

# Semi-Lossless Text Compression

Yair Kaufman and Shmuel T. Klein

Department of Computer Science

Bar Ilan University, Ramat Gan, Israel

{kaufmay,tomi}@cs.biu.ac.il

One widespread partition when coming to classify data compression methods is into lossless and lossy methods. *Lossless* methods include usually those applied on text files or other data for which no loss of information can be tolerated, *lossy* techniques are generally applied to image files as well as to video and audio data, for which the overall knowledge a user might extract does not seem significantly reduced even if a part of the data is omitted.

The basic idea behind lossy compression is thus the fact that even if not all of the available data is presented, the human brain can often make up for the missing parts and guess, at least partially, whatever has been omitted, so that overall one has the feeling that nothing has been lost. We try, in this work, to transfer this paradigm also into the framework of text compression, to which usually only lossless techniques have been applied.

A hint to the fact that strict losslessness might be relaxed can be found by anybody who tries to read a newspaper, and mostly succeeds in understanding all the required information in spite of occasional typing errors and other mistakes. It turns out that we are able to understand English text even if there are many more errors, as suggested by the following paragraph, which circulated recently on the Internet

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mtttaer in waht  
oredr the ltteers in a wrod are in; the olny iprmoetnt tihng is taht frist and  
lsat ltteer be at the rghit pclae. The rset can be a toatl mse and you can  
sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed  
ervey lteter by istlef, but the wrod as a wlohe.

If indeed it is true that under certain constraints the exact letter order can be altered without impairing our understanding of the information contained in English text, it follows that the order of the characters induced by English grammar and syntax may contain more redundancy than one thought so far, and eliminating this redundancy might yield improved compression. Being a hybrid of the two classes of compression methods mentioned above, we call the type of compression suggested by these ideas *semi-lossless*: the original text will not be fully reconstructed, just as a decompressed JPEG image is not identical to the original, and thereby the method will be lossy; on the other hand, again similarly to the decompressed image for which our eyes and brain fill in the omitted parts, here it is the knowledge of English that will enable the extraction of the full information of the original text, so that at least from the information point of view, if not from the physically stored file, the method can be considered as lossless.