Words' Burstiness in Language Models

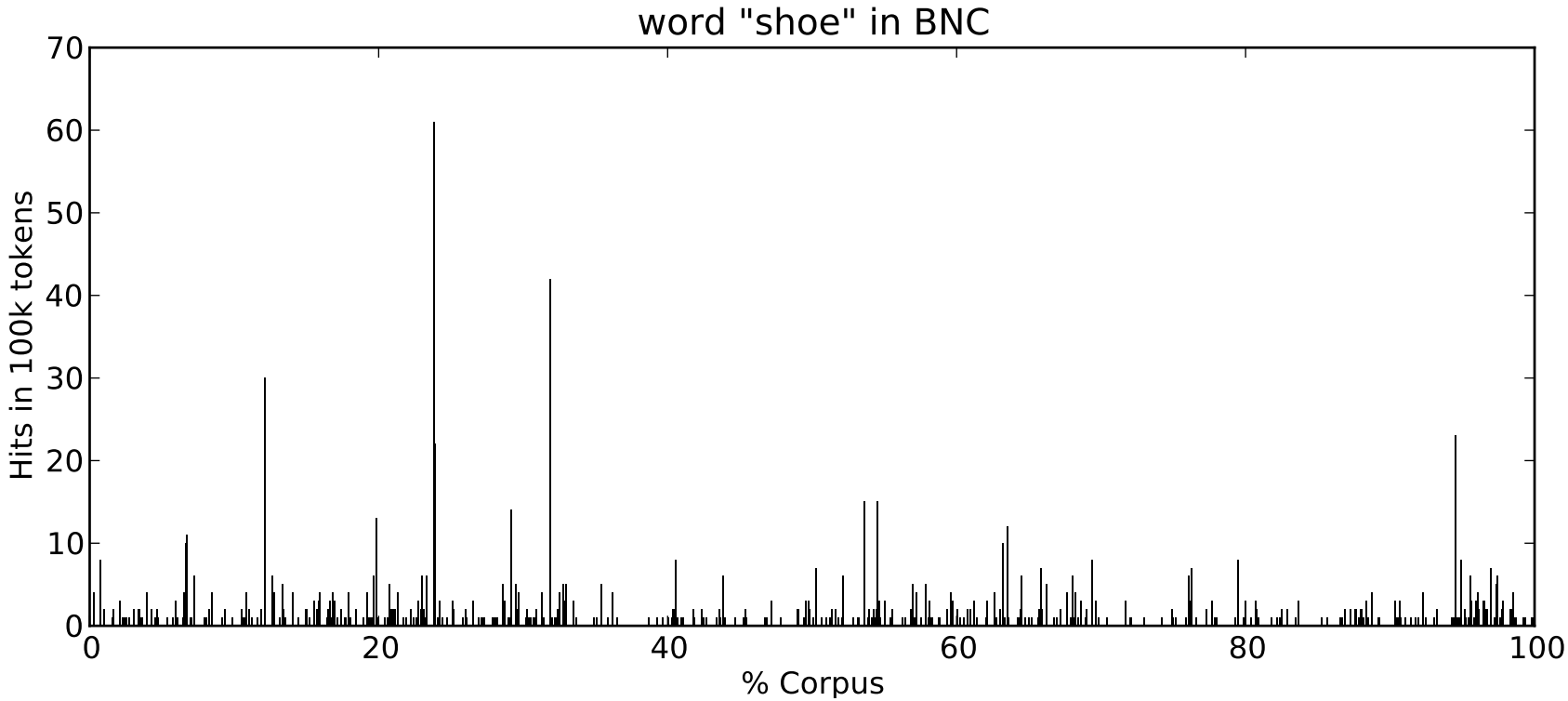
RASLAN 2011

Pavel Rychlý

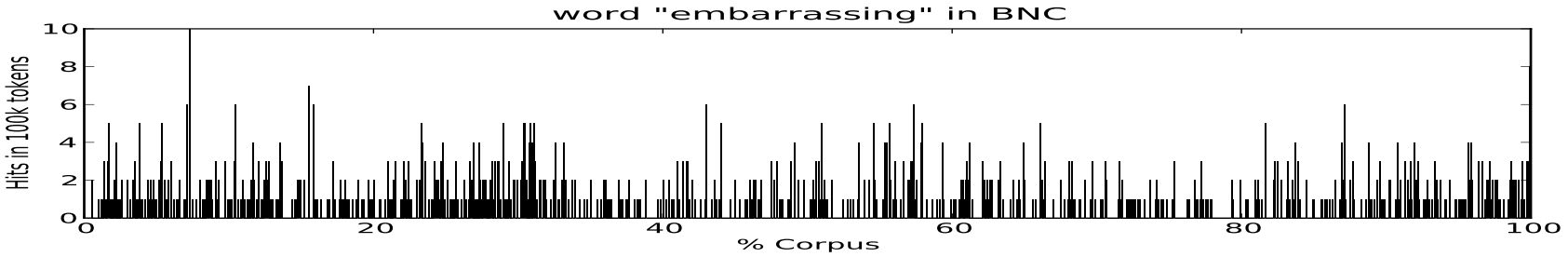
NLP Centre
Faculty of Informatics, Masaryk University
pary@fi.muni.cz

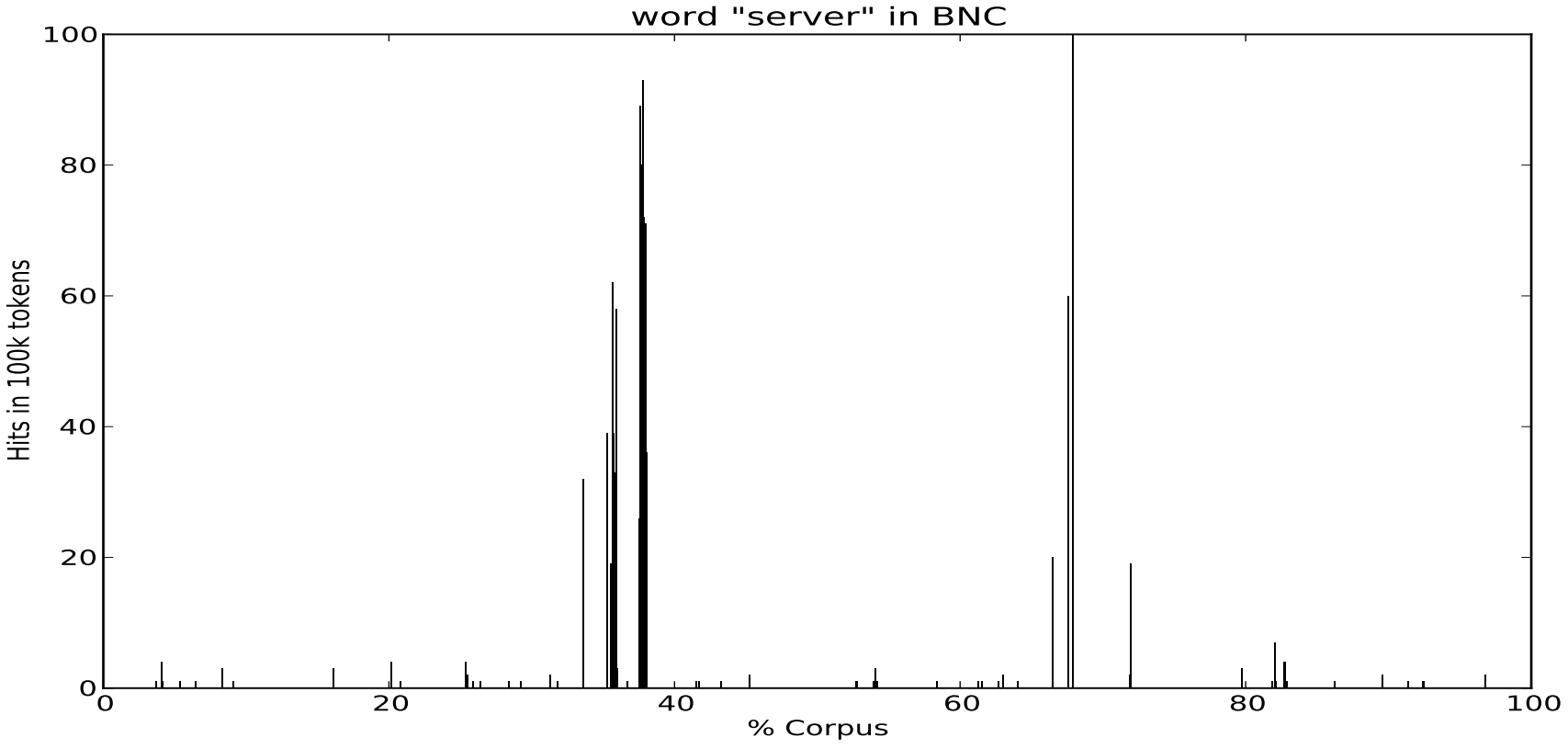
Contents

word “shoe” – 1035 hits in BNC



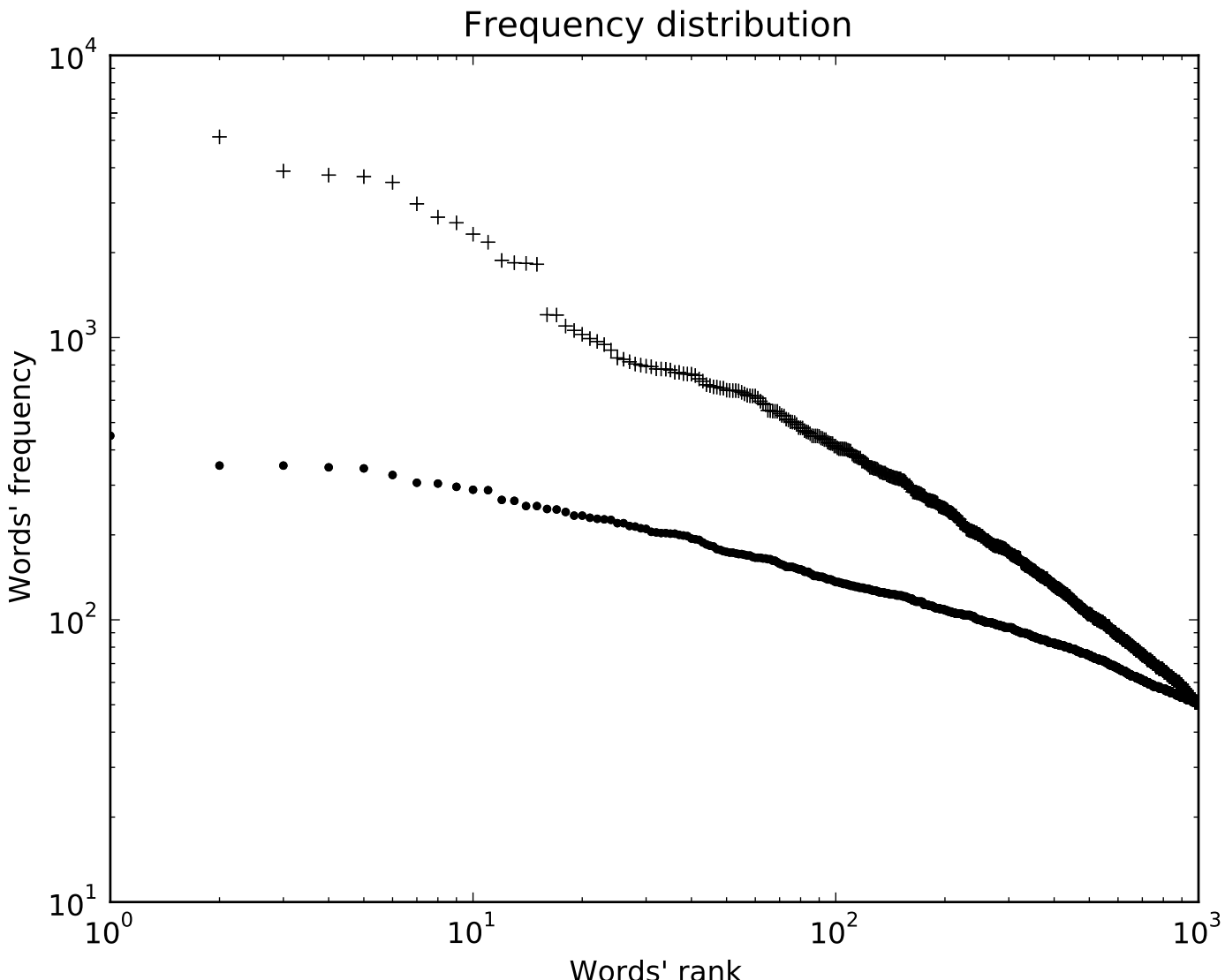
selected words – 1035 hits in BNC





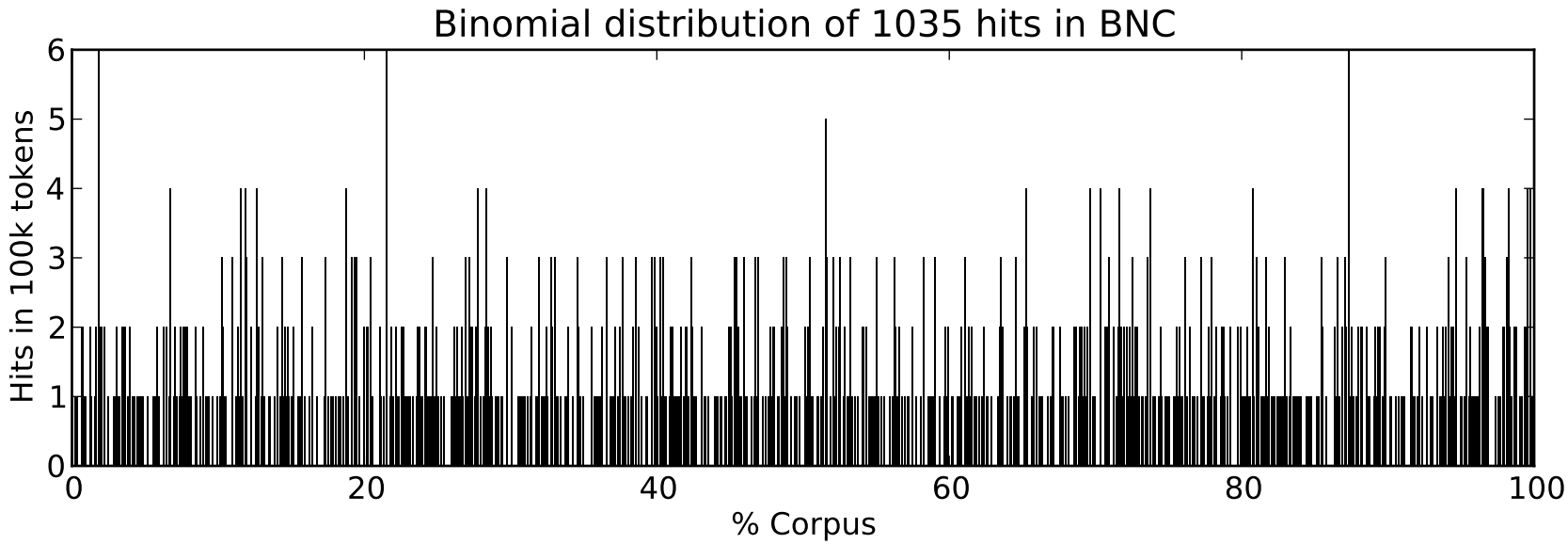
Burstiness do not depend on frequency

Frequency distribution of 1000 words with biggest (upper +) and lowest (lower .) *freq*/*ARF* ratio in BNC.



Binominal distribution – 1035 hits in BNC

Randomly generated:



Bursting Language Model

- words occurs in clusters
- separating probability of clusters and probability of words within a cluster
- three parameters for each word:
  - $P(cluster_w) = \frac{ARF_w}{\hat{N}}$
  - $P(w\ in\ cluster) = \frac{C_w^2}{ARF_w \hat{N}}$
  - $clustersize_w = \frac{N}{10 C_w}$

Evaluation

- training on BNC, cross-entropy on the corpus Susanne
- 10-fold cross-validation on Word Street Journal corpus (WSJ)

	BNC	WSJ
Cross-entropy of unigram model	10.71	10.17
Cross-entropy of bursting model	10.39	9.97
Perplexity of unigram model	1676	1149
Perplexity of bursting model	1337	1006