

A Corpus Model of Structure Formation in Hypertext Types

Alexander Mehler¹

¹Bielefeld University (Germany)

Short Abstract

This presentation describes a web genre corpus model. Its starting point is a graph model of the logical document structure of hypertext types and of the linkage of their constituents. We describe an XML-based serialization of this model and provide a database mapping which retains a wide range of web genre data. This will be exemplified by three web genres.

Long Abstract

We argue in favor of a corpus model of structure formation in hypertext types or web genres, respectively. Our starting point is a layered graph model which distinguishes document networks, document units and their constitutive modules down to the level of elementary linguistic components. Consequently, we will distinguish web pages and websites as manifestation units of web genres. This model of the logical document structure of hypertext types focuses on the linkage of web genre constituents and, thus, investigates the distinction of traditional and web genres from a structural point of view. One advantage of this model is that it allows specifying the correlation of functional and structural varieties in web-based communication in terms of poly-functionality, polymorphism and vagueness. We will present a structural model which integrates this informational uncertain mapping and will describe a serialization thereof in terms of the Graph eXchange Language (GXL). Finally, we will provide a database mapping which allows retaining a wide range of web data for corpus linguistic and computational linguistic genre studies. This will be the starting point for a classification experiment which sheds light on Biber's structural hypothesis about the correlation of functional and structural varieties in the area of web mining. In order to do that we will exemplify three web genres, i.e. personal academic homepages, conference websites and city websites. Each of these genres will be mapped onto our graph model and we will provide statistical data which describes the distribution of document structures within these genres. A feature which makes classification of functional constituents of websites a very hard task is the Zipfian nature of the distribution of hyperlink-based document structures. In other words: the absolute majority of sites consist of single pages which hide, so to speak, the functional or thematic structuring of the site. At the same time we observe a rapid transition to more elaborate sites with kernel hierarchical structures superimposed by graph-inducing up, down, and across links. This Zipfian transition (which is mapped by a power law) is affected by the life cycle of the web genre in question: younger instances are more likely less structured than older ones. We argue that both observations have a high impact on web genre classification especially in those cases where the simple classification model of unstructured web genre tags is replaced by a model which is more sensitive to document structures.