

# Genre Identification

Mikael Gunnarsson

mikael.gunnarsson@hb.se

<http://www.adm.hb.se/~mg/>

Swedish School of Library and Information Science  
Swedish National Graduate School of Language Technology

May 20, 2006

Core areas of Library and Information Science and library practices are those of bibliographic classification and indexing, usually seen as acts of “subject analysis” aimed at identifying the topic(s) of documents. This intellectual effort focuses mainly on the question of *what a document is about*, treating the questions around what a document is or is intended to do as less important.

Documents, if we restrict ourselves to talk about textual documents, are compounds of linguistic utterances, and are thus objects of interest for linguistics as well. Following the speech act theory of Austin (1975), we may say that linguistic utterances are *constative* or *performative*. The constative character is related to the aboutness question referred to above, whereas the performative is what utterances do. It is then tempting to assume that documents may as well be characterised as constative and performative and that in most cases the performative character is no less important for the use of documents.

The performative character of documents relate to the rhetoric and linguistic notions of *genre* and *text types*, scarcely investigated within Library and Information Science, though there are some recent attempts to apply these notions of genres and text type for classification problems in more depth. For instance, Crowston and Kwasnik (2004) applies a faceted classification approach in order to identify what “... clues do people use to identify genre when engaged in information access activities”. Other areas, such as Computational Linguistics in particular, demonstrate an increasing interest for automated genre identification and classification, as surveyed by Santini (2004).

The aim of my thesis work is to model the notions of genres and text types for algorithmic classification and clustering tasks. The understanding of genres and text types applied is an attempt to integrate empirical linguistic research on genres, register and text types with the sociocultural understandings of genre expressed by e.g. Swales (1990), Miller (1994), Mayes (2003) and recent document theory within Library and Information Science.

The models investigated focus on feature extraction for similarity measuring, where the features are taken as indicators of text types. Special emphasis is on features derived from document structure as it is conveyed by markup, assuming that document structure variation expresses a genre repertoire.

**Keywords:** Genres, Text types, Automated classification, Document structure, Markup

## References

- Austin, J. L. (1975). *How to do things with words: the William James Lectures delivered at Harvard University in 1955*. Oxford University Press, Oxford, 2 edition.
- Crowston, K. & Kwasnik, B. H. (2004). A framework for creating a faceted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Hawaii International Conference on System Sciences*. IEEE.
- Mayes, P. (2003). *Language, social structure and culture*. John Benjamins, Amsterdam.
- Miller, C. R. (1994). Genre as social action. In Freedman, A. & Medway, P., editors, *Genre and the New Rhetoric, Critical Perspectives on Literacy and Education*, pages 23–42. Taylor & Francis, London.
- Santini, M. (2004). State-of-the-art on automatic genre identification. Technical Report ITRI-04-03, ITRI, Univ. of Brighton.
- Swales, J. M. (1990). *Genre analysis; English in academic and research settings*. Cambridge Univ. Press, Cambridge.