The Frame Problem in Text Analysis

Maki Miyake (mmiyake@dp.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

Hiroyuki Akama (akama@dp.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

Masanori Nakagawa (nakagawa@nm.hum.titech.ac.jp)
Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

General Formatting Instructions

The aim of this research is to evaluate the extent to which Thesaurus allow us to modify a researcher's knowledge frame. A well calculated Thesaurus has the power to overwrite existent knowledge frames, or habitual "heuristics" for the humanities, if word occurrence data are rigorously manipulated by the algorithms of statistical linguistics. However, even if all targets and means of data gathering and analyzing are readily available, there remains behind various interpretations of subjects a sort of Frame question of how we partition off the texts and documents to avoid arbitrary text segmentation. Frames are needed before we can gather and interpret the data for a word occurrence computation.

If the problem of text segmentation remains unresolved, any experiment in quantitative text analysis will be still far from being realized. We need a sort of "TextTiling" methodology enabling an objective segmentation based on objective criteria.

This holds true for the frame setting in parallel and variant texts as the synoptic Gospels. We have to point out that the basic frame for Biblical research was taken from second hand data called the Parallel Synoptic Table (in abbreviation, PST), which shows the order and arrangement of the "pericopes" belonging to the Synoptic Gospels. The frame traditionally prepared was built only by a "Form Criticism", which divided the texts into parts by the arbitrary unities coming from tradition or reduction.

The purpose of the PST framework has been the resolution of the problem of "who quoted whom" in writing respectively the first three Gospels. However we propose an alternative and more objective way of segmenting the parallel texts by using our web-based biblical software, named "Tele-Synopsis", which is designed to gather information of the word usage under various conditions and to help further statistical approach to the origin of the variant texts. A quantitative analysis (factor analysis) is applied to the lexical datasets obtained by changing framing conditions in order to verify some traditional hypotheses made to explain the mutual relationship of the synoptic Gospels. Our framing principle is that the entire texts can be classified into the following 7 categories which are A: common part of the three Gospels, B: part common to Matthew and Mark, C: part common to Mark and Luke, D: part common to Matthew and Lukas, E: part peculiar to Matthew, F: part peculiar to Mark, G: part peculiar to Luke. It is natural that the category to which each instance of word has to belong for constructing a biblical Thesaurus varies according to the way in which we partition off the parallel texts. But the results of the factor analysis applied to the multiple datasets showed high robustness in the sense that the types of loading matrix are more or less similar *except* for the dataset depending on the traditional PST.

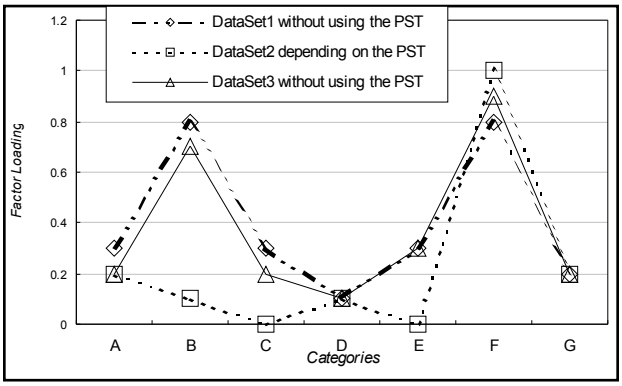

Figure 1: One of the factor patterns for the identical variables with different datasets.

Acknowledgments

This paper was made possible largely through grants from the 21st Century Center of Excellence Program "Framework for Systematization and Application of Large-scale Knowledge Resources". We would like to acknowledge here the generosity of this Center.

References

Minsky, M.L., (1975). A Framework for representing knowledge, *The Psychology of computer vision*, pp.211-277.

Hearst, Marti A.(1997). TextTiling: "Segmenting text into multi-paragraph subtopic passages", *Computational Linguistics* 23, pp.33-64.

Kurt Aland. (1989). *Synopsis of the Four Gospels*, German Bible Society Stuttgart.