

# Genre Identification on the Web

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

September 29, 2006

**Abstract: One of the problems with the web and its tools for browsing and searching its contents is that its genre heterogeneity is not sufficiently accounted for. Recent years have demonstrated some successful efforts in algorithmic genre identification, but the area is still in its infancy. This paper presents the general conception of genre and how document structure may be used as an additional discriminative feature in algorithmic genre identification.**

## 1 Introduction

The web is nowadays an almost self-evident source in any information seeking task. It is sometimes described as a huge digital library, a statement which is refuted by others who object that it is not a library since it is not carefully organized. The description, indexing and classification of its contents that characterizes an ordinary library does not characterize the web. However, it cannot be denied that the web is a huge repository of documents and in that respect resembles an ordinary library. Therefore, methods and techniques that apply to library collections may be (and is) adopted for the web.

One of the main properties of the web that distinguishes it from a physical library is its heterogeneity with respect to genre. Genre as a reflection of typifications in documentary and communicative practices is generally not taken care of in all the tools available for browsing and searching the web. Most efforts in improving information seeking tools focus on topical retrieval and seem to consider genre adherence as a secondary matter.

Genre is generally not fully considered in libraries either, but the classification schemes used in libraries to organize its collections do incorporate aspects on genre, though genre is often counted among the so called form subdivisions of classification schemes. (Wilson & Robinson, 1990) But it has been emphasized by Crowston & Kwasnik (2003) that the understanding of genre within Library and Information Science and library practices is shallow, which sometimes is explained by the fact that genre within library collections is not considered a problem. They also observed that library collections demonstrate a rather restricted genre repertoire in comparison to the web. What is more is that library classification schemes have a bias for library collections and do not transfer well for application to extensive collections of web documents, at least not to the extent that it can be used to cover its entirety.

The economic unfeasibility of having humans to classify the complete contents of the web is apparent, and if we want to take advantage of the web's encompassing coverage we have to rely on algorithmic approaches to the description, indexing and classification of its contents. With respect to its contents in terms of topic, theme or subject matter, much work has been and is being done. More scarce are the attempts to algorithmically organize contents from the perspective of genre.

Recent years have seen an increasing interest in algorithmic genre identification, mainly within computational linguistics and computer science, and mainly with respect to textual documents. Such attempts demonstrate that genre is not a straight-forward concept. Many consider genre as a matter of artefactual style only (e.g. Bogdanov & Worring, 2001; Dewdney et al., 2001; Folch et al., 2000), while others take a more sociocultural stance and consider genre as a reflection of communicative purpose (e.g. Kessler et al., 1997; Wolters & Kirsten, 1999; Rauber & Müller-Kögler, 2001).

## 1.1 Genre

Though genre may be best known in the context of literary studies, genre is referred to in many areas. Swales (1990) gives an extensive overview of its occurrence. To this may be added the recent attention given to the concept as a framework for the study of organizational communication. (e.g. Orlikowski & Yates, 1994; Honkaranta, 2003)

From the perspective of the so called new genre theory, genre is usually considered a triple of 1) communicative purpose in particular situations, 2) a discourse community in which communication and documentation activities take place, and 3) a set of typified artefacts. (Cf. Swales, 1990; Mayes, 2003; Bazerman, 1994) For Carolyn Miller, in her seminal article of the 1980s, genre is understood as “typified rhetorical actions based in recurrent situations” (Miller, 1994) and thus a sociocultural and sociolinguistic phenomenon.

The new genre theory reflects an inductive approach in which empirical studies is carried out in order to describe genre artefacts and the conceptions and expectations of particular genres. Some genres have gained more interest than others. The research article is one of these, and Chapter 7 of (Swales, 1990) demonstrates this, where the research article is investigated against the background of the common sequential IMRD structure — Introduction, Method, Results and Discussion — with an extensive exposé of earlier studies. At least in some domains it is expected that all articles obey the IMRD formula in order to be accepted. There are other structural typifications as well. For instance, it is common to start a research article from a general reasoning and narrow it down to a more specific problem while at the end of the discussion moving out from the particular to the general.

Genre theory provides background knowledge for the identification of genre if it can be reasonably assumed that their findings reflect the ways that humans recognize genre. One aspect of particular impact is that genre relates to human practices in a way that both “shapes the schematic structure of the discourse and influences and constrains choice of content and style” (Swales, 1990, p. 58). Neither communicative purpose, communities of practices, nor artefacts can be given evident primacy over any other aspect on genre.

Linguists in general, in addition to the term genre, operate with terms that have similar connotations. One of these terms is *register*, preferred by e.g. Biber & Finegan (1994) that compared with genre theory are more directed towards the details in how language is used. Swanson (2003, p. 21f), who studies the use of anaphoric and cataphoric constructs within different genres, argues for a distinction where genre is

associated with cultural context and the word register with situational context. The register, for Swanson, is a mediator of the realization of a genre.

In any case, genre is inherently an abstract construal of documentary practices. This abstraction is determined by the three very general variables referred to above, but for the identification of genre we must rely on the third observable variable — the artefact. Here is a tension to be found, between individual variation and genre typification. Within one and the same communicative situation and discourse community different people may realize their artefacts in slightly different ways. Genre identification must identify the non-variable aspects of the artefacts and at the same time bear in mind that individual variation over time may change the genre. In addition, genres may be identified on different levels of granularity. (Cf. Santini, 2006) The genre of research articles is a very broad one. It is obvious that research articles are used within many different scholarly communities where authorial conventions differ. It may also be supposed that research articles are not always produced with the same intentions or purposes. The communicative purpose is a cumbersome variable, as has been admitted by Askehave & Swales (2001), and it is not always evident that the purpose of a research article is the intention of reporting on research, which may be a common conception of its purpose.

Genres need not to have commonly recognized names. (Cf. Santini, 2005; Ferguson, 1994, p. 22) Maybe it is reasonable to look at the term research article as a common denominator for a set of unlabeled genres. The indeterminacy of genre necessitates that any algorithm for genre identification refrain as far as possible from any presupposition of a finite and predetermined space of labeled genres.

## 1.2 Genre as a Reflection of Speech Acts

The idea of genre theory is that genre is a conflation of language use and action, an otherwise common focus for any sociolinguistic theory influenced by certain philosophies of language. For instance, Austin (1975, p. 5) stated that a linguistic utterance “is, or is part of, the doing of an action”. In his influential 1955 lectures Austin started out by observing that many utterances do not conform to the then common philosophical conception of a statement, they do not constating anything that can be characterized as being true or false. He termed these non-constating utterances performatives, because they are used in order to achieve some particular goal, to perform some job, without which the goal cannot be achieved. Common and distinct examples are to be found in the context of marriage and baptism ceremonies where the uttering of certain phrases are the necessary requirements for the fulfilment of the ceremony. Consider a more subtle example, a research article in which an utterance starts out

We define classification as being the assignment ...

*Example 1.1*

This is obviously different from

Classification is an assignment of ...

*Example 1.2*

The theme or topic for both Example 1.1 and 1.2 is classification and they both say

something about the theme, which is considered the rheme. According to the systemic-functional grammar of Halliday & Matthiessen (2004) where words “get their meaning from activities in which they are embedded” (Halliday & Hasan, 1989, p. 5), this is the textual meaning of both utterances. The example in 1.1 is different from the one in 1.2 with respect to its degree of apparent performativeness. In 1.1 the author does not state anything that is to be considered true or false. The author is simply stating in what sense he or she is going to use the word classification, a commissive act of speech. According to the systemic-functional grammar there are differences with respect to ideational and interpersonal meaning. The ideational meaning of 1.1 is conveyed by the commissive verb “define” as opposed to the existential “is” and the interpersonal meaning by the choice of the pronominalization “we” of the author(s) in 1.1 as opposed to the implicitness of the utterer in 1.2.

In the end of his lectures, after trying to identify a lexical catalog of verbs and verb forms that mark the occurrence of performatives, Austin ends up by observing that the distinction between performatives and constatives is not so clear as it first appeared to be. Almost any utterance implies an enactment of a speech act. The difference between constatives and performatives is only a slight shift in balance between locutionary and illocutionary forces — between what an utterance is saying and what it is intended to do. This marks the difference between what is focused in identification of topic versus genre.

The kind of variation exemplified above is important for language use as action, as opposed to simple sayings, and with respect to genre the act of saying matters equally much as what is said. Certain ways of saying something are indicative of the situation and the sociocultural setting in which the act takes place. The bulk of empirical and theoretical research on language use, that embarks from views on language use as exemplified with Austin’s speech act theory, provide us with the background knowledge for modeling genre identification.

### **1.3 Modeling Genre Identification**

Characteristic for algorithmic applications assisting human tasks is that they must rely on some model of human action (or cognition). If genre is typified action and recognized as such, its embedded artefacts can be assumed to demonstrate this typification as typified language use. We must ask which clues do people use in deciding upon genre, just as has been done with respect to topical identification, highly connected to the identification of theme. This is what has been done by e.g. Crowston & Kwasnik (2004) with respect to heuristic genre classification.

As for topical identification the task is greatly simplified by the mere existence of technological conventions, such as titles, tables of contents and back of the book summaries. These conventionalized text units make it unnecessary to read the complete text and sum up all the themes of clauses and paragraphs into one overarching theme. These text units have developed over centuries and are part of many kinds of documents.

In automated genre identification clues derived from the linguistic contents is almost self-evident, since typification in language use is considered indicative of genre. The question whether technological and other extra-linguistic properties may be used in genre identification is to a large extent uninvestigated — both as indicative of genre and as feature weighting indications. Let us return to this question later on and first consider some more aspects on how texts are composed into artefacts, by reference to what has been termed text grammars.

In the text grammar of Werlich (1976) a text is a structure “marked by both *coherence* among the elements and *completion*” [emphasis in orig.]. It is a structure because it is constituted by clauses linked together to larger text units that form completed texts by being linked in their turn. What seems to be the most attended aspects of a text for Werlich is how the linking between text units is accomplished and how the semantics of each unit refers to contextual phenomena. On this framework Werlich plots out five categories of text units based on their “dominant contextual foci” and how their units are linked according to these foci.

Werlich adapts the fairly traditional term *text type* to signify these categories. There are five text types in Werlich’s typology: descriptions, narrations, expositions, argumentations, and instructions. These are categories on an abstract level reflecting authorial strategies and should not be taken as directly reflecting genre. For instance, text units of the narrative category are characterized by a dominance of temporal linking, often realized with extensive use of temporal adverbials, whereas a description on the other hand is realized with extensive use of spatial adverbials.

With respect to genre identification, text type occurrences may very well be indicative. It is for instance not common to find narrations and instructions in scholarly writings, if not as illustrations. According to Werlich (1976, p. 46) text types are conventionally manifested as text forms — which is just another word for the common conception of genre. The binary feature of a text being narrative or not has been used by Kessler et al. (1997) for algorithmic genre identification as a “facet”.

Topical identification (which must be distinguished from the subject analysis of library practices — a more encompassing activity) is modeled according to assumptions on word frequency distributions over sets of documents. The algorithms differ mostly in the ways word frequency distributions and collocations are accounted for in the algorithms. In heuristic topical identification, this task is assisted by the existence of technological conventions.

The question is what assumptions algorithmic genre identification may rely upon. Let us consider what is usually relied upon.

The number of features used in genre identification is generally high. The number of features used by Biber (1988) in his seminal work of the 1980s was 67, but recent efforts tend to regard even more features. For instance, Finn & Kushmerick (2003) used 152 features. (Stamatatos et al., 2000) is probably one of the crudest approaches. They used word counts and investigated how certain words were discriminative for certain genres. Words that within a group of documents with the same topic demonstrated a high frequency within one genre but a very low frequency in other genres were considered discriminative. Their result was then a set of 30 discriminating words. To this kind of features they added figures on the occurrence of punctuation characters. Kessler et al. (1997) in addition to words (lexical features) and punctuation counts used what they refer to as “derivative” features, which is a combination of lexical features and character-level features — sometimes similar to traditional text complexity measures, such as the proportion of long words. A fourth group of features, referred to as structural cues, is used by some: Karlgren (2000); Dewdney et al. (2001); Argamon et al. (1998); Wastholm et al. (2005) all used part-of-speech tags in their experiments. An obvious drawback is that this kind of features requires part-of-speech tagging or parsing which presupposes a computationally expensive preprocessing. A fifth group of features used by, for instance, Santini (2005), Lim et al. (2005), Elsas & Efron (2004) and Rauber & Müller-Kögler (2001), are counts of certain HTML tags. Lim et al. (2005) put particular interest in the URLs of the hyperlinking tags and take account of

if they refer to documents within the same domain or not.

Besides these five groups of features, there are some who focus on the appearance of documents, and nothing else. Bogdanov & Worring (2001) and Ihlström & Åkesson (2004) do this, but it should be remembered that these two investigations do not primarily aim at classification for information seeking tasks. (Power & Scott, 1999; Hu et al., 1999) are two other examples.

The results of all research give no significant clues to whether some set of features are better than others, but most report on the fact that combinations from different groups of features yield better results in classification tasks.

Algorithmic genre identification does not generally consider technological conventions. For instance, the occurrence of certain text unit types, such as bibliographies, quotations and the like, is not at all considered. This may be seen as surprising, since some text unit types are particular for certain groups of genres. A bibliography is for instance common for many scholarly genres, whereas it is not for non-scholarly works, at least not with linking to inline citations to the same extent.

The work that is only tentatively presented here, given the restricted space, focuses on the identification of these text units to improve genre identification. These text units map to a set of text unit types (which will be referred to as genre modules, from the point of view of genre) where each member is distinguished by common functions in documentary practices.<sup>1</sup> Genre modules should not be taken as cognitive categories, the model does not assume anything of authorial intention or readerly perception that otherwise distinguish many text typologies.

## 2 Methodology

Space does not allow for a complete description of the method here. It must suffice to point out some directions with relation to what has been said above on algorithmic genre identification.

### 2.1 The document seen from the perspective of genre

Formally we may say that from the point of view of genre, a document  $d_i$  can be defined as an ordered set of micro-acts realized as distinct text units (Equation 1).

$$d_i = \langle t_1, t_2, \dots, t_n \rangle \quad (1)$$

Thus, from the point of view of document structure typification, a text unit  $t_j$  is of a certain type. Let us term this type a *genre module* and denote a particular genre module with the symbol  $gm_k$ , where  $k$  denotes the type (given as an index number in the following). For a given set of documents  $D$  we have a set of genre modules as in Equation 2.

$$Gm_D = \{gm_1, gm_2, \dots, gm_n\} \quad (2)$$

Having identified the text units  $T_{d_i}$  of a certain document  $d_i \in D$ , each text unit  $t_j \in T_{d_i}$  has to be identified as being of type  $gm_k \in Gm_D$ . A classified text unit is denoted by  $t_j(gm_k)$ . A document can then be described as a set of classified text units as in Equation 3.

---

<sup>1</sup>The term genre module has been introduced by Rehm (2002) with a similar meaning

$$d_i = T_{d_i} = \langle t_1(gm_x), t_2(gm_y), \dots, t_n(gm_z) \rangle \quad (3)$$

Representing a document as a flat sequence of text units is a simplification, because we ignore the fact that some text units are part of other text units. Lists are easier to work with than trees, and the position of any text unit in a tree can be encoded as a feature of the unit.

## 2.2 The document seen from the perspective of the artefact

Seen from the point of view of genre, text units and genre modules are both abstractions in the sense that they require human heuristics to be identified. From the appearance of documents a text unit may be identified from e.g. its surrounding white space. The type of such a text unit may then be interpretatively identified by a human mind. Algorithmic applications, on the other hand, need to be developed from apparent data on a technical level from which the identification and interpretation proceeds.

Since the domain which this work focuses on is the web, algorithms must be developed with respect to the technology of the web, which HTML markup is part of. From the technological perspective of markup a document  $d_i$  can be defined as a set of nodes  $N_{d_i}$  (Equation 4).

$$d_i = N_{d_i} = \langle n_1, n_2, \dots, n_n \rangle \quad (4)$$

Figure 1 of the appendix depicts an HTML snippet consisting of the representation of an endnote. The nodes of an HTML document are of different types. According to markup terminology they may be of any one of seven different types: element nodes, text nodes, attribute nodes, comment nodes etc. Some of these types are of less interest here<sup>2</sup>, and we shall look upon a document as a set of text nodes. The endnote in Figure 1 is realized as 7 text nodes, 6 element nodes and 3 attribute nodes.

Now, each text node  $n_l$  in the set of text nodes in  $N_{d_i}$  of a document  $d_i$  is assumed to correspond with a part of a text unit  $t_j \in T_{d_i}$ , or being identical to it. Each text node of Figure 1, for instance, is part of the text unit of type endnote.<sup>3</sup> This assumption can be stated as in Equation 5

$$t_j(gm_k) \equiv \{n_1(gm_k), n_2(gm_k), \dots, n_n(gm_k)\} \vee n_l(gm_k) \quad (5)$$

Because of this equivalence, we may substitute the task of text unit classification with a text node classification. If the task is considered a function that maps a text unit onto a space of genre modules, we assume the following equivalence as well.

$$\mathcal{F}_n : n_l \rightarrow gm_k \equiv \mathcal{F}_t : t_j \rightarrow gm_k \quad (6)$$

## 2.3 Text Unit Sequences as Document Features

The purpose of this decomposition of documents is to find a way to represent a document structure used as a feature in the process of genre identification. Actually, we do not want to operate on text nodes, we only intend to use sequences of classified text

<sup>2</sup>For instance, the root node is always html, and comment nodes contain text or code that do not often relate to the linguistic content of the document.

<sup>3</sup>It may be observed that the element node with its contained text node at line 10 is a navigational commodity and could be treated as a particular type of text unit in itself.

units as document features. The equivalence assumed in Equation 5 thus requires a composition from text nodes to text units. A fairly simple algorithm is enough for this.

Taking the ordered set of classified text nodes,  $\langle n_1(gm_x), n_2(gm_y), \dots, n_n(gm_z) \rangle$ , the algorithm maps each sequence of text nodes with the same class assigned, i.e.  $x = y$ , to one and the same text unit. Where there is a transition from one class to another in the sequence, there is a shift of text unit.

Two problems may occur with respect to an idealized conception of a text unit. First, two text units of the same type will be collapsed into one, and genre modules that are naturally contained in other genre modules will split its container into two text units. Since the resulting structure will be used as a document feature only and this approximation is consistently applied, this is not treated as a problem.

### 3 Closing Words

Research on algorithmic genre identification has not reached the maturity that applies to information retrieval research. There is a lack of common and large text collections that allow for benchmarking. In addition, different research initiatives focus on fairly different genre spaces. As a consequence, results in terms of accuracy give nothing more than vague indications on how research efforts best proceed, and there are no common understandings of what features are the most valuable features for genre identification.

The work presented here aims to support the area with an attempt to model document structures for use as additional features in genre identification.

### References

- Argamon, S., M. Koppel, & G. Avneri (1998). Routing document according to style. In *Proceedings of the First International Workshop on Innovative Internet Information Systems*.
- Askehave, I. & J. M. Swales (2001). Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics*, 22(2):195–212.
- 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.
- Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In A. Freedman & P. Medway (editors), *Genre and the New Rhetoric*, Critical Perspectives on Literacy and Education, pp. 79–101. Taylor & Francis, London.
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.
- D. Biber & E. Finegan (editors) (1994). *Sociolinguistic Perspectives on Register*. Oxford Studies in Sociolinguistics. Oxford University Press, Oxford.
- Bogdanov, A. D. & M. Worring (2001). Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*.

- Crowston, K. & B. H. Kwasnik (2003). Can document-genre metadata improve information access to large digital collections. *Library Trends*, 52(2):345–361.
- Crowston, K. & B. H. Kwasnik (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.
- Dewdney, N., C. VanEss-Dykema, & R. McMillan (2001). The form is the substance: Classification of genres in text. In *ACL Workshop on Human Language Technology and Knowledge Management*, pp. 142–149.
- Elsas, J. & M. Efron (2004). Html tag based metrics for use in web page type classification.
- Ferguson, C. A. (1994). Dialect, register, and genre: Working assumptions about conventionalization. In D. Biber & E. Finegan (editors), *Sociolinguistic Perspectives on Register*, pp. 15–30. Oxford University Press.
- Finn, A. & N. Kushmerick (2003). Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*.
- Folch, H., S. Heiden, B. Habert, G. Illouz, S. Fleury, P. Lafon, J. Nioche, & S. Prévost (2000). TyPTex: Inductive typological text classification by multivariate statistical analysis for nlp systems tuning/evaluation. In *LREC 2000*.
- Halliday, M. A. K. & R. Hasan (1989). *Language, Context and Text; Aspects of Language in a Social-semiotic Perspective*. Oxford University Press, Oxford, 2 edition.
- Halliday, M. A. K. & C. M. I. M. Matthiessen (2004). *An Introduction to Functional Grammar*. Arnold, London, 3 edition.
- Honkaranta, A. (2003). *From genres to content analysis: experiences from four case organizations*. PhD thesis, Department of Computer Science and Information Systems. University of Jyväskylä. Jyväskylä Studies in Computing ; 31.
- Hu, J., R. Kashi, & G. T. Wilfong (1999). Document classification using layout analysis. In *DEXA Workshop*, pp. 556–560.  
[citeseer.ist.psu.edu/article/hu99document.html](http://citeseer.ist.psu.edu/article/hu99document.html)
- Ihlström, C. & M. Åkesson (2004). Genre characteristics: a front page analysis of 85 swedish online newspapers. In *Proceedings of 37' Hawaii International Conference on Systems Science*. IEEE Press.
- Karlgren, J. (2000). *Stylistic experiments for information retrieval*. PhD thesis, Department of Linguistics. Stockholm Univ.
- Kessler, B., G. Nunberg, & H. Schütze (1997). Automatic detection of text genre. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and the 8th meeting of the European Chapter of the Association for Computational Linguistics*, pp. 32–38, San Francisco. Morgan Kaufmann Publishers.
- Lim, C. S., K. J. Lee, & G. C. Kim (2005). Multiple sets of features for automatic genre classification of web documents. *Information Processing and Management 41 (2005) 1263–1276*, 41:1263–1276.

- Mayes, P. (2003). *Language, social structure and culture*. John Benjamins, Amsterdam.
- Miller, C. R. (1994). Genre as social action. In A. Freedman & P. Medway (editors), *Genre and the New Rhetoric*, Critical Perspectives on Literacy and Education, pp. 23–42. Taylor & Francis, London.
- Orlikowski, W. & J. Yates (1994). Genre repertoire: the structuring of communicative practices in organizations. *Administrative science quarterly*, 39:541–574.
- Power, R. & D. Scott (1999). Using layout for the generation understanding or retrieval of documents : Papers from the 1999 AAAI fall symposium. Notes.
- Rauber, A. & A. Müller-Kögler (2001). Integrating automatic genre analysis into digital libraries. In *JCDL '01, Roanoke, Virginia, USA*. ACM.
- Rehm, G. (2002). Towards automatic web genre identification. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4*. IEEE Computer Society.
- Santini, M. (2005). Genres in formation? an exploratory study of web pages using cluster analysis. In *Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.
- Santini, M. (2006). Common criteria for genre classification: Annotation and granularity. In *Workshop on Text-Based Information Retrieval (TIR-06)*, Riva del Garda, Italy.
- Stamatatos, E., N. Fakotakis, & G. Kokkinakis (2000). Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, pp. 808–814, Luxembourg.
- Swales, J. M. (1990). *Genre analysis; English in academic and research settings*. Cambridge Univ. Press, Cambridge.
- Swanson, W. (2003). *Modes of Co-reference as an Indicator of Genre*. Linguistic Insights: Studies in Language and Communication; Vol. 12. Peter Lang, Bern.
- Wastholm, P., A. Kusma, & B. Megyesi (2005). Using linguistic data for genre classification. In *Advances in Artificial Language in Sweden. The Annual Swedish Artificial Intelligence and Learning Systems Event, SAIS-SSLS, April 2005*, pp. 173–176.
- Werlich, E. (1976). *A Text Grammar of English*. Quelle & Meyer, Heidelberg.
- Wilson, P. & N. Robinson (1990). Form subdivisions and genre. *Library Resources & Technical Services*, 34(1):36–43.
- Wolters, M. & M. Kirsten (1999). Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics*, pp. 142–149. Association for Computational Linguistics.

## Appendix

```
1 <li>
2   <small>
3     <a name="note55">The</a> book in question is Dekius Lack,
4     <i>Biljard klockan noll i cyber-cyber</i>
5     (Lund: Bakhåll, 1998).
6     Alternative endings at:
7     <a href="http://www.novapress.se/cyber">
8     http://www.novapress.se/cyber
9     </a>&nbsp;
10    <a href="#55">Åter till texten</a>
11  </small>
12 </li>
```

Figure 1: HTML encoding of an endnote