
Automatic Text Analysis: Gradations of Text Types in Web Pages

MARINA SANTINI

University of Brighton

Lewes Rd, Brighton, UK

Marina.Santini@itri.brighton.ac.uk

ABSTRACT. In this paper we present a first implementation of a linguistic expert system, which combines Bayesian inference and the knowledge of a domain expert, the linguist. The system is capable of automatically deriving the gradations of four text types -- *descriptive/narrative*, *explicatory/informational*, *argumentative/persuasive*, *instructional* -- from web pages. The evaluation of the system against human assessment shows that the system is effective and promising.

1 Introduction

In this paper we present a first implementation of a computational model for automatic text analysis. This model is capable of automatically deriving the gradations of the different text types from web pages. It is a simple but effective *linguistic expert system*, which combines Bayesian inference and the knowledge of a domain expert, the linguist. The linguist decides the co-occurrence and the weight of the features relevant to a text type, following previous linguistic and rhetorical studies; the inference is based on the evidence found in a collection of web pages. As no benchmark exists for evaluating such a system, two tests comparing the judgment of the system with the judgment of human subjects have been set up. This prototype includes only a restricted number of widely acknowledged text types: *descriptive/narrative*, *explicatory/informational*, *argumentative/persuasive*, *instructional*.

Automatic Text Analysis: Gradations of Text Types in Web Pages

By text types we refer to rhetorical/discoursal/linguistic patterns, which convey the purpose or the function of a text¹. For instance, a text can be written to instruct, explain, describe, narrate, persuade, support an argument, and so on. These purposes or functions are consciously or unconsciously enacted by text producers, and identified by text receivers. The identification of text types is deeply rooted in our culture (cf. Faigley and Meyer 1983), and it is considered to be a basic skill, crucial in all activities involving text production and comprehension. It is so much so that the BBC has set up a website where web users can learn how to identify different types of text (descriptive, informative, persuasive, instructive)². This is not an isolated case, for example the UK Adult Literacy website also focuses on the importance of identifying the purpose of a text³. Furthermore, many universities have online writing labs where students are taught how to deal with different types of texts, from argumentation to definition, description and so forth⁴.

A number of text types are universally recognized and rely on standardized linguistic devices. For example, instructions are often delivered as sequences of paragraphs containing imperatives, conditions, a purpose, as in *If the radiator is still not hot, try turning it off and wait to see whether it heats up.* An argumentative text usually makes use of logical connectives and concessive clauses, as in *Although the official policy was to help reconciliation, people did not trust the measures.* However, a single text rarely contains only one purpose or function, i.e. a single text type. Most texts are mixed. For example, a DIY guide is predominantly instructional, but it can also contain descriptions of tools, together with tips and suggestions. The mixture of different purposes or functions is often exacerbated on the Web. Web pages are mostly multi-purpose or multi-functional documents. Thanks to the hypertextual structure, which gives greater freedom, a single web page can contain separate sections, each having a its own communicative function, namely a specific text type.

However, detecting the gradations of text types is only an intermediate step. The final version of the system presented here is geared towards automatic genre identification of web pages. Genres are conventionalized types of documents bearing a socio-cultural connotation and showing standardized rhetorical/discoursal/linguistic patterns. One important aspect of genres is that

¹ "Text type" is an ambiguous term. In this paper, we follow the rhetorical and corpus-linguistic tradition. Some scholars use the label "text types" to indicate instrumental or practical genres, as opposed to literary genres (e.g. Görlach, 2004). Others use "text types" and "genres" interchangeably, as synonyms (e.g. Stubbs, 1996).

² The BBC Skillwise, *Types of text* website is at: <http://www.bbc.co.uk/skillswise/words/reading/typesoftext/index.shtml>

³ For example, see <http://members.aol.com/twittwoo/grpdfs/purpose.pdf>

⁴ For instance, see www.ucalgary.ca/applied_history/write/Contenteg.html, <http://owl.english.purdue.edu/>, and <http://www.longleaf.net/ggrow/modes.html>

they raise expectations. For instance, in a personal home page we expect to find a narration of the 'self', a description of interests, hobbies, etc. We suggest that knowing the gradations of different text types within a single web page is important not only for text analysis itself, but also because it helps identify genres. In fact, a genre often shows a predominant text type together with other minority text types. The combination of different text types and their gradations might contribute to the identification of a specific genre. For example, an online tutorial is predominantly instructional and to a lesser extent descriptive, while a tourist guide is mainly descriptive and to a lower extent instructional. New or different gradations of text types might signal novel genres.

The paper is organized as follows: in section 2 two previous approaches to text type analysis are compared; section 3 describes the linguistic expert system and includes the evaluation of the results; in section 4 conclusions are drawn and a number of future tasks are suggested.

2 Two Approaches to Text Type Analysis

Nearly all classifications of texts by purpose or function derive directly or indirectly from Aristotle's Rhetoric. However, there is no agreed palette of text types. The range of text types suggested by different linguists depends on the goal of the text typology and the approach with which this typology has been worked out (for an overview, see Lee 2001). The system presented in this paper was inspired by two authors with opposite stances: Egon Werlich and Douglas Biber.

Werlich belongs to the tradition of German text linguistics which has not filtered into the English linguistic tradition. Germany experienced a boom of research in text typology in the 70s. Unfortunately most of these studies remain untranslated into English. Werlich is unusual in this respect because he wrote a *Text Grammar of English* in English (Werlich, 1976). His approach is neither corpus-based nor quantitative; he follows the descriptive tradition based on selected examples extracted from exemplar texts. Werlich's five text types are very intuitive and reflect cognitive processes: perception in space (*description*), description in time (*narration*), comprehension of general concepts (*exposition*), creation of relations among concepts through the extraction of similarities, contrasts etc. (*argumentation*); planning of future behaviour (*instruction*). He has the merit of methodically listing many linguistic and textual features that interact and co-occur in each single text type. His text analysis is based on presuppositions about a well-formed text in the adult's mind. We call his approach *deductive*, because the interpretation of the actual text depends on premises about cognitive processes.

Biber has instead an *inductive* approach. His text typology (Biber, 1988 and 1989) is based on his working corpus (the LOB corpus plus the London-Lund

Corpus of Spoken English). He thoroughly reviews previous (socio-)linguistic studies and selects 67 linguistic features that can be interpreted functionally. For example, nominalizations co-occurring with passives convey abstract information (Biber 1988: 227). His basic assumption is simple and powerful: if certain features co-occur consistently, then it is reasonable to think that they share an underlying function that encourages their use. In this way the functions of a text, instantiated in text types, are not posited on an *a priori* basis; rather they are required to account for the observed co-occurrence patterns among linguistic features. He uses factor analysis to identify these co-occurring patterns. His approach was novel and suggestive, but, leaving aside the statistical doubts that factor analysis might raise, it has two main weaknesses. First, his text typology is too corpus-dependent. For instance, the instructional text type is not included in his typology because no instructional genres were included in his working corpus (cf. Biber, 1988: 67). Second, some of the text type labels remain opaque, such as “situated reportage” or “intimate interpersonal interaction” (Biber, 1989). They are indeed interesting interpretations of the linguistic phenomena found in a corpus, but they are more similar to learned language analysis than intuitive types of text.

3 A “Linguistic Expert System”

The computational model presented here is a *linguistic expert system* which is based on an intuitive text typology, four sets of co-occurring features related to four text types, and computationally tractable features. This system tries to combine the advantages of both deductive and inductive approaches. It is deductive because it starts from a limited number of widely acknowledged text types. Moreover, the co-occurrence of features in each 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. In fact, like all expert systems, it contains the knowledge of experts (linguists) along with an inference procedure based on a large pool of data used to predict some events (text types of web pages). Inferences are based on the calculation of the probability value for a hypothesis (a text type) given one or more pieces of evidence (the frequencies of some features).

This expert system is based on the technique introduced in PROSPECTOR (Duda and Reboh, 1984). PROSPECTOR uses a form of Bayes’ theorem called “odds-likelihood” to work out the probability of an event happening based on prior probability and new evidence. Odds and probabilities contain exactly the same information, and are interconvertible. Odds, also called likelihood ratio, is a number that tells us how much more likely one hypothesis is than the other. Two

confidence measures are used in this approach: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist, while LN is used when evidence is known NOT to exist. These two measures are weights which can be tuned according the data and the expert's knowledge.

3.1 Four Intuitive Text Types

In this first implementation of the system the aim is to analyze web pages according to four intuitive text types: *descriptive/narrative*, *explicatory/informational*, *argumentative/persuasive* and *instructional*. These text types are coarse conglomerates, but they can be easily refined in future.

The descriptive/narrative type includes features that relate primarily to phenomena in space and in time. In narration these two aspects are tightly related (for example, in a news story), while in description the time dimension can be omitted (for example, in a technical description). The descriptive/narrative text type conflates the descriptive type of text from the BBC Skillwise website (see footnote 2), the descriptive and narrative types from Werlich (1976), and imaginative narrative from Biber (1989).

The explicatory/informational text type includes features related to nouns. A high frequency of noun indicates high density of information because nouns bear most of the referential meaning in a text. This type conflates the informative type of text from the *BBC Skillwise* website, the explicatory type from Werlich (1976), and informational interaction from Biber (1989).

The argumentative/persuasive text type includes features that relate to logical reasoning and emotional appeal. It conflates the persuasive type of text from the *BBC Skillwise* website, argumentation from Werlich (1976), and involved persuasion from Biber (1989)

The instructional text type includes features related to activities and verbs (especially imperatives and necessity modals). It conflates the instructive type of text from the *BBC Skillwise* website, and instruction from Werlich (1976).

Web pages contain many other text types, not only these four. However, we start identifying the gradations of this restricted but intuitive text typology to test whether the system reliably matches human judgment in this task.

3.2 Web Pages

The web corpus used in the system is the SPIRIT collection, which is a random crawl of the Web carried out in 2001 by a Canadian university (Clarke et al. 2002). It represents a genuine unclassified slice of the real Web, "frozen" in a collection for research purposes. 1000 web pages written in English were extracted from this collection and included in the system. Since the SPIRIT collection does not bear any text type classification, a different source had to be

Automatic Text Analysis: Gradations of Text Types in Web Pages

used for evaluation. The evaluation set is made of 78 web pages downloaded primarily from websites suggested by the *BBC Skillswise* website (tutor page⁵) as sources of additional examples. The evaluation set includes 20 BBC articles selected as representatives of the explicatory/informative text type⁶; 20 BBC pages about historic figures chosen as representatives of the narrative/descriptive text type⁷; 20 BBC DIY pages downloaded as representatives of the instructional text type⁸; 18 between leaders and comments downloaded from *The Guardian* website as representatives of the argumentative/persuasive text type⁹. The final working corpus contains a total of 1078 web pages.

3.3 Features

More than 100 features were employed for the four text types. Two different types of features were expressly devised: linguistic facets and syntactic patterns. Both can be interpreted in a functional way, and both are parser-dependent.

Linguistic facets are word-based, and are represented by lexical items or functional tags. A facet shows an aspect of a text. For example, the facet "communication verbs" is widely used in reported speech, while the facet "nominals" includes nouns, adjectives, appositions, prepositional phrases, etc. and highlights the density of information in a text (cf. also Biber 1988: 104).

Syntactic patterns include a limited number of subclause patterns and six main clause patterns. The set of subclause patterns includes clauses that give a hint about the type of a text, such as "verbless temporal clause with until in final position", represented by the pattern IMP[ERATIVE] OBJ *UNTIL* ADJ (as in *Beat the mixture until fluffy*), which is common in instructional texts (Quirk et al. 1985: 1079. Main clause patterns were suggested by Werlich (1976: 216 ff.) to give account for the referential context in a text. Werlich suggested six phenomenon sentences for his five text types.

3.4 Methodology

The system performs the following steps:

1. Feature extraction and normalization by document length of 1078 web pages.

⁵ <http://www.bbc.co.uk/skillswise/words/reading/typesoftext/tutor.shtml>, under the section *BBCi Text sources*.

⁶ <http://www.bbc.co.uk/science/hottopics/>

⁷ http://www.bbc.co.uk/history/historic_figures/

⁸ <http://www.bbc.co.uk/gardening/basics/techniques/archive.shtml> and http://www.bbc.co.uk/homes/diy/diy_guide/

⁹ <http://observer.guardian.co.uk/comment/0,6903,156041,00.html> and <http://observer.guardian.co.uk/leaders/0,6903,00.html>

2. Conversion of the normalized frequencies into z-scores. Z-scores represent the deviation from the “norm” coming out from the working corpus. The concept of gradation is based on these deviations from the norm.
3. Conversion of z-scores into probabilities.
4. 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):
$$\text{prOdds(H)} = \text{prProb(H)} / 1 - \text{prProb(H)}$$
5. Calculation of multipliers (M) for the pieces of evidence (E). LS and LN were empirically set to 1.25 and 0.8, respectively:
if $\text{Prob(E)} \geq 1$:
$$\text{M(E)} = 1 + (\text{LS} - 1)(\text{Prob(E)} - 0.5) / 0.25$$

if $\text{Prob(E)} \leq 1$:
$$\text{M(E)} = 1 - (1 - \text{LN})(0.5 - \text{Prob(E)}) / 0.25$$
6. Calculation of a posteriori odds:
$$\text{Odds(H)} = \text{PrOdds(H)} * \text{M(E}_1) * \text{M(E}_n)$$
7. Calculation of the probability of H from odds:
$$\text{Prob(H)} = \text{Odds(H)} / 1 + \text{Odds(H)}$$

It is important to stress that in this system, the contributing features for each text type are decided by the linguist on the basis of previous studies. Therefore, the multipliers at step 5 are computed only for pre-determined sets of features, a specific set of features per text type. Finally, the system returns a summary table containing a list of web pages together with the gradations of the four text types.

4 Evaluation of the Results

The gradations of text types for the evaluation set are reported in Figures 1, 2, 3 and 4. The highlighted cells under the headings *narrative*, *inform(ational)*, *argum(entative)* and *instruct(ional)* show the predominant text type for a web page. Gradations are currently returned in terms of probability values. For example, the first file in the table below shows a predominant gradation of narrative type, a lower gradation of informational type, an even lower degree of instructional type and the argumentative type is definitely minor.

Automatic Text Analysis: Gradations of Text Types in Web Pages

file_name	narrative	inform	argum	instruct
bbc_eval_infor_01	0.8394	0.7280	0.4525	0.6230
bbc_eval_infor_02	0.0864	0.5702	0.4050	0.5234
bbc_eval_infor_03	0.8734	0.0985	0.1704	0.2517
bbc_eval_infor_04	0.0203	0.8675	0.4931	0.3369
bbc_eval_infor_05	0.5167	0.8898	0.8025	0.5752
bbc_eval_infor_06	0.1899	0.4051	0.1910	0.6262
bbc_eval_infor_07	0.6122	0.7895	0.5729	0.3967
bbc_eval_infor_08	0.8899	0.7722	0.8375	0.9223
bbc_eval_infor_09	0.2578	0.5217	0.3201	0.4094
bbc_eval_infor_10	0.4070	0.8486	0.3012	0.7493
bbc_eval_infor_11	0.0599	0.3297	0.2759	0.4460
bbc_eval_infor_12	0.8146	0.8984	0.9112	0.5851
bbc_eval_infor_13	0.8227	0.9735	0.5634	0.2971
bbc_eval_infor_14	1.0000	0.9853	0.9883	0.6791
bbc_eval_infor_15	0.9810	0.9889	0.9278	0.6116
bbc_eval_infor_16	0.1082	0.3917	0.2281	0.5296
bbc_eval_infor_17	0.5275	0.7930	0.7047	0.2831
bbc_eval_infor_18	0.6450	0.6605	0.5534	0.4698
bbc_eval_infor_19	0.3878	0.5949	0.3440	0.8528
bbc_eval_infor_20	0.2120	0.9306	0.2206	0.5660
TOTAL	3	11	1	5

Table1. Informative web pages

file_name	narrative	inform	argum	instruct
bbc_eval_instr_01	0.7340	0.5705	0.5757	0.9770
bbc_eval_instr_02	0.9352	0.7129	0.3396	0.9796
bbc_eval_instr_03	0.0591	0.8221	0.9454	0.9992
bbc_eval_instr_04	0.9052	0.7460	0.8550	0.9794
bbc_eval_instr_05	0.2661	0.9106	0.9449	0.9955
bbc_eval_instr_06	0.5543	0.6453	0.7809	0.9553
bbc_eval_instr_07	0.7203	0.9836	0.9968	0.9986
bbc_eval_instr_08	0.0518	0.1169	0.1362	0.9976
bbc_eval_instr_09	0.0875	0.0941	0.3855	0.9998
bbc_eval_instr_10	0.6811	0.6345	0.4201	0.9995
bbc_eval_instr_11	0.6460	0.9965	0.9995	0.9831
bbc_eval_instr_12	1.0000	0.9987	0.9276	0.9944
bbc_eval_instr_13	0.9992	0.6410	0.6534	0.9946
bbc_eval_instr_14	0.9082	0.9506	0.4682	0.9755
bbc_eval_instr_15	0.9763	0.7738	0.3913	0.9942
bbc_eval_instr_16	0.5792	0.9331	0.3655	0.9942
bbc_eval_instr_17	0.9994	0.9867	0.9309	0.9676
bbc_eval_instr_18	0.6504	0.9761	0.9981	0.9035
bbc_eval_instr_19	0.8272	0.7041	0.1779	0.9957
bbc_eval_instr_20	0.9994	0.9594	0.8246	0.9316
TOTAL	4	11	2	14

Table2. Instructional web pages

file_name	narrative	inform	argum	instruct
bbc_eval_narra_01	0.6705	0.9974	0.9635	0.6807
bbc_eval_narra_02	0.7914	0.9944	0.8534	0.2817
bbc_eval_narra_03	0.9303	0.9845	0.2210	0.4018
bbc_eval_narra_04	0.8160	0.9918	0.5155	0.3254
bbc_eval_narra_05	0.9604	0.9516	0.5908	0.7797
bbc_eval_narra_06	0.9976	0.9999	0.9998	0.7050
bbc_eval_narra_07	0.9999	0.9836	0.7936	0.8836
bbc_eval_narra_08	0.0633	0.9858	0.5944	0.2722
bbc_eval_narra_09	0.0027	0.9804	0.0568	0.1018
bbc_eval_narra_10	0.9998	0.8678	0.5300	0.7185
bbc_eval_narra_11	0.999867	0.999866	0.9969	0.8435
bbc_eval_narra_12	0.9998	1.0000	0.9996	0.7888
bbc_eval_narra_13	0.9986	0.9777	0.3887	0.7224
bbc_eval_narra_14	0.1349	0.9746	0.3608	0.3712
bbc_eval_narra_15	0.9964	0.9059	0.5768	0.2561
bbc_eval_narra_16	0.0243	0.9805	0.7081	0.3772
bbc_eval_narra_17	0.9947	0.9997	0.9918	0.5064
bbc_eval_narra_18	0.9997	0.9998	0.9991	0.7016
bbc_eval_narra_19	0.9907	0.9945	0.5973	0.4691
bbc_eval_narra_20	0.7394	1.0000	0.9997	0.5932
TOTAL	6	14	3	1

Table3. Narravive web pages

file_name	narrative	inform	argum	instruct
gua_eval_argum_01	0.9961	0.9672	0.9838	0.3696
gua_eval_argum_02	0.9904	0.9673	0.9930	0.2932
gua_eval_argum_03	0.9996	0.9801	0.9682	0.4271
gua_eval_argum_04	0.9991	0.9371	0.9901	0.2777
gua_eval_argum_05	0.0620	0.1622	0.2887	0.3720
gua_eval_argum_06	0.9868	0.9979	0.9827	0.4187
gua_eval_argum_07	0.9732	0.9998	0.9960	0.5217
gua_eval_argum_08	0.9998	0.9960	0.9771	0.2833
gua_eval_argum_09	0.9883	0.9847	0.9531	0.4738
gua_eval_argum_10	0.9918	0.7511	0.4492	0.4150
gua_eval_argum_11	0.9992	0.9985	0.9932	0.2886
gua_eval_argum_12	0.9972	0.9792	0.9631	0.4852
gua_eval_argum_13	0.7149	0.8911	0.6663	0.3430
gua_eval_argum_14	0.2591	0.9964	0.9877	0.2394
gua_eval_argum_15	0.9063	0.9602	0.8439	0.3392
gua_eval_argum_16	0.8750	0.9945	0.9988	0.4476
gua_eval_argum_17	0.2046	0.9544	0.7683	0.4473
gua_eval_argum_18	0.6535	0.9874	0.9995	0.4528
TOTAL	8	6	3	1

Table 4. Argumentative web pages

The best performance is for informative and instructional web pages (see Tables 1 and 2). For argumentative and narrative web pages (Tables 3 and 4), weights needs to be calibrated because their gradations are sometimes too undifferentiated, and the predominant text type has a very thin margin over the others. However, it might also be the case that these undifferentiated gradations are realistic. The goal of the system is not only to show the predominant type, but also to give a reliable depiction of the gradations of the four text types in a single web page. In order to evaluate the gradations, two tests with four human subjects were set up.

In both tests, the gradations of the four text types were converted into a rank in descending order of importance. For example, the text types of the last file in Table 4 were ranked as follows: 1) argumentative; 2) informational; 3) narrative; 4) instructional.

In Test 1, two subjects were asked to rank the four text types in 12 web pages selected from the evaluation set (see the highlighted file names in Tables 1, 2 3 and 4). Then the ranks returned by subjects were compared with the ranks

derived from the system. The Gamma test, also called Goodman and Kruskal's gamma, was used to compare the ranks. Gamma is a measure based on “greater than” and “lesser than” operations, and assesses the difference between concordant and discordant pairs. It is a symmetric measure ranging from +1 to -1. It takes a positive value if the number of concordant pairs is larger; it is negative, if the opposite occurs; it is zero if the number of concordant and discordant pairs is the same. The Gamma values for the evaluation set are shown in Table 5, where column 2 shows the level of agreement between the two subjects; columns 3 and 4 show the level of agreement between each subject and the system. The level of agreement among the two subjects and the system is very coherent. Only two cases are negative. Sometimes, the subjects agree better with the system than with each other.

In Test 2, the ranks of 10 web pages belonging to the SPIRIT collection were evaluated. Two different human subjects were asked to rank them following the same criteria as in Test 1. Once again we compared the level of agreement between the two subjects and between each subject and the system. The Gamma values for the SPIRIT set are shown in Table 6. This time too the level of agreement among human subjects and the system is very coherent. There are only 3 negative cases. These two subjects tend to agree more between them, and subject 1 seems to be particularly tuned with the system.

From these two tests it appears that the gradations returned by the system are largely reliable.

File name	Sub1 vs Sub2	Sub1 vs Sys	Sub2 vs Sys
bbc_eval_inform_01	0.33	0.33	0.33
bbc_eval_inform_08	0.67	0.33	0
bbc_eval_inform_12	0.33	0.33	0
bbc_eval_inform_14	0.33	0.33	0.67
bbc_eval_instr_09	0	1	0
bbc_eval_instr_12	0.33	-0.33	0.33
bbc_eval_instr_18	0.33	0	-1
bbc_eval_narra_15	0.67	1	0.67
bbc_eval_narra_20	0.67	0.33	0.67
gua_eval_argum_08	0.33	0.33	0.33
gua_eval_argum_16	1	0.67	0.67
gua_eval_argum_17	1	0.33	0.33

Table 5. Agreement on the evaluation set

File name	Sub1 vs Sub2	Sub1 vs Sys	Sub2 vs Sys
SPRT_010_049_112_0055534	0.67	0.67	1
SPRT_010_049_112_0055579	1	0.33	0.33
SPRT_010_049_112_0055582	0.33	0.67	0.67
SPRT_010_049_112_0055796	0.67	0.33	0.67
SPRT_025_058_105_0052118	0.67	-0.33	0
SPRT_025_058_105_0052122	0.67	0	-0.33
SPRT_025_058_105_0052245	0.33	0.33	-0.33
SPRT_025_058_105_0052384	0.67	0.67	0.67
SPRT_025_058_105_0052394	0.67	0.67	0.33
SPRT_025_058_105_0052410	0.67	0.67	0.33

Table 6. Agreement on the SPIRIT set

5 Conclusions and Future Work

The system presented in this paper returns gradations of text types in web pages that are largely consistent with those returned by human subjects. Considering that this is just a first and rough implementation, considering that features and weights have not been entirely tuned and calibrated, the results appear to be extremely encouraging. Yet, lots remains to be done. More fine-grained text types should be included in the system. Web pages are highly innovative in their textuality, and a device that captures and measures the level of textual innovation should also be incorporated. Gradations in terms of probabilities should be turned

into proportions and returned in terms of percentages, to make them more transparent. The full potential of the automatic text analysis presented here will be deployed when the system will be enlarged to incorporate automatic genre analysis of web pages.

Acknowledgements

Many thanks to Richard Power and Ling Yin.

7 References

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

Biber D. (1989), A typology of English texts, *Linguistics*, Vol. 27, 3-43.

Clarke C., Cormack G., Laszlo M., Lynam T., and Terra E. (2002), The Impact of Corpus Size on Question Answering Performance, *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in IR*, Tampere, Finland.

Duda R. and Reboh R. (1984). AI and decision making: The PROSPECTOR experience. In Reitman W. (Ed.), *Artificial Intelligence Applications for Business*. Norwood, NJ.

Faigley L. and Meyer P. (1983), Rhetorical theory and readers' classification of text types, *Text*, Vol. 3, 305-325.

Lee D. (2001), Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle, *Language Learning and Technology*, Vol. 5, Num. 3, 37-72.

Quirk R., Greenbaum S., Leech G., Svartvik J. (1985), *A Comprehensive Grammar of the English Language*, Longman.

Stubbs M. (1996), *Text and Corpus Analysis*, Blackwell Publishers, Oxford.

Tapanainen P. and Järvinen T. (1997), A non-projective dependency parser, *Proc. of the 5th Conference on Applied Natural Language Processing*.

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