Identifying Genres of Web Pages

Marina Santini

University of Brighton
marina.santini@itri.brighton.ac.uk

Résumé

Dans cet article nous présentons un modèle déductif-inductif pour l'identification des typologies textuelles et des genres dans les pages Web. Dans ce modèle, les typologies textuelles sont déduites en utilisant une forme modifiée du théorème de Bayes, tandis que les genres sont dérivés au moyen de simples règles « si-alors ». Étant donné que le système des genres sur le Web est complexe et que les pages Web sont plus imprévisibles et individualisées que les documents traditionnels, nous proposons cette approche déductive-inductive comme une alternative aux méthodes statistiques supervisées et non-supervisées. En effet, le modèle déductif-inductif permet une classification qui peut s'accommoder des genres non complètement standardisés. Il est aussi plus respectueux à l'égard de la vraie nature de la page Web, qui est en fait mixte et ne correspond presque jamais à un type idéal ou à un prototype précis, mais présente plutôt un mélange de genres, ou pas de genre du tout. L'évaluation de ce modèle reste un problème à résoudre.

Mots-clés : genre, typologies textuelles, pages Web, modèle déductif-inductif, identification automatique, théorème de Bayes.

Abstract

In this paper, we present an inferential model for text type and genre identification of Web pages, where text types are inferred using a modified form of Bayes' theorem, and genres are derived using a few simple *if-then* rules. As the genre system on the Web is a complex phenomenon, and Web pages are usually more unpredictable and individualized than paper documents, we propose this approach as an alternative to unsupervised and supervised techniques. The inferential model allows a classification that can accommodate genres that are not entirely standardized, and is more capable of reading a Web page, which is mixed, rarely corresponding to an ideal type and often showing a mixture of genres or no genre at all. A proper evaluation of such a model remains an open issue.

Keywords: genre, text types, Web pages, inferential model, automatic identification, Bayes' theorem.

1. Introduction

In this paper, we present a model for automatic text type and genre identification of Web pages based on an inferential approach, where text types are inferred using a modified form of Bayes' theorem, and genres are derived using a few simple *if-then* rules.

Automatic identification of text types and genres represents a great advantage in many fields because manual annotation is expensive and time-consuming. Apart from the benefits that it could bring to information retrieval, information extraction, digital libraries and so forth, automatic identification of text types and genres could be particularly useful for problems that natural language processing (NLP) is concerned with. For example, parsing accuracy could be increased if parsers were tested on different text types or genres, as certain constructions may occur only in certain types of texts. The same is true for Part-of-Speech (POS) tagging and word sense disambiguation. More accurate NLP tools could in turn be beneficial for

automatic genre identification, because many features used for this task are extracted from the output of taggers and parsers, such as POS frequencies and syntactic constructions.

The inferential model presented in this paper is an alternative to more traditional methods and has a two-fold motivation.

On the one hand, the Web is a complex reality and genre systems and genre repertoires on the Web inherit this complexity. The Web is a relatively new communication medium, not fully standardized, and still changing. It is populated by traditional genres, transplanted to the Web without much adaptation apart from the electronic form; adapted genres, *i.e.* genres showing signs of adaptation to the functionality of the Web; and novel genres, which show loose similarities or no similarities at all with paper genres. This partition is not only supported by the dynamics behind genre evolution suggested by genre analysts (*e.g.* Görlach, 2004), but also by some surveys (*e.g.* Crowston and Williams, 1997; Roussinov *et al.*, 2001). Also, the terms « Web genre » and « cybergenre » (Shepherd and Watters 1998, 2004) were specifically coined to refer to those genres created by the combination of the use of the computer and the Internet.

On the other hand, Web pages are a new kind of document, much more unpredictable and individualized than paper documents (Santini, 2006). While the linear organization of most paper documents is still reflected in traditional electronic corpora, such as the British National Corpus (BNC), Web pages have a visual organization that allows the inclusion of several functions or several different texts with different aims in a single document. The effect of hyperlinking (Crowston and Williams, 1999; Haas and Grams, 1998), interactivity and multi-functionality (Shepherd and Watters, 1999) can deeply affect the textuality[1] of Web pages. Web pages tend to be more mixed than traditional paper documents, bearing several communicative purposes at the same time. These different communicative purposes are represented by several textual snippets (navigational buttons, menus, ads, search boxes, etc.) that are visually dislocated in different areas of a single page.

In this scenario, we propose a classification model aiming at fulfilling two complementary requirements. First, the classification should accommodate genres that are not entirely standardized. Second, it should be more respectful of the actual nature of a text, which is mixed, rarely corresponding to an ideal type, and often showing a mixture of genres or no genre at all. Previous experiments with genres and Web pages showed that both unsupervised and supervised approaches have a number of drawbacks. On the one hand, the unsupervised approach (cluster analysis), used to highlight textual novelties detected on the Web (Santini, 2005a, 2005b), had shortcomings as for cluster stability, interpretation and evaluation. On the other hand, the supervised approach (discriminant analysis and machine learning) showed that classification accuracy tends to be good when documents belonging to a genre are selected having a genre exemplar in mind. In addition, a genre classification system based on a supervised approach is very static and conservative, because it represents a close-world situation, based on a relatively small number of examples and a limited number of genres. It does not have the capacity to deal with genres that are not entirely standardized or with genre-mixed Web pages, a contingency that is common in the current state of evolution of the Web.

Given these limitations, we suggest a more flexible model based on inference, which allows zero-to-multi-label classification. More specifically, in addition to the traditional single-label classification, a zero-label classification is useful when, for example, a Web page is so peculiar from a textual point of view that does not show any similarity with the genres

[1] By textuality we broadly refer to the way a text is written and organized.

included in the model. Instead, a multi-label classification is useful when Web pages show several genres at the same time. Both zero-labelled and multi-labelled Web pages might develop into new Web genres in future (for a broader discussion, cf. Santini, forthcoming). As there is no standard evaluation metrics for a comprehensive assessment of such a model, we defer to further research the evaluation of the model as a whole. In this paper, we report a partial evaluation based on the comparison of the single-label classification accuracy of the inferential model with the accuracies of standard classifiers.

From a theoretical point of view, the inferential model makes a clear-cut separation between the concepts of « text types » and « genres ». Text types are rhetorical/discourse patterns dictated by the purpose of a text. For example, when the purpose of a text producer is to narrate, the narration text type is used. On the contrary, genres are cultural objects created by a society or a community, characterised by a set of linguistic and non-linguistic conventions, which can be fulfilled, personalized, transgressed, colonized, etc., but that are nonetheless recognized by the members of the society and community that have created them, raising predictable expectations. For example, what we expect from a personal blog is diary-form narration of the self, where opinions and comments are freely expressed, and so on. The model presented here is capable of inferring text types from Web pages using a modified form of Bayes' theorem, and derive genres through *if-then* rules.

The paper is organized as follows: section 2 briefly summarizes previous work on inference and text types; section 3 describes the composition of the Web corpus used to build the model; section 4 illustrates the methodology and describes the modified version of Bayes' theorem on which the model is based on; section 5 presents results and a partial evaluation; finally in section 6 we draw conclusions and point out some open issues.

2. Previous Work

While previous work on genre classification of Web pages mainly rely on a supervised approach (Santini, forthcoming for a comprehensive review), in this section we would like to focus on the two main novelties proposed in this paper: the inferential approach and the relation between genres and text types.

Inference methodology. As far as we know, an inferential approach has never been applied be-fore for automatic genre identification. Instead, inferential approaches have been used extensively in Artificial Intelligence. In particular, the inferential model described in this paper, the form of Bayes' theorem called odds-likelihood or subjective Bayesian method, was suggested by Duda and Reboh (1984) to handle uncertainty in the rule-based system PROSPECTOR for classifying mineral exploration prospects. However, although we make use of this statistical model, the approach presented here has nothing of PROSPECTOR's complexity. PROSPECTOR is a sophisticated rule-based system and the inference is applied to assess the certainty or reliability of the rules (which could be modified interactively), while in our approach the inference is merely applied to the normalized frequencies of features, converted in terms of probabilities (see section 4.1). The main reason for choosing the odds-likelihood form of Bayes' theorem is that the model is very simple and odds allows more complex reasoning but keeps the simplicity of the Bayes' method.

Relation between genres and text types. The idea that genres favour a predominant text type or a combination of text types come from text linguistics, and in particular from Werlich (1976). In his *A Text Grammar of English*, Werlich creates a hierarchy of all the components constituting a text. In his view, text types (description, narration, argumentation, exposition, instruction) are « idealized norms » (Werlich, 1976: 39), while genres at different granularity

(« text forms » and « text form variants » according to his terminology) are « conventional manifestations of a text type » (Werlich, 1976: 46). For example, a « comment » is considered to be the dominant manifestation of subjective argumentation (*ibidem*). Although his *Text Grammar* is not corpus-based, it is rich of examples showing the relationship between text types and genres. More recently, the idea of linking text types to genres has been supported by genre analysts studying professional genres. For example, the leader or editorial is normally realized by expository or persuasive text type (Vastergaard, 2000: 102), even if this one-to-one relation can be disturbed by other factors. Other views on the relation between genre and text types can be found in Biber (1988) and Lee (2003).

3. The Web Corpus

As pointed out earlier, the Web is a changing reality. Currently, it is almost impossible to work out what is the composition of a representative corpus/sample of the Web as a whole (the multi-lingual Web), or only of a single language, such s English. There are estimates about the number of indexed Web pages (in April 2005 Google could search 8,058,044,651 Web pages; cf. Kilgarriff and Grefenstette, 2003 for previous estimates), which is a daily growing number, but we do not know anything about the proportions of the different types of text on the Web (cf. also Kilgarriff and Grefenstette, 2003).

The Web can be seen from many different angles, but from a textual point of view, it is a huge reservoir of documents. On the Web virtually everything can be seen as a « document » or better a « Web page », static or dynamic. From a statistical point of view, when the composition of population is unknown, the best solution is to extract a large random sample and make inference from that sample. But it is hard to work out how large a random sample should be. The solution that we suggest is to approximate one of the possible compositions of a random slice of the Web, statistically supported by reliable standard error measures. Following these guidelines, we built a Web corpus with the composition reported in table 1.

Genres	**Number of Web Pages**	**Proportions**
Random Web pages from the SPIRIT collection	1000	40.32 %
Blogs	200	8.065 %
Eshops	200	8.065 %
FAQs	200	8.065 %
Front pages	200	8.065 %
Listings	200	8.065 %
Personal Home Pages	200	8.065 %
Search Pages	200	8.065 %
BBC Editorials	20	0.806 %
BBC DIY mini-guides	20	0.806 %
BBC Short Biographies	20	0.806 %
BBC Features	20	0.806 %
Total	**2480**	**100 %**

Table 1. Web Corpus Composition

Four BBC Web genres (editorials, Do-It-Yourself (DIY) mini-guides, short biographies, and feature articles) and seven novel Web genres (blogs, eshops, FAQs, front pages, listings, personal home pages, and search pages) represent the known part of the Web, *i.e.* 59.86 % of

the sample. The SPIRIT collection (Joho and Sanderson, 2004) (unclassified Web pages) amounts to 40.32 % and represents the unknown part of the Web. The four BBC genres represent traditional genres adapted to the functionalities of the Web, while the seven novel Web genres are either unprecedented or showing a loose kinship with paper genres (for a description, see Santini, *Forthcoming*). Proportions are purely arbitrary and based on the assumption that at least half of Web users tend to use recognized genre patterns in order to attend felicitous communication. The Web pages included in the Web corpus were not annotated manually. In order to avoid the problem of manual annotation of raw Web pages, a task that can be difficult and controversial (cf. Santini, 2005a), Web pages were randomly downloaded from genre-specific archives or portals freely available on the Web.

4. Methodology

The inferential model tries to combine the advantages of both deductive and inductive approaches. It is deductive because the co-occurrence and the combination of features in text types is decided a priori by the linguist on the basis on previous studies, and not derived by a statistical procedure, which is too biased towards high frequencies (some linguistic phenomena can be rare, but they are nonetheless discriminating). It is also inductive because the inference process is corpus-based, which means that it is based on a pool of data used to predict some text types. A few handcrafted *if-then* rules combine the inferred text types with other traits (mainly layout and functionality tags) in order to suggest genres.

The four text types included in this implementation are: *descriptive_narrative, expository_informational, argumentative_persuasive,* and *instructional*. The selection of linguistic features (fully described in Santini, 2005d) for these text types come from previous (corpus-)linguistic studies (namely, Werlich 1976; Biber, 1988; Biber *et al.*, 1999). For each Web page the model returns the probability of belonging to the four text types. For example, a Web page can have 0.9 probabilities of being argumentative_persuasive, 0.7 of being instructional and so on. Probabilities are interpreted in terms of degree or gradation. For example, a Web page with 0.9 probabilities of being argumentative_persuasive shows a high gradation of argumentation. Gradations/probabilities are ranked for each Web page (Santini, 2005c).

The first hypothesis tested in this experimental setting is that the combination of two predominant text types, *i.e.* the top-ranked text types, is sufficient to derive BBC Web genres, more traditional in their textuality. The second hypothesis is that the combination of two predominant text types, *i.e.* the top-ranked text types plus a combination of additional traits (for example, layout or functionality tags) is sufficient to derive novel Web genres, which show a textuality more influenced by the interaction allowed by the Web. For both hypotheses, the individual Web page as a whole is taken as unit of analysis, without removing any textual component.

The computation of text types as intermediate step between features and genres is useful if we see genres as conventionalised and standardized cultural objects raising expectations. For example, what we expect from an editorial is an « opinion » or a « comment » by the editor, which represents, broadly speaking, the view of the newspaper or magazine. Opinions are a form of « argumentation ». Argumentation is a rhetorical pattern, or text type, expressed by a combination of linguistic features. If a document shows a high probability of being argumentative, *i.e.* it has a high gradation of argumentation, this document has a good chance of belonging to argumentative genres, such as editorials, sermons, pleadings, academic papers, etc. It has less chances of being a story, a biography, etc., which are narrative genres.

We suggest that the exploitation of this knowledge about the textuality of a Web page can lead to a more flexible classification model.

4.1. Inferring with Odds-Likelihood

The inferential model is based on a modified version of Bayes' theorem. This modified version uses a form of Bayes' theorem called *odds-likelihood* or *subjective Bayesian method* and is capable of solving more complex reasoning problems than the basic version. Odds is a number that tells us how much more likely one hypothesis is than the other. Odds and probabilities contain exactly the same information and are interconvertible. The main difference with original Bayes' theorem is that in the modified version much of the effort is devoted to weighing the contributions of different pieces of evidence in establishing the match with a hypothesis. These weights are confidence measures: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist (larger value means greater sufficiency), while LN is used when evidence is known NOT to exist (a smaller value means greater necessity). LS is typically a number > 1, and LN is typically a number < 1. Usually LS*LN=1. In this implementation of the model, LS and LN were set to 1.25 and 0.8 respectively, on the basis of previous studies and empirical adjustments. Future work will include more investigation on the tuning of these two parameters. Here is the list of steps performed by the model to infer text types:

- Feature extraction and normalization by document length of 2,480 Web pages.
- Conversion of the normalized frequencies into z-scores. Z-scores represent the deviation from the "norm" coming out from the Web corpus. The concept of "gradation" is based on these deviations from the norm.
- Conversion of z-scores into probabilities.
- Calculation of prior odds from prior probabilities of a text type. The prior probability for each text type was set to 0.25 (all text types were given an equal chance to appear in a Web page):
  `prOdds(H)=prProb(H)/1-prProb(H)`
- Calculation of multipliers (M) for the pieces of evidence (E):
  `if Prob (E)>=0.5 then: M(E)=1+(LS-1)(Prob(E)-0.5)/0.25`
  `if Prob (E)<0.5 then: M(E)=1-(1-LN)(0.5-Prob(E))/0.25`
- Calculation of a posteriori odds:
  `Odds(H)=PrOdds(H)*M(E1)*M(`E_n`)`
- Calculation of the probability of H from odds:
  `Prob(H)=Odds(H)/1+Odds(H)`

Once text types have been inferred, *if-then* rules are applied for determining genres. Example 1 is a simplified and readable example of *if-then* rules.

IF text_type_1 equals to narrative AND
text_type_2 equals to expository AND
functionality greater than 0.5 probabilities AND
layout greater than 0.5 probabilities [...]
THEN good_personal_home_page_candidate

Example 1. Example of an if-then rule to derive a genre

5. Evaluation of the Results

The output of the inferential model is rich of information and it is not only a list of correct and incorrect guesses. As no standard evaluation measures exits hitherto for a zero-to-multi-label classification scheme, we report the preliminary evaluation of single-label classification in order to have an idea of the effectiveness of the model.

5.1. Evaluation of BBC Web Genres

The first hypothesis tested with the inferential model says that a combination of predominant text types, tested with *if-then* rules (three rules per text types) is sufficient to derive the four BBC genres included in the Web corpus. Figure 1 shows the accuracy results of 20 BBC DIY mini-guides as outputted by the model. The first row indicates that 19 out of 20 DIY Web pages were considered to be good candidates for the DIY genre. The second row says that out of 20 DIY Web pages, none was considered to be a good candidate for the editorial genre. The third row suggests that one DIY Web page could be a good candidate for the short biography (bio) genre. Finally, the fourth row reports that six DIY Web pages have been judged good candidates for the feature genre. As you can see, the membership in different genres is not mutually exclusive, as a Web page can be a DIY mini-guide and bear some traits of the feature genre, together with traits of biography. But the current evaluation only confirms that it is true that the 19 Web pages can be objectively considered as DIY mini-guides.

```
1 BAD DIY;19 GOOD DIY; ********accuracy=95%  ********error rate=5%
20 BAD editorial; ********accuracy=100%  ********error rate=0%
19 BAD bio;1 GOOD bio; ********accuracy=95%  ********error rate=5%
14 BAD feature;6 GOOD feature; ********accuracy=70%  ********error rate=30%
```

Figure 1. Accuracy results for 20 BBC DIY mini-guides

Similarly, 15 BBC editorials out of 20, *i.e.* 75 %, were acknowledged as editorials by the model; 17 good guesses for short biographies, *i.e.* 85 %; finally, 12 features out of 20, *i.e.* 60 %, were classified correctly by the model. According to these results, the general hypothesis that there is a main combination of text types specific to genres is then confirmed. One of the advantages of this model is that automatic text type analysis and genre classification can be carried out on a very small sample, not suitable for machine learning (80 Web pages for four genres). Since this model is not entirely data-driven, it does not need a great amount of labelled data for working out genre patterns, as many of the traditional classification methods would require. This is because the co-occurrence of the features in a text type, and the combination of text types in *if-then* rules are decided once for all at the beginning by the analyst on the basis of previous studies. In this implementation of the model text types were built following Werlich (1976) and Biber (1988), and the combination of text types in *if-then* rules for BBC Web genres was decided following Werlich (1976:112-113) and Vastergaard (2000: 97-113). In this way, the inferential model integrates as much as possible of the qualitative studies on genre analysis in order to gain generality, since automatic feature selection methods are too corpus-dependent and hardly exportable to other corpora.

5.2. Evaluation of Seven Novel Web Genres

The second hypothesis says that the combination of the first two text types plus a combination of other traits, such as layout and functionality tags, tested with *if-then* rules (in average 6-7 rules per Web genre), is sufficient to derive the seven Web genres included in the Web corpus. For the seven Web genres we compare the classification accuracy of the inferential model with the accuracy of classifiers. Two standard classifiers – SVM and Naïve Bayes from Weka Machine Learning Workbench (Witten and Frank, 2005) – were run on the seven Web genres. The stratified cross-validated accuracy returned by these classifiers for one seed is ca. 89 % for SVM, and ca. 67 % for Naïve Bayes. The accuracy achieved by the inferential model is ca. 86 % (see table 2).

Web genres	SVM Classifier	Naïve Bayes Classifier	Inferential Model
Blogs	96 %	92 %	91 %
Eshops	88 %	76 %	83 %
FAQs	94.5 %	67 %	88.5 %
Front Pages	100 %	98 %	97 %
Listings	80 %	29 %	75.5 %
Pers. Home Pages	79 %	27 %	77 %
Search Pages	85 %	82 %	88 %
TOTAL	**ca. 89 %**	**ca. 67 %**	**ca. 86 %**

Table 2. Accuracies of standard classifiers and accuracy of the inferential model

An accuracy of 86 % is a good achievement for a first implementation, especially if we consider that the standard Naïve Bayes classifier returns an accuracy of about 67 %. But this is not all, because the inferential model returns much more than that. Figure 2 shows a snippet of the output: for each Web page, probabilities values for the four text types are returned (second, third, fourth and fifth column), together with a ranking of the two predominant text types (sixth column). A list of candidacies is also returned for each Web page (seventh, eighth, etc. columns).

File Name	Narrat.	Exp.	Argum.	Instr.	Ranked text types	BLOG	ESHOP	[...]
blog_02	0.80	0.57	0.71	0.69	narrative_descriptive1 + argumentative_persuasive 2	GOOD	BAD	[...]

Table 3. A snippet from the output of the inferential model

6. Conclusions

From a technical point of view, the inferential model presented here is a simple starting point for reflection about all the complex issues connected with the identification of genres of Web pages. As highlighted in the Introduction, the Web represents a challenge with its high level of peculiarity and hybridism of Web pages. The zero-to-multi-label classification model, based on text analysis, is proposed as an alternative to more traditional approaches because it

is capable of dealing with the textual complexity of Web pages. Even if parameters need a better tuning and text type and genre palettes need to be enlarged, it seems that this model is effective, as shown by the preliminary evaluation reported in sections 5.1 and 5.2, which was carried out taking solely the traditional single-label classification accuracy into account. A comprehensive evaluation of the model remains an open issue. Two main points require further and intensive discussions in the future: genre evaluation of highly hybrid or individualized Web pages not belonging to any specific genre (zero-label classification), and genre evaluation of Web pages belonging to multiple genres (multi-label classification).

References

BIBER D. (1988). *Variations across speech and writing*. Cambridge University Press, Cambridge.

BIBER D., JOHANSSON S., LEECH G., CONRAD S., FINEGAN E. (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.

CROWSTON K., WILLIAMS M. (1997). "Reproduced and Emergent Genres of Communication on the World-Wide Web". In *Proceedings 30th Hawaii Internat. Conference on System Sciences*.

CROWSTON K., WILLIAMS M. (1999). "The Effects of Linking on Genres of Web Documents". In *Proceedings 32nd Hawaii Internat. Conference on System Sciences*.

DUDA R., REBOH R. (1984). "AI and decision making: The PROSPECTOR experience". In W. Reitman (ed.), *Artificial Intelligence Applications for Business.* Norwood, NJ.

GÖRLACH M. (2004). *Text Types and the History of English.* Mouton de Gruyter, Berlin-NY.

HAAS S., GRAMS E. (1998). "Page and Link Classifications: Connecting Diverse Resources". In *Third ACM Conference on Digital Libraries*: 99-107.

JOHO H., SANDERSON M. (2004). "The SPIRIT collection: an overview of a large Web collection". In *SIGIR Forum* 38 (2).

KILGARRIFF A., GREFENSTETTE G. (2003). "Introduction to the Special Issue on the Web as a corpus". In *Computational Linguistics* 29 (3): 333-347.

LEE D. (2001). "Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle". In *Language Learning and Technology* 5: 37-72.

ROUSSINOV D., CROWSTON K., NILAN M., KWASNIK B., CAI J., LIU X. (2001). "Genre-Based Navigation on the Web". In *Proceedings 34th Hawaii Internat. Conference on System Sciences*.

SANTINI M. (2005a). "Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis". In *Proceedings CLUK 05*.

SANTINI M. (2005b). "Clustering Web Pages to Identify Emerging Textual Patterns". In *Proceedings TALN-RECITAL 2005*: 703-708.

SANTINI M. (2005c). "Automatic Text Analysis: Gradations of Text Types in Web Pages". In *Proceedings10th ESSLLI Student Session*: 276-285.

SANTINI M. (2005d). "Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features". In *Technical Report ITRI-05-02.* University of Brighton.

SANTINI M. (2006). "Web pages, text types, and linguistic features: Some issues". In *ICAME Journal* 30.

SANTINI M. (forthcoming). *Automatic Identification of Genres in Web Pages*. PhD thesis, University of Brighton.

SHEPHERD M., WATTERS C. (1998). "The Evolution of Cybergenre". In *Proceedings 31st Hawaii Internat. Conference on System Sciences*.

SHEPHERD M., WATTERS C. (1999). "The Functionality Attribute of Cybergenres". In *Proceedings 32nd Hawaii Internat. Conference on System Sciences*.

VASTERGAARD T. (2000). "That's not News: Persuasive and Expository Genres in the Press". In A. Trosborg (ed.), *Analysing Professional Genres.* John Benjamins, Amsterdam-Philadelphia.

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

WITTEN I., FRANK E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Publishers, Amsterdam.