

Classic textual compression methods work over the alphabet of characters or alphabet of words. For languages with rich morphology as well as for compression of smaller files it can be advantageous to use an alphabet of syllables. For some compression methods like the ones based on Burrows-Wheeler transformation the syllable is a reasonable solution also for large files - even for languages having quite simple morphology. Although the main goal of our research is the compression over the alphabet of syllables, all implemented methods can compress also over the alphabet of words. For small files we use the LZW method and Huffman coding. These methods were improved by the use of initialized dictionary containing characteristic syllables specific for given language. For the compression of very large files we implemented the project XBW allowing combination of compression methods BWT, MTF, RLE, PPM, LZC, and LZSS. We have also tried to compress XML files that are not well-formed. When compressing over a large alphabet, it is necessary to compress also the used alphabet. We have proposed two solutions. The first one works well especially for small documents. We initialize the compression method with a set of characteristic syllables whereas other syllables are coded when necessary character by character. The second solution is intended for compression of larger documents. The alphabet of used syllables is encoded as a compressed trie what significantly reduces the space necessary for encoding of the alphabet.