This paper extends the work of Cavnar and Trenkle N-gram text categorization [2], enhances the study of statistics application on document language recognition as simplier variant of categorization. Proposed program shows qualities like modular design or running on one universal character set. As an enhancement of the original work is presented an automatic text sample filtration algorithm altogether with Internet text extraction and iterative improvement for this purpose. Presented paper studies accuracy development, concentrating on short samples. Similar work was not found in available literature, as categorization (and in corollary language recognition) usually assumes long enough input. In conclusion, a discussion about using the learned data and algorithms created here to mark foreign phrases. To be specific, we study the application on Prague Dependency Treebank [8], where the foreign phrases are not recognized, only their occurences specified.