Tables 2, 5, 8, and 7 show that Gzip does not take advantage of the rearrangements. The same is for Bzip.

This is probably because both compression algorithms are not based on statistical coding and therefore the rearrangements we have applied are not an improvement with respect to the original situation.

This does not necessarily mean that semi-lossless compression does not work for those algorithms, but only that the specific rearrangements we have tried are not suitable to improve those algorithms.

It is interesting to note that for both Gzip and Bzip the *Random* rearrangement gives the worst performance, because it breaks the correlation between parts of the word and reduce the effectivity of the two algorithms.

From the data compression point of view, the most effective algorithm in this case study is Gzip: because of the small length of the input files Bzip (that is generally more performing than Gzip), does not achieve the compression that often gets on larger files.

4 Conclusion

In semi-lossless text compression the decompressed text will not be identical to the original text, but our brain will adjust the text data to make it usable and understandable.

In this paper we have experimented with semi-lossless compression on a case study of five small text files in Italian language.

We have reported the experimental results for Huffman coding, Gzip and Bzip on the files where the inner letters of each word have been rearranged, respectively, with the *Random*, *Alphabetic*, *Probabilistic* or *Frequency* rearrangement methods.

All the resulting files were enough readable: the text was still fully understandable, with a clear slow down in the reader's speed in understanding the words.

The results we have obtained show a slight compression improvement when using Huffman Coding and *Probabilistic* or *Frequency* rearrangements, but no improvement (actually a worsening) when using Gzip or Bzip.

This does not necessarily mean that semi-lossless compression does not work for those algorithms, but only that the specific rearrangements we have tried are not suitable to improve those algorithms.

Future research will involve testing on different rearrangement methods that might be more suited for non-statistical compression algorithms and also testing with files in different languages.

Acknowledgements

I wish to thank my students: Marco Lettieri, Alessia D'Andria, Antonella Masi, Antonella Palladino and Isabella Ruggiero that have conducted a preliminary set of experiments on semi-lossless compression of Italian texts.

References:

[1] Y. Kaufman and S. T. Klein, "Semi-lossless text compression", *International Journal of Foundations of Computer Science*, Vol. 16, No. 6, 2005, pp. 1167-1178.

[2] I. H. Witten, A. Moffat, and T. Bell, *Managing Gigabytes*. NY: Van Nostrand Reinhold, 1994.

[3] *The Gzip home page*, www.gzip.org.

[4] *Bzip2: home*, www.bzip.org.