Linguistic Research 30(2), 217-242

Speech in written form? A corpus analysis of computer-mediated communication*

Tim Marchand
(J. F. Oberlin University)

Marchand, Tim. 2013. Speech in written form? A corpus analysis of computer-mediated communication. *Linguistic Research* 30(2), 217-242. This paper investigates the nature of computer-mediated communication (CMC) and examine whether CMC more closely resembles written or spoken language in its structure and organization. The CMC in this paper refers to messages posted on the BBC's *Have Your Say* website over a two year period, and begins by describing how the 1.5 million word corpus was constructed from these postings. It then discusses how the characteristics of the corpus can be analyzed, with reference to research undertaken by Biber et al. (1999) into *lexical bundles.* Biber et al. compared the distribution of lexical bundles across typical written discourse (academic writing) with typical spoken discourse (conversation) and found there to be a marked contrast in the form and function of the most predominant chunks of language in these two registers. This study uses a similar methodology to determine the degree to which each kind of discourse more closely matches the CMC corpus by examining the statistical composition of various lexical bundle types in the CMC corpus. The paper concludes that while CMC shares several characteristics of both written and spoken language, it is in fact far more formulaic in its structure than either, and so properly deserves to be considered as having a register type of its own. **(J. F. Oberlin University)**

Keywords computer-mediated communication, genre analysis, lexical bundles, academic register, spoken register

1. Introduction

There have long been calls for the use of authentic materials in the language classroom, and for using topics that may engage with the learners in a motivational way (see for example Breen 1985, Little et al. 1988). One potential source of such

[*] This paper was presented at the 2012 KACL (Korean Association of Corpus Linguistics). I am grateful to Ramesh Krishnamurthy who checked an earlier version of the paper, and to the two anonymous reviewers who provided several valuable comments that helped me to clarify some key points.

material is the Internet, and in particular forums and message boards (terms which will be used interchangeably in this paper) where people can leave their opinion on contemporary issues.

One example of this *computer-mediated communication* (CMC) is the *Have Your Say* page on the BBC website (BBC 2001-2013), which I have used in some language classes for university students in Japan. The advantage of using this website is that it can be readily adapted for classroom use, and has been done so successfully in a variety of contexts (Guarda and Dalziel, 2013; Marchand and Rowlett, 2013). Over time, I started building a database of these *Have Your Say* pages and began to wonder whether the language I was collating more closely resembled *written* or *spoken* English. I was not alone in this speculation: Crystal writes 'the heart of the matter seems to be [CMC's] relationship to spoken and written language.' (Crystal 2001: 24). He goes on to cite Elmer-Dewitt (1994) who referred to Internet language as *written speech*, Hale and Scanlon (1999: 75) who advised people to 'write the way people talk', and a study by Davis and Brewer (1997: 2) which found that 'electronic discourse is writing that very often reads as if its spoken'.

As blogging and social networking are modes of communicating that many language learners use in their daily lives, digital technology is now considered to be an effective way of connecting with the current population of students (Alm, 2006; Erbaggio et al., 2010). As such, CMC is finding itself placed centrally in more and more curricula (Belz and Thorne, 2006; Belz and Vyatkina, 2008; Sun and Chan, 2012; Guarda and Dalziel, 2013; Marchand and Rowlett, 2013) and is being used evermore as a source of input for learners. So the question of its relationship to written and spoken forms of language is becoming increasingly important. As I already had the beginnings of a corpus, I decided that I wanted to use some kind of corpus analysis to answer this question, and so I formulated two research questions:

1) Does the language on the *Have Your Say* website more closely resemble written or spoken English?
2) What would be the best way to build and analyse a corpus that may answer question one?

In this paper, the Background Reading will briefly review the differences between written and spoken English before attending to some principles of corpus design, and the most suitable course for analysing the data. The Method section then outlines the *how* of the corpus construction and its analysis, while the *what* and the *why* of this analysis will be looked at in the Results and Discussion sections respectively. The paper concludes with the suggestion that the highly recurrent nature of the CMC in this study is rather more than just an intermediary between written and spoken English, and in fact can be considered as being a unique register of its own.

2. Background reading

There are many writers in the field of applied linguistics who have compared and contrasted the lexical and grammatical composition of the spoken and written registers of English (for example Crystal 1995, Jahandarie 1999, Biber et al. 1999). Very often they set up the comparison by listing a number of dichotomies, with spoken English seen to occupy one pole and the written form the opposite end. For example Crystal (1995) lists, among others, *time-bound/space-bound, spontaneous/contrived,* and *face-to-face/visually decontextualized*; Jahandarie (1999) includes *involved* versus *detached*, *fuzzy* versus *precise* and *contextualised* versus *autonomous* in his comparison; while Biber et al. (1999) contrast *impersonal* with *non-impersonal* style and *involved* with *informational* production. CMC, on the other hand, does not seem to sit consistently on one side or the other and it is for this reason that it can be considered as an intermediary register between the written and spoken. For example while forum language is clearly not *face-to-face* and therefore *visually decontextualized*, it is also more likely to be *involved* rather than *detached* seeing as it serves as a medium to air one’s opinions.

These thematic poles manifest themselves in grammatical and lexical patterns that can be expected to leave traces in a corpus of any given register. *Face-to-face* communication for instance, would be marked by a high frequency of deictic expressions, such as ‘that one’ or ‘in here’ (Crystal 2001: 26); *involved* production sees a tendency to use private verbs and second-person pronouns, which contrasts with *informational* production which consists of a large number of nouns and prepositions (Biber et al 1999: 144). So one way to determine whether the *Have* *your say* texts were more akin to spoken or written language would be to trawl a corpus for certain grammatical or lexical forms, and then compare their frequency of occurrence with the two respective registers. However two problems may arise with this approach. First, according to Biber et al 'it is not possible to reliably distinguish among registers by considering the relative distribution of individual linguistic features' (1999: 144) as there are too many to consider and impossible to know a priori which ones will be significant. For example, the past tense is used relatively rarely in conversation and academic writing, but can be frequently found in fiction.

The second problem with analysing the corpus this way relates to the not insignificant question of corpus design and construction. Studying a corpus according to certain linguistic parameters implies the use of texts that have been suitably tagged or parsed, which in turn suggests a corpus small enough to make that a feasible proposition. However according to Hunston 'the question of corpus size can be a contentious one.' (2002: 26). While some researchers such as Carter and McCarthy contend that for studying grammar in spoken language, a relatively small corpus is sufficient (1995: 143), Sinclair (1992) and Hundt et al. (2007a), among others insists that it is preferable to select from a large amount of data. As Biber et al. write, lexicographic studies require particularly large corpora since many collocations have a low frequency. They go on to say that it is not just the number of words that matters, but also the number of texts themselves: enough texts must be included to encompass variation across speakers and writers, and corpora composed of proportional samples are 'rarely useful' as they would be 'relatively homogenous' (1998: 246). In the case of building a corpus from the BBC pages then, this would suggest that a large lexical corpus, comprised of a sufficient diversity of texts (and in this case, for *texts* we can read forum *topics*) would offer the best hope of providing meaningful data for analysis.

An alternative to a tagged corpus as a means of analysis has in fact been used to examine other registers that were assumed to be intermediary between written and spoken English (see Biber et al. 2004, Biber and Barbieri 2007). In the Longman Grammar of Written and Spoken English, Biber et al. (1999) introduced the linguistic feature *lexical bundles,* which in fact resemble the *chunks, multi-word sequences,* and *word clusters* of other studies (Nattinger and DeCarrico (1992), Butler (2003), Carter and McCarthy (2006) respectively). Their operational definition of a lexical bundle (henceforth LB) is a recurring sequence of three or more words in an uninterrupted combination. They go on to define recurrent as a lexical sequence that occurs at least 10 times per million words which is spread across at least 5 different texts (to negate individual speaker/writer idiosyncrasies). The Longman grammar then discusses some of the findings for three-word and four-word LBs in their corpus of academic and conversation register where they found some stark contrasts in the structural and lexical composition of their respective LBs. According to Biber et al., the academic register (which in their study consists of academic prose taken from book extracts and research articles) is exemplary of written English, while conversation (taken from tape-recorded conversational interactions of English speakers from the UK and the USA) is prototypical of spoken language. This assertion can be supported by the fact that the two registers both seem to sit on the extremes of the thematic poles outlined at the beginning of this section. Therefore when Biber and his colleagues wanted to examine other presumed intermediary registers, they used the LB structural profiles of academic and conversational English as their yardsticks for comparison (Biber et al. 2004, Biber and Barbieri 2007). So in order to answer my first research question, I chose to build a large un-tagged corpus and use lexical bundles as the means to examine the nature of the CMC register.

3. Method

The data collected for the corpus was obtained using the Internet Archive (2008). The Internet Archive is a non-profit organisation that has since 1996 taken snapshots of websites and stored the data on to servers. Using the archive's search engine it was possible to find previous pages on the BBC's *Have Your Say* site, from where it was just a simple matter of copying and pasting the text into a Word file. I then stripped the posts of all superfluous text, such as captions accompanying pictures and details of who had originally written the comments. In this way I could ensure the anonymity of the contributors who naturally had not known their words would be used for research purposes at the time of submitting them. However I did seek and receive permission to use the data from the BBC website enquiries page.

In order to maximise the number of texts and avoid any problems of representativeness through faulty sampling, I decided to utilise all the pages from the year 2001. This was the latest year from the BBC website to have complete coverage in the Internet Archive, as some pages from all subsequent years contained links to other news discussions that were no longer accessible through the Internet Archive search engine. In all, this led to 346 different 'texts' being incorporated, ranging from topics such as 'Bin Laden: Guilty as charged?' to 'What's your favourite poem?'

Once all the data for 2001 had been collected, the Word file was converted into plain text and loaded into the Wordsmiths Tools concordance programme (Scott, 2004). A four-word cluster analysis was done, yielding the most frequent lexical bundles according to the parameters set.

The final part of the procedure was to sort the four-word lexical bundles into the same structural categories as formulated by Biber et al. in the Longman Grammar (see Table 1). This was by far the most time consuming part of the process, as there seemed to be a number of LBs that did not neatly fit into the categories. Problems arose because some bundles could be read in two ways, while others overlapped two structural categories.

Table 1. Major structural patterns of lexical bundles

Structural patterns associated with the conversation register	**Examples from the *Have Your Say* corpus**
personal pronoun + lexical verb phrase	*I do not believe*
pronoun/noun phrase + be	*it would be a*
active verb	*have a problem with*
yes/no /wh- question fragment	*what is the point*
wh- clause fragment	*when I was a*
Structural patterns associated with the academic register	**Examples from the *Have Your Say* corpus**
noun phrase + post modifier	*the way in which*
preposition + noun phrase fragment	*on the basis of*
anticipatory it	*it is clear that*
passive verb + prep. phrase fragment	*should be allowed to*
that- clause fragment	*and the fact that*
Structural patterns associated with both registers	**Examples from the *Have Your Say* corpus**
to- clause fragment	*nothing to do with*
others	*now is the time*

A common example of a lexical bundle that could be read in two ways was the *wh-question fragment* as opposed to a *wh-clause fragment*. This was resolved by manually looking at example of these bundles in the corpus, and deciding which category they predominantly belonged to. For example the bundle WHAT IS HAPPENING IN was found to be almost always a *wh-clause fragment*, as in:

> Trying to keep up tco date on **what is happening in** the world of today, I have been monitoring the different networks.

Where as the bundle WHAT IS WRONG WITH tended to be used as a question:

> For example the fact that we can't use pounds and ounces anymore hasn't changed the quantities of food that people buy for themselves and just **what is wrong with** a bendy cucumber?

Lexical bundles of these types where categorised by the most prevalent structural organisation found in the corpus.

Another problem emerged as some lexical bundles clearly overlapped two categories. For example IT IS HARD TO could be considered as an *anticipatory it + adjective phrase* or as an *adjective with to-clause fragment*. However, a closer reading of Biber el al.'s original study (1999) reveals a preferential organisation for classifying lexical bundles. The example above fits into the category of *adjectival predicates taking extraposed to-clauses* which are attested for in the semantic domains of *necessity/importance*, *ease/difficulty*, or *evaluation* (ibid.: 720). Therefore just as the Longman study categorises it is interesting TO as an *anticipatory it + adjective phrase* bundle, so too should it is hard to belong in the same grouping.

One final problem was with the lexical bundles that did not seem to fit neatly into any category, and so these were placed into the catchall *other expressions* group. Interestingly enough these bundles tended to be more idiomatic or clichéd in meaning, for example: my heart goes out / eye for an eye / innocent until proven guilty.

Altogether problematic lexical bundles accounted for less than five percent of the total found in the corpus, and so even allowing for error, they should not affect the broader results unduly.

4. Results

Looking at Figure 1 we can see that the frequency of four-word lexical bundles in the *Have Your Say* corpus (henceforth HYS) far exceeds the frequencies Biber et al. found in both the academic and conversation registers. Four-word bundles were found to occur 12,672 times per million words in the HYS corpus, compared with over 8,500 times per million in conversation and over 5,000 times per million in academic prose (note that the Longman Grammar did not actually provide exact figures for these counts). The HYS corpus in fact has a greater occurrence of typically *written* lexical bundles (6,778 per million words) than is found in academic writing, and a slightly fewer number of typically *spoken* lexical bundles than was found in conversation (a frequency of 4,673 per million as opposed to around 7,000 in Biber's register). So at first blush it looks as if the HYS corpus is composed of a mix of both typically written and spoken forms as we may have expected from the outset.

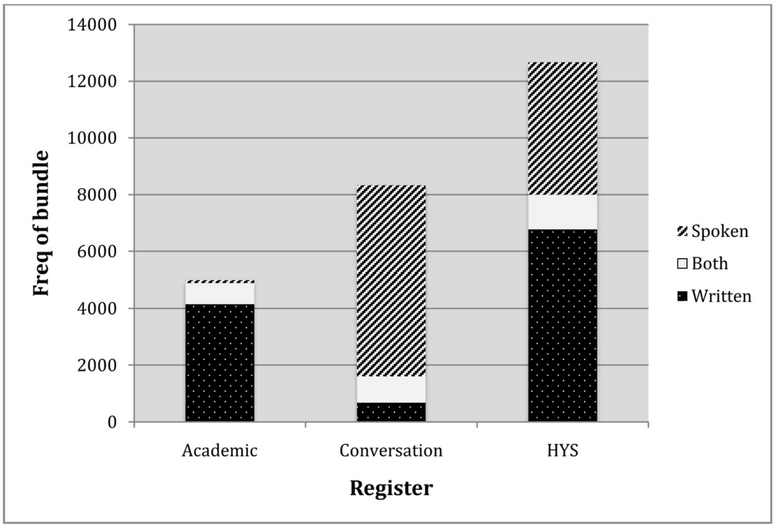

Figure 1. Frequency of lexical bundles per million words by corpus

Table 2. Distribution of lexical bundle structural patterns by corpus

Lexical Bundle structural pattern	Distribution of patterns in 'academic' register	Distribution of patterns in HYS	Distribution of patterns in 'conversation' register
spoken lexical bundles			
personal pronoun + lexical VP	0%	11%	44%
pronoun/NP + be	2%	8%	8%
active verb	0%	7%	13%
yes/no/wh- question fragment	0%	1%	12%
wh- clause fragment	0%	1%	4%
written lexical bundles			
NP + post modifier	30%	18%	4%
preposition + NP fragment	33%	21%	3%
anticipatory it	9%	3%	0%
passive verb + PP fragment	6%	1%	0%
that- clause fragment	5%	6*%	1%
common lexical bundles			
to- clause fragment	9%	8%	5%
others	6%	15*%	6%
Pearson Correlation r^2	0.76		0.08

Table 2 shows how the lexical bundle structural patterns are distributed by percentage in the three registers, with the figures for the HYS corpus aligning left or right according to which Longman register they more closely resemble. Had the HYS corpus indeed been a perfect blend between conversation and academic writing, all the figures would have aligned centrally, whereas in fact only three of the patterns do. Another three align towards the conversation register, but the majority of patterns more closely resemble academic writing in distribution. This is corroborated by the Pearson correlation scores: an r^2 of 0.76 indicates some kind of correlation between the lexical bundle structural distributions in academic writing and the HYS corpus, whilst an r^2 of 0.08 between HYS and conversation shows that there exists no statistical correlation between their respective lexical bundle distributions.

Table 3. Number of lexical bundle structural pattern types by corpus

Lexical Bundle structural pattern	Types in academic writing	Types in HYS	Types in conversation
spoken lexical bundles			
personal pronoun + lexical VP	0	77	187
pronoun/NP + be	5	61	33
active verb	0	56	56
yes/no fragment	0	7	49
wh- clause fragment	0	8	17
(and +) NP	0	41	9
quantifier expressions	0	3	4
adverbial clause fragment	4	30	10
meaningless sound	0	0	4
written lexical bundles			
NP + of-phrase fragment	69	78	16
NP + other post modifier	15	38	0
prep. phrase with of-phrase fragment	56	49	6
other prep. phrase fragment	35	90	7
anticipatory it	24	27	0
passive verb + PP fragment	16	8	0
copula be + NP/adjective P	11	49	0
that- clause fragment	13	26	3
common lexical bundles			
to- clause fragment	24	52	20
others	5	14	3
total	277	714	524
Pearson Correlation r^2	0.54		0.41

This result indicates that on a grammatically structural level at least, the HYS corpus appears closer in form to written English rather than spoken English. This is perhaps to be expected given the likely conditions under which most of the texts were constructed, but it does not provide much detail regarding the actual lexical make-up of the individual bundles.

Table 3 shows that the HYS corpus in fact uses a larger stock of lexical bundles than can be found in both academic writing and conversation, with 714 patterns identified which nearly matches the total of the other two registers combined.

Table 4. Percentage of bundle patterns shared between HYS and Longman corpora

Lexical Bundle structural pattern	Occurrence* of LBs in HYS	Occurrence* of LBs in HYS also found in Longman	Percentage
spoken lexical bundles			
personal pronoun + lexical VP	1351	393	29%
pronoun/NP + be	981	0	0%
active verb	874	30	3%
yes/no/wh- question fragment	99	0	0%
wh- clause fragment	131	0	0%
written lexical bundles			
NP + post modifier	2323	445	19%
preposition + NP fragment	2618	369	14%
anticipatory it	442	52	12%
passive verb + PP fragment	123	0	0%
that- clause fragment	857	46	5%
common lexical bundles			
to- clause fragment	969	171	18%
others	1904	0	0%
Total	12672	1507	12%

*(times / million words)

Furthermore the HYS corpus is composed of very few of the lexical bundle patterns that Biber et al. identified in both academic writing and conversation. Table 4 shows the percentage of lexical bundles that were found in both the HYS and the Longman corpora. Overall only 12% of the lexical bundles identified as being recurrent in the HYS corpus were also to be found in the Longman corpus.

This suggests that the nature of recurrent language in CMC is actually quite different to either that found in academic writing or conversation, and it's possible to examine where this difference lies by looking at table 3 and 4.

As we saw in Table 4, nearly one third of the lexical bundles of the type *personal pronouns + lexical verb phrase* found in the HYS corpus were also found in the Longman corpus, suggesting this grouping shared the most number of identical word clusters. In fact the frequency *of personal pronoun + lexical verb phrase* in conversation far exceeds that found within the BBC forum. Conversation bundles also outnumber HYS bundles in occurrence for *question fragments*, and to a lesser extent *wh- clause fragments* and *active verb fragments*, while word clusters from the BBC corpus exceed those found in the conversation register in the categories of *pronoun/noun phrase + be*, and especially *to- clause fragments* and *others* (with notably more bundles of *(and) + noun phrase* and *adverbial clause fragments*).

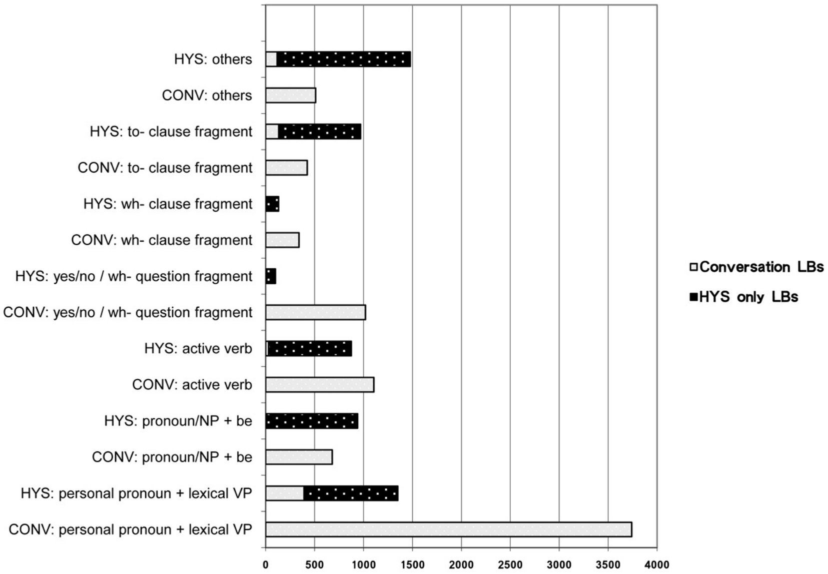

Figure 2. Overlap of "spoken" bundle types in HYS and Conversation corpora

Looking at the data for lexical bundles of a typically written nature, we see a slightly different picture. The only category where the academic register outnumbers the HYS corpus in terms of word cluster frequency is in the occurrence of *passive verb + prepositional phrase fragment*. Lexical bundles of *anticipatory it* clauses recur at approximately the same frequency, although like almost all typically written word clusters, they actually share very few of the same bundles themselves. Both *noun phrase + post-modifier* bundles and *preposition + noun phrase fragment* bundles are significantly more prevalent in the HYS corpus, and clusters of *that-clause fragments* and *to- clause fragments* are even more recurrent proportionally in the CMC-based register than the Longman academic corpus.

So in summary it appears that the HYS corpus uses a great deal more recurrent language than that found in either academic writing or conversation, and while structurally CMC seems to more closely resemble academic writing, in fact the BBC corpus displays many differences between the number and relative proportions of various LB patterns when compared to both spoken and written English. In the next section we will look at some individual lexical bundles in context to see whether it is possible to deduce why these differences occur.

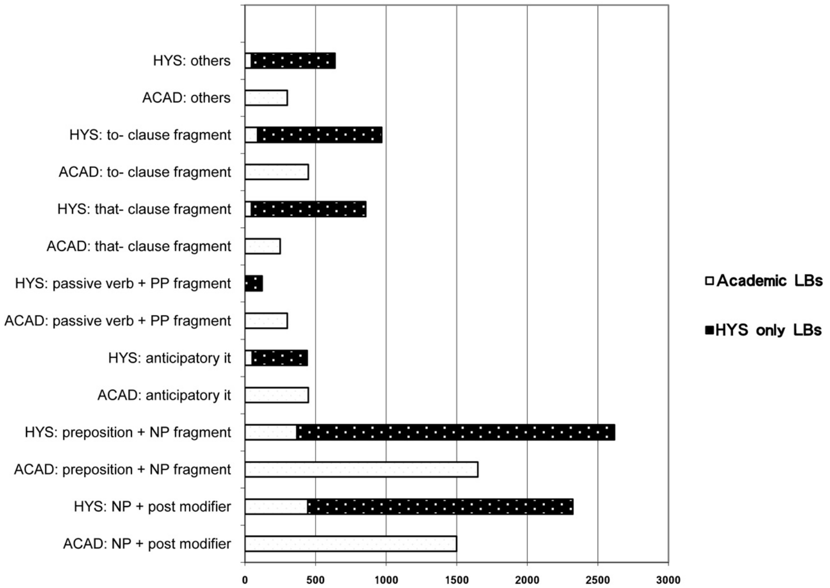

Figure 3. Overlap of "writing" bundle types in HYS and Academic corpora

5. Discussion

In the interests of brevity, I shall only focus on two lexical bundle types here: *noun phrase* + *post-modifier* ('typical' of the academic register), and *personal pronouns* + *lexical verb phrase* (a typically conversational lexical bundle). These choices are somewhat arbitrary, but should illuminate some patterns that can be generalised for other structural types too.

5.1 Analysis of *noun phrase* + *post-modifier* lexical bundles

In Biber et al.'s original study on lexical bundles, they list common LBs by structural category and discuss their findings (1999: 9). In the case of *noun phrase* + *post-modifier* LBs, Biber states, 'lexical bundles in this category cover a wide range of meanings. A few functions, however, are especially important' (ibid.: 1015). Examples of these functions from the Longman corpus can be found in the left-hand column of Table 5, with the corresponding examples and sub-classification for the HYS corpus on the right. Several patterns seem to emerge when you contrast the two sets of findings.

First, the HYS bundles contain a large number of lexically complete units, whereas all the academic LBs come from parts of extended noun phrases. For example sentences (1a) and (1b) contain the bundles THE PEOPLE OF AFGHANISTAN and THE HOUSE OF COMMONS, both of which are fully formed noun phrases, and contrast to the patterns in (Ia) and (Ib) where the bundles THE EXISTENCE OF A and THE PRESENCE OF THE are incomplete structural units that serve as discourse building blocks for nominal chunks of language. In fact Biber finds that most LBs in both academic and conversational English form such 'recurrent discourse building blocks, with the following slot being used to express the content specific to each individual situation' (p. 991). While this is also true for a large number of the lexical bundles found in the BBC corpus, the common occurrence of complete lexical units is significant, and further exemplified by looking at certain LBs reflecting news events. Sentence (2a) contains one such example, others include the war on terrorism, the fight against terrorism and even two years before the invasion of Iraq, WEAPONS OF MASS DESTRUCTION.

The bundle in sentence (2b) more closely resembles the function of describing processes or events lasting over a period of time as can be found in the academic corpus (IIa and IIb), however the 'slot' following THE DESTRUCTION OF THE was frequently filled by 'Twin Towers' or 'World Trade Centre' reflecting the topical nature of the event being discussed.

By comparing the abstract language used in academic prose with the more metaphorical bundles found in HYS, one can see that the tendency for the final slot to be filled in a more predictable manner in the BBC corpus is again repeated. Sentences (3a) and (3b) contain bundles that are completed in fairly predictable ways to form clichéd or metaphorical phrases such as 'the will of the British public' and 'the eyes of the law'. Other bundles in the corpus that are also such idiomatic building blocks include THE FACE OF THE and THE WISHES OF THE, whilst THE CYCLE OF VIOLENCE is both metaphorical and structurally complete. In contrast the abstract qualities described by the LBs identified by Biber (IIIA and IIIB) are far more generic, and so they are more likely to be used in a wide range of contexts.

Table 5. Examples of the functional use of *noun phrase with of- fragment*

	Longman 'Academic' Register	HYS Corpus
NP with *of-* fragment	Simple existence or presence: (Ia) Principles (1) and (2) lead us to interpret this regular correlation as an indication of **the existence of** a local reality. (Ib) A reheat system incurs some penalty in pressure loss due to **the presence of the** burners and flame stabilising devices.	Existing people, objects or institutions: (1a) Ask **the people of Afghanistan** what they had gone through since the Soviet occupation of that country to the multiparty/groups government and the lawlessness of their regime which was the main reason of Taleban coming to power. (1b) MP's won't even use electronic voting in **the House of Commons.**
	Processes or events lasting over a period of time: (IIa) They contributed very slowly to **the development of an** additional depletion zone.	News events: (2a) Equating **the events of September 11** to the military action in Afghanistan is complete nonsense. (2b) As shocking as **the destruction of**

	(IIb)An appropriate design should be developed so that the results do not vary beyond acceptable limits during **the course of the** program.	**the** World Trade Centre was - and I'm still struggling to believe it - if America wants to prove itself as the 'good' in the world that it so confidently boasts about, then life should continue as normal to show that the country has not been affected.
	Abstract qualities: (IIIa) The amount of rainfall considered necessary has therefore but local significance, for it depends on **the nature of the** country. (IIIb)**The use of a** constant inner diameter is often found in industrial units.	Metaphoric language: (3a) His role is to listen to **the will of the** British public but sadly we know that Mr Blair is locked into 'Tony knows best' mode. (3b) If it was reasonable in **the eyes of the** law then we would have much harsher verdicts delivered upon convicted burglars.
	Physical description of place, size and amount: (IVa) Rotating stall may lead to aerodynamically induced vibrations resulting in fatigue failures in **other parts of the** gas turbine. (IVb) He asks what each legislator might do to reduce **the total number of** incidents of injustice or unfairness.	Physical description, especially of amount and size as it relates to groups of people: (4a) Businesses will be unable to recruit people at levels below 'director with stock options' and will relocate to **other parts of the** country - or to Europe. (4b) He is obviously running from justice because he knows it is true as reflected in the minds of **the vast majority of** people throughout the world.
NP with other post-mod fragment	Identifying relationships among entities: (Va) They also varied with respect to the loci of changes envisaged, and **the relationship between the** project initiatives and the past, current, and proposed developments. (Vb) **The difference between the** two	Identifying a group within a larger group: (5a) If we were to actually make environmental issues a priority **here in the US**, we might have to start cutting back on all the waste and excessiveness that has become so ingrained in our lifestyles.

	weights is equivalent to the weight of the equal volume of water.	(5b) However, there are many **people in this country** who follow different religions.
	How a process occurs (often using *way*): (VIa) This concerned **the way in which** electrons were ejected from metals by an incident beam of light. (VIb) **The extent to which** sodium enters the melilite structures is not precisely known.	Stance bundle (often using *way*): (6a) **The only way to** counter the 'market forces' is to pay a fair price for a certified coffee that will be returned to the grower. (6b) Chancellor Schroeder's kite flying should be seen as a **step in the right** direction.

The third pattern that emerges is the centrality of *people* in many of the lexical bundles. While *noun phrase + of- fragment* bundles in the Longman corpus are often used for physical description in a variety of forms (place, size and amount − see IVa and IVb), the corresponding HYS bundles tend to function mainly by describing the amount or size of groups of people, as in sentences (4a) and (4b). Similarly sentences (5a) and (5b) are examples of *noun phrase + other post-modifier* bundles which serve to identify or highlight one group of people, often in order to contrast their situation or stance towards a topic with another.

Expressing stance is also the principle way that the word *way* is used in the HYS corpus, as in sentence (6a), the other common bundle of this sort being THE BEST WAY TO, whilst STEP IN THE RIGHT (6b) also serves a similar function This contrasts with academic prose, where the word *way* in bundles tends to help describe how a process occurs (for example VIa), as do other common bundles that Biber identifies such as 'the extent to which' in (VIb).

5.2 Analysis of *personal pronouns + lexical verb phrase* lexical bundles

At first blush, several distinctions between the two sets emerge. For one, the BBC set of LBs in this category seems to contain a wider spread of lexical verbs. Looking at the reporting in the first person of both negative and affirmative personal states, we can see that the BBC corpus contains a greater variety of main verbs than

Biber's oral register. For example when expressing a negative opinion, the bundle I DON'T THINK, as in (VIIa), was almost always used. However in the HYS data this common bundle was also joined by others such as I DO NOT SEE, I DON'T AGREE WITH (as in 7b and 7c), as well as I DO NOT BELIEVE, I FAIL TO SEE, and I DON'T SEE WHY. This is perhaps because contributors to the BBC website have more time to compose and express their ideas in a varied way than usually goes on during conversation, where repetition of lexis has often been accounted for (Carter 2004). Meanwhile affirmative LBs expressing first person personal states in the Longman corpus almost exclusively used *thought* (VIIIa and VIIIb) or *want* (VIIIc) as the main verb, but again the CMC contained a far more diverse set of verbs. Sentences (8a) and (8b) offer two examples of this variety, other bundles include I THINK THAT, I SEE NO REASON, I WONDER HOW MANY, and I HOPE THAT THE. Despite this, it is worth contrasting sentences (VIIIc) and (8c). I WOULD LIKE TO was the most common affirmative bundle of this type in HYS, and seems to function in place of any bundle containing 'I want to', a common conversational LB often found in the oral register. This perhaps suggests a slightly higher level of formality of utterance can be found in the BBC-mediated communication than in ordinary conversation.

Table 6. Examples of the functional use of *personal pronouns + lexical verb phrase*

	Longman 'Conversation' Register	HYS Corpus
Personal pronoun + lexical verb phrase	Reporting negative personal states in the first person: (VIIa) **I don't think I** could handle it. (VIIb) **I don't know what** she's got. (VIIc) Oh, **I don't want to** hear this.	Reporting negative personal states in the first person: (7a) I don't think Saa has done the right thing and **I don't think that** Argentina's anger about habitual internal corruption and mismanagement will be so easily quelled. (7b) **I do not see** the need for multiple sexual partners. (7c) **I don't agree with** taking drugs but the current drug campaign doesn't work.

Reporting affirmative personal states in the first person: (VIIIa) **I thought he was** going for three weeks. (VIIIb)**I thought I would** warn you, though. (VIIIc)Yeah **I want to go** and see that..	Reporting affirmative personal states in the first person: (8a) **I believe that the** world is big enough for all people, and compassion should be shown to all those poor women and children in desperate situations everywhere in the world. (8b) I feel sorry about the people who died, but mostly **I feel sorry for** the future of my life. (8c) **I would like to** know just how a man is supposed to defend himself against a greater number of fitter and younger attackers.
Interrogative or conditional clauses with *you + want*: (XIa) Do **you want me to** send them today? (XIb) Craig do **you want to do** something? (XIc) If **you want to come** along.	Negative conditional clauses with *you*: (9a) Ultimately I think the situation has to be that **if you don't want to** invest in your own future you have no right to expect the state to. (9b) However, smoking is an individual's choice, and when a person starts they are fully aware that they will be paying more in tax than will ever be spent on their healthcare - **if you don't like it**, don't smoke.
Reporting the speech of the first person with *say (said)*: (Xa) **I said to him,** you need it. (Xb) Anyway, **I said I would** make inquiries. (Xc) **And I said well** why don't we call	Expressing personal stance with modal or semi-modals (+ *say)*: (10a) All **I can say is**, there are plenty of destinations beyond Europe that one can go to, and for a lot less money, and far better service. (10b) As someone who flies regularly

	it that.	on business, **I have to say** that this was a bonus. (10c) As a proud young British Asian, I was horrified by the events in Oldham, but **I can't help but** feel a sense of double standards by the media and press.
	Reporting the actions of the third person with *say (said)*: (XIa) Because **she said to me**, is your brother called Anthony? (XIb) **And she said oh** I didn't know what to do.	Reporting the actions or state of the third person plural: (11a) If Britain wants to attract tourists **they are going to** have to offer more affordable prices and cleaner and more inviting eating establishments. (11b) Dictatorships are weak governments because **they do not have** the popular support of the majority of the people. (11c) Everyone wants the benefits of a consumer society, but **they don't want to** deal with the consequences of concomitant environmental degradation.
	Directives with *you* + modal or semi-modal verbs: (XIIa) Oh **you have to go** back down there. (XIIb) **You've got to have** one for a change. (XIIc) Well if you're not working **you might as well** go.	Directives with *you* or *we* + modal or semi-modal verbs: (12a) **You don't have to** fork money out to be romantic. (12b) **You only have to** look at BBC (and ITV) children's TV presenters to see where they get their loud, rude behaviour from. (12c) Perhaps **we need to get** the balance right.
	Discourse markers with *I mean* or *you*	Invoking commonality with *we*:

	know: (XIIIa) Well I – **I mean I don't** know, but I just – (XIIIb) My first year at Grange Hill, right, **you know when I** used to wear boxer shorts with no knickers on underneath. (XIIIc) **I mean you know** you say all these people are coming.	(13a) **We all know that** record companies are a 20th century institution, and that music was around long before they showed up on the scene. (13b) For God's sake **we are talking about** young human lives and not cars... (13c) Can't we just be grateful that **we live in a** country that other people actually want to come to?

Another difference is that although there appears to be a greater lexical variety of main verb for *personal pronoun + lexical verb phrase* bundles in the HYS corpus, from a functional perspective these clauses are rather more limited than their conversational counterparts. One lexical bundle subset that Biber identified was the common occurrence of *you + want* in both interrogative and conditional clauses (for example XIa-c). There is no equivalent subset in the BBC data, and in fact no interrogative clauses with *you* at all, with only some negative conditional ones such as sentences (9a) and (9b). Furthermore the verb *say* in its past form was identified as frequently being used to report the speech of the first or third person in conversation (see examples Xa-c and XIa-b). This is not the case in the HYS corpus, where - as with all main verbs for these LBs - *say* was never found in the past tense 'reporting' actions, and in fact is often used as part of a clause expressing personal stance as in (10a) and (10b). Also in sentence (9a), we can see the negative conditional clause being part of a larger piece of discourse expressing personal stance, which mirrors the function of the vast majority of all *personal pronoun + lexical verb phrase* bundles in the HYS corpus.

In Biber's study, only the verb *said* was found to be used in the function of reporting actions of the third person, combined with both the pronoun *she* of examples (XIa-b) and *he*, the other singular pronoun. There is no direct equivalent of this reporting of an individual's speech in the BBC data set, and in fact the HYS corpus contains no occurrence of singular third personal pronouns in this grouping at all. There are however several examples of *they* being used in bundles, where the *they* tends to refer to a generalised group of people, as in sentences (11a-c). This contrast of generalised versus particular usage of pronouns is repeated when comparing directives found in conversation and on the web. The *you* pronouns from sentences (XIIa-c) are most likely referring to an individual or a small number of people specified by the speaker and relevant only to the respective contexts of each conversation. However in sentences (12a-b), the *you* pronouns are of a more generalised nature, referring to people as a whole in (12a) or to an unspecified imaginary reader, as in (12b). Directives in the HYS corpus also commonly used negative conditional clauses, such as sentence (9b), as part of their construction, and one further contrast is that the HYS corpus contained some examples of directives with the *we* pronoun (see 12c).

In fact, *we* pronoun occurrence outnumbered the *you* pronoun six to four in this classification, making it the second most common personal pronoun found after *I*. This contrasts with conversation where LBs containing *you* pronouns far outnumber those with *we*, again perhaps reflecting the more immediate relationship between interlocutors in the spoken register than is found in CMC. This is also reflected in the final sub-category highlighted by Biber in his examination of *personal pronoun + lexical verb phrase* bundles, the prevalence of the discourse markers *I mean* and *you know* in conversation (see XIIIa-b). According to Carter and McCarthy (2006 539), these types of clusters reflect the 'interpersonal meanings⋯ created between speakers and listeners' and 'show how speakers are constantly monitoring...[the] assumptions about common ground between themselves and their listeners'. These discourse markers are yet another function that is not replicated in the BBC data, although the use of the *we* pronoun, it could be argued, serves to do something similar: namely, invoking a sense of commonality between the writer and their peers. In (13a), for example, the writer uses the bundle WE ALL KNOW THAT to assert what they perceive to be common knowledge, while in (13b) the contributor to the online discussion reminds people what is surely the most important point for everyone with the bundle WE ARE TALKING ABOUT. Meanwhile the writer of (13c) assumes not only common knowledge, but also common citizenship by suggesting that 'we' should be grateful that others want to live in the UK.

So it seems that the HYS corpus tends to be more context-specific and less generic than academic English in its usage of *noun phrase + post-modifier*. At the same time, according to the analysis of *personal pronoun + lexical verb phrase* bundles, the BBC corpus appears to be more generalising and less functionally diverse than conversational English. The noun phrase bundles are quite often *lexically whole* and pertain to specific topics recurrent in the news at the time, or else have their 'slots' filled more predictably or in an idiomatic fashion. This gives the impression that these LBs are more concrete, colourful and personal than their academic counterparts, whose bundles tend to reflect the measured and objective writing common to academic discourse. The frequent occurrence of stance bundles is a pattern common to both *noun phrase* and *personal pronoun + lexical verb phrase* bundles in the BBC corpus. In comparison to conversation (as per Appendix D), other functions are far less prevalent with perhaps bundles being used as *directives* being the only other significantly common function. Pronoun choices are also telling, with the oral register often using *you, he* or *she* to refer to specific people in reporting their speech or inquiring after their situation, compared to the more generalised usage of *we, you* and *they* in the HYS data.

6. Conclusion

In summary, by looking at a couple of lexical bundle categories prevalent in the academic and conversational registers respectively we could take a peek at how the HYS corpus differs from those collated in the Longman study. What stands out is that despite the diverse composition of the corpus (over 340 topics representing hundreds of thousands of individual postings), there is a great deal more recurrent language than that found in either academic writing or conversation. The frequency distribution of lexical bundles in terms of grammatical structure more closely resembles that of academic writing, but in fact the BBC corpus shares very few of the individual bundles that compose the different structural sub-categories of either registers in the Longman corpus. HYS texts also use a significantly larger stock of lexical bundle patterns, and looking at how the BBC corpus differs from the two Longman corpora reveals that HYS site tends to have significantly fewer recurrent *personal pronoun + lexical verb* bundles and *question fragment* bundles than found in conversation, and fewer *passive verb fragments* than can be found in academic writing. At the same time a great deal many more *to- clause fragments*, *adverbial clause fragments, that- clause fragments* and any kind of *noun phrase fragments* are found to occur in the HYS corpus than either in conversation or academic writing.

The results of looking at individual bundles in the discussion section suggest that CMC resembles neither writing nor speech in particular. In fact these findings go some way to justify Claridge's conclusion after building a similar corpus from message boards: 'it is necessary to regard forums as constituting a text type of their own, whose investigating should go beyond the quasi-contrastive analysis vis-à-vis speech.' (2004: 100). While the LB analysis suggests that the HYS corpus does indeed represent a unique register of its own, I would argue that using LBs as a means to explore the nature of the BBC website was reasonably effective since lexical bundles highlighted prevalent features of the register that could be broken down into manageable parts for analysis.

Future avenues of research could include comparing the HYS corpus with other 'intermediary' registers such as the classroom teaching explored by Biber in his later studies, or examining the importance of corpus size and the issue of representativeness. In spite of choosing to build a corpus of the BBC page for the entire year of 2001, traces of the dominant news stories from that year were readily apparent in some of the lexically complete sequences of recurrent language, such as THE PEOPLE OF AFGHANISTAN and THE EVENTS OF SEPTEMBER 11. As Leech (2004: 155) asserts, 'Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus − and cannot be extended to anything else'.

However, this repetitiveness has interesting implications in pedagogical terms. Krashen has put forward the case for "narrow reading" (Krashen, 2004), arguing that comprehensible input is aided by repeated exposure to familiar lexis and grammatical structures. The highly recurrent nature of the CMC in this study, together with the topicality of the news topics covered, suggest that websites such as *Have Your Say* make suitable sources of language learning materials in the narrow reading vein.

References

Alm, A. 2006. CALL for autonomy, competence and relatedness: Motivating language learning environments in Web 2.0. *The JALT CALL Journal* 2(3): 29-38.

BBC (2001-2013) *Have Your Say* [online] Available at http://www.bbc.co.uk/news/have_your_say/ [Accessed 13 January 2013].

Belz, J., and Thorne, S. (eds). 2006. *Computer-mediated intercultural foreign language education.* Boston, MA: Heinle & Heinle.

Belz, J., and Vyatkina, N. 2008. The pedagogical mediation of a developmental corpus for classroom-based language instruction. *Language Learning & Technology* 12(3): 33-52.

Breen, M. P. 1985. Authenticity in the language classroom. *Applied Linguistics* 6(1): 60-68.

Biber, D., Conrad, S., and Reppen, R. 1998. *Corpus Linguistics Investigating Language Structure and Use.* Cambridge: CUP.

Biber, D., Johanssonn, S., Leech, G., Conrad, S., and Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Biber, D., Conrad, S., and Cortes, V. 2004. If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371-404.

Biber, D., and Barbieri, F. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263-286.

Butler, C. S. 2003. Multi-word sequences and their relevance for recent models of Functional Grammar. *Functions of Language* 10: 179-208.

Carter, R. 2004. *Language and Creativity: The Art of Common Talk.* London: Routledge.

Carter, R., and McCarthy, M. 1995. Grammar and the spoken language. *Applied Linguistics* 16: 141-158.

Carter, R., and McCarthy, M. 2006. *Cambridge Grammar of English: A Comprehensive Guide: Spoken and Written English Grammar and Usage.* Cambridge: CUP.

Claridge, C. 2007. Constructing a corpus from the web: message boards. In Hundt, M., N. Nesselhauf, and C. Biewer. (eds), (2007b). *Corpus Linguistics and the Web.* New York: Rudopi.

Crystal, D. 1995. *The Cambridge Encyclopaedia of the English Language.* Cambridge: CUP.

Crystal, D. 2001. *Language and the Internet.* Cambridge: CUP.

Davis, B. H., and Brewer, J. P. 1997. *Electronic Discourse: Linguistics Individuals in Virtual Space.* Albany: State University of New York Press.

Elmer-Dewitt, P. 1994. Bards of the Internet. *Time,* 4 July, 66-67.

Erbaggio, P., Gopalakrishnan, S., Hobbs, S., and Liu, H. 2012. Enhancing student engagement through online authentic materials. *International Association for Language Learning Technology* 42(2): 27-51.

Guarda, M., and Dalziel, F. C. 2013. Focus on form in CMC: Using written learner data to foster language and pragmatic skills in communicative contexts. Paper presented at the Compiling and Using Learner Corpora Conference, Padua (Italy) 16-17 May 2013.

Hale, C., and Scalon, J. 1999. *Wired Style: Principles of English Usage in the Digital Age.* New York: Broadway Books.

Hunston, S. 2002. *Corpora in Applied Linguistics.* Cambridge: CUP.

Hundt, M., N. Nesselhauf, and C. Biewer. 2007a. Introduction. In Hundt, M., N. Nesselhauf, and C. Biewer. (eds), (2007b). *Corpus Linguistics and the Web.* New York:

Rudopi.

Hundt, M., N. Nesselhauf, and C. Biewer. (eds). 2007b. *Corpus Linguistics and the Web.* New York: Rudopi.

Internet Archive. 2008. Available at http://www.archive.org/index.php [Accessed 14 March 2009].

Jahandarie, K. 1999. *Spoken and Written Discourse: A Multi-disciplinary Perspective*. Stanford: Ablex.

Krashen, S. 2004. The case for narrow reading. *Language Magazine* 3(5): 17-19.

Leech, G. 2007. New resources, or just better old ones? In Hundt, M., N. Nesselhauf, and C. Biewer. (eds), (2007b). *Corpus Linguistics and the Web.* New York: Rudopi.

Little, D., S. Devitt, and D. Singleton. 1988. *Authentic Texts in Foreign Language Teaching: Theory and Practice.* Dublin: Authentik.

Marchand, T., and Rowlett, B. 2013. Course design in the digital age: learning through interaction with news-based materials. *Language Education in Asia* 4(1).

Nattinger, J., and DeCarrico, J. 1992. *Lexical phrases and Language Teaching.* Oxford: OUP.

Scott, M. 2004. WordSmith Tools version 4, Oxford: Oxford University Press. ISBN: 0-19-459400-9.

Sinclair, J. M. 1992. The automatic analysis of corpora. In Svartik, J. (ed), (1992). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82.* Berlin: Mouton de Gruyter.

Sun, Y., and Chang, Y. 2012. Blogging to learn: becoming EFL academic writers through collaborative dialogues. *Language Learning & Technology* 16(1): 43-61.

Svartik, J. (ed). 1992. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82.* Berlin: Mouton de Gruyter.

Tim Marchand
J. F. Oberlin University
3758 Tokiwa-machi, Machida-shi,
Tokyo 194-0294, Japan
E-mail: timmarchand@gmail.com

Received: 2013. 03. 26
Revised: 2013. 08. 06
Accepted; 2013. 08. 06