Marina Santini
ITRI (Information Technology Research Institute)
University of Brighton (UK)
Marina.Santini@itri.brighton.ac.uk

Annotated Corpora vs. Raw Web Page Collections.
Text Types, Web Pages and Linguistic Features: Some Issues

We consider whether the traditional text typologies suggested by (corpus-)linguists might apply to web pages. By traditional text typologies we mean both those text types worked out by text linguists (Werlich 1976), and those derived using a statistical and corpus-based approach (Biber 1988). We illustrate the problems that arose when we tried to derive text types from a sample of web pages belonging to the SPIRIT collection (Clarke et al. 1998) using simple heuristics and a set of linguistic features. The SPIRIT collection is multilingual and without any meta-information, except a short header including the original URL, the date and time when the pages were crawled from the Web, and few other details. It is not provided with any morphological or syntactic annotation, nor is balanced with respect to genres or registers. This collection represents a genuine random slice of the real Web, "frozen" for research purposes. Two hundred web pages written in English were randomly chosen and used in this experiment. The analysis of the results led to the following conclusions:

1) Parsing a raw web page (i.e. a web page saved in a text-only format) may return an unreliable output. Some pre-processing might be needed, especially when ending punctuation marks are missing. For example, headings, which usually are not marked by an ending period, are often joined together with the following sentence when parsed.
2) Linguistic features alone are not enough to derive a text typology of Web pages. Layout and interactive functionalities, such as hyperlinks and search facilities, should be taken into account (cf. Amitay 1999, and Shepherd and Watters 1999). Therefore, HTML markup tags should be interpreted in a functional way and combined with linguistic features for the identification of text types of web pages.
3) A web page is often a mixed document, bearing several purposes at the same time. Around the main body of the page, there is often peripheral text, such as navigational links, ads, text boxes, and other textual material. How are we going to deal with this new, complex and pervasive types of documents?
4) Web pages require a broader range of text types. For example, what is the text type of an ‘alumni directory’ or a ‘course schedule’? The repertoire should be enlarged and discussed.

The results of our experiment highlight a range of problems related to a corpus-based approach to the text typology of a raw collection of web pages. These issues will entail useful and fruitful discussion within the corpus-linguistic community.

References

Amitay E. (1999), Anchors in context: a corpus analysis of Web pages authoring conventions, in L. Pemberton and S. Shurville (eds.), *Words on the Web*, Intellect Books, UK.

Biber D. (1988), *Variation across speech and writing*, Cambridge University Press, Cambridge.

Clarke C., Cormack G., Laszlo M., Lynam T., Terra E. (1998), The Impact of Corpus Size on Question Answering Performance, *Proc. of the 25 Annual International ACM SIGIR Conference on Research and Development in IR*, Tampere, Finland.

Shepherd M., Watters C. (1999), The Functionality Attribute of Cybergenres, *Proceedings of the 32 Hawaii International Conference on System Sciences*.

Werlich E. (1976), *A Text Grammar of English*. Quelle & Meyer, Heidelberg (Germany).