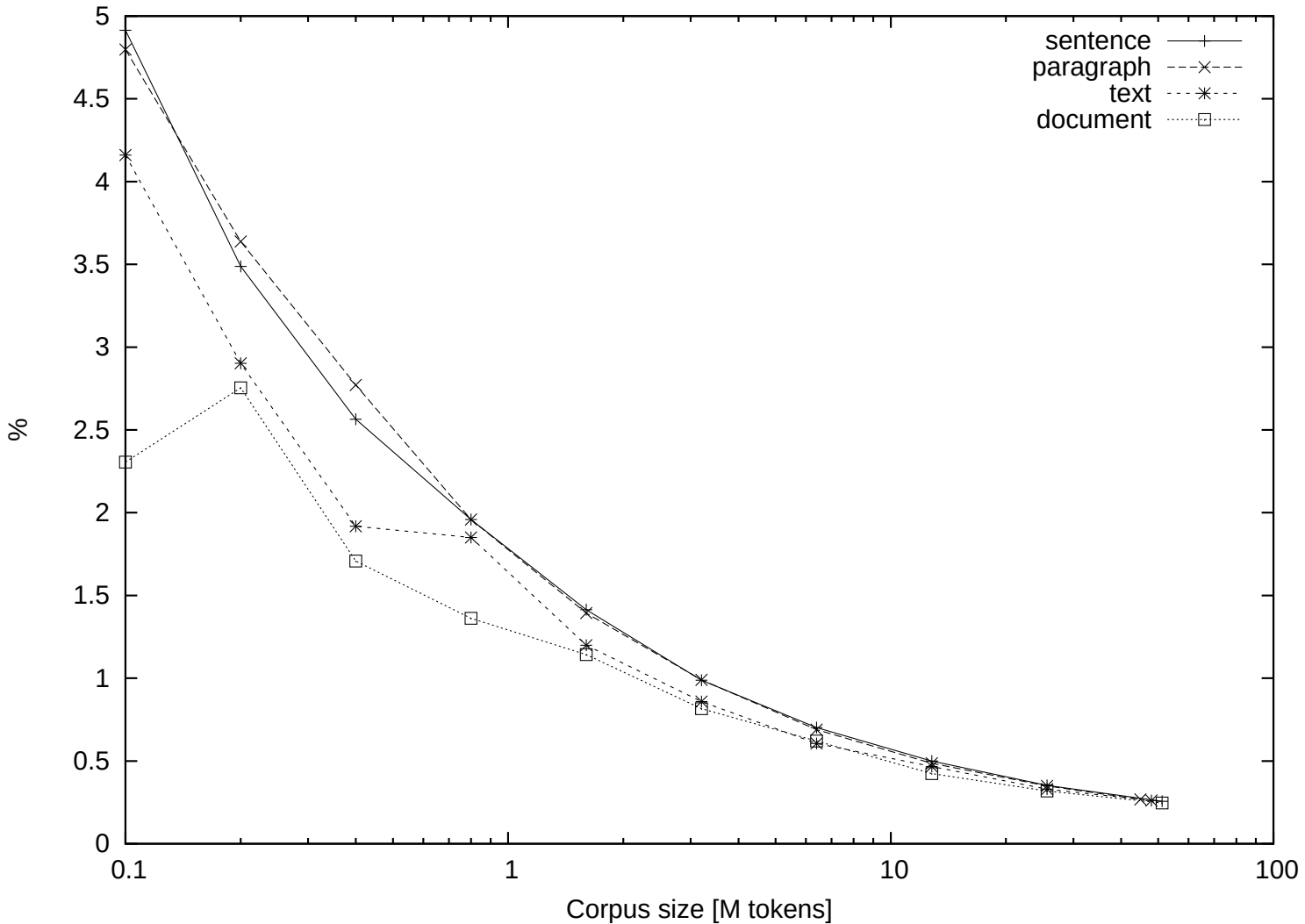
Percentage of text covered by hapax legomena



Percentage of word types of hapax legomena

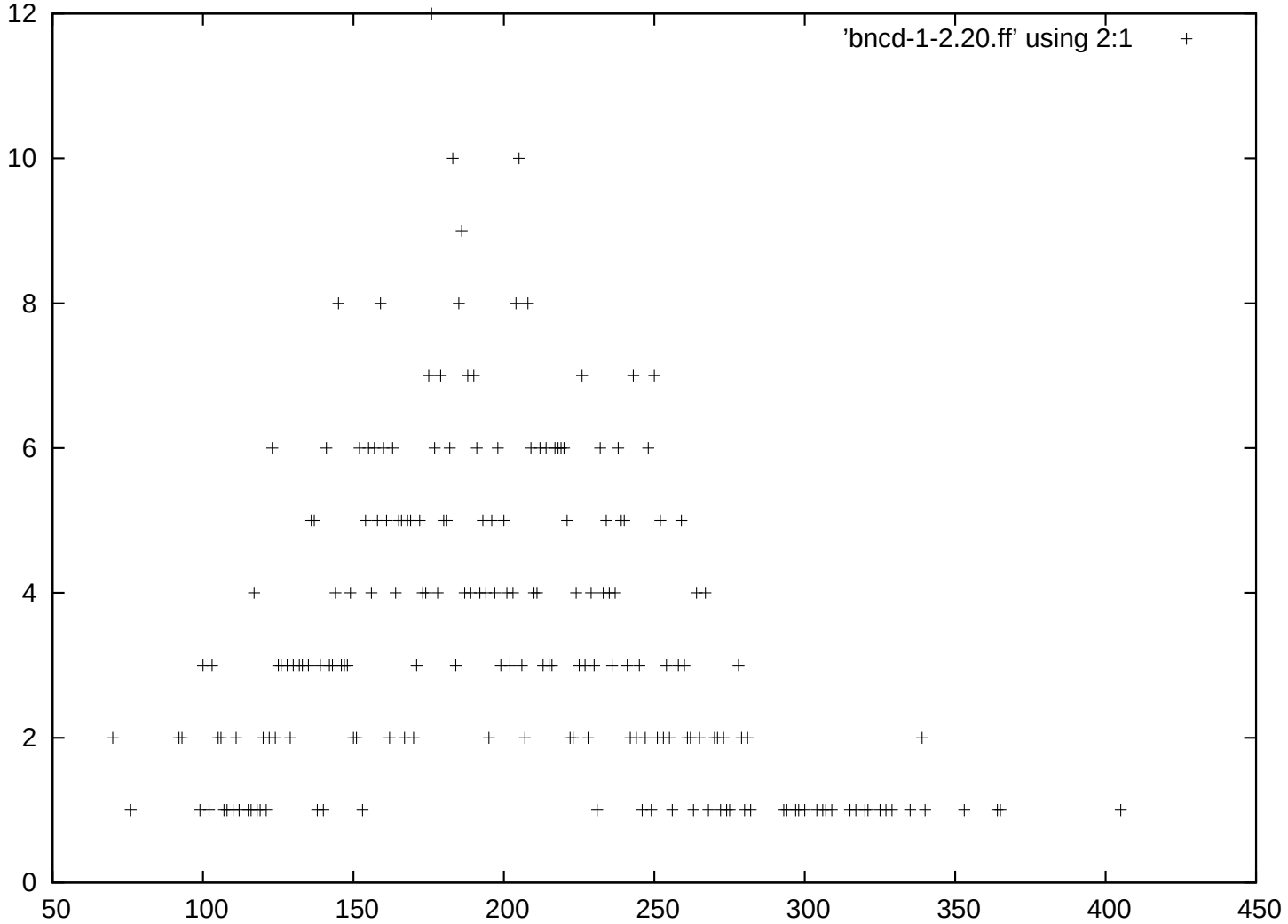


Percentage of word types of dis legomena

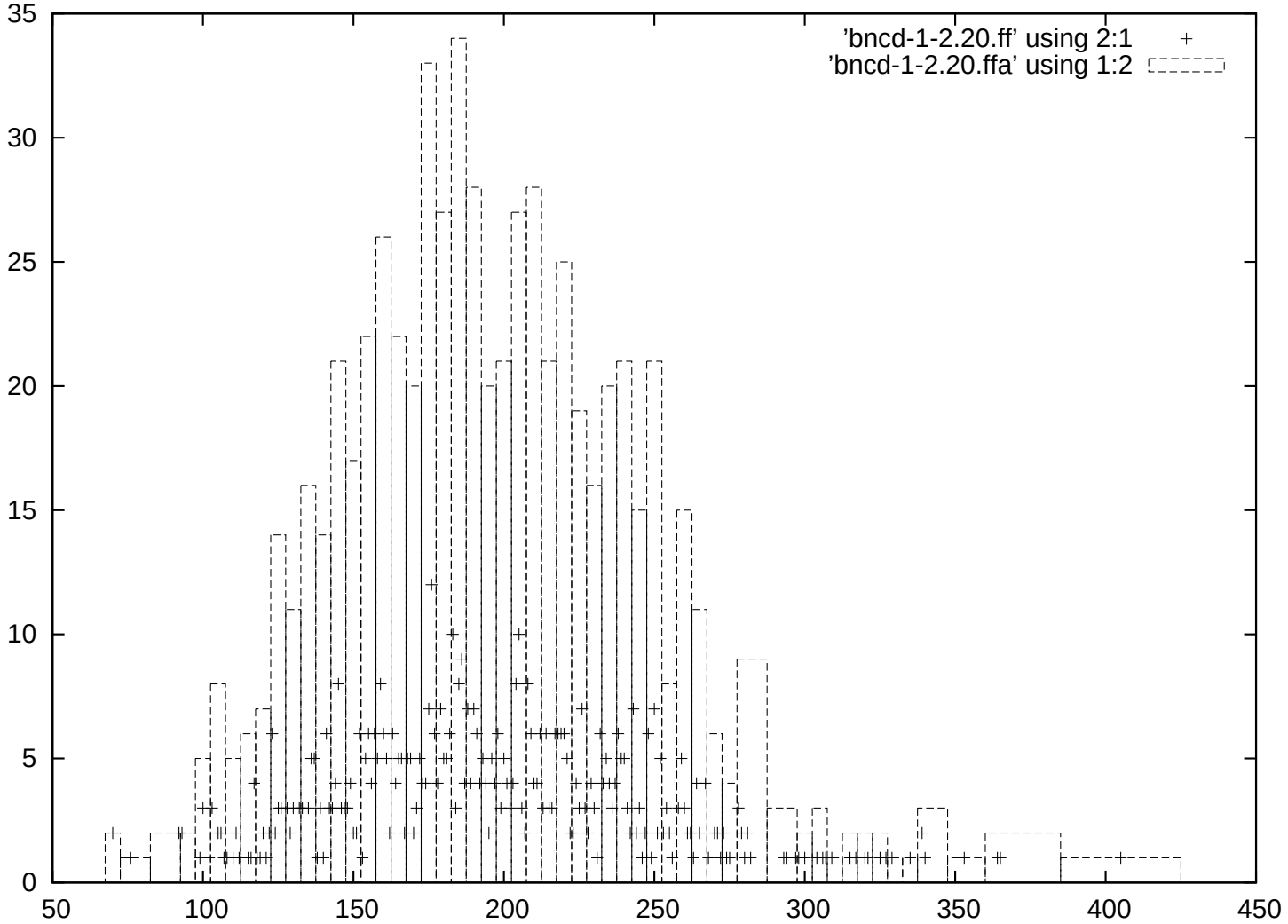


Contents

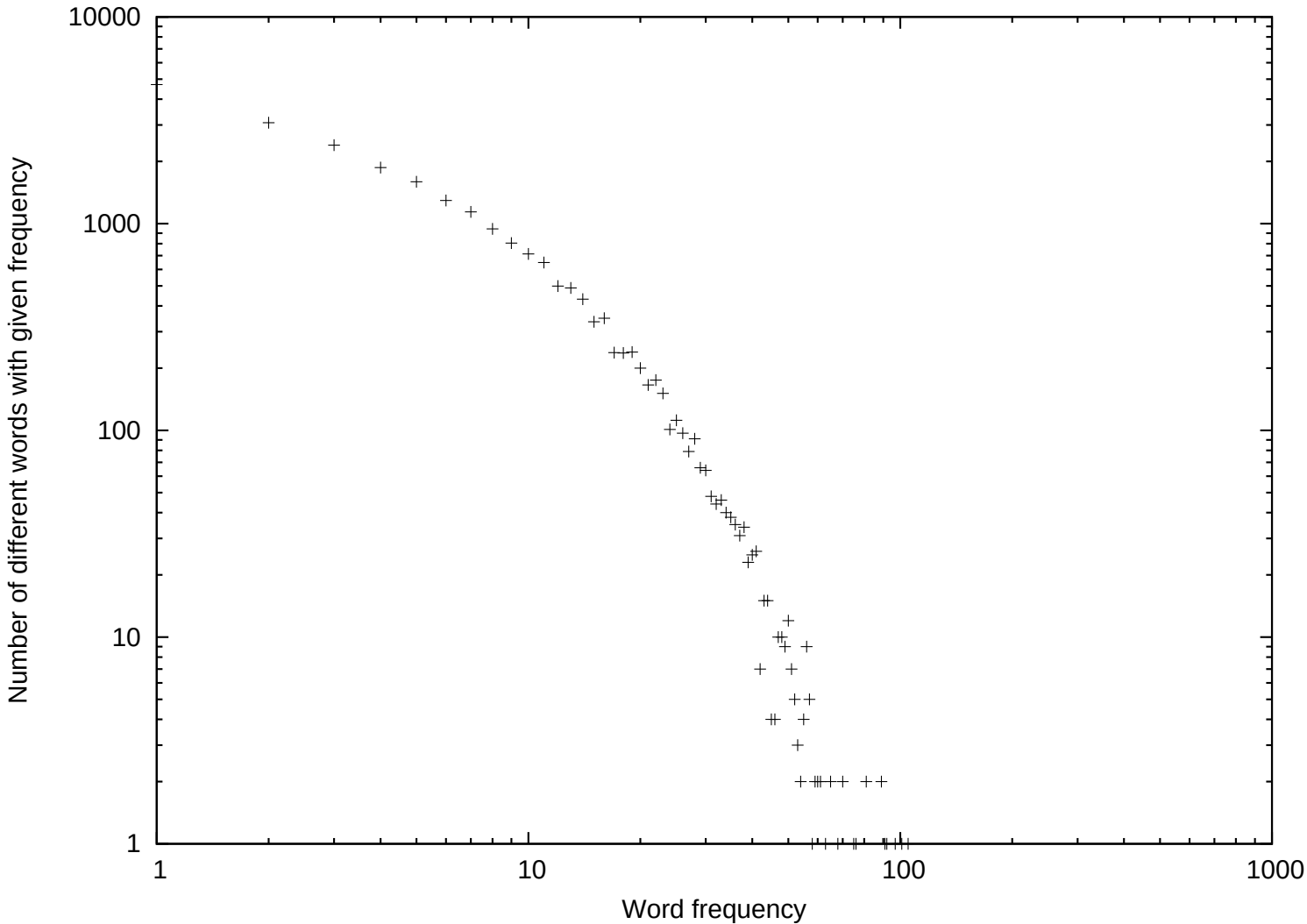
Stability 10M-100M, freq=20



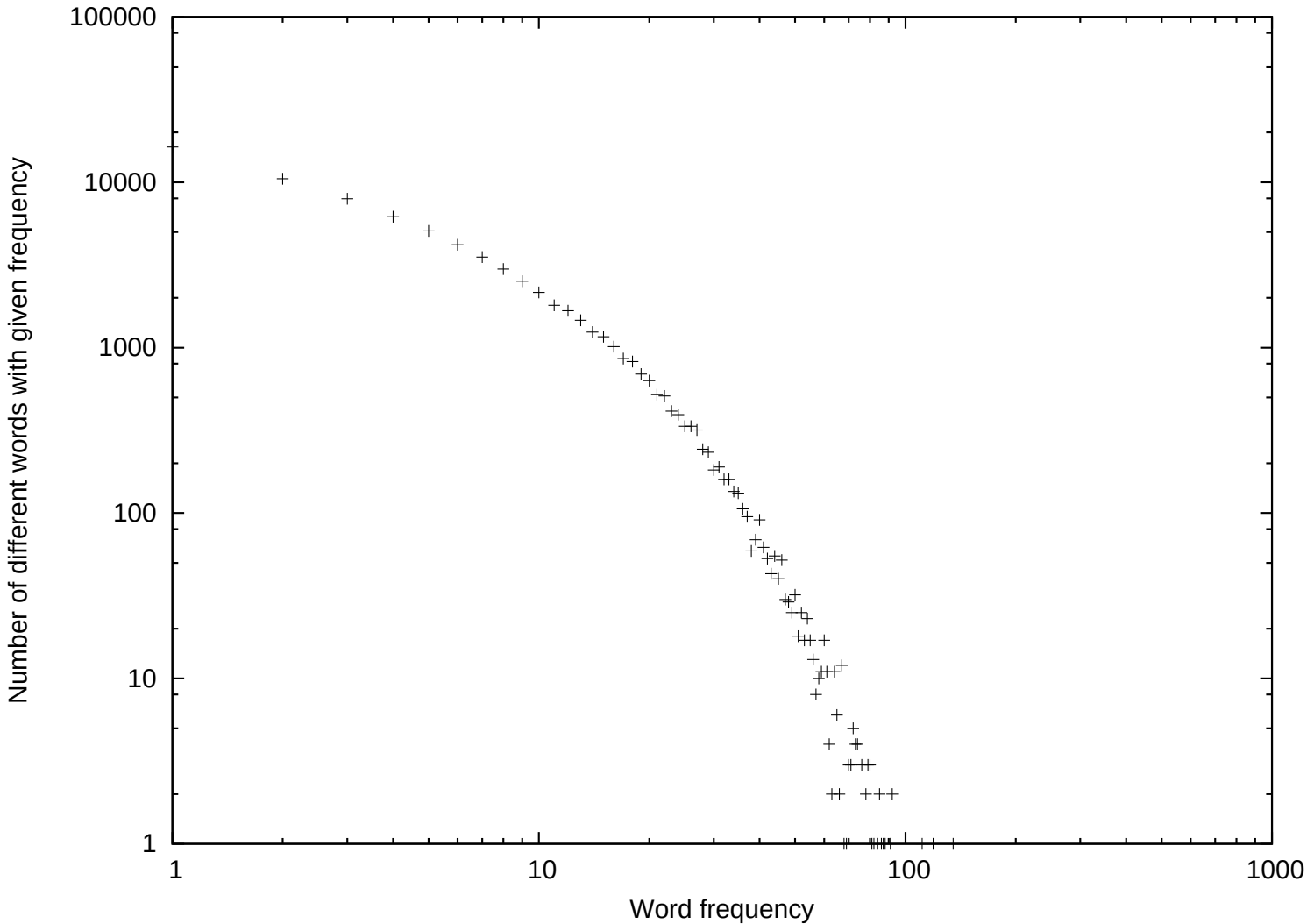
Stability 10M-100M, freq=20, agregated(5)



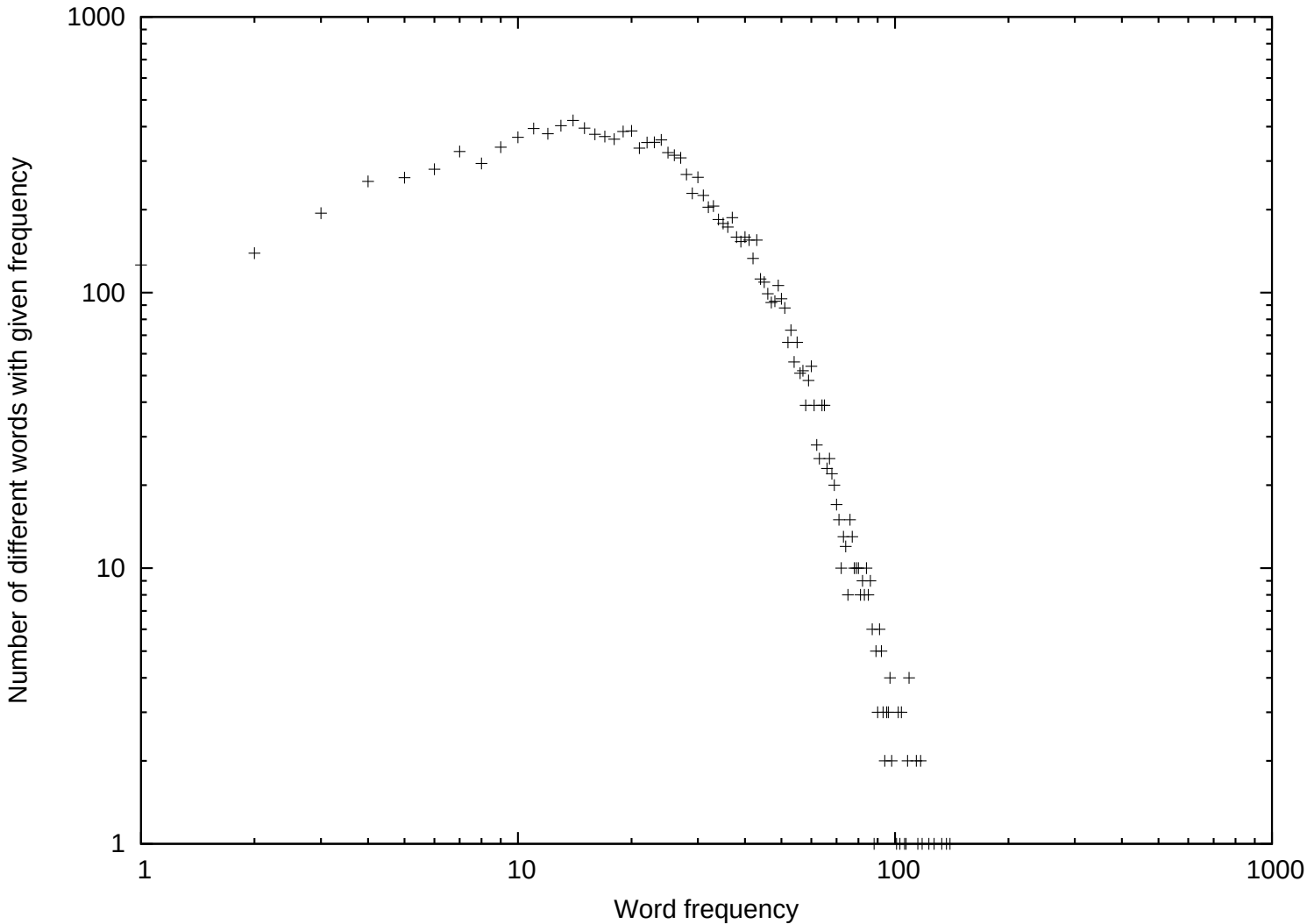
Hapax Legomena, 1M-10M



Hapax Legomena, 10M-100M

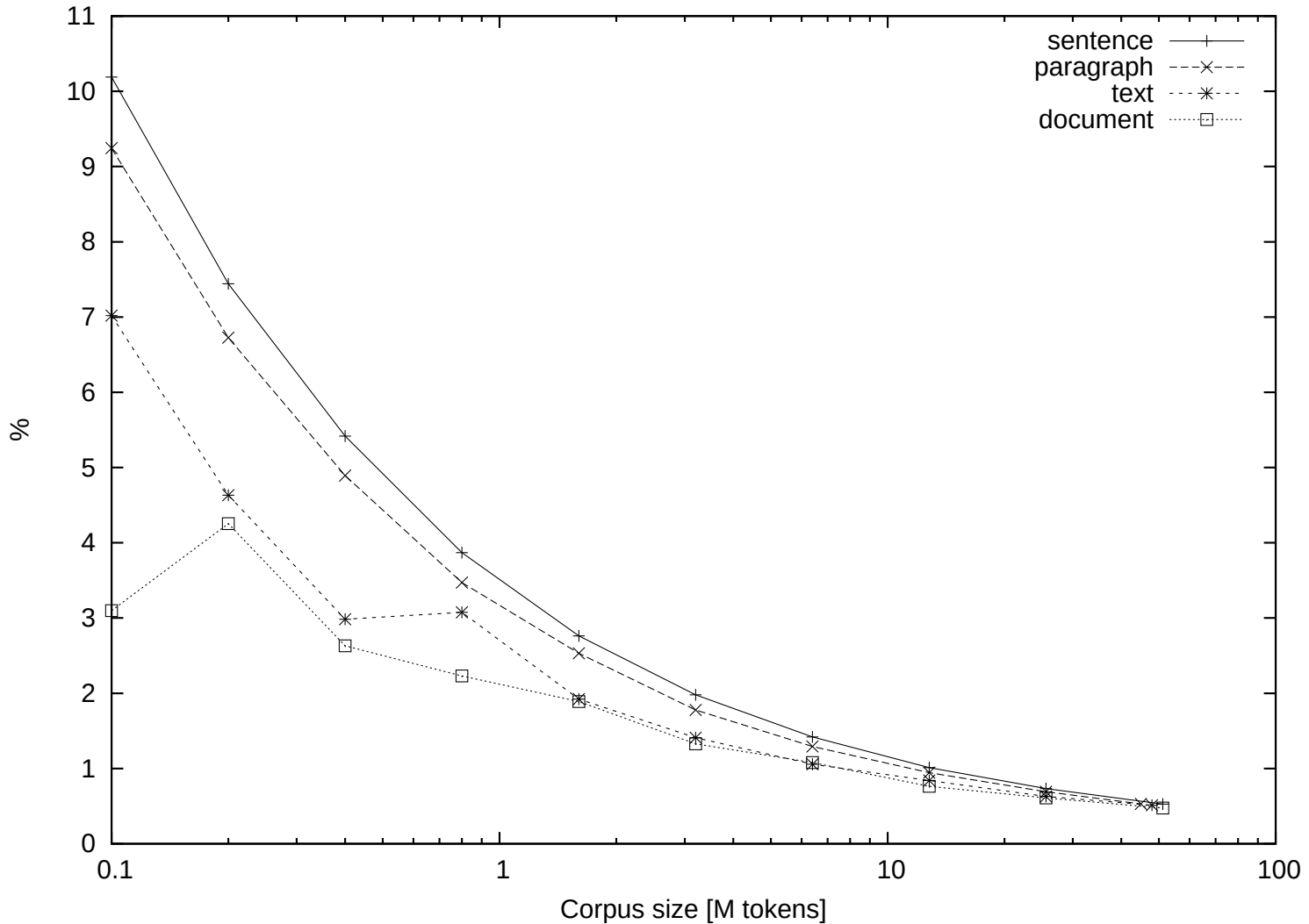


Tris Legomena, 10M-100M

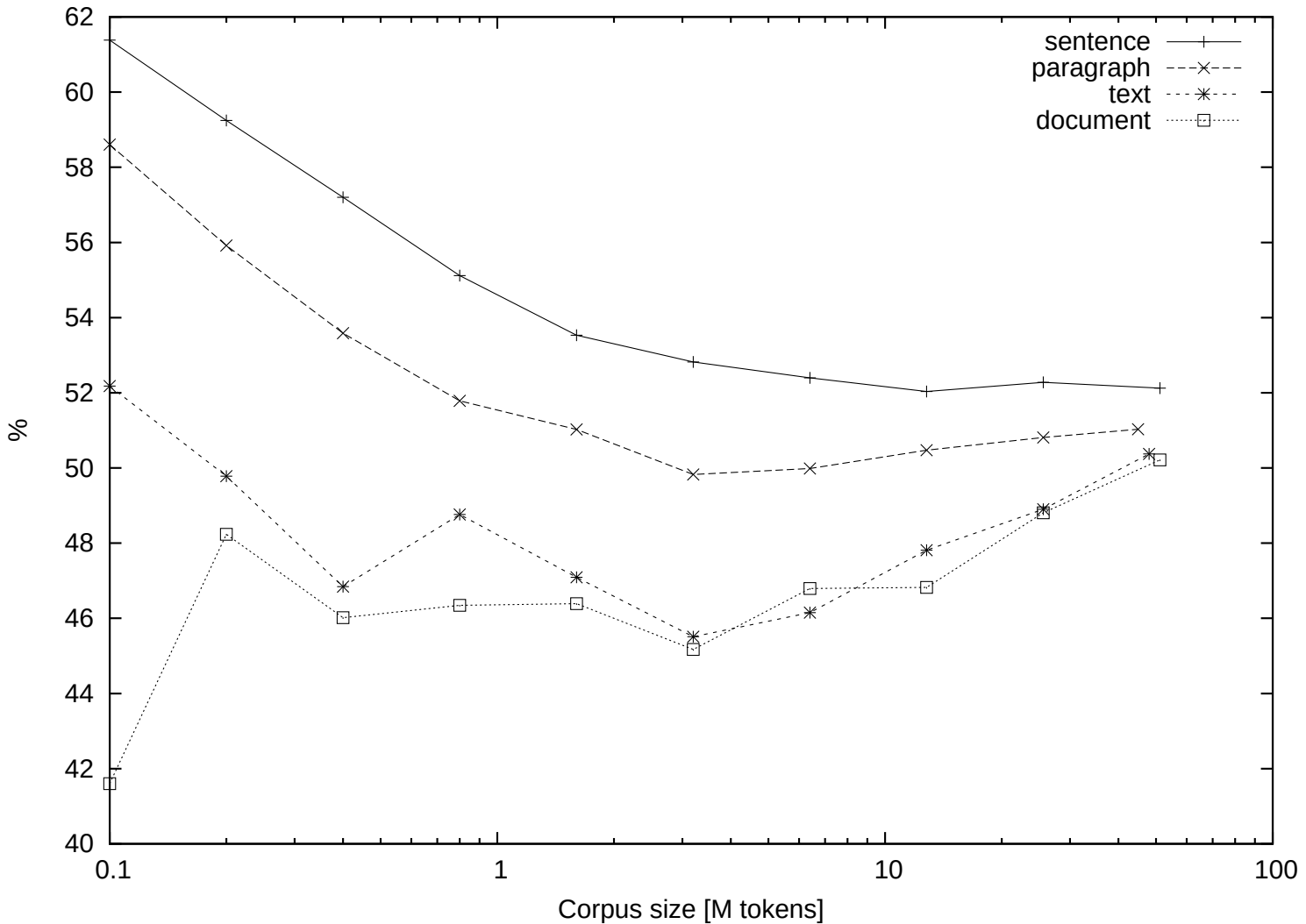


Contents

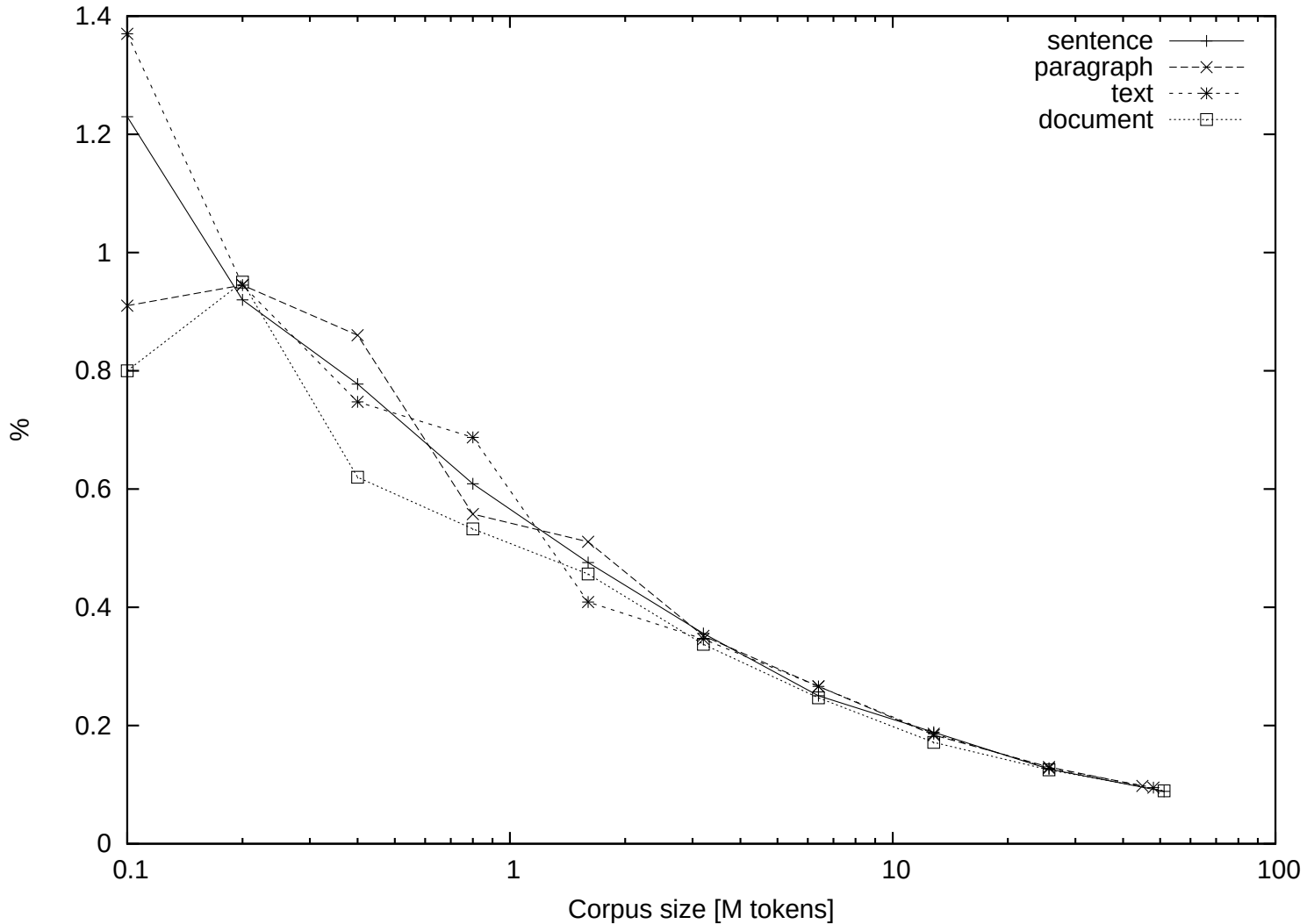
Percentage of text covered by hapax legomena



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

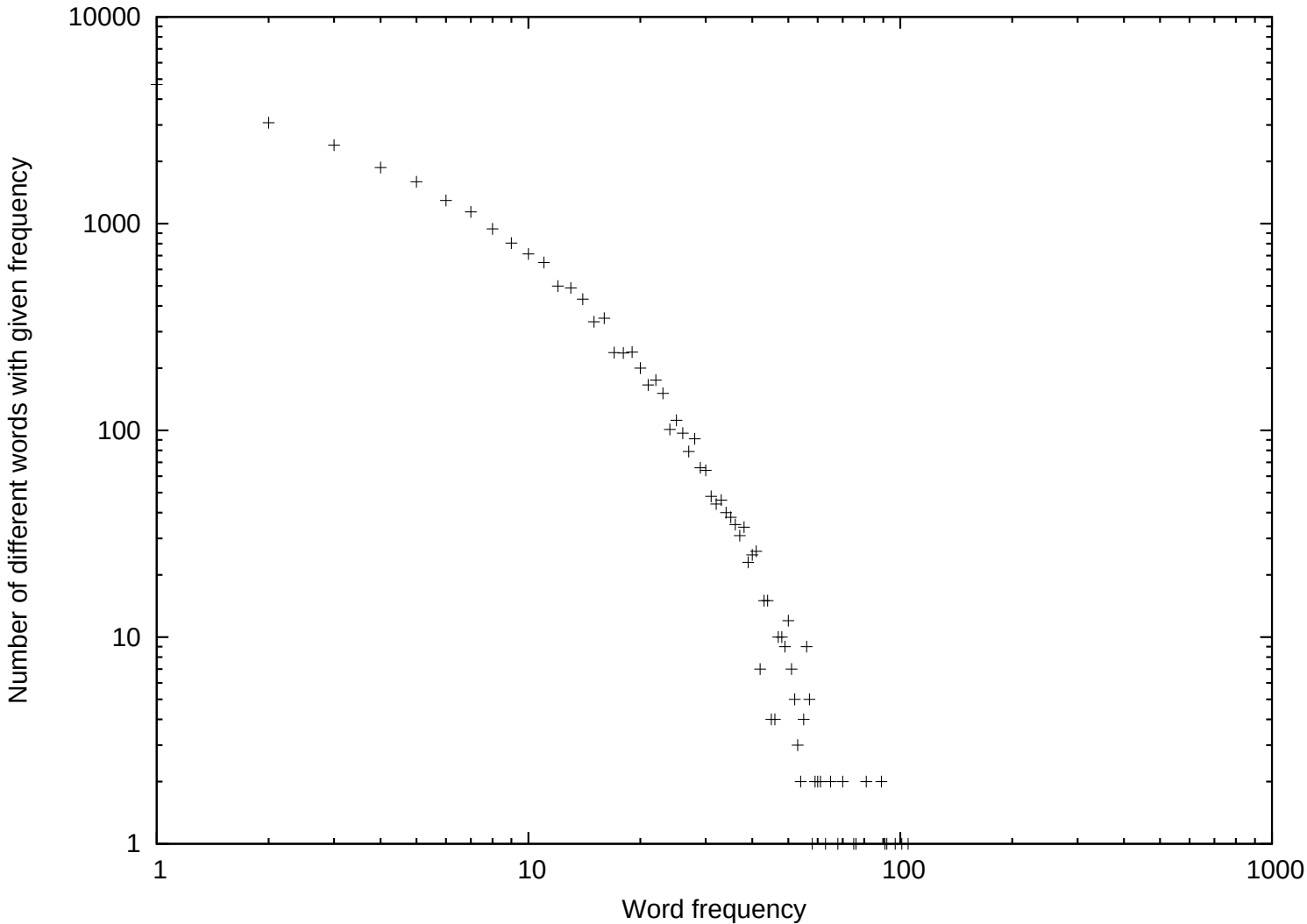


Percentage of text covered by words occurring exactly 10 times in the corpus

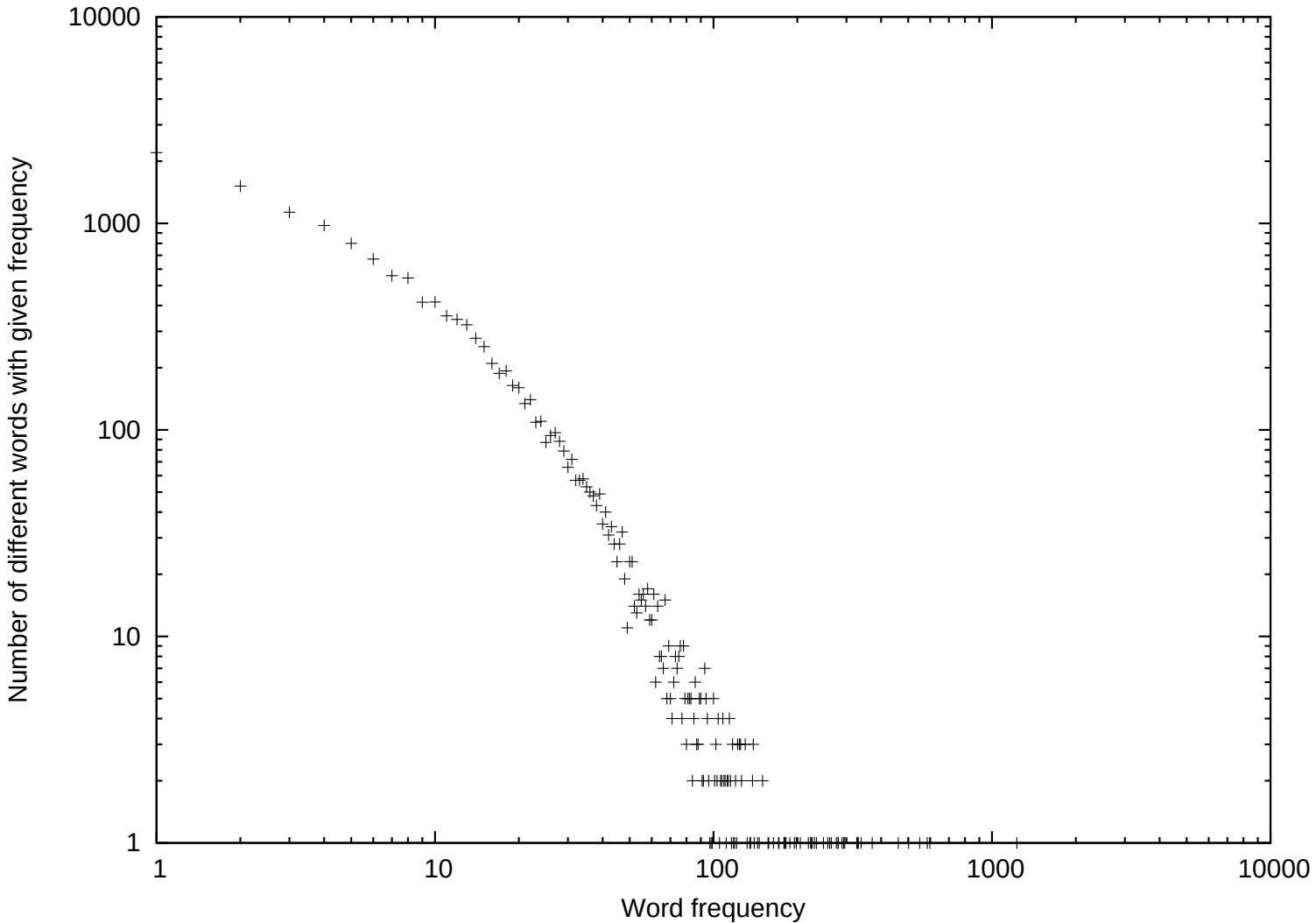


Contents

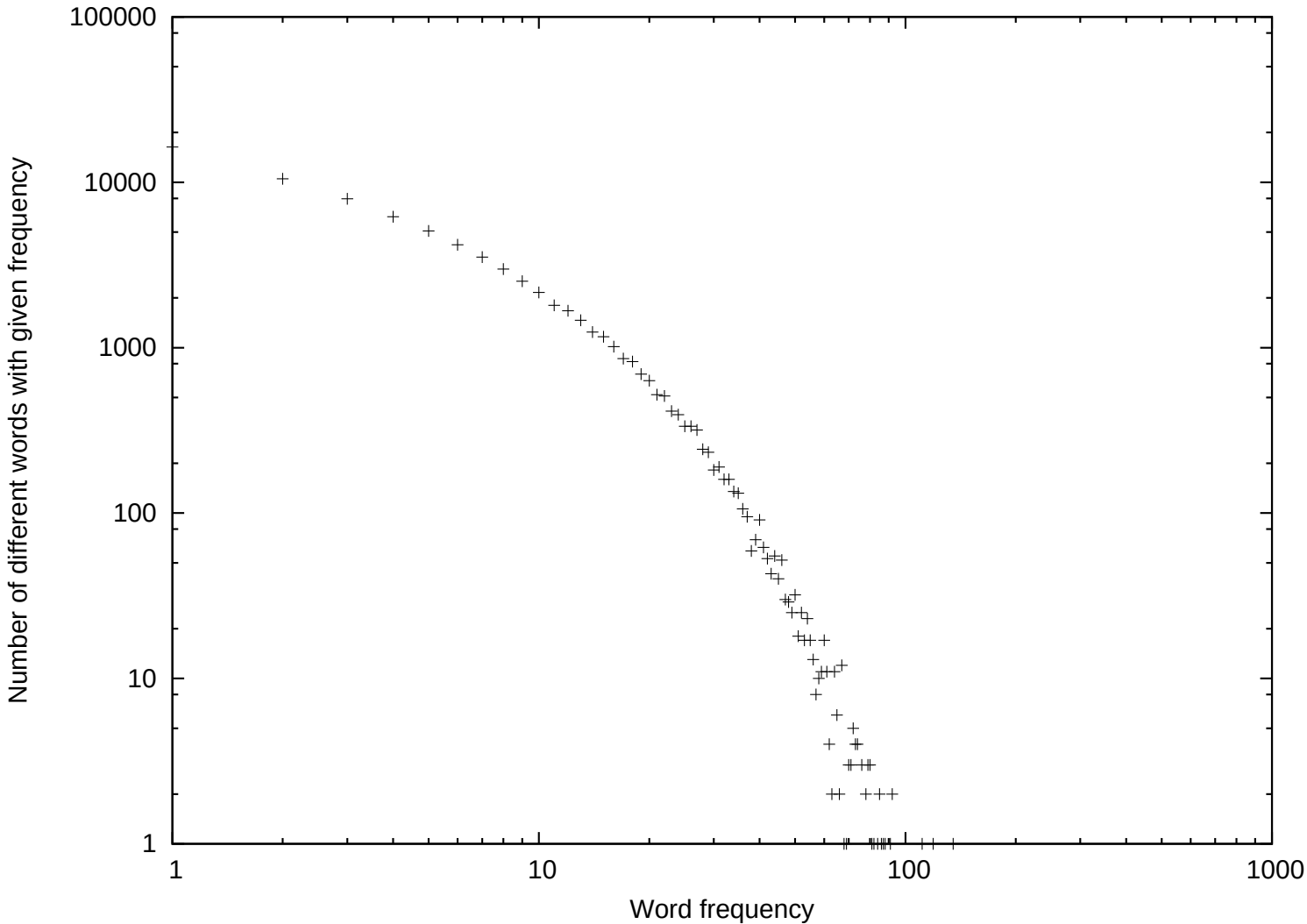
Hapax Legomena, 1M-10M, sentences



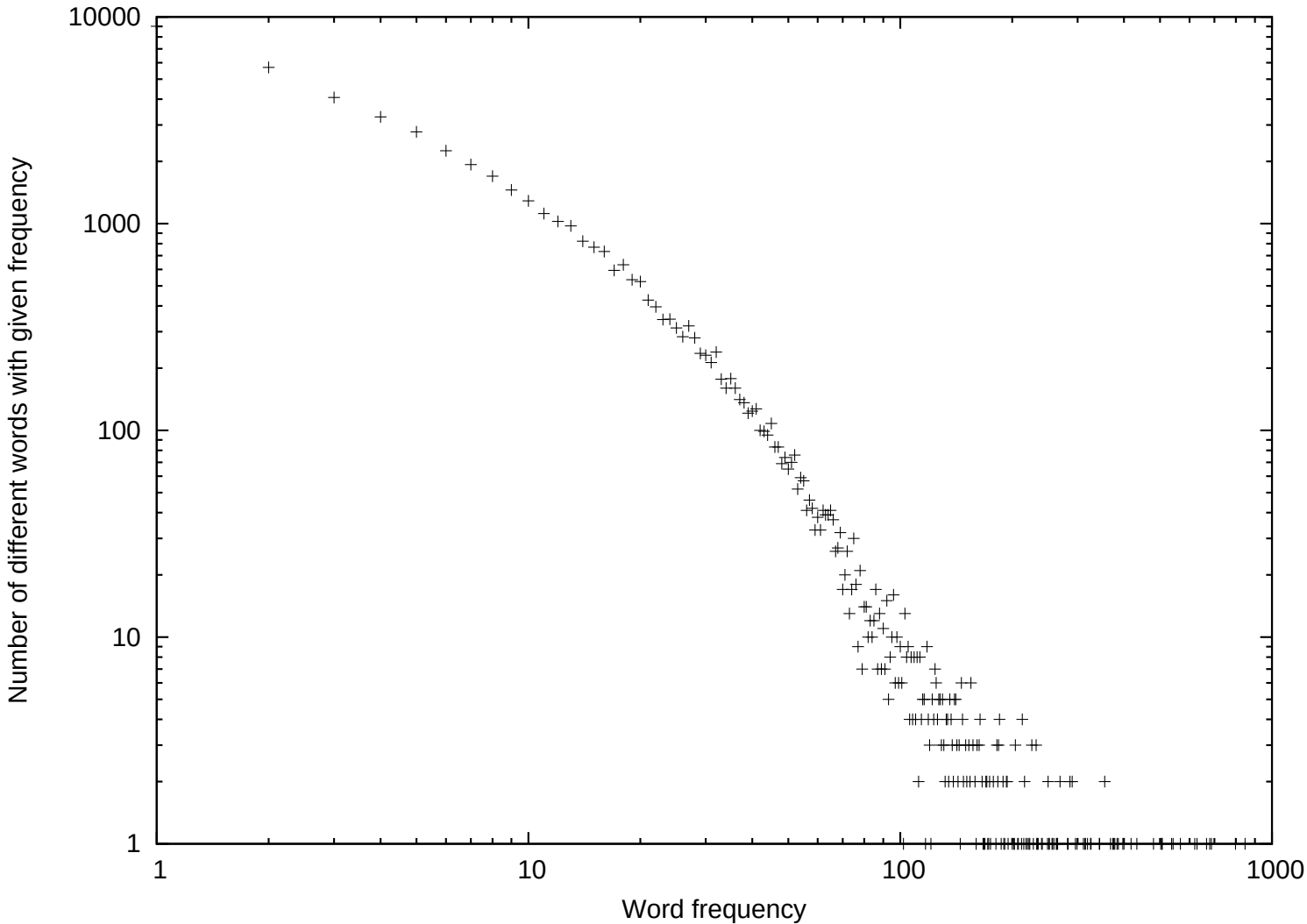
Hapax Legomena, 1M-10M, documents



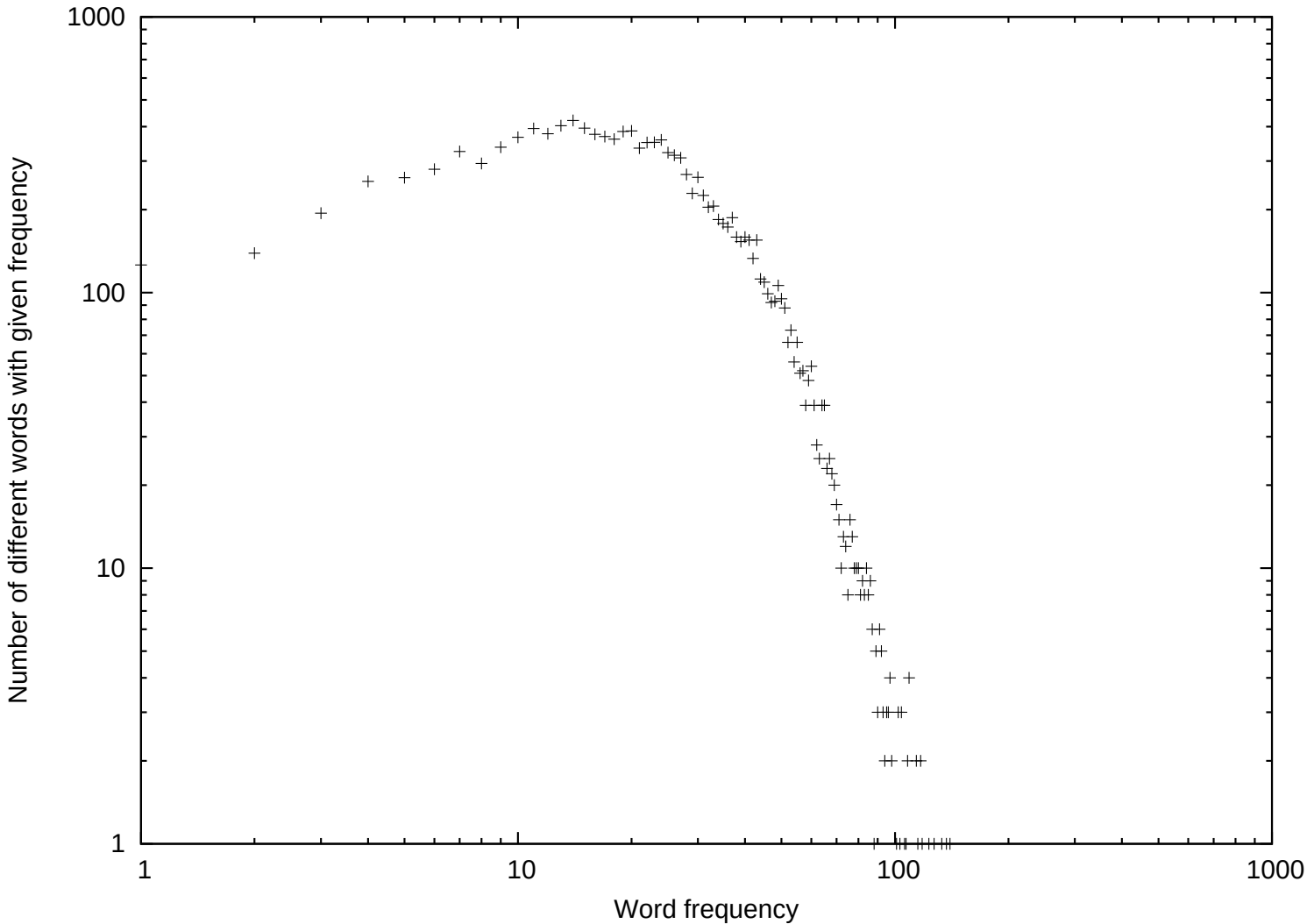
Hapax Legomena, 10M-100M, sentences



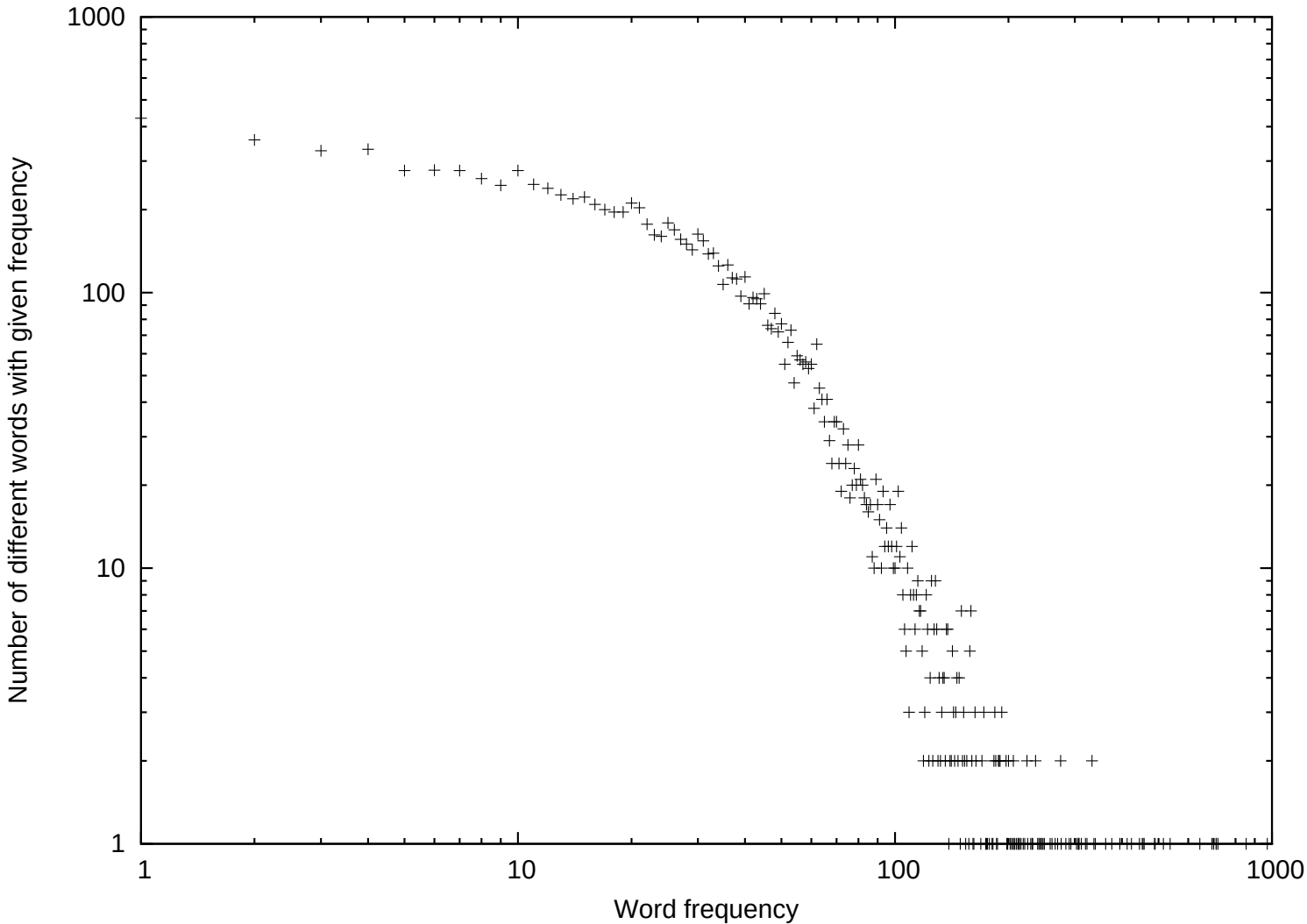
Hapax Legomena, 10M-100M, documents



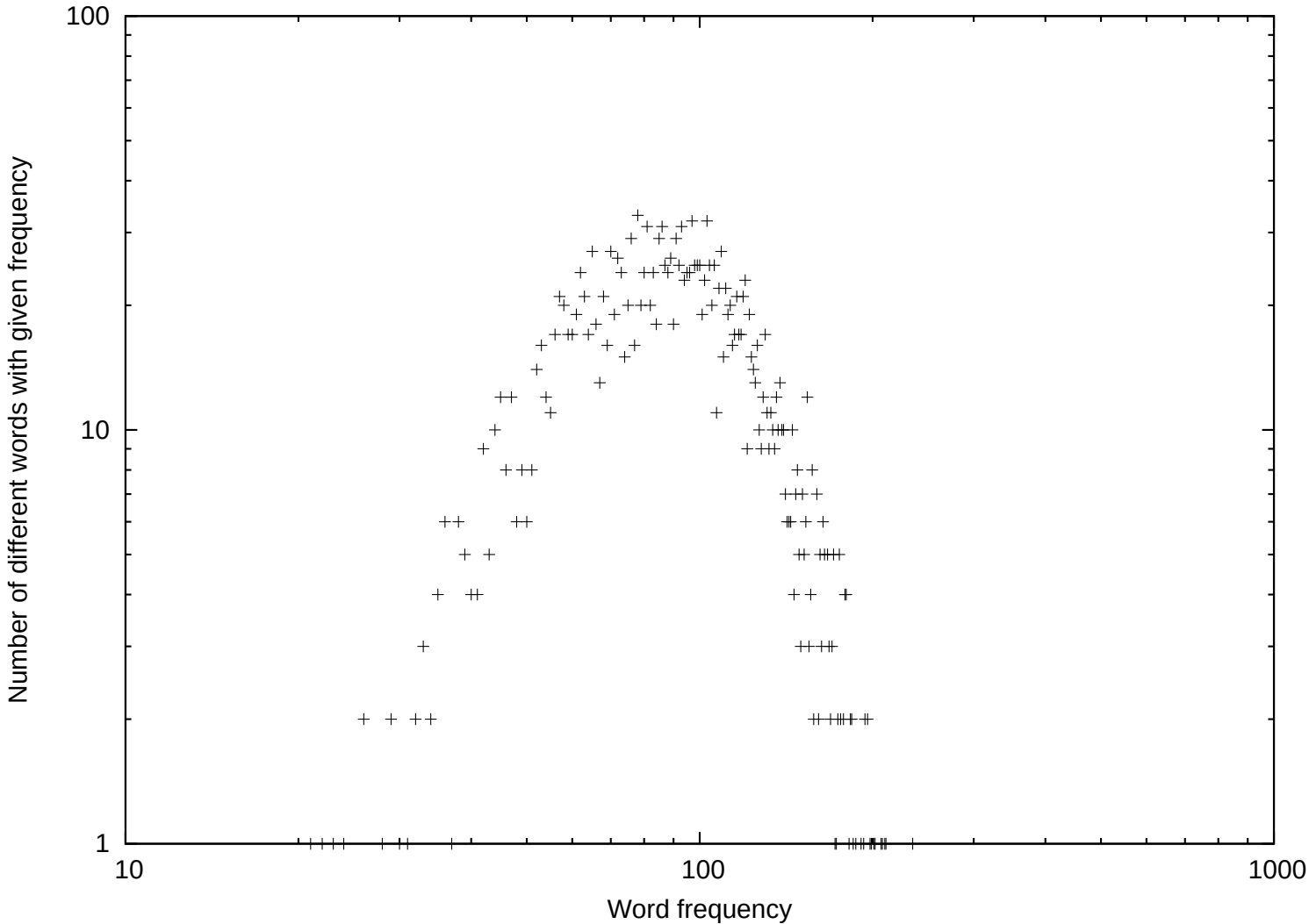
Tris Legomena, 10M-100M, sentences



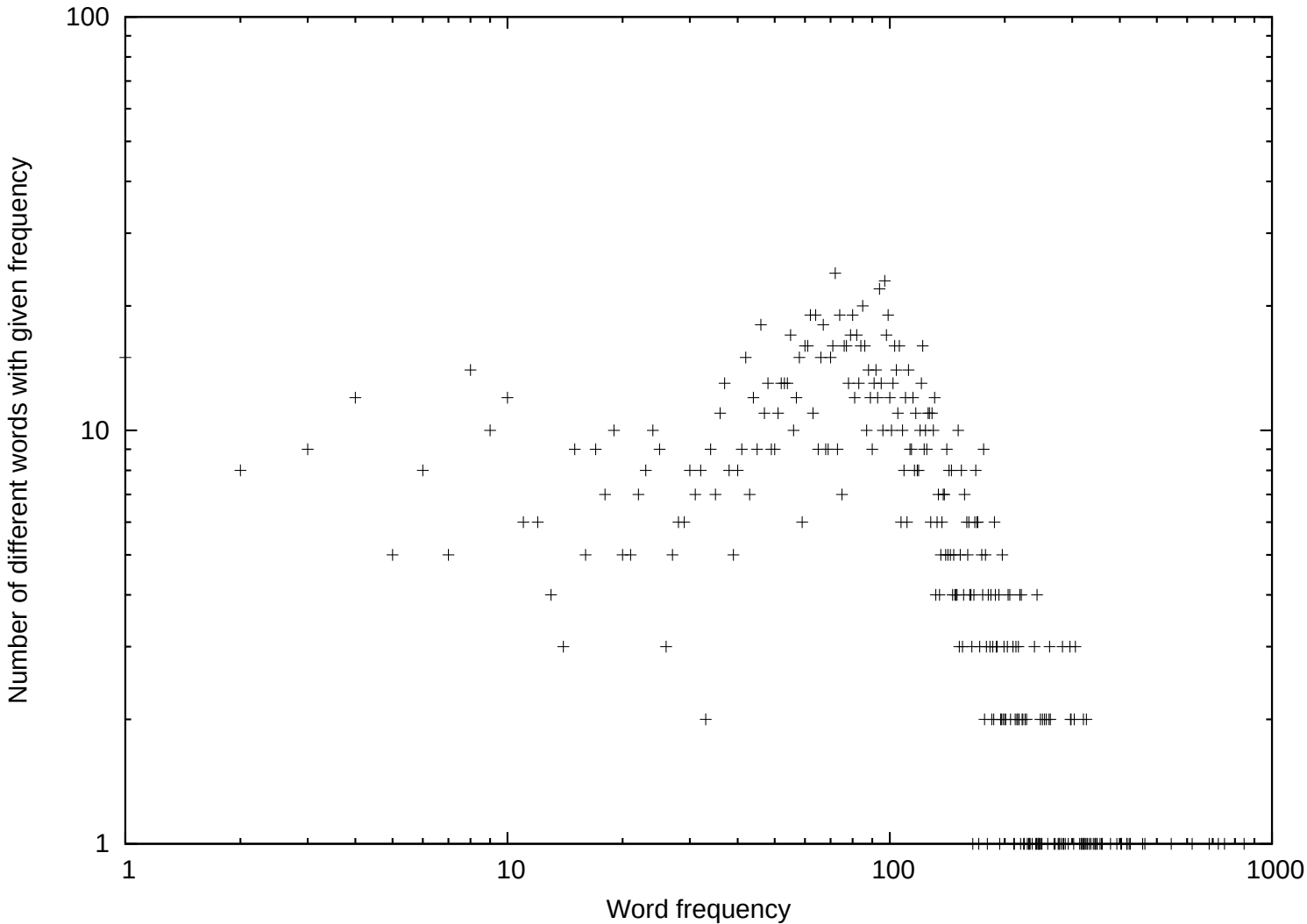
Tris Legomena, 10M-100M, documents



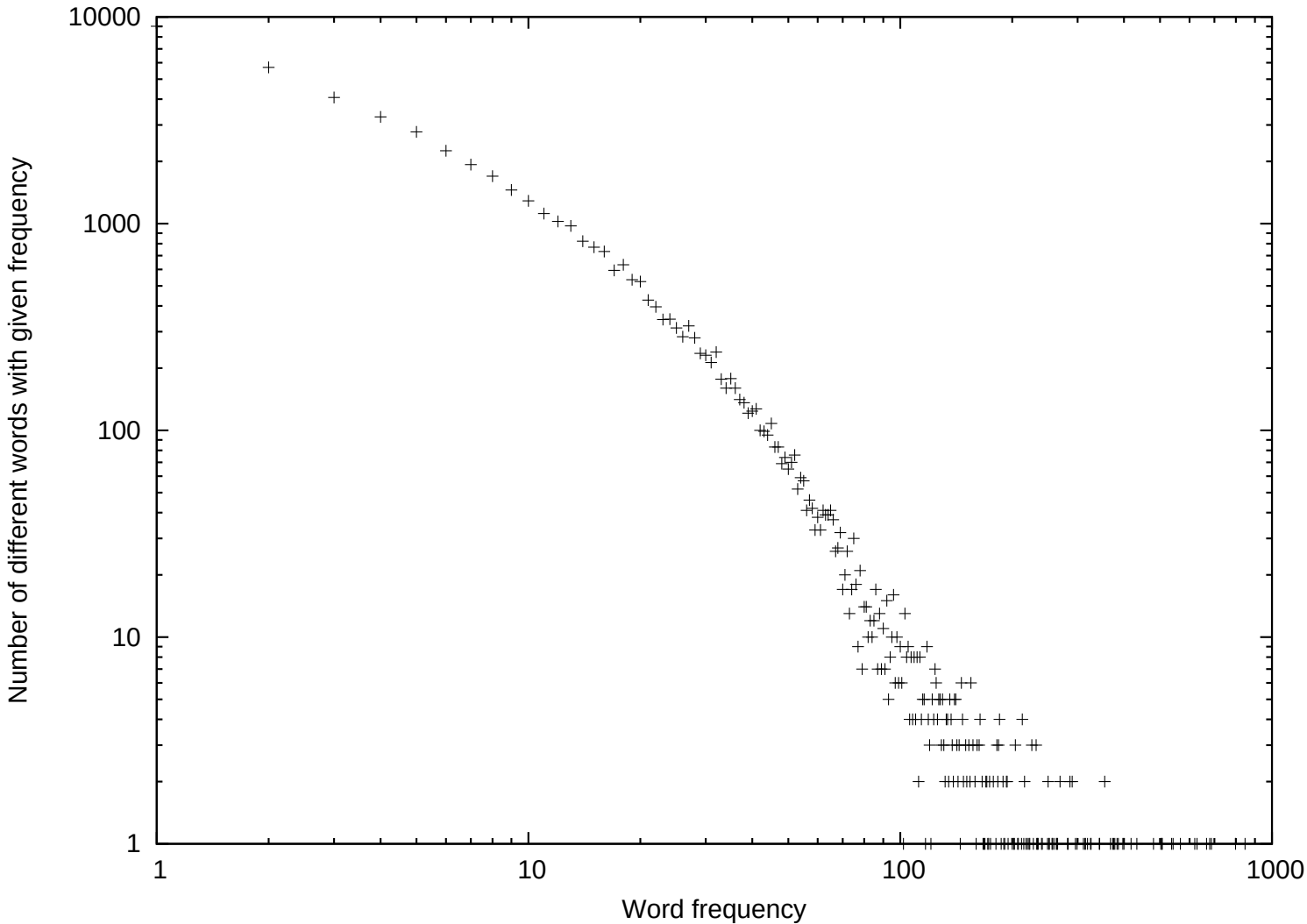
freq=10, 10M-100M, sentences



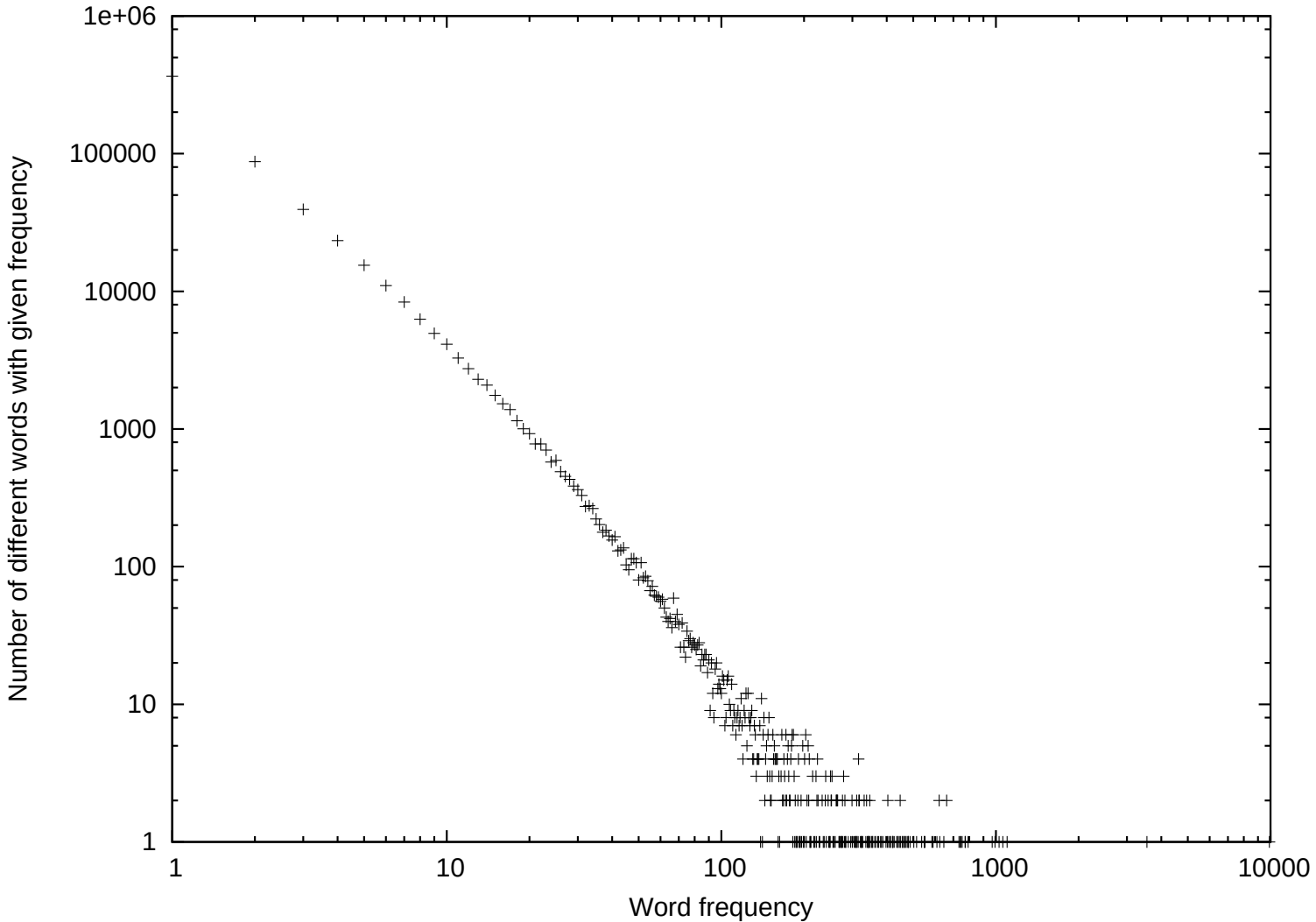
freq=10, 10M-100M, documents



Hapax Legomena, 10M-100M, documents



freq=0, 10M-100M, documents



Conclusion

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