By looking at spoken and written corpora collected from a wide range of users and everyday contexts, we can make fairly reliable statements about how many words are 'in circulation' in everyday communication among native speakers.Conversely, the software often fails to perceive obvious similarities such as young/younger/youngest, so, without considerable manual reprocessing of tens of thousands of words, lemmatised counts can be unreliable: the form bit as in a little bit / a bit small, etc.professional, technical and scientific words, colloquialisms and dialect words), but it will certainly enable us to list the vocabulary of common usage which users are likely to encounter in their daily conversations and in their routine reading of newspapers, magazines, novels, internet texts, etc.This is not to say that a corpus of, say, ten million words can capture all the words of a language, and some users will always know obscure and rare words, for all sorts of reasons, but it will certainly enable us to list the vocabulary of common usage which users are likely to encounter in their daily conversations and in their routine reading of newspapers, magazines, novels, internet texts, etc.However, the computer software also allows us to bring together semi-automatically the inflected forms of words and treat the combined totals as 'lemmas' (i.e. LOOK would be one lemma composed of the total occurrences of look(s), looked and looking).Firstly, the process of lemmatisation tends to bundle together all the forms that the computer judges to look similar, for example the quite different nouns man and mane, or the noun bit, which will be conflated with the past-tense form of the verb to bite.The frequency curve does not decline at a regular rate across the whole of the vocabulary; there is a continental shelf of high-frequency, core items, after which the curve takes a nose-dive into the vast depths of tens of thousands of (relatively) low-frequency words.There are, however, reasons for hesitating in taking lemmas as our benchmark.

Summarized by © lakhasly.com