"Yes, user!": compiling a corpus according to what the user wants

Rachel Aires[1,2]
Diana Santos[2]
Sandra Aluísio[1]
[1]NILC
University of São Paulo, Brazil
[2]Linguateca
SINTEF ICT, Norway
raires@icmc.usp.br
diana.santos@sintef.no
sandra@icmc.usp.br

Abstract

This paper describes a corpus of webpages, named "Yes, user!". These pages were classified in order to satisfy different types of users' needs. We introduce the assumptions on which the corpus is based, show its classification scheme in detail, and describe the process used to build this corpus. We also present the results of a questionnaire inquiring about the general clarity and understanding of our classification and those proposed by other researchers. We describe both the corpus and a metasearch prototype which was built with those classifiers and make it accessible for other researchers to use.

1 Introduction

In today's world, users are faced with information explosion: there is too much information available to process, in particular, information available on the World Wide Web (WWW, referred to as "the Web"). The field of Information Retrieval (IR) is focusing on different ways of organizing information, including descriptions of a) what a particular text is about, b) how it is written and c) why.

In order to explain how texts are written, several researchers have proposed the use of style categorizations and quality information. However, there has been no prior work that focuses on why, namely, trying to separate webpages according to their goal. Based on a thorough qualitative analysis of the logs of TodoBr, a major Brazilian search engine (www.todobr.com.br), and inspired by Broder's (2002) work, we selected seven types of webpages, classified according to users' needs who were trying to find:

1) Definitions of objects or learning how or why something happens;
2) How to do something;
3) Comprehensive presentations or surveys about a given topic;
4) News on a specific subject;
5) Information about a specific person, company or organization;
6) A specific webpage with prior visits;
7) URLs of specific online services.

Although there exists some correlation between these classes above and textual genres (such as scientific, informative, instructional), our interest is to focus on "goals" rather than genres. One of the motivation for this is that Web texts, because of medium, may have in common properties hitherto uncovered, such as those derived from interactiveness, attempts to expand user/reader involvement, specialized design features, hypertextuality and multimedia content. Thus, the traditional genre distinctions may be somehow obfuscated by all these other details. Conversely, people who focus on “Web genre” would not be able to distinguish among different features related to different users’ goals.

Instead of using text categorizations that have indirectly to do with user's goal (see Section 3 for an overview of those), we wanted to study whether “page purpose” was something tangible and therefore whether it was possible to capture it and use it to automatically separate the pages.

As part of this research we built a corpus of 500 texts (640,000 words) classified according to the users’ needs and organized based on the previously introduced seven types as mutually exclusive. We were able to show that it was possible to reliably discriminate among pages that provide information and those that offer services (with a success rate of 90.95%) in Aires et al. (2004).

Although these results were encouraging, there were two outstanding concerns:

1. Was it easy to understand those seven users’ needs and thus becoming useful in a practical setting?
2. Was it wise to postulate mutually exclusiveness among the classes?

In order to respond to these concerns, we reclassified the original corpus from scratch, adding new texts and relaxing the common classification prohibition. Special care was put into documenting the decisions made with regards to inclusion/exclusion of specific Web pages, and with aims to extending and replicating them in the future. This process is described in Section 3. In addition, we tested other types of classification for the same texts, in order to compare our results with those of indirect approaches (Section 2). We also created a set of special interest binary corpora, i.e., corpora classified by different users according to their particular interests.

Specifically, to address the issue of the seven users’ needs, we developed questionnaires for potential users which check their understanding of our classification schemes and compared our schemes with others proposed in the literature. Our findings are shown in Section 4.

Section 5 describes "Yes, user!", the newly, enlarged corpus. It also presents the methods used for developing the classifiers that underlie a prototype for a desktop metasearcher, named Leva-e-traz: it classifies resulting pages into different classes.

2 Overview of related work

2.1 Text categorization

Various methods for text categorization have been proposed and carried out for different purposes. Furthermore, a distinction has been made between text classification and text categorization: the

former describes a response to an arbitrary query (such as text retrieval), while the latter describes "the assignment of texts to one or more of a pre-existing set of categories" (Lewis, 1991: 312). However, in practice, this distinction has not been carefully followed. In fact, Jackson & Moulinier (2002: 119) began by presenting precisely the opposite definition: 'Text categorization' refers to sorting documents by content, while 'text classification' is used to include any type of documents' assignments into classes". Later, they drop the distinction altogether and use the two terms interchangeably. Lewis himself later rephrased the two definitions, turning text categorization into a type of text classification: "*Text classification* algorithms assign texts to predefined classes. When those classes are of interest to only one user, we often refer to text classification as filtering or routing. When the classes are of interest to a population of users, we instead refer to *text categorization*." (Genkin et al., 2004:3).

We use text categorization to specify "the assignment of documents to predefined categories". These may be content categories (such as a topic detection task, Yang & Liu (1999)) or stylistic categories, as those proposed by Karlgren (2000) and Stamatatos et al. (2000).

Our work was originally inspired by Karlgren's (2000) studies on stylistic relevance for IR. He states that "texts have several interesting characteristics beyond topic" and investigates whether "stylistic information, [distinguishable using simple language engineering methods,] can be used to improve information retrieval systems". While we may share the same goal, our domain language is Portuguese, with the Brazilian Web as our source of texts and user's behaviour. We have extended the evaluation of the hypotheses reported by Karlgren, as well as adapted one of the genre classifications proposed by Stamatatos et al. (2000).

Our work has been influenced by Biber's (1988) attempts to go beyond genre as an unanalysed primitive and finding a set of textual dimensions, through the principal components analysis method. Biber (1995) has claimed that it was possible to adapt the underlying ideas to other languages besides English, and this is somehow validated by our work. Similar approaches, using easy-to-compute linguistic features, have been used or advogated by other researchers in many areas, such as the detection of stylistic inconsistencies in collaborative writing (Glover & Hirst, 1996), machine translation, computer aided teaching (Nilsson & Borin, 2002) and gender studies.

There are many studies in the machine learning (ML) community on text categorization (Sebastiani, 2002), our work being just one example. We performed supervised learning by feeding a ML program with human-encoded category labels and a set of features per text, and our program inductively learned to classify related texts (see Witten and Frank (2000) for the technology used). Our contribution is the introduction of an original set of categories.

2.2 User adaptation

Amilcare, an Information Extraction (IE) system was developed to allow for several types of personalization based on user expertise and willingness. It integrates machine learning techniques when extracting information from the Web (Ciravegna & Wilks, 2003). While Amilcare was intended as an annotation helper for the Semantic Web it shares some concepts with our work. Future versions of Leva-e-traz could be used to extract specific types of information as opposed to entire webpages.

2.3 Web corpora for text categorization

Compiling corpora from the Web following genre or text types has been the focus of other researchers, after Kessler et al.'s (1997) seminal work in automatic genre detection having named the Web as the most powerful reason to start studying style (again).

Karlgren (2000) mentions a training corpus of 1358 pages obtained by running TREC queries in Web search engines and selecting the top ten hits. Additional pages were obtained from history files collected from colleagues. 184 pages were classified as Error messages, which, as Karlgren himself notes, were not a very probable genre a user would choose for result.

Stamatatos et al. (2000) created a "genre-based corpus" with ten categories and 25 full texts per category. This Modern Greek corpus was created from scratch, using on-line newspapers, magazines and radio stations, and institutional pages, and it was used to train their text categorizer.

Pekar et al. (2004) compiled a corpus with a specific service for classifying webpages and e-mail messages from conferences, jobs, resources and trash, consisting of about 400 pages.

Hahn & Wermter (2004) distinguish between medical and non-medical documents, and attempt a fine-grained categorization within the medical ones, e.g., distinguishing between surgery vs. pathology reports. They compiled test and training sets of German documents. The test set included 270 medical documents and 232 non-medical (from newswire material), the training one included about 22Mb medical and 19Mb non-medical.

Lee and Myaeng (2002) collected 7,828 Korean documents and 7,615 English documents manually from more than twenty portals. Each document was analysed by at least two people, and inserted into one of seven categories (reportage, editorial, research articles, reviews, homepages Q&A and Spec).

We believe that none of these various corpora have been made freely available for further training or use by other researchers.

3 Re-creating our corpus

In this section we present the revised instructions and some statistics about the corpus. Let us stress that in the final corpus the category assignment of each text was checked by two different persons, in order to secure some consistency.

3.1 Guidelines for assigning texts to categories

In order to reclassify the webpages included in our first corpus, we devised these strategies:

- Choose only pages written in Brazilian Portuguese
- Look at all the contents of the page, not just the titles or whatever appears in larger font or highlighted

- Ignore text genre, what matters is to satisfy the users' needs
- Ignore quality or quantity matters: if a page has little information to satisfy type A and a lot of information to satisfy type B, it should be categorized as type AB.
- Classify only the current page, ignoring its links to other pages which may satisfy other types of users' needs. For instance, if a page has a link specifying: "got to the subscription page", this page itself does not provide a subscription service. But if the page has a link indicating "Download" to download something, it is of the "service" type.

In addition, the following instructions were provided for each type of users' need, together with plenty of examples and counterexamples (which we obviously omit here):

User need 1: *A definition of something or to learn how or why something happens. For example, what are the northern lights? To satisfy this need, the best results would be found in dictionaries and encyclopaedias, or textbooks, technical articles and reports and texts of the informative genre.*

A page responds to need 1 when it defines what something is, it explains what something is or why something happens. It is not important whether it describes one subject or many. That is, as long as the page describes how something happens or defines something, it is classified as Type 1. For instance, a page explaining how life began, even though lacking formal definitions, should be considered satisfying needs of Type 1.

User need 2: *To learn how to do something or how something is usually done. For example, find a recipe of a favourite cake, learn how to make gift boxes, or how to install Linux on a computer. Typical results are texts of the instructional genre, such as manuals, textbooks, recipes and possibly some technical articles or reports.*

A page responds to need 2 if it teaches/explains how to do something (for instance, by providing instructions) or it explains how something is or was done.

User need 3: *A comprehensive presentation or survey about a given topic, such as a panorama of 20th century American literature. In this case, the best results would be found in texts of the instructional, informative and scientific genres, e.g. textbooks, area reports and long articles.*

A page responds to need 3 if it provides a description/gathering/panorama of a specific subject. The text could be classified as responding to need 3:

- If, independently of the main topic, there is a description/gathering/panorama of a specific subject. For instance, the topic could be a specific book (a review that discusses its impact on the marketplace, the thorough research carried out in its writing, a description of the topics introduced, information about the writer) or the literature in the 20th century.
- Independently of the size of the text. A description/gathering/panorama on a specific subject can be long or short. If it provides additional information, beyond those addressed in types 1 (what it is, why/how it happens), 2 (how to do something or how it is done), 4 (news provider), it corresponds to type 3.
- Independently of text genre/type. Note that we did not include only news reports or newspapers as Type 3. We assume that any text genre/type could satisfy any one of the needs. For instance, a magazine editorial describing different aspects of literature is classified as type 3, while an interview of a writer published in a newspaper may provide

information about the person and an overview of related literature, and should be classified as Type 35.

Of course an overview would ideally have an extensive coverage, but we cannot include this as a criterion, for two reasons: deciding whether it is extensive depends crucially on the judge's prior knowledge about the subject matter; and because we are not attempting to make quality judgements, just judgements about what the user is looking for.

Guides about a place, a country or an activity, like tourist guides, should also be classified as Type 3.

User need 4: *To read news about a specific subject. For example, what is the current news about the situation in Israel, what were the results of the soccer game on the previous day or to find about a terrible crime that has just occurred in the neighbourhood. The best answers in this case would be found in texts of the informative genre, e.g. news in newspapers and magazines.*

Instructions: A page responds to need 4 if it contains news, independently of the subject described in the news. For instance, pages that provide news about something that happened to a famous person (gossip) are news. If the news is about the release of a new book, even when the text style makes it appear as a review, it is considered news. Conversely, not all pieces that appear in newspapers or magazines are considered type 4. A page such as providing advice to undergraduate students, even if it published as news for young people, is not type 4.

User need 5: *To find information about a person, a company or an organization. For example, the user wants to know more about his/her blind date or to find the contact information of a person she met at a conference. Typical answers here are found in personal, corporation and institutional webpages.*

A page responds to need 5 if it provides information about a person, a company or an institution or organization. Examples are personal homepages, pages with contact information (such as a resume, telephone, and address), company/organization pages (e.g., this ONG was founded in … with the purpose of …)

Biographies are examples of type 5 because they provide information about a specific person. Special care was taken in the case of biographies, in order to verify whether the data about the person also included panorama or descriptions. For instance, biographies some times also include a description of a specific past time frame. In that case, the page would have to be classified as also responding to Type 3.

It is irrelevant whether a page contains a short history or whether it presents a lot of information, if it provides information about people or companies it is Type 5. Information about a group of people, such as research groups or rock bands, is also considered Type 5.

Some pages include an "empty description" with no content at the beginning, a sort of presentation as to what the page itself contains, and not a description about a store or person. This empty (or self-referential) description was not considered as providing any type of information that could be classified in one of our seven user needs.

User need 7: *To find URLs where he can have access to a specific online service. For example, s/he wants to buy new clothes or to download a new version of software. The best answer to this kind of request is found in commercial pages (companies or individuals offering products or services).*

Instructions: A page responds to need 7 if it offers online service(s) or is a service provider. For instance, the postal service page offers the possibility of tracking a package; there are online services that provide software downloads, and stores that sell their products online.

The service provided must be done by that page. Various types of pages were not considered as satisfying type 7. Among these were (1) pages that only publicize a service but do not provide access to it, (2) pages with simple lists such as one with a list of lyrics to a song, and (3) in-site services such as specialized search tools within a site, or contact forms

3.2 Category distribution in the resulting product

Our original corpus size was based on a heuristic based on population size. The data should include five times as many texts as linguistic features in order to be examined within a factor analysis (Gorsuch (1983: 332, apud Biber 1988: 65). For our seven categories, we used a corpus of 511 texts extracted from the Web, 73 for each type of need (except for type 6, which can fit into any category) plus additional 73 texts that would not fit into any of those six types: the "others" category.

Table 3 shows the number of texts and words in each category, for the original corpus (OC) and the current corpus, Yes, User! (YU). Note that by selecting the same number of texts for each type, our results showed considerable differences in the number of words in the corpus. This difference was not considered a problem because the training cases used are texts, not their specific words.

After improving and augmenting the original corpus we arrived at a total number of words of 1,801,962, and 1,703 texts. In Table 3, and due to the fact that we have texts which are classified as satisfying more than one category (see Table 4), we provide the total number of words used for each category (in words YU), and the number of word pertaining to texts classified only in pure categories. None of these values, if added, would give us the total number of words in the corpus.

Type	**1**	**2**	**3**	**4**	**5**	**7**	**others**	**total**
Words OC	76,841	51,959	19,6450	39,533	67,601	39,951	168,295	640,630
Texts OC	73	73	73	73	73	73	73	511
Words YU	752,882	571,625	983,542	452,674	273,074	270,085	137,563	
Words2 YU	61,068	72,303	149,387	87,030	65,852	53,639	137,563	
Texts YU	704	547	859	473	308	397	76	1,703

Table 3: Corpus size per type of users' need: OC – original corpus; YU – Yes, user! The same text in YU if classified in two categories is counted twice, the same occurs for its words.

Table 4 describes the number of additional categories found besides the 7 used. We found pages that could be classified as 12, 13, 14, 23, 24, 27, 34, 35, 37, 45, 123, 134, 135, 137 and 1237.

User need(s)	**Texts**	**Words**
T1	78	61,068
T12	77	102,008

T123	77	122,003
T1237	80	69,999
T13	72	134,954
T134	80	89,999
T135	79	64,656
T137	79	51,925
T14	80	56,270
T2	77	72,303
T23	76	88,724
T24	79	72,791
T27	79	44,497
T3	77	149,387
T34	75	81,785
T35	79	77,767
T37	80	52,943
T4	75	87,030
T45	79	64,799
T5	69	65,852
T7	77	53,639
Others	76	137,563
Total	1,703	1,801,962

Table 4: Number of texts with common classifications in Yes, User!

It is important to emphasize that both corpora were used to train classifiers with a set of shallow parsing features (inspired by Biber's work (1988)) and lexical general-content words, thus comparing several machine learning techniques, as reported in Aires et al. (2005) and Aires (forthcoming).

4 Questionnaires about general applicability of the classification schemes

In order to determine the needs of potential users, we developed a questionnaire given to undergraduate students in Computer Science, Linguistics, Medicine, and to graduate Photography students, available from http://nilc.icmc.usp.br/nilc/download/AiresQuestUtilClass.pdf.

Basically, we asked them about the

1. Clarity of the seven users' needs: how clear are they?
2. Genre classification scheme(s): which schemes did they find easy to use?
3. Time: would they spend one day collecting text samples in order to generate a corpus that would target to commonly troubling tasks?

Sixty three students answered the questionnaire and 41 of those reported encountering problems in their Web searches.

The questionnaire had four sections inquiring about:

1. Personal data
2. Search experience information. This included (i) the place in which it was carried out (home, job, university...); (ii) the purpose or tasks in which the search was carried out; (iii)

their opinion and experience with regards to problems and errors in their searches; (iv) their goal when doing searches specifically for their work; (v) their willingness to invest an entire day in order to personalize a search system; (vi) specific examples of such systems

3. Description of the seven needs in our model, asking the subjects to apply them to seven different cases. They were also asked questions whether they found them useful, any doubts they had and their suggestions (such as merging or adding new types)
4. Text genre: three schemes were shown (see Figure 2) in a table in parallel (Lácio-web genres (Aluísio et al., 2003), Stamatas et al.'s (2000) classification grid and Karlgren's genre palette (2000:103)). The users were asked to mark the names of the categories that they found unclear, and to judge whether they found each scheme to be helpful. In addition, a list of "text types" was provided, and the users were required to mark whether they were clear and whether they were useful, and assess this new classification scheme in terms of helpfulness; the users were prompted to suggest new genres (e. g., contracts or chronicles). At the end, the users were asked which classification scheme(s) they would consider most useful and easy to use, among the seven types, the three genre and the "types of text".

Figure 2 shows how the different genre schemes were shown. Subsequently the users were asked whether they considered them useful in their searches. Lácio-Web's genre palette was translated into "text types", shown via examples. For instance, the "instructional" genre was changed to "text book, culinary recipe, course notes" and so on. This was done to determine whether the answers differed significantly when changing their labels (the classes were the same as Genre 1, with the minor adjustment that we did not consider relevant to distinguish among literary subgenre). If such were the case, one might conclude that the terminology was opaque, but the reasoning behind it – and the classes obtained – remained sound.

Genre 1	**Genre 2**	**Genre 3**
Scientific	Press editorial	1 Informal, Private
Law	Press reportage	2 Public, Commercial
Technical Management	Academic	3 Searchable indices
Reference	Of.cial documents	4 Journalistic materials
Instructional	Literature	5 Reports
Informative	Recipes	6 Other running text
Prose (fiction)	Curricula vitae	7 FAQs
Drama	Interviews	8 Link Collections
Poetry	Planned speeches	9 Listings, Tables
	Broadcast news	10 Discussions
		11 Error Messages

Text Types

Paper, dissertation, technical report, ...
Law, legislation, sentence, ...
Letter, memo, manual, CV, ...
Encyclopaedia, dictionary, lexicon, ...
Text book, culinary recipe, course notes, ...
News report, editorial, ...
Biography, short story, novel, poem, play...

Figure 2: Three genre schemes presented in the questionnaire (shown in parallel) followed by the text types list

Twelve subjects did not consider any of the genres schemes to be useful while only three indicated lack of interest in the "text types".

Table 5 shows the number of respondents that considered each categorization not useful, negative answers:

Classification schema	No. of subjects
Seven users' needs	2
Text types	3
Genre 1	8
Genre 2	12
Genre 3	13

Table 5. How many subjects considered the classification not useful

Table 6 shows the results of those classification(s) found to be most useful:

Classification schema	No. of subjects
Seven users' needs	25
Text types	19
Genre 1	15
Genre 2	13
Genre 3	6

Table 6. How many subjects considered the classification easiest and most useful

The data shows that there was a majority of subjects who preferred our classification scheme and/or found it to be useful. In practice, we will confirm these results by carrying out user-oriented evaluations of a prototype based on our hypothesis that categorization according to user needs helps Web search.

An overwhelming majority of respondents believed that there was much to be gained by using an a priori webpage classification. This classification being a set of users' needs, text genres or even a more personalised one (e.g., binary). All 41 users who reported having frequent problems in their searches indicated their willingness to spend a day creating personalised schemes. This confirms the feasibility of providing an option for users to create their own training corpora, with relevant and irrelevant pages.

5 A corpus of Brazilian Portuguese webpages classified according to users' goal

"Yes, user!" is a corpus of 1,703 Brazilian Portuguese webpages classified following general goal/users' needs. The selection of the texts in this corpus was done by five different people. They were allowed to include websites already familiar to them. They saved only the text of the pages, which could be HTML, pdf or doc (converted to plain text). After careful examination of 1,760 webpages initially selected, some were removed because they were duplicates, pages written in other variants of Portuguese, resulting in the final 1,703 pages, multiply categorized.

The corpus is currently available, in text format at www.linguateca.pt/Repositorio/yesuser.html, together with specific information about pages and collection dates. We also provide additional files with training sets, for researchers interested in applying other machine learning algorithms to them.

We also make available a set of binary domain-specific corpora of 200 pages each (100 positive and 100 negative), developed independently by different users. These users willingly tried out the corpus collection task in order to evaluate later on the classifiers produced by the training with a ML algorithm. We presently have a corpus with legal texts, divided into legal texts for lawyers and those for the general public, but we intend to have an open-ended "Yes, user!" with additional binary corpora as soon as they are compiled and integrated into Leva-e-traz.

Figure 3 shows a prototype of the desktop metasearcher Leva-e-traz. Web results can be seen categorized by the several different classifications – genres, text types and users' needs. A specific classifier has been built, for each one of them. The first experiments are described in Aires et al. (2004), more advanced ones in Aires et al. (2005) and Aires (forthcoming).

Figure 3. Main screen of Leva-e-Traz, a personalized desktop metasearcher

Acknowledgements

This work was partially supported by grant POSI/PLP/43931/2001 from Fundação para a Ciência e Tecnologia (Portugal), co-financed by POSI.

References

Aires, Rachel (forthcoming) O uso de características lingüísticas para a apresentação dos resultados de busca na Web de acordo com a intenção da busca do usuário – uma instanciação para o português. PhD Dissertation, Computer Science Department (ICMC), University of São Paulo.

Aires, Rachel, Aline Manfrin, Sandra Maria Aluísio and Diana Santos (2004a) What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of LREC'2004, Fourth International Conference on Language Resources and Evaluation*, (Lisboa, 26-28 May 2004), 1943-1946.

Aires, Rachel, Aline Manfrin, Sandra Maria Aluísio and Diana Santos (2004b) Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs? Technical report, NILC-TR-04-09. Available on-line at http://www.nilc.icmc.usp.br/nilc/download/airesetaltr0409.pdf (accessed June 13th, 2005)

Aires, Rachel, Sandra Aluísio and Diana Santos (2005) User-aware page classification in a search engine, in *Proceedings of Stylistic Analysis Of Text For Information Access, SIGIR 2005 Workshop* (Salvador, Bahia, Brazil, August 19, 2005), to appear.

Aluísio, Sandra M., G. Pinheiro, Marcelo Finger, M. Graças Volpe Nunes and Stella E. Tagnin (2003) The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. *Proceedings of Corpus Linguistics (2003)*, Lancaster, UK. v. 16, 14-21.

Berglund, Ylva & Oliver Mason (2003) "But this formula doesn't mean anything...!?": Some reflections on parameters of texts and their significance, in A. Wilson, P. Rayson & T. McEnery (eds.), *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech* (Frankfurt: Peter Lang), http://www.english.bham.ac.uk/staff/omason/publications/cl2001/berglund-mason.html

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

Biber, Douglas (1995) *Dimensions of Register Variation* (Cambridge: Cambridge University Press).

Broder, Andrei (2002) A Taxonomy of Web Search. *SIGIR Forum* 36 (2), Fall 2002, 3-10.

Ciravegna, Fabio & Yorick Wilks (2003) Designing Adaptive Information Extraction for the Semantic Web in Amilcare, in S. Handschuh and S. Staab (eds), *Annotation for the Semantic Web*, (Amsterdam: IOS Press). Available on-line at http://www.dcs.shef.ac.uk/~fabio/paperi/AmilcareAnnotation.pdf. (accessed June 13th, 2005)

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1-47.

Genkin, Alexander, David D. Lewis and David Madigan (2004) Large-Scale Bayesian Logistic Regression for Text Categorization. Available on-line at http://www.stat.rutgers.edu/~madigan/PAPERS/shortFat-v13.pdf. (accessed June 13th, 2005)

Glover, Angela & Graeme Hirst (1996) Detecting stylistic inconsistencies in collaborative writing, in Mike Sharples and Thea van der Geest (eds.), *The new writing environment: Writers at work in a world of technology* (London: Springer-Verlag), 147-168.

Hahn, Udo and Joachim Wermter (2004) Pumping Documents Through a Domain and Genre Classification Pipeline, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of LREC'2004, Fourth International Conference on Language Resources and Evaluation* (Lisboa, 26-28 May 2004), 735-738.

Jackson, Peter & Isabelle Moulinier (2002) *Natural Language processing for Online Applications: Text Retrieval, Extraction and Categorization* (Amsterdam/Philadelphia: John Benjamins).

Karlgren, Jussi (2000) *Stylistic Experiments for Information Retrieval*, Doctoral dissertation, Department of Linguistics, Stockholm University, 2000 [Swedish Institute of Computer Science, SICS Dissertation Series 26]. Available on-line at http://www.sics.se/~jussi/Artiklar/2000_PhD/. (accessed June 13th, 2005)

Kessler, Brett, Geoffrey Nunberg and Hinrich Schütze (1997) Automatic Detection of Text Genre, in Proceedings of the 35th annual meeting on Association for Computational Linguistics (Morristown, NJ, USA: ACL), 32-38.

Lee, Yong-Bae and Sung Hyon Myaeng (2002) Text Genre Classification with Genre-Revealing and Subject-Revealing Features, in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA: ACM Press), 145-150.

Lewis, David D. (1991) Evaluating text categorization, in *Proceedings of Speech and Natural Language Workshop*, February 1991 (San Mateo, CA: Morgan Kaufmann), 312-318.

Nilsson, Kristina and Lars Borin (2002) Living off the land: The Web as a source of practice texts for learners of less prevalent language, in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), (ELRA), 411-418.

Pekar, Viktor, Richard Evans and Ruslan Mitkov (2004) Categorizing Web Pages as a Preprocessing Step for Information Extraction, in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of LREC'2004, Fourth International Conference on Language Resources and Evaluation* (Lisbon, 26-28 May 2004), 723-726.

Stamatatos, Efstathios, Nikos Fakotakis and George Kokkinakis (2000) Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics* **26**(4), December 2000, 471-495.

Witten, I.H.& E. Frank (2000) *Data Mining: Practical machine learning tools with Java implementations* (San Francisco: Morgan Kaufmann).

Yang, Y. & X. Liu (1999) A re-examination of text categorization methods, in *Proc. 22th ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR'99)* (Berkeley, CA, 1999), 42-49.