Assembling a Balanced Corpus from the Internet

Johan Dewe, Telia Research,
Jussi Karlgren, SICS, and
Ivan Bretan, Telia Research

Address for correspondence:
Jussi Karlgren, SICS, Box 1263, 164 29 Kista, Sweden
Fax: + 46 8 751 72 30
Jussi.Karlgren@sics.se

Balanced Corpora for Textual Research

For empirically oriented textual research it is crucial to have materials available for extraction of statistics, training probabilistic algorithms, and testing hypotheses about language and language processing in general. In recent years, the awareness that text is not just text, but that texts comes in several forms, has spread from more theoretical and literary subfields of linguistics to the more practically oriented information retrieval and natural language processing fields. As a consequence, several test collections available for research explicitly attempt to cover many or most well-established textual *genres*, or *functional styles* in well-balanced proportions (Francis and Kucera, 1982; Källgren, 1990).

The creation of such a collection is a complex matter in several respects. Our research area is to build retrieval tools for the Internet, and thus, for our purposes, the choice of genres to include is one of the more central problems: there is no well-established genre palette for Internet materials. To find materials to experiment with, we need to create them in a form suitable for our purposes. This is a double edged problem, involving both vaguely expressed user expectations and establishing categories using large numbers of features which taken singly have low predictive and explanatory power. This paper gives an outline of the methodology we use for determining which genres to include.

Stylistic Variation and Genre

Texts exhibit considerable variation. While the variation in topic or content is quite obvious and the basis for most categorization enterprises in information retrieval research variation in style is as noticeable, and forms a second basis for categorization: poetry, prose, non-fiction, reference materials, and so forth are all stylistic categories or genres.

Stylistic variation shows through *stylistic items*: observable choices of linguistic items. Stylistic items can be observed on any level of linguistic abstraction: lexical, for the choice between words of similar meaning but different connotations; syntactic, for the choice between equivalent constructions with different communicative import; textual, for decisions of textual organization.

Each stylistic item is of little import, but taken together they are indicative of systematic differences. A set of documents with a perceived consistent tendency to make the same stylistic choices is called a genre or, specifically, if it has an established communicative function, a functional style (see e.g. Enkvist, 1973; Vachek, 1975).

Stylistic variation between genres or language varieties can be detected reliably using a large battery of quite simple stylistic items such as pronoun counts or relative frequencies of certain types of constructions such as agentless passives (Biber, 1988, 1989; Karlgren and Cutting, 1994), utilized for authorship determination by simple calculations of average word length distributions (Mendenhall, 1887), and with some success predictively for information retrieval (Karlgren, 1996; Karlgren and Straszheim, 1997; Stralkowski et al, 1996).

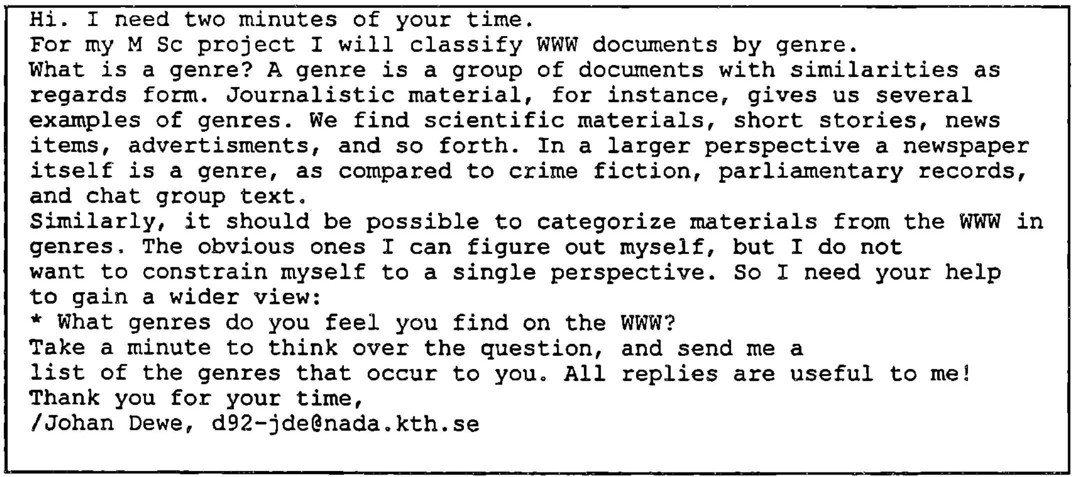
Establishing Genres

Method

In previous similar studies, we have used introspective methods: we have established genres mainly based on personal experience (Ben Cheikh and Zackrisson, 1994; Hussain and Tzikas, 1995). Other text collections organized by genre, genre is largely equated with *source*. Texts from some organization are categorized together with texts from similar organizations, without regard for text usage: e.g journalistic press archives, personal letters, technical documentation. (e.g. Källgren, 1990). For this study, we wished to have a better foundation for our genre palette. Our basic souce of knowledge is interviewing users about their perceptions of what types of material they find and interact with online. We collate the impressions and try to define genres that are both reasonably consistent with what users expect and observable and conveniently computable using measures of stylistic variation as outlined in the previous section. Cf. Figure 1, see last page of this paper.

Questionnaire

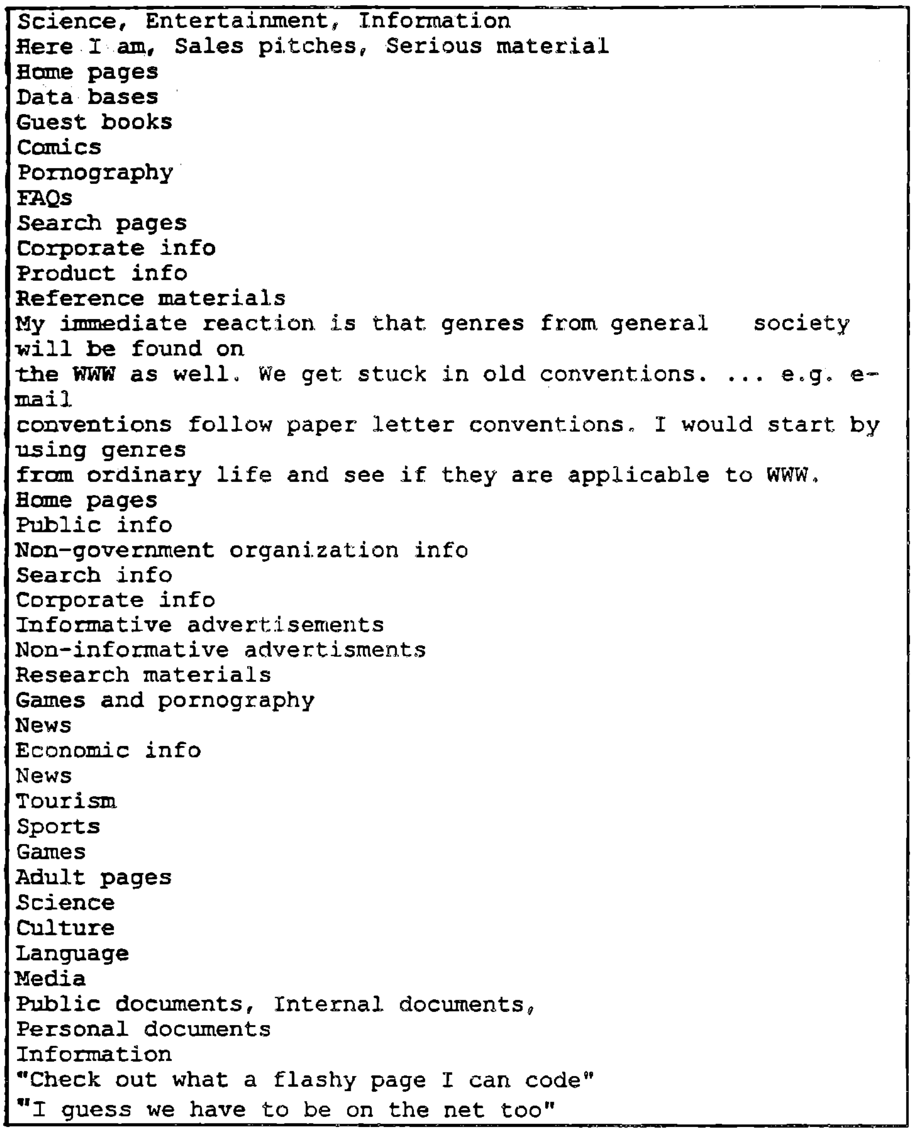
The questionnaire in Figure 2 was sent to 648 computer users - students, researchers, and teachers at Stockholm University and the Royal Institute of Technology. We received 7 error messages and 67 responses, which gives a response rate of 10 per cent.



```
Hi. I need two minutes of your time.
For my M Sc project I will classify WWW documents by genre.
What is a genre? A genre is a group of documents with similarities as
regards form. Journalistic material, for instance, gives us several
examples of genres. We find scientific materials, short stories, news
items, advertisments, and so forth. In a larger perspective a newspaper
itself is a genre, as compared to crime fiction, parliamentary records,
and chat group text.
Similarly, it should be possible to categorize materials from the WWW in
genres. The obvious ones I can figure out myself, but I do not
want to constrain myself to a single perspective. So I need your help
to gain a wider view:
* What genres do you feel you find on the WWW?
Take a minute to think over the question, and send me a
list of the genres that occur to you. All replies are useful to me!
Thank you for your time,
/Johan Dewe, d92-jde@nada.kth.se
```

Figure 2. The genre questionnaire (This is an English translation. The Swedish original can be found at http://www.stacken.kth.se/~dewe/dropjaw/enkat.txt)

Compiling the results



```
Science, Entertainment, Information
Here I am, Sales pitches, Serious material
Home pages
Data bases
Guest books
Comics
Pornography
FAQs
Search pages
Corporate info
Product info
Reference materials
My immediate reaction is that genres from general   society
will be found on
the WWW as well. We get stuck in old conventions. ... e.g. e-
mail
conventions follow paper letter conventions. I would start by
using genres
from ordinary life and see if they are applicable to WWW.
Home pages
Public info
Non-government organization info
Search info
Corporate info
Informative advertisements
Non-informative advertisments
Research materials
Games and pornography
News
Economic info
News
Tourism
Sports
Games
Adult pages
Science
Culture
Language
Media
Public documents, Internal documents,
Personal documents
Information
"Check out what a flashy page I can code"
"I guess we have to be on the net too"
```

Figure 3. Some translated excerpts from the answers to the questionnaire. (The answers in their entirety can be found athttp://www.stacken.kth.se/~dewe/dropjaw/enkatsvar.txt).

The answers ranged from very short to extensive discussions - some examples are shown in Figure 3. It was very clear to us from that most readers conflated genre and form on the one hand with content and topic on the other: "tourism", "sports", "games", "adult pages". This is not surprising. Genre and topic are not independent dimensions of variation, and a typical library categorization reflects both dimensions simultaneously. Several respondents did give examples of more cleanly form-oriented genres as well: "home pages", "data bases", "FAQs", "search pages", "reference materials". Some respondents gave explicit references to *paper genres* - one lengthy quote is given among the examples in Figure 3. The *intention* of the information provider showed up as a genre formation criterion in several responses: "here I am", "sales pitches", "serious material"; or, as an alternative formulation of the same criterion, the type of author:

"commercial info", "public info", "non-governmental organization info". Some responses explicitly brought up *quality*: "boring home pages" and text *ecology* or intended environment: "public documents", "internal documents", "personal documents".

We have attempted to systematize some of the user perceived distinctions, namely those that are predictable enough to be modeled with simple metrics, in the genre palette shown in Figure 4.

- Informal, Private
  - Personal home pages.
- Public, commercial
  - Home pages for the general public.
- Searchable indices
  - Pages with feed-back: customer dialogue; searchable indexes.
- Journalistic materials
  - Press: news, reportage, editorials, reviews, popular reporting, e-zines.
- Reports
  - Scientific, legal, and public materials; formal text.
- Other running text
- FAQs
- Link Collections
- Other listings and tables
- Asynchronous multi-party correspondence
  - Contributions to discussions, requests, comments; Usenet News materials.
- Error Messages

Figure 4. The current genre palette.

When trying to assign textual materials to the various categories automatically we expect to find that some genres are not as useful as they may seem at first sight; we will find that some of these categories may have to be adjusted - merged, split, or redefined - as the collection is evaluated using statistical methods. The categories shown in Figure 4 are starting points for research, not final results.

Finding Samples

We use three methods to collect data from the World Wide Web.

Firstly, we take queries used for the Text Retrieval Conference (Harman, 1996) (TREC queries nos. 251-300; fields "topic" and "description") and run them through Altavista, a search service on the Internet. We use the top ten hits for each query to retrieve about 500 documents.

Secondly, we take sixty queries from Magellan, another search service on the Internet. Magellan provides a "voyeur page" (http://mckinley.voyeur.com/voyeur.cgi#voyeur?1) which displays real user queries in real time. We run the sixty queries through Magellan, and similarly obtain about 600 documents.

Thirdly we use history files from local Netscape users to retrieve about 700 additional documents.

URL source	TREC	Magellan	History	Total
	via	Voyeur	List	
	Altavista			
01 Informal, Private	11	67	50	128
02 Public, Commercial	23	87	87	197
03 Searchable indices	4	14	55	73
04 Journalistic materials	50	28	16	94
05 Reports	106	5	2	113
06 Other running text	73	49	38	160
07 FAQs	0	4	8	12
08 Link Collections	31	50	67	148
09 Listings, Tables	17	138	70	225
10 Discussions	16	0	8	24
11 Error Messages	55	36	93	184
Total	386	478	494	1358

Figure 5. The current composition of the corpus.

Evaluating the choice of genres

To evaluate the genre palette we sent out the list of genres we settled on to the same recipients we originally solicited the genre distinctions from, with a question if they understood what the genres represented and if any obvious genre was missing. We received 102 responses. Most respondents claimed to understand what type of text our genre labels were intended to cover, and while most categories got some comments of one form or another, most comments were caused by our giving too few examples of what the genres were intended to cover. Most comments concerned the category "Interactive pages". Many respondents were annoyed by the fact that the category was not of the same type as the other types. Some respondents objected or did not understand the labels -- e.g. "FAQ" or "Listings, tables" or "Error messages"; many asked for a download page or ftp database category; some wondered about the all-inclusiveness of "Other running text"; several asked for a specific category for "Search engines"; several suggested more content based genres.

Many pointed out that some of the categories were less suitable for search in that they did not imagine themselves ever searching for "Error messages" or "Interactive pages" specifically. Several respondents pointed out that the categories were not mutually exclusive. In summary, the most central objections were either such that would be remedied in an interactive situation where examples are readily available, or requests for more flexible genre assignment.

Recognizing genres automatically

The genre palette, besides being intuitively understandable, needs to be workable for automatic analysis. We calculate a quite large number of textual features for each individual text and work them together for a categorization decision using a machine learning algorithm. The pioneering work by Douglas Biber(1988, 1989) on computational corpus-based stylistics has been descriptive rather than predictive, aiming to find distinctions between different registers or varieties of spoken and written language. It has made use of large numbers of stylistic features collected from previous, non-computational work and weighing them together using standard methods from multivariate statistics. We use this work as a basis for ours. Most of Biber's features we use here are rather lexical in nature, for ease of processing: the relative frequency of certain classes of words such as personal pronouns, emphatic expressions, or downtoning expressions, for instance. We add more general textual and genre specific features: relative number of digits, or average word length, for instance (Karlgren, 1996; Karlgren and Straszheim, 1997). Others yet are vectored specifically to the Internet material we have been using for experimentation: number of images or number of HREF links in the document, for instance. We normalize the measurements by mean and standard deviation, and combine them -- 40 of them, at present -- into simple if-then categorization rules using C4.5, a non-parametric categorization tool (Quinlan, 1993).

```
If- there are more "because" than average,
- longer words than average,
- type-token ratio is above average,
then
- the object is of class Textual
with
- a certainty of 90.0%.
```

Figure 6. An example classification rule.

We have a few dozen rules to categorize texts into one of the eleven genres defined in the above sections. The genres partition into two major hypercategories: textual (04, 05, 06, 07, 10) and non-textual (01, 02, 03, 08, 09, 11); each of them in turn splits to one of five or six sub-categories. These splits are of varying quality: the first does quite well, something like a ninety per cent success rate, while the subsplits make the wrong choice somewhere between once in three or four times. With additional features and a better defined genre palette results will improve. However, to get really useful results the categorization should not be exclusive. Every object should potentially be of several genres.

Conclusions

Internet users have a vague sense of genres among the documents they retrieve and read. The impressions users have of genre can be elicited and to some extent formalized enough for genre collection. The names of genres should be judiciously chosen to be on an appropriate level of abstraction so that mismatches will not faze readers.

References

1. Douglas Biber. 1988. Variation across speech and writing. Cambridge University Press.

2. Douglas Biber. 1989. "A typology of English texts", Linguistics, 27:3-43.

3. Naoufel Ben Cheikh and Magnus Zackrisson. 1994. "Genrekategorisering av text för filtrering av elektroniska meddelanden" (Genre Classification of Texts for Filtering of Electronic Messages) Stockholm University Bachelor's thesis in Computer and Systems Sciences, Stockholm University.

4. Nils Erik Enkvist. 1973. Linguistic Stylistics. The Hague: Mouton.

5. Donna Harman (ed.). 1996. The Fourth Text REtrieval Conference (TREC-4). National Institute of Standards Special Publication 500-236. Washington.

6. Fahima Polly Hussain and Ioannis Tzikas. 1995. "Ordstatistisk kategorisering av text för filtrering av elektroniska meddelanden" (Genre Classification of Texts by Word Occurrence Statistics for Filtering of Electronic Messages) Stockholm University Bachelor's thesis in Computer and Systems Sciences, Stockholm University.

7. Jussi Karlgren. 1996. "Stylistic Variation in an Information Retrieval Experiment" In Proceedings NeMLaP 2, Bilkent, September 1996. Ankara: Bilkent University. (In the Computation and Language E-Print Archive: cmp-lg/9608003).

8. Jussi Karlgren and Douglass Cutting. 1994. "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis", Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), Kyoto. (In the Computation and Language E-Print Archive: cmp-lg/9410008).

9. Jussi Karlgren and Troy Straszheim. 1997. "Visualizing Stylistic Variation." In the Proceedings of the 30th HICSS, Maui.

10. W. N. Francis and F. Kucera. 1982. Frequency Analysis of English Usage, Houghton Mifflin.

11. Gunnel Källgren. 1990. The First Million is Hardest to Get: Corpus Tagging. Proceedings of the 13th International Conference on Computational Linguistics (COLING-90) Hans Karlgren (ed.), Helsinki.

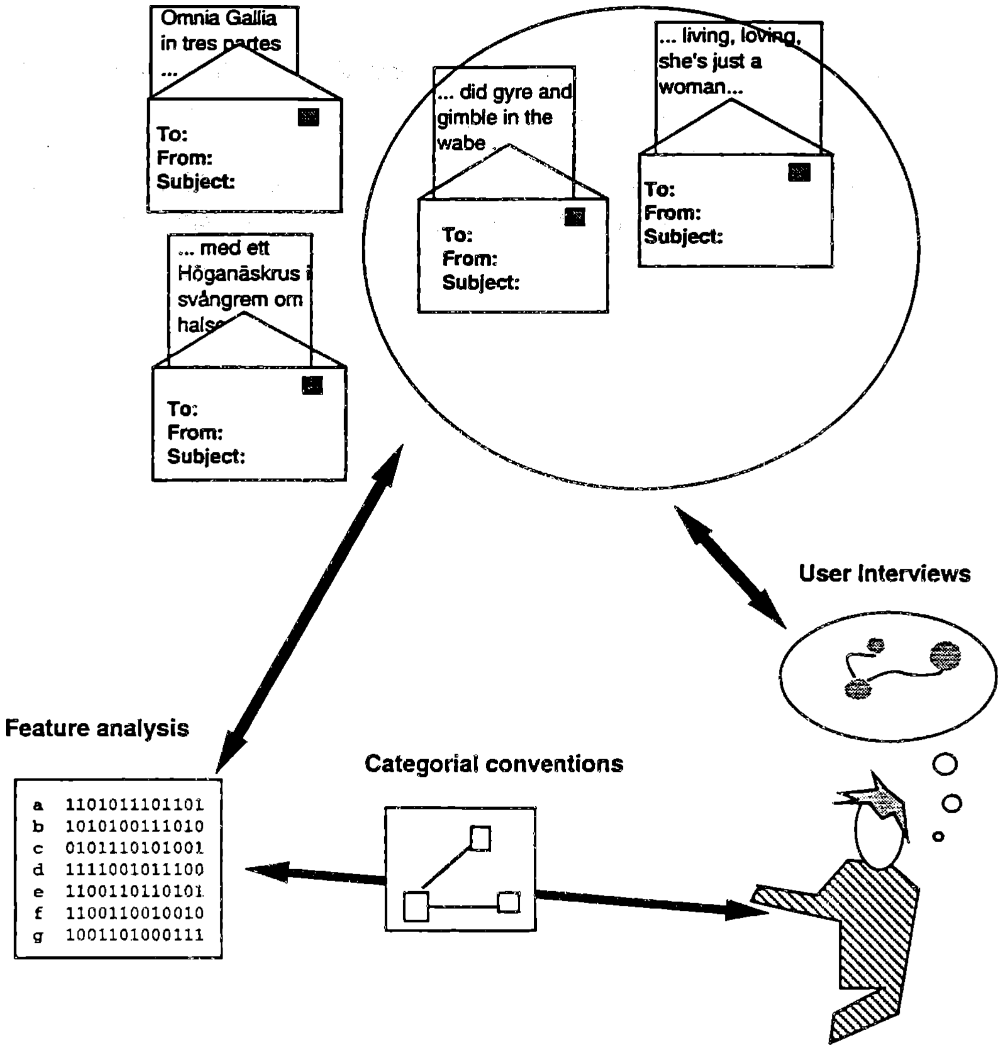
12. T.C. Mendenhall. 1887. "The Characteristic Curves of Composition." Science 9: 237-49.

13. J. Ross Quinlan. 1993. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

14. Tomek Strzalkowski, Louise Guthrie, Jussi Karlgren, Jim Leistensnider, Fang Lin, Jose Perez-Carballo, Troy Straszheim, Jin Wang, Jon Wilding. 1996. "Natural Language Information

Retrieval: TREC-5 Report" Proceedings of The Fifth Text REtrieval Conference (TREC-5). Donna Harman (ed.). National Institute of Standards Special Publication. Washington.

15. Josef Vachek. 1975. "Some remarks on functional dialects of standard languages". In Styleand Text - Studies presented to Nils Erik Enkvist. Håkan Ringbom. (ed.) Stockholm: Skriptor and Turku: Åbo Akademi.



a	1101011101101
b	1010100111010
c	0101110101001
d	1111001011100
e	1100110110101
f	1100110010010
g	1001101000111

Figure 1. A snapshot of the methodology shows the interplay between vaguely expressed user expectations and observable and conveniently computable categories.