Applying Zipf's Law on Hamshahri Collection

31. Mai 2015

We depicted the Zipf's distribution of the processed Hamshahri collection together with the distribution of American National Corpus (ANC) as a representation of English language in Figure . While the distributions seem similar, we observe that more frequent words in Persian occur more often than English ones while the frequencies of the less often words are less than English. Our hypothesis is that this behavior is due to the fact that each word (specially verbs) can appear in many forms and thus there are more words with very low frequencies.

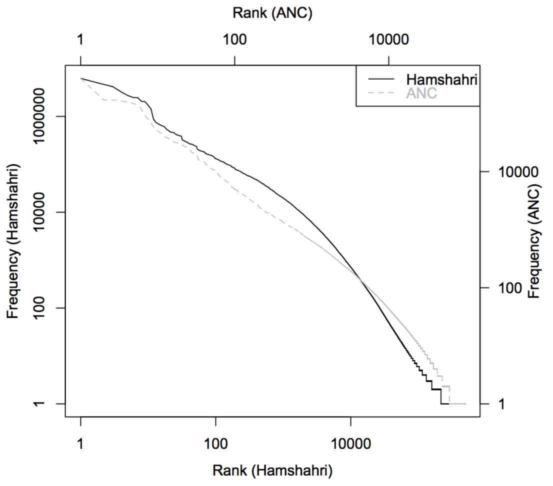

Abbildung 1: Zipf distribution of Hamshahri and ANC corpora