Revista Alicantina de Estudios Ingleses 19 (2006): 387-404

Towards an Analysis of Academic Weblogs

Keith Stuart
Polytechnic University of Valencia
kstuart@idm.upv.es

ABSTRACT

This paper analyses both the communicative purposes and formal linguistic features of academic weblogs. An initial analysis of 496 weblogs in use within tertiary level institutions was reduced to an in-depth analysis of 39 academic weblogs (a corpus of 16 million words). The objective was to see how much variation there might be between traditional academic genres and academic weblogs. The initial hypothesis is that academic weblogs are far less formal and would favour greater use of lexico-grammatical realisations belonging to the interpersonal function of language. In order to quantify this possible variation between traditional academic genres and academic weblogs, some well-known linguistic features which have been considered to be typical of academic discourse such as the agentless passive and nominalisations were investigated.

1. Introduction

This paper takes a genre-based and linguistic view of academic weblogs. The reality of the situation of academic weblogs can only be captured by developing a complex and dynamic picture. They are a social phenomenon that is breaking down boundaries across and within spoken and written discursive practices and changing the way knowledge is communicated. They display similarities as well as overlaps with more traditional genres, but like other Internet genres constitute a hybrid genre that draws from various online and offline sources.

Weblogs have to be situated within the context of the broader genre ecology of the Internet as one of the latest forms of digital communication. They have elements that have developed out of previous Web genres (e.g., online journal, personal home page, hotlist) and have antecedents in previous offline genres (e.g., diaries, newsletters, editorials) (Herring et al., 2005). It is not uncommon to find mixed content within a single weblog.

Weblogs can be considered an example of the appropriation of generic resources from a previously established genre to give shape to a more dynamic and innovative form, motivated by new technology and social need. Bhatia (2004: 87-88) has called this appropriation of generic resources as the "invasion of territorial integrity" and claims that, in the context of "the present-day interdisciplinary and dynamic world of work", it is difficult to maintain individual generic boundaries intact, particularly because of the explosion of information technology. He also states that "this tendency to appropriate generic resources is becoming increasingly common in all areas of academic and professional discourse" (Bhatia, 2004: 87).

A genre is identified by its socially recognized *communicative purpose* and shared characteristics of *form* (Swales, 1990). The communicative purpose of a genre refers to the social motivation, the discursive practices and discourse domains, which are constructed and recognized in the communication. Form refers to observable aspects of the communication, such as *structural features* (e.g., text formatting devices such as lists, posts and comments in the case of weblogs) and *linguistic features* (e.g., level of formality, specialized vocabulary or terminology).

This paper analyses both the communicative purposes and formal linguistic features of academic weblogs. An initial analysis of 496 weblogs[1] in use within tertiary level institutions was reduced to an in-depth analysis of 39 academic weblogs (a corpus of 16 million words).

2. Communicative Purpose of Weweblogs

If we analyse the purpose for publishing a weblog, we can speak of weblogs on a horizontal level (general purpose weblog) and on a vertical level (a more specific purpose or specialist weblog). A diary would fall within the first category, while academic weblogs would enter the second.

At the same time, it is possible to classify weblogs depending on which sector of the economy they are being used in. A high level generic classification would include journalism, political, corporate, professional, personal and academic /educational weblogs. This very general classification illustrates the way in which the weblog phenomenon has permeated all socio-economic environments, from the personal sphere to politics, passing through professional, organizational and institutional environments. Looking at corporate weblogs, we can divide them into six main types:

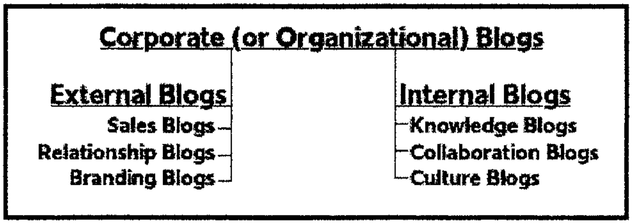

Figure 1. Corporate Weblogs[2]

For example, sales weblogs market or sell products and services while relationship weblogs can help a corporation to create, maintain and strengthen customer relations.

If we analyse weblogs from the point of view of who is responsible for these publications, we have to talk about individual, group, collective and corporate weblogs. Weblogs create communities but they also may be used to gain prestige within a specific profession such as journalism. However, a weblogzine may be used to praise more the know-how contributed by a community of expert collaborators.

3. Structural Features of Weweblogs

As this is a relatively new genre, the most important structural features (*title, posts, comments, links, weblogrolls, archives, feeds, about page*) of weblogs are briefly summarised.

> **Universal Acid**
> Biology, science, politics, and society

The title of a weblog (the weblog's name) reflects the purpose and tone of the site.

> **Brokeback Mountain**
> I want to pose a question about a scene in Brokeback Mountain, but to avoid spoilers (it's not much of one, actually), I'll pose it in a comment to this post.
> posted by Andrew @ 10:30 AM 4 comments links to this post

A weblog is a "frequent, chronological publication" whose main component is the "post" which is dated and given a title. Those who comment on others' weblogs are like the listeners who frequently call in to radio stations to voice their opinions. The comments section is often the most interesting area of a weblog. It is where the weblogger's one-sided opinion is tested and turned into debate and dialogue.

Another almost-obligatory element of weblogs is that they are link-driven. Just as technology has made every weblog post instantly available to people anywhere in the world with a computer and Internet access, the other defining trope of weblogging is its engagement, via hyperlinking, with the larger Internet conversation. At its best, weblogging uses links to support critical, well-considered arguments with supporting facts. Links, in those cases, serve as bibliographical references, directing readers to additional sources of relevant material. Links to relevant information can often be found not only within the content of a particular weblog post but also within a permanent sidebar. A "blogroll" contains links to other weblogs and websites, and can typically be found running down the left- or right-hand side of the page. In the archives section of a weblog, a link is provided to previous posts.

Web feeds are widely used by the weblog community to share the latest entries' headlines or their full text between each others' weblogs and other Internet sources. RSS (short for Really Simple Syndication), in particular, delivers this information as an XML file called an RSS feed. In addition to facilitating syndication, weblog feeds allow a weblog's frequent readers to track updates on the site using an aggregator.

One final component in this anatomy of a weblog is the About page. The About page is a place for a weblogger to introduce him/herself to their readers, to provide contact information, and to describe the purpose of the weblog. So, an academic weblog sometimes includes the author's curriculum vitae. However, for obvious reasons, many webloggers prefer to remain anonymous.

Weblogs as a genre may also be understood from the point of view of the functionality of their structural features, for example, their *hypertextual nature* and *conversational character*. The intensive use of hypertextual links is a defining characteristic of the blogosphere, which means that they are automatically in linguistic terms intertextual (weblogs quote copiously). One possible way of measuring a weblog is in terms of the degree of "hypertextuality" of a weblog. To measure the degree of intensity in the utilization of those links, we should be able to define the Hypertextual Density of a weblog as the relationship between the number of hypertextual links and the average number of words per link.

Weblog Discipline	Weblog Name	Hyperlinks	Words	Hypertextual Density
Biology	Universal Acid	1,0253e+43	5,9067e+69	57,62
Biology	Pharyngula			486,67
Biology	Aetiology			427,75
Chemistry	The Culture of Chemistry			377,67
Chemistry	Mass Spectrometry Weblog			250,41
Chemistry	Corante			413,96
Computer Science	The Geomweblog			8,47
Computer Science	Computational Complexity			30,56
Computer Science	Musings			287,22
Sociology	East Ethnia			71,63
Sociology	Pub Sociology			129,17
Sociology	Jeremy Freese's weweblog			111,05

Table 1. Hypertextual Density of a weblog.

As can be seen from our corpus examples, there is no apparent tendency but one can note that larger weblogs have a lower hypertextual density. Another perspective on linking would be to work out whether the links to one's own weblog predominate (intra-weblogging) or whether there is a predominance of external links (extra-weblogging[3]) or some degree of equilibrium between the two.

From the perspective of the interactive nature of the blogosphere, one could define the Conversational Capacity of a weblog as the average number of comments per post. In a wider sense, we could try to define the Capacity (or Index) of Influence of a particular weblog as the relation between sites the weblog links to and weblogs that link to it. In other words, a weblog that has a lot of inbound links becomes a hub and is likely to have some influence in the conversational dynamics of its corner of the blogosphere.

4. An Overview of Academic Weblogs

Initial research of 496 weblogs revealed the following communicative purposes of weblogs within academic organisations (these organisations include prestigious institutions such as Harvard University and Massachusetts Institute of Technology):

- Academic Community forum, e.g. Crooked Timber, Rhetorica
- Academic critique
- Academic debate
- Announcement of author appearances (book promotions, conferences, talks: radio / television, workshops)
- Article / book reviews
- Weblog research / meta-weblogging
- Weblogs as ePortfolio
- Book promotion / selling
- Class weblog (announcements, discussion questions and topics, readings, schedules etc.)
- General information about news, conferences, publications and academic jobs
- Personal journal with viewing/comment by teacher/tutor
- Personal knowledge management
- Posting student work for viewing/comment by peers
- Promoting the university (more particularly, degree courses)
- Publication of texts, links or commentary to seed discussion
- Research diary
- Self-organized collaborative learning (reflective learning)
- Student weblog as coursework
- Student group discussion tool

This list of the communicative purposes of weblogs within academia is not all inclusive.

Nevertheless, they can be classified into four general categories and can be considered to form a subset of genres within the larger system of academic and educational genres which are intertextually and interdiscursively linked within the boundaries of academic and educational activity. Bazerman (1994: 97) proposes the concept of 'the system of genres' as referring to all 'the interrelated genres that interact with each other in specific settings'. The genre subset proposed here consists of the following: PhD weblogs (research diaries); Student weblogs (coursework); Educational weblogs (class weblogs); Academic weblogs (research work) (see table below). It is the latter that this paper focusses on.

Author	Student	Teacher
Researching	PhD / Masters Weblogs	Academic announcements, critique, debate
Learning / Teaching	Student weblog as coursework Student group discussion tool	Class weblog Publication of texts, links or commentary to seed discussion

Table 2. Use of Weblogs in Academic Institutions.

There are some anomalies that do not fit into this classification such as when universities use weblogs to promote degree courses[4] or, for that matter, the university itself. More and more universities[5] in the United States are using freshman weblogs as a marketing tool to boost prospective students' interest in their institution by supporting and promoting a few hand-picked freshman bloggers. Fairclough (1993) has termed this the 'marketization' of academic discourse, while Bhatia (1995) mentions several instances (such as promotion within academic course descriptions) where one may find an increasing use of promotional strategies in genres which are traditionally considered non-promotional in their communicative purpose. It is this kind of genre mixing and hybridization that is found in academic weblogs. Academic weblogs are used for all the traditional academic activities (article / book reviews, announcements of and links to conferences, presentation of research work in progress) but there is also a marked promotional nature to them. They often announce the publication of the weblog author's book (with positive reviews of the book below the announcement) as well as informing readers of the author's appearance at bookshops, workshops and conferences.

Academic weblogs invite criticism and debate by practising the art of criticism through critiques of fellow academics' work and are openly evaluative. In a more traditional genre such as academic journal articles, there are discussion and conclusion sections but they do not normally express an overt point of view. As Blood (2002: 59) says, "personal detail is not necessary [...] but every good weblog has a point of view". The evaluative nature of weblogs runs against the grain of academia. Academia is not generally speaking about publishing your 'ideas', but about scientific inquiry and publishing results. This may explain the relatively fewer weblogs in the engineering and hard science disciplines, although there are a larger number of weblogs in newer fields such as biotechnology and nanotechnology. Precisely, because they are emerging fields, weblogs may not seem such an anathema to them and their promotional value can be seen.

However, scholars do tackle serious research questions and sometimes write carefully honed lengthy treatises on their weblogs. What comes down in favour of academic weblogs is the fact that there are a growing number of users among academics and academic institutions (among them some very prestigious institutions such as the Harvard Law School[6]). The difference in the long run between the personal publishing of your ideas 'blogging' or getting them published in a reputable journal may not be so great. It is a matter of prestige and recognition. Blogging has become an integral part of some academics' research processes which may or may not be converted into an academically reputable published product.

An end note for this overview of academic weblogs is that one defining characteristic of academic writing is the rigorous and formal citation practice. On the surface, references used in weblogs seem to be random linking to sources. However, they frequently refer as explicitly as do academic texts, though more simply by linking to a text at another weblog or web page (for example, an article in a newspaper's website or to a book's page at Amazon). Like in academic writing, it is also very common to quote other writers, but what you write in your weblog can be quoted and discussed freely in any forum.

5. Methodology

5.1. Data collection

Various techniques were used to track down the 39 academic weblogs in the corpus. This included using seeds with Google. A seed such as Law Professors Weblog would immediately bring up http://www.lawprofessorweblogs.com/. This makes the researcher's task fairly easy to obtain data but this does not always work. This group weblog also reflects the increasing institutionalisation of weblogs and their use for promotional purposes - the weblog site is sponsored and welcomes advertising. However, the weblog meets the main criterion of being written and edited by Law Professors (from Cincinnati University, Ohio State University, among others).

Besides using Google to track down possible candidate weblogs, one can also use Directories. In the case of academic weblogs, there are two main sources: http://crookedtimber.org/academic-weblogs/ (Group weblog) and http://rhetorica.net/professors_who_weblog.htm (Andrew Cline, Missouri State University). The first weblog site uses the following academic policy link.

> The qualifications are fairly straightforward. First, if you either have an academic position at a university type institution, or are a Ph.D. student or equivalent at same, you qualify (Henry Farrell, George Washington University).

The result is rather a mixmatch bag of weblogs although they have been classified by disciplines. This is the most useful starting point for tracking down academic weblogs, as academic bloggers refer you to their colleagues. There are also well-known general directories of weblogs that include science and technology subcategories. These can be found by introducing into Google seeds such as Weblog Directories.

Another method is to use weblog award sites such as "The 2005 Koufax Awards: Best Expert Weblog". Three of our eventual candidate weblogs were tracked down from this weblog site. Although we were interested in popularity (that is, weblog sites that have a regular readership and receive a steady flow of comments), this was not an overriding criterion as there are several academics who weblog and are extremely popular but whose purpose is far from academic. An example is Instapundit edited by Glenn H. Reynolds, a professor of law at the U. of Tennessee at Knoxville or weblogs such as the Volkhov Conspiracy or Oxweblog. Instapundit receives over 100,000 hits a day and acts as a news filter.

Having identified possible candidates, an open source offline browser (WinHTTrack) was used to create a representative corpus of 39 science and social science weblogs (a total of over 16 million running words). The table below provides the most important statistics of the academic weblogs selected:

	text file	file size	tokens	types	type/token	standardise
0	overall	109.197.290	16.070.696	126.009	0,78	42,56
1	Biology1.txt	370.894	56.805	7.860	13,84	45,76
2	Biology2.txt	9.564.459	1.559.576	42.148	2,70	43,89
3	Biology3.txt	2.324.893	355.994	15.437	4,34	43,53
4	Biotech1.txt	1.591.858	229.514	8.025	3,50	45,87
5	Biotech2.txt	1.732.477	233.872	7.087	3,03	44,60
6	Biotech3.txt	3.363.068	514.989	18.830	3,66	44,12
7	Business law.txt	4.839.752	674.512	10.220	1,52	41,96
8	Clinic law.txt	731.826	99.588	4.411	4,43	37,97
9	Compliance law.txt	2.034.027	280.650	7.652	2,73	42,77
10	Computerscience1.	961.730	156.320	12.340	7,89	43,09
11	Computerscience2.	3.952.631	192.322	11.187	5,82	43,43
12	Computerscience3.	3.461.350	466.485	12.873	2,76	45,80
13	Chemistry1.txt	153.005	23.615	4.375	18,53	45,96
14	Chemistry2.txt	658.368	74.366	3.498	4,70	23,66
15	Chemistry3.txt	4.902.955	542.289	18.477	3,41	44,70
16	Economics1.txt	2.836.904	453.840	24.472	5,39	46,46
17	Economics2.txt	7.688.212	1.109.071	23.519	2,12	44,15
18	Economics3.txt	841.337	134.181	10.398	7,75	45,13
19	History1.txt	3.689.356	545.327	11.011	2,02	43,38
20	History2.txt	856.462	113.156	8.867	7,84	44,21
21	History3.txt	1.304.179	212.409	8.867	4,17	33,89
22	Maths1.txt	14.334.249	2.338.922	31.049	1,33	38,35
23	Maths2.txt	1.149.890	165.004	5.435	3,29	28,52
24	Maths3.txt	39.900	6.804	1.549	22,77	38,53
25	Nanotechnology1.t	3.168.518	483.914	11.002	2,27	41,11

26	Nanotechnology2.t	260.150	34.730	2.792	8,04	38,75
27	Nanotechnology3.t	4.766.380	688.206	22.572	3,28	46,61
28	Physics1.txt	4.701.318	742.883	13.248	1,78	37,60
29	Physics2.txt	258.813	40.385	6.029	14,93	46,89
30	Physics3.txt	845.695	133.452	10.686	8,01	46,00
31	Politicalscience1.tx	1.209.983	136.883	12.886	9,41	48,64
32	Politicalscience2.tx	8.432.797	1.369.701	25.853	1,89	43,76
33	Politicalscience3.tx	2.085.068	331.062	17.116	5,17	46,73
34	Psychology1.txt	2.193.159	351.301	10.267	2,92	35,76
35	Psychology2.txt	1.945.979	299.111	12.905	4,31	45,54
36	Psychology3.txt	833.050	120.616	11.159	9,25	45,33
37	Sociology1.txt	1.487.359	210.358	14.083	6,69	44,74
38	Sociology2.txt	720.116	115.276	11.108	9,64	46,66
39	Sociology3.txt	2.905.123	473.207	22.555	4,77	45,71

Table 3. Corpus of Academic Weblogs (CAW), analysis carried out with WordSmith

As can be seen from the table, the corpus consists of 16,070,696 running words with 126,009 different words. The type/token ratio is 0.78 and the standardised type/token ratio is 42.56. The type/token ratio (TTR) varies very widely in accordance with the length of the text, or corpus of texts, which is being studied. A text of 1,000 words may have a type/token ratio of 40%; 4 million words will probably give a type/token ratio of about 2%; while in our case the corpus is even larger so there is a very low type/token ratio of 0.78%. Such type/token information is rather meaningless in most cases, though it is supplied in the analysis that WordSmith carries out. What is more useful is the standardised type/token ratio (STTR) which is computed every n words as WordSmith goes through each text file (by default, n = 1,000). In other words, the ratio is calculated for the first 1,000 running words, and then calculated afresh for the next 1,000 and so on to the end of the text or corpus. A running average is computed, which means that you get an average type/token ratio based on consecutive 1,000-word chunks of text. This makes it possible to compare type/token ratios between corpora as we have done in our linguistic analysis below (see table 5).

5.2. Messiness of the data

There are considerable methodological difficulties in analysing weblogs. The main impediments are precisely those structural features outlined above and, above all, the different kinds of tags that weblogs use so they can be visualised in web browsers. This meant a considerable amount of boilerplate removal. This was done initially automatically but some manual removal was also needed. For automatic removal of tags, commercial software (Detagger[7]) was used. Among the useful features of this software is the fact that it controls what happens to any hyperlinks in the original document. Since text files do not support hyperlinks, the options are to ignore the link entirely, only use the display text, or to turn the link into a reference and add a reference table at the end, listing the URLs the links pointed to. This software does the latter and provides useful information about who the author of the weblog is linking to. In an academic weblog, a list of hyperlinks has a similar function to the bibliography or references section of a journal article. However, for corpus analysis purposes, these hyperlinks then have to be removed manually or they would distort the results. One of the ways to get at text in academic weblogs is by going to the archives section of the weblog and downloading only the files found in the archives section. One can illustrate the tedious work that has been carried out by comparing the size of the corpus text file with that of the original weblog that has been spidered.

Text file	Corpus file size	Weblog size
Biology1.txt	370.894 kb	9,59 mb
Biotech1.txt	1.591.858 kb	41,3 mb
Computerscience1.txt	961.730 kb	36,7 mb
Chemistry1.txt	153.005 kb	4,57 mb
Economics1.txt	2.836.904 kb	82,2 mb
History1.txt	3.689.356 kb	21,5 mb
Maths1.txt	14.334.249 kb	56,7 mb
Nanotechnology1.txt	3.168.518 kb	16,3 mb
Physics1.txt	4.701.318 kb	30,1 mb
Politicalscience1.txt	1.209.983 kb	6,25 mb
Psychology1.txt	2.193.159 kb	17,7 mb
Sociology1.txt	1.487.359 kb	32,2 mb

Table 4. Before and after the corpus clean-up process.

The main problem that arises with such messy data is that the clean-up process does not result in a well-balanced corpus. Furthermore, the cleaning-up process is problematic and involves many hours of work. The size of the corpus to some extent compensates for this and, as Tognini-Bonelli (2001: 1) states, "what we are witnessing is the fact that corpus linguistics has become a new research enterprise and a new philosophical approach to linguistic enquiry" driven by massive amounts of data. "It is strange to imagine that just more data and better counting can trigger philosophical repositionings, but [...] that indeed is what has happened" (Tognini-Bonelli, 2001: 48). Empirical data about language has the ability to confirm or deny what up to that point time may have only been hypothesized.

6. Results: Linguistic Data

In this section, empirical data about language used in academic weblogs is presented. The main interest was to compare linguistic features in the Corpus of Academic Weblogs with results from Biber's earlier study of Academic Prose (Biber, 1995; Biber et al., 1998). His study used the Lancaster-Oslo-Bergen Corpus of British English (a corpus of approximately one million words) (Johansson et al., 1978). The corpus comprises fifteen genres including academic prose. The academic prose part of the corpus is made up of 80 texts of about 2,000 words each (a total of 160,000 words) and is about one hundred times smaller than our corpus.

The objective was to see how much variation there might be between traditional academic genres and academic weblogs. The initial hypothesis is that academic weblogs are far less formal and would favour greater use of lexico-grammatical realisations belonging to the interpersonal function of language. In order to quantify this possible variation between traditional academic genres and academic weblogs, some well-known linguistic features which have been considered to be typical of academic discourse such as the agentless passive and nominalisations were investigated. However, one or two lexico-grammatical realisations do not provide sufficient evidence to describe a text genre or to be able to state whether a text has high informational density and exact informational content (academic discourse) or affective, interactional and generalized content (conversational discourse) (Biber, 1995: 107). Nevertheless, a linguistic profile of the texts in a corpus can be built up by analysing a number of linguistic features. In this study, the results of analysing 26 linguistic features are presented.

The linguistic features have not been chosen randomly but rather have been chosen as typically representative of either more interactional discourse (first and second person pronouns, hedges, amplifiers, possibility modals, private verbs) or more informational discourse (agentless passive, nominalisation, prepositions). Results for the mean frequency of the 26 linguistic features that have been studied are given in the table below. The mean frequency in the Corpus of Academic Weblogs was calculated by normalizing the frequency counts of all linguistic features to a text length of 1,000 words. Therefore, the mean frequency for first person pronouns would be:

$$302{,}646 \ / \ 16{,}070{,}696 \times 1{,}000 = 18.83$$

In our corpus, there are 302,646 first person pronouns which we divide by the total number of words in the corpus. Then, we multiply the result by 1,000 to get a mean frequency of 18.83 first person pronouns for every 1,000 words in the corpus.

This process is fairly easy for first person pronouns as there are only a few members of this word class. The subjects of cognitive verbs are usually first person pronouns, indicating that mental processes are a personal matter often associated with high ego-involvement. In the case of emphatics, calculations had to be made for the following words and phrases.

Really, just, for sure, such a, real + adj., so + adj., do + verb, most, more, a lot

Emphatics are characteristic of informal, colloquial discourse, marking involvement with the topic (Chafe, 1985). While for nominalisation, all words ending in *-tion, -ment, -ness,* or *-ity* (plus plural forms) had to be calculated. Biber (1986) finds that they tend to co-occur with passive constructions and prepositions and thus interprets their function as conveying highly abstract information.

Linguistic Feature	Total Frequency (Academic Weblogs)	Mean frequency (Academic Weblogs)	Mean Frequency (Academic Prose)
Agentless passive	205,628	12.79	17.0
Amplifiers	27,793	1.73	1.4
Analytic negation	135,612	8.44	4.3
Causative clauses	16,552	1.03	0.3
Concessive clauses	9,523	0.59	0.5
Conditional clauses	46,013	2.86	2.1
DO as pro verb	10,344	0.64	0.7
Downtoners	28,284	1.76	2.5
Emphatics	59,942	3.72	3.6
Existential 'there'	24,070	1.5	1.8
First person pronouns	302,646	18.83	5.7
Hedges	16,752	1.04	0.2
Indefinte Pronouns	42,097	2.61	0.2
Necessity modals	23,849	1.48	2.2
Nominalisation	359,843	29.64	35.8
Possibility modals	95,383	5.94	5.6
Predictive modals	76,942	4.78	3.7
Prepositions	1,540,332	95.84	139.5
Private verbs	184,132	11.46	12.5
Pronoun IT	116,551	7.25	5.9
Public verbs	108,489	6.75	5.7
Second person pronouns	110,972	6.91	0.2
Seem / appear	16,621	1.03	1
Stranded prepositions	9,800	0.61	1.1
Synthetic negation	29,753	1.85	1.3
Third person pronoun	170,782	10.63	11.5
Type/token ratio	-	42.56	50.6
Word length	-	4.72	4.8

Table 5. Comparison of linguistic features in the Corpus of Academic Weblogs with Biber's study of Academic Prose.

After having examined the results for the 26 linguistic features, there are less surprises than one would have initially hypothesised. There are no great differences in the frequency of use of the modals. Typical features of spoken discourse such as amplifiers, downtoners, emphatics and hedges do not show any marked variation from Biber's results of his analysis of Academic Prose. These are normally considered indicators of personal involvement. There is less use of nominalisation and agentless passives are not so common in the Corpus of Academic Weblogs. However, if we look at Biber's analysis of nominalisation in other written genres such as Biographies (20.6), Press Reviews (21.6) and Press Editorials (27.6), they all have a lower mean frequency than Academic Weblogs. Again, if we compare the use of agentless passives in Academic Weblogs with these three genres, we find that Biographies (9.9), Press Reviews (8.6) and Press Editorials (11.7) all have a lower mean frequency use of the agentless passive than Academic Weblogs.

However, there is one area in which there is marked variation between the two genres. This is in the use of the first and second person pronouns. This reflects the diary/journal aspects of weblogs and the fact that weblog authors are very aware of their audience. The use of these pronouns is re-inforced by the comment sections of weblogs where debate and discussion are held and there is likely to be frequent use of this word class. It is interesting to note that, in the case of third person pronouns, there is no significant difference in the frequency of their use. Below are three excerpts from a post called 'Who's afraid of reproductive cloning?' (December 21st, 2004; Universal Acid) which illustrate the use of pronouns in academic weblogs.

Example 1:

> To be clear on terms: cloning is when **you** take a nucleus from an adult cell, put it into a de-nucleated egg cell, and then stimulate the egg to divide and form an embryo that is (for all intents and purposes) genetically identical to the adult that donated the nucleus. In therapeutic cloning, the purpose is to create a ball of cells after a few days, from which **you** can harvest embryonic stem cells. In reproductive cloning, the purpose is to create a full adult organism that would be genetically identical to the individual that donated the transplanted nucleus.
>
> The most obvious reason to ban reproductive cloning is safety. Reproductive cloning hardly works at all in animals, and if **we** tried it in humans, **we**'d almost certainly create babies with horrible birth defects. Even if **we** got it working, it probably wouldn't be as safe as regular fertilization, and even the tiniest risk would be a significant reason not to do it. **I** suppose the whole argument is moot for now, as the safety concern trumps all. So **I** still have time to figure out what **I** think.
>
> **You** are not made up of **your** genes; **you** are made up of a very complex and unique pattern of protein, lipids, sugars, etc; **your** identity is coded in the pattern of neurons and synapses firing in **your** brain. This pattern is not entirely determined by **your** genes (Biology1.txt).

Besides the use of pronouns, we find that analytic negation (*not*) is used twice as much in Academic Weblogs as in Biber's analysis of Academic Prose. According to Tottie (1983), there is twice as much negation overall in speech as in writing, a distribution that he attributes to the greater frequency of denials, rejections, questions and mental verbs in speech.

In our results, there is one striking anomaly: the mean frequency of prepositions in the two corpora. As the difference was so great, the results were checked against the British National Corpus (100 million words; 90% written text, 10% spoken text). In the British National Corpus, there are an average 94.11 prepositions for every 1,000 words. This result is very similar to our own (95.84) and either suggests that in traditional academic genres there is a significantly higher use of prepositions as an important device for packing information in extended nominal groups or that in Biber's smaller corpus there is a slight distortion.

There are some limitations to this study in that, for example, only modal verbs were analysed (as in Biber's study, so as to make results comparable) and there are many other lexico-grammatical realisations of modality, including adverbials such as *probably* and *possibly*. A quick comparison of their mean frequency in the Corpus of Academic Weblogs with the British National Corpus gives the following results.

	Total Frequency (CAW)	Mean Frequency (CAW)	Total Frequency (BNC)	Mean Frequency (BNC)
Probably	5	0.32	26,531	0.26
Possibly	7	0.07	1,193	0.07

Table 6. Comparative frequencies of *probably / possibly* in two corpora.

As can be seen, there is no significant variation or greater use of these two expressions of modality in the Corpus of Academic Weblogs. Lexico-grammatical realisations of evaluation are another aspect of language worthy of study with respect to academic weblogs. Overt evaluation is not considered to be typical of academic genres. An initial analysis of the twelve most common evaluative adjectives from the Corpus of Academic Weblogs compared with their use in the British National Corpus did not reveal any significant variation, although there was a tendency towards a greater use of these adjectives in the Corpus of Academic Weblogs and this needs further study.

In this study of linguistic features, we have tried to observe if there are language choices being made in academic weblogs that suggest greater interactivity. Lexico-grammatical realisations of interpersonal elements are seen as modifications of a basic message or content for greater communicative effectiveness, standing out against unmarked forms. In comparing linguistic features with Biber's results for Academic Prose, we have found that there are differences but the level of variation is not as great as initially hypothesised. The generic integrity of a genre is often perceived in the literature in terms of typical lexico-grammatical and discourse patterns, simply because these are the most obvious surface-level linguistic features of textual genres and this has been the approach taken here. However, writing is more than the generation of a text-linguistic product and cannot easily be separated from the broader institutional and socio-historical context which inform those particular occasions of writing. This is equally applicable to the analysis of academic weblogs.

To conclude this section, a few examples of posts and comments from academic weblogs are presented for analysis. The first example is clearly impeccable academic discourse.

Example 2:

> Breakthroughs in adult stem cell research in Asia
> Of the 74 patients, 5 suffered from cerebral infarctions, 23 from Buerger's disease [194], 11 from femoral head avascular necrosis [195] and 35 from unhealed bone fractures. After

> undergoing self-derived stem cell implantation, 64 of them showed significant improvement without developing any negative side effect. 21 out of 23 patients suffering from Buerger's disease improved significantly after the cell therapy. In the case of cerebral infarction, an apparent improvement occurred in 3 cases out of the total 5, while 7 patients with femoral head avascular necrosis out of the total 11 and 33 out of the 35 patients with non-union of bone fracture were considered successful cases.
>
> Meanwhile, researchers at the National Centre for Cell Science in Pune, India, who studied the status of bone marrow stem cells in experimental-diabetic mice, have conclusively established the reversal of experimental diabetes [196] by multiple diabetic bone marrow transplants. While a single injection of 1 million bone marrow stem cells taken from the siblings of the laboratory mice and injected into the experimental-diabetic mice resulted in the reduction and stabilization of moderate hyperglycemia, multiple injections at regular intervals led to the restoration of stabilized normogylcemia and ultimately, diabetes was reversed. (Biotech1.txt)

Example 3:

> Size and selection times: Fitts's Law
>
> Fitts discovered that movement time was a logarithmic function of distance when target size was held constant, and that movement time was also a logarithmic function of target size when distance was held constant. Mathematically, Fitts's law is stated as follows:
>
> $$MT = a + b \log2(2A/W)$$
>
> where
> - MT = time to complete the movement
> - a,b = parameters which vary with the situation ('regression coefficients')
> - A = distance of movement from start to target center
> - W = width of the target along the axis of movement (also equivalent to the degree of permissible error in movement target)
>
> Fitts's Law is an example of a principle in psychology which was developed from information theory (you can read more about this here [1251] [1]). Although the basic message is obvious (big things are easier to select) it is the precise mathematical characterisation that is exciting, and that this characterisation includes a logorithmic function... (Psychology2.txt)

In this second example, the text starts with what appears to be impeccable academic discourse. However, in the second paragraph, the clauses that have been underlined display interpersonal elements which are traditionally less common in academic discourse. The use of 'you' to directly address the reader, the use of the expression 'basic message' is found in both written and spoken English but, as with the explicit explanation in brackets of what the word 'message' refers to, they are not conventional academic discourse practice. Similarly, to describe a 'precise mathematical characterisation' as 'exciting' shows an enthusiasm which is not appropriate to academic discourse. The last sentence is structurally typical of spoken discourse as it is made up of a string of clauses.

Example 4:

> It's also worth noting that most, and usually the most interesting, challenges to prevailing wisdom come from within the "elite", not from outside. The example that springs to mind is mitochondrial recombination, which came from as solid a neo-Darwinian pedigree as one could imagine.
>
> I wouldnt' bother trying to dissect what ignoramuses say. They believe their own tripe, and they will continue to do so because they feel they don't have to account for it in the court of public opinion. It's a sad state, surely. (Biology2.txt)

In example 4, there are two paragraphs from the comments section of an academic weblog. This first comment maintains some formality 'it's also worth noting that...', but uses more colloquial language such as 'springs to mind' and 'as one could imagine'. However, it does not descend into overtly dismissive evaluation as in the second paragraph. It is this type of language (although the writer shows some sophistication) that can be found in academic weblogs which clearly does not belong to academic discourse.

7. Conclusion

Texts are multidimensional constructs requiring multiple perspectives for their understanding. The fundamentally textual approach taken here sees textual variation and similarity in terms of lexico-grammatical and discursive patterning as realisations of particular genres. A more interactive approach sees text as a product of interaction between writer and readers and, therefore, some perspectives not taken into account here such as face threatening acts (Brown and Levinson, 1987), implicature and the maxims of interpretation and politeness (Grice, 1975), mechanisms for conveying newness of information and appealing to shared knowledge (Prince, 1992) may need to be used to produce a richer interpretation of the communicative purposes of academic weblogs.

Generally speaking, blogging offers academics a way to speedy scholarly discussion, the opportunity to reach and interact with diverse groups of readers both inside and outside academia, and the freedom to adopt a voice and point of view which are considered appropriate in this genre whereas traditional academic discourse surpresses this kind of greater ego-involvement. So, even if academics do not receive any credit for blogging as such, they will get more satisfaction because their ideas are being more widely spread and they may eventually get a better academic reputation as a result, precisely because more people will be noticing their ideas. It is the personal, promotional functions complementing serious rational argument that characterises these weblogs as an emerging hybrid academic genre.

Notes

[1] Our starting point was http://crookedtimber.org/academic-weblogs/ but various sources were finally used to track down the 39 academic weblogs in the corpus.

[2] Source: http://www.corporateweblogging.info/.

[3] Source: www.electricvenom.com/2003/05/05/weblogging-thoughts-and-philosophies/

[4] For example, the MBA Admissions Weblog at The Wharton School of the University of Pennsylvania (http://adcomweblog.wharton.upenn.edu/).

[5] For example, Alfred University, New York (http://www.alfred.edu/real_life/index.cfm?fuseaction=diary.listlatest&site=1).

[6] http://weblogs.law.harvard.edu/

[7] Copyright © 2005 JafSoft Limited.

References

Bazerman, Charles (1994): "Systems of genres and the enhancement of social intentions". In A. Freedman and P. Medway, eds., *Genre and New Rhetoric*. London: Taylor and Francis, 79-101.

Bhatia, Vijay (1995): "Genre-mixing in professional communication: the case of 'private intentions' v. 'socially recognised purposes'. In P. Bruthiaux, T. Boswood and B. Bertha, eds., *Explorations in English for Professional Communication*. Hong Kong: City University of Hong Kong, 1-19.

_______. (2004) *Worlds of Written Discourse*. London: Continuum.

Biber, Douglas (1986) "Spoken and written textual dimensions in English: resolving the contradictory findings". *Language* 62: 384-414.

_______. (1995) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad and Randi Rippen (1998): *Corpus Linguistics*. Cambridge: Cambridge University Press.

Blood, Rebecca (2002): *The Weblog Handbook*. Cambridge, Ma.: Perseus Publishing.

Brown, Penelope and Stephen Levinson (1978): "Universals in language usage: Politeness phenomena". In E. Goody, ed., *Questions and Politeness: Strategies in Social Interaction*. Cambridge: Cambridge University Press, 56-310.

Chafe, Wallace (1985): "Linguistic differences produced by differences between speaking and writing". In D. Olson, N. Torrance and A. Hildyard, eds., *Literature, Language and Learning: The Nature and Consequences of Reading and Writing*. Cambridge: Cambridge University Press, 105-123.

Fairclough, Norman (1993): "Critical Discourse Analysis and the Marketisation of Public Discourse: The Universities". *Discourse & Society* 4(2): 133-168.

Grice, H. P. (1975): "Logic and conversation". In Peter Cole and J. L. Morgan, eds., *Syntax and semantics: Speech acts*. Volume 3. New York: Academic, 41-58.

Herring, Susan, L. Scheidt, E. Wright and S. Bonus (2005): "Weblogs as a bridging genre". *Information, Technology & People* 18(2): 142-171.

Stig Johansson, Geoffrey Leech, and Helen Goodluck (1978): *Manual of information to accompany the Lancaster-Oslo-Bergen corpus of British English for use with digital* *computers*. Bergen: Norwegian Computing Centre for the Humanities.

Prince, Ellen (1992): "The ZPG letter: Subjects, definiteness and information structure". In Sandra Thompson and William Mann, eds., *Discourse Description: Diverse Analyses of a Fundraising Text*. Philadelphia: John Benjamins, 295-325.

Swales, John (1990): *Genre Analysis. English in academic and research settings*. Cambridge: Cambridge University Press.

Tognini-Bonelli, Elena (2001): *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tottie, Gunnel (1983): *Much about 'not' and 'nothing': a study of the variation between analytic and synthetic negation in contemporary American English*. Lund: CWK Gleerup.