I don't want to and *don't get me wrong*: Lexical bundles as a window to subjectivity and intersubjectivity in American blogs

Federica Barbieri

Swansea University

Abstract

Blogs are one of the most prominent genres of Web 2.0; yet, research on their linguistic characteristics is limited. This study contributes to addressing this research gap by investigating lexical bundles in American blogs. Lexical bundles are units of discourse structure which can reveal a great deal about the unique linguistic characteristics and communicative functions shaping registers. Extraction of four-word bundles in a corpus of American blogs reveals, firstly, that lexical bundles are relatively uncommon in blog writing. Analyses of discourse function and grammatical patterns show that blogs rely mainly on stance expressions, which often encapsulate first person reference (e.g., *I don't want to*), thus reflecting the focus on self-expression and subjectivity which characterizes this register. Like in conversation, bundles in blogs tend to be verb-phrase based. But blogs also rely substantially on referential (e.g., *a lot of people*) and narrative expressions (e.g., *I got to see*), and thus share characteristics of literate registers and fiction writing. In sum, lexical bundles in blog writing are characterized by a unique combination of features which reflect two underlying forces: mode and communicative purpose.

1 Introduction

Research on the linguistic characteristics of internet genres or registers is in its infancy. Given the challenges inherent to the identification of web-based

registers (Biber, Egbert, and Davies 2015), this is unsurprising. Blogs, however, are amongst the oldest genres of the internet, and, along with wikis, Facebook and Twitter posts, internet forums, and chat, one of the most perceptually salient to users (Biber et al. 2015). The rise of blogging is also generally regarded one of the most acclaimed features of Web 2.0, the internet 'phase' which relies on user participation and 'collective intelligence' (O'Reilly 2009).

Typically defined as 'frequently modified web pages in which dated entries are listed in the reverse chronological sequence' (Herring et al. 2005: 142), blogs as we know them today first appeared in the mid-/late 1990s,[1] but the milestone date marking their exponential rise in popularity is 1999, when the first free blogging software (e.g., *Blogger, LiveJournal, Xanga*) became available (Blood 2002; Herring et al. 2005; Baron 2008). And it was in the early 2000s that major news stories were first broken on blogs. The 2000s is also reportedly when blogs, a clipping of *web-log*, actually started to be called 'blogs' (Grieve et al. 2010: 304). By 2006, blogs had become the fastest growing genre of the internet (Herring and Paolillo 2006: 440). Biber et al. (2015) found that blogs accounted for about a quarter of their corpus, a random sample (over 48,500 documents) of the Corpus of Global Web-

[1] While most scholars recognize sites such as Jorg Barger's 'Robot Wisdom', Dave Winer's 'Scripting News', Cameron Barrett's 'CamWorld' as the earliest weblogs, some 'purists' trace the birthdate of blogs back to 1991, when Tim Berners-Lee's (father of the WWW) launched 'What's New', a webpage that listed (and linked to) all existing websites at the time (Blood 2002; Baron 2008; Herring et al. 2005; Myers 2010).

based English (GloWbE), a corpus comprising 1.9 billion words in 1.8 million web documents. As they put it, blogs might just be "the quintessential register of the searchable web" (p. 40).

Yet, while multi-dimensional analyses of internet corpora have begun to paint a picture of the linguistic make-up of web-based registers (Titak and Roberson 2013; Biber and Egbert 2016), including blogs (Grieve et al. 2010; Hardy and Friginal 2012), we know relatively little about the lexico-grammatical characteristics of even these early internet registers. Corpus-driven approaches to phraseology have proven especially effective in the study of natural discourse. Thus, this paper aims to address the current research gap by investigating patterns of formulaic language in blogs, specifically lexical bundles – expressions such as *I don't want to, is going to be, those of you who.*

Lexical bundles (sometimes also referred to as 'n-grams', 'clusters', 'chunks', 'formulaic sequences', or simply 'bundles') are simply the most frequent recurring sequences of three or more words in a register (Biber et al. 1999). As such, they reflect a purely corpus-driven approach. Although they typically do not represent structurally complete units, are not idiomatic in meaning, and are not particularly perceptually salient, lexical bundles serve important discourse functions in texts. They are 'building blocks of discourse' (Biber, Conrad, and Cortes 2004: 401) carrying basic communicative functions, which provide frames for the expression of new

information in the larger phrases or clauses that follow them. A related point is that lexical bundles carry traces of the lexico-grammatical characteristics and communicative purposes of texts. For example, in a seminal study, Biber et al. (2004) showed that university classroom talk is characterized by a wider range of types and higher frequency of lexical bundles than casual conversation and academic prose. The study also showed that American classroom talk is characterized by an approximately equal distribution of the three main functional categories of bundles, namely stance expressions, referential expressions, and discourse organizers. These findings reflect the complex communicative purposes of classroom teaching, which combines the informational focus typical of academic prose with the expression of personal stance and interpersonal meanings typical of casual conversation. Partington and Morley (2002) compared lexical bundles in White House press briefings with news interviews, showing how press briefings are more formulaic and repetitive, and how lexical bundles can reveal the metaphors or discourses of this speech event. In other words, lexical bundles are a powerful tool for the understanding of the unique characteristics of registers, that is situated language varieties (Biber 1988).

Over the past two decades or so, lexical bundles have been investigated in a wide range of registers, but especially intensely in academic writing, particularly in research articles (Cortes 2004; Hyland 2008; see Hyland 2012, for a review) and in the production of novice (Cortes 2004) and

second language writers (Cheng and Baker 2010; Ädel and Erman 2012; see Paquot and Granger 2012, for a review). Bundles have also been examined in a wider range of academic registers, including several spoken registers (Biber 2006; Biber and Barbieri 2007), and university classroom teaching has been studied particularly intensely (Biber et al. 2004; Nesi and Basturkmen 2006; Csomay 2013). This work has made the crucial contribution of demonstrating that use of lexical bundles cannot be explained merely by speech and writing differences; rather, it reflects also the communicative purposes of the register.

Thus, over the past decade the study of lexical bundles has been extended to a wide range of non-academic spoken and written registers, from highly specialized written registers, such as legal genres (Breeze 2013) and Early Middle English medical genres (Kopaczyk 2013), to registers of wider consumption, such as popular television series (Bednarek 2011) and hotel websites (Fuster-Márquez 2014). Fuster-Márquez showed how, in this genre of computer-mediated B2C (business-to-customer) interaction, bundles convey stance, as well as reference to textual elements or physical or abstract entities, to a far higher extent than academic written registers.

Fuster-Márquez's (2014) detailed work on the phraseology of hotel websites, coupled with some of the contributions in this volume, however, stands out in the general dearth of research on the phraseology of web-based genres, and to my knowledge no study has looked at formulaic sequences in

blog writing. Lexical bundles have been shown to be units of discourse structure which can reveal a great deal about the unique linguistic characteristics and communicative functions shaping registers. As Biber and Barbieri put it, “each register employs a distinct set of lexical bundles, associated with the typical communicative purposes of that register” (265). It seems likely, therefore, that lexical bundles might also uncover unique linguistic features and discourse functions of blog writing. Accordingly, the present study begins to tackle the current research gap by investigating lexical bundles in American blogs. American blogs are a good place to tap into blog writing because blogging arose in the US, and at least up to the mid-2000s, reading and writing blogs were eminently American practices (Baron 2008: 109).

Commentators and grassroots bloggers alike have consistently described blogging as a thoroughly individualistic, intimate form of self-expression (Herring et al. 2005): ‘an outbreak of self-expression’, in Blood’s (2000) words; a place ‘to let off steam’ and ‘get it out there’, in the words of the bloggers in Nardi et al.’s (2004) ethnography. Thus, an important goal in this study is to explore the extent to which lexical bundles encode salient communicative purposes of blogs, such as writer stance and self-expression. In doing so, I draw on the notions of subjectivity and intersubjectivity as pragmatic constructs which capture ‘the complex dynamic nature of self-expression’ (Fitzmaurice 2004: 428).

2 Background: the language and discourse of blogs

While blogs are generally easily recognized by end-users (Biber et al. 2015), there is still limited consensus, among scholars, as to the nature of blogs, that is whether they are a genre or a medium (Miller and Shepherd 2009), as well as on whether blogs are 'an emergent, or a reproduced genre' (Herring et al. 2005: 157), that is whether they reproduce or adapt some 'off-line antecedent' (144) or whether instead they are "native" to the web.[2] Myers (2010) points out that 'blogs are not like personal home pages, because they are regularly updated, and they are not like diaries, because they are built around links, and they are not like wikis, which involve many authors collaborating on one text' (2). Herring et al. (2005), in contrast, note that blogs are characterized by features typical of other web-based genres, including personal homepages and community blogs, and propose that rather than evolving from a single genre, blogs are actually 'a hybrid of existing genres', in other words a unique combination 'of the source genres they adapt' (160).

[2] A discussion of the different positions in the debate on whether blogs are a reproduced, adapted, or distinctive new genre is beyond the scope of this paper. This theme is covered in Herring et al. (2005), Mauranen (2013), Miller and Shepherd (2009) and several other works cited here. Readers interested in reproduction, adaptation, and emergence of (new) genres on WWW can turn to Crowston and Williams (1997).

Given that blogs are an emerging internet genre,[3] it is not surprising that early studies were mostly concerned with situating the genre of blogs within CMC and classifying blogs into types, typically from the perspective of genre analysis and content analysis. Thus, Krishnamurthy (2002, cited in Herring et al. 2005) proposed a taxonomy including four main types along two dimensions: personal vs. topical, and individual vs. community blogs. Puschmann (2009) draws a distinction between 'ego blogging' and 'topic blogging'. In a content analysis of 203 randomly-selected blogs, Herring et al. (2005) found that personal journal blogs and filter blogs (i.e., blogs containing links to other webpages, annotated with commentary by the blogger/editor)[4] were the most common types, but they also found types not represented in Krishnamurthy's taxonomy, such as k-logs (knowledge logs), that is blogs consisting in information and observations centered around a particular topic, project, or product. Perhaps the most striking finding of Herring et al.'s (2005) study is that, despite the fact that online journals on *LiveJournal*, *DiaryLand*, and *Xanga* were deliberately excluded from the sample, personal journal blogs were by far the most common blog type,

[3] While blogs were indisputably an 'emerging genre' in the mid-/late 2000s, they can arguably still be considered an 'emerging genre' today, a decade later – if anything, in relative terms, that is, compared to more established, fixed, and unified genres, such as the novel, the research article, the recipe, the memo, the sports broadcast, the travel guide, etc. But also because like many internet genres, blogs are inherently fluid and ever-evolving in response to the affordances of the medium, as shown, for example, by the rise of sub-genres (j-blogs, video blogs, photo blogs, audio blogs, etc.).

[4] According to Blood (2000), this is the format of early weblogs, which provided a valuable filtering function for readers, as the web had basically been "pre-surfed" for them.

representing over 70% of the sample. This finding points to the quintessentially personal nature of blogs. Grieve et al.'s (2010) multi-dimensional analysis of blogs corroborates these findings, showing that blogs in the corpus tended to fall into two main sub-types: personal blogs and thematic blogs.

The focus on defining the status of blogs and classifying them vis-a'-vis pre-existing genres has continued in recent years, though shifting to domain-specific blogs. Thus, Mauranen (2013) examined two research science blogs (looking both at posts and threads) to illustrate their evolving genre, situating it within the array of genres in the sciences, and tracking its different features to pre-existing genres. Mauranen argues that science blogs are best regarded as a 'genre cluster' rather than one individual genre, because they comprise different features which fulfill different purposes, and these different features can be traced back to pre-existing genres (e.g., commentaries in blogs can be traced back to pamphlets, editorials, and opinion columns). She concedes, however, that blogs have introduced new practices.

Early studies focusing on classifying and positioning blogs as a unique genre have been recently followed by a new line of inquiry examining blogs within particular communities of practice (e.g., academic blogs, executive blogs) or disciplinary communities (e.g., popular science blogs), or domain-specific blogs (e.g., science). For example, Ruiz-Garrido and Ruiz-Madrid

(2011) and Puschmann (2010) looked at executive and corporate blogs respectively. Luzón (2011) and Luzón (2013a) explored different aspects of academic blogs; Luzón (2013b) investigated rhetorical and discursive strategies that science bloggers use to convey and recontextualize scientific discourse, and to engage the diverse readership of science blogs.

These studies have typically focused on rhetorical or discursive strategies, and discourse modes, and with few exceptions (e.g., Puschmann 2010) have not focused on lexico-grammatical features. Nonetheless, these studies reveal important features of blogs representing particular disciplines or communities of practice. A key theme emerging from this body of research is the salience of self-disclosure, even in disciplinary-specific blogs or blogs from specific communities of practice. For example, Luzón (2013a) found that narratives are pervasive in academic blogs (both by individuals, and community blogs). Academic bloggers use both narratives of personal experience and narratives focusing on the discipline. Luzón claims that academic bloggers use self-disclosure and proximity with the reader to create 'participatory narratives' in which the writers' voices 'mingle with the stories of others who share their academic interests' (191). Luzón (2011) examined discursive strategies of 'social behavior' and 'anti-social behavior' in 11 academic blogs from different disciplines. These strategies are actually an assorted set of rhetorical and more genuinely linguistic features. Notably, strategies of social behavior include strategies such as

expression of oral discourse, self-disclosure, inclusive pronouns, and others which have typically been covered under the rubrics of involvement (Chafe 1982; Biber 1988; Barbieri 2015) or engagement (Hyland 2005). Self-disclosure has been found to be a salient feature of executive blogs as well, where it is used as a rhetorical and persuasive strategy (Ruiz-Garrido and Ruiz-Madrid 2011). Executives do self-disclosure through a range of lexical and discursive strategies: an informal conversational style, a 'positive tone', self-mention, and lexical features contributing to supporting 'the interactional objective of blogs', namely hedges, boosters, attitude markers (120).

Self-disclosure and proximity in English are strictly related to first and second person pronoun use. And indeed first person references are a salient feature of executive blogs (Ruiz-Garrido and Ruiz-Madrid 2011); inclusive pronouns have been found to be common in both academic (Luzón 2011) and science blogs (Luzón 2013b), while corporate blogs use first and second person pronouns more frequently than multi-genre corpora such as the BNC (Puschmann 2010). Further, Bondi and Diani's (2015) cross-linguistic comparison of evaluative semantic sequences in English and Italian in multi- domain-specific corpora, comprising blogs on business, entertainment, politics, and sports amongst other topics, shows that first person reference and stance verbs (e.g., *think, guess, love*) are amongst the

top keywords in English, highlighting 'bloggers' propensity for subjectivity and self-expression' (120).

Taken together, these studies suggest that the linguistic 'make up' of blogs, regardless of specific domain, is strongly characterized by features associated with self-disclosure, subjectivity, the expression of personal feelings, as well as lexico-grammatical features typically associated with casual conversation and informality. Only few of these studies though have investigated actual lexico-grammatical features, while most have looked at discursive or rhetorical strategies, usually in a limited number of blogs.

However, these indications of the salience of this dimension in blogs are corroborated by findings from large-scale corpus-based studies looking at a comprehensive set of lexico-grammatical features. These studies apply multi-dimensional (MD) analysis (Biber 1988), a model of linguistic analysis which relies on factor analysis to identify systematic patterns of co-occurring features in a corpus representing different registers or texts. The patterns of co-occurring features (which are referred to as 'factors') are interpreted functionally as 'dimensions' associated with particular communicative functions, based on the assumption that "linguistic co-occurrence patterns reflect underlying communicative functions" (Conrad and Biber 2001: 24). Grieve et al.'s (2010) MD analysis of American English blogs showed that nearly 95% of blogs in the corpus can be classified into two clusters (i.e., groups of texts which are maximally

similar) which are strongly to moderately characterized by features associated with involvement and personal focus (shown, e.g., by first person pronouns, discourse particles, hedges), focus on addressee (e.g., second person pronouns) and with narrative discourse modes (e.g., past tense, place and time adverbials). These results are magnified in Hardy and Friginal's (2012) cross-varietal comparison of blogs and opinion columns in American and Filipino English, which shows that on all four dimensions identified in Grieve et al. (2010), American blogs have higher scores than Filipino blogs. Thus, American blogs display a stronger personal focus, focus on addressee, thematic variation, and narrative orientation than do Filipino blogs (and both American and Filipino opinion columns).

Perhaps even more compelling are findings from two MD analyses of internet registers. Titak and Roberson (2013) showed that compared to other web registers (e.g., emails, reader comments, online newspaper articles, FB/Twitter posts), blogs – more than any other register – are characterized by a personal, narrative focus (cfr. high scores on Dimension 1), and – second only to emails – by an involved interactive style (cfr. high scores on Dimension 2). In the most comprehensive study of internet registers to date, Biber and Egbert (2016) conducted a MD analysis of a corpus including 27 user-identified registers (and 8 macro-registers), which included different kinds of blogs: news blogs, personal blogs, travel blogs, informational blogs, personal opinion blogs, and religious blogs. Consistent with the

findings in Grieve et al. (2010) and Titak and Roberson (2013), Biber and Egbert's MD analysis revealed that personal blogs have high scores on three dimensions which are similar in distinguishing oral from literate web-registers: blogs are characterized by features of oral, highly involved production (Dimension 1), oral elaboration (Dimension 2), and oral narrative (Dimension 3).

In sum, findings from MD analyses reveal that blog writing in American English is strongly characterized by linguistic features associated with oral language, as well as features marking 'personal focus' and 'focus on addressee' (Grieve et al. 2010; Titak and Roberson 2013). The discourse functions 'personal focus' and 'focus on the addressee' reflect what other researchers have called self-disclosure and proximity, and are best captured by the related notions of subjectivity and inter-subjectivity (Fitzmaurice 2004).

Lexical bundles have been shown to be 'important building blocks in discourse' (Biber and Barbieri 2007: 270) which provide a window on the linguistic and communicative profile of registers. It seems likely, therefore, that lexical bundles might help identify distinctive communicative functions of American blogs, and reveal traces of discourse functions identified in previous research, such as stance, subjectivity, and intersubjectivity. Subjectivity is broadly understood here as the linguistic marking of self-expression, and intersubjectivity as 'the representation of speaker stance as

addressee stance' (Fitzmaurice 2004: 429). Subjectivity and intersubjectivity are encoded most explicitly by first and second person reference. Accordingly, the present study examines person reference within lexical bundles, as well as their discourse function and grammatical structure.

3 Methods

3.1 Corpus for analysis

The present study is based on a 2.2 million word corpus of American blogs representing blogging in American English at the turn of the century (Grieve et al. 2010). The corpus indeed includes texts, from 2003 to 2005, representing 500 personal and thematic blogs by bloggers from the 50 US states, identified via the index, globeofblogs.com. Specifically, the corpus comprises 500 texts, and each text in the corpus includes several blog posts extracted from a single blog. Each text thus represents a single blog, and may be regarded as 'sub-corpus'. Blogs average 4,500 words in length, with the shortest amounting to 1,099 words and the longest 9,864 words (305).

In order to ensure balanced regional and demographic distribution, blogs were selected so as to evenly represent all 50 US states (i.e., 10 blogs were selected for each state), as well as both female and male writers and different age-groups. Topic was not controlled (Grieve et al. 2010: 306).

3.2 Identification of lexical bundles

Lexical bundles were extracted using the text analysis freeware AntConc 3.4.1 (Anthony 2013). The study here focused on four-word bundles occurring at least 20 times in at least 5 different texts in the corpus. The study focused on four-word bundles, following Biber et al. (1999) and most subsequent lexical bundles studies, because five-word sequences tend to be much more infrequent than four-word bundles, while three-word bundles tend to be included in longer bundles (e.g., *I went to* is included in *I went to the*) and are harder to interpret (Cortes 2004; Csomay 2013). Following Biber et al. (2004) and many others, contracted words (e.g., *don't*) were considered one word.

In previous research on lexical bundles, frequency and dispersion criteria for cut-off points have been somewhat arbitrary (Biber and Barbieri 2007: 267; Fuster-Márquez 2014: 92). Biber et al. (2004) – in a study based on corpora comparable in size to the current one – took a rather conservative approach,

limiting the analysis to bundles occurring at least 40 times per million words, in at least 5 different texts. Such a conservative cut-off point has been deemed necessary for spoken registers, such as conversation or university classroom talk, which rely extensively on lexical bundles (see Biber et al. 2004); studies focusing on written registers, however, have relied on far less conservative cut-off points, ranging from 20 to 25 per million words (Cortes 2004; Chen and Baker 2010; Ädel and Erman 2012).

After exploring the overall frequency of lexical bundles in the American blogs corpus, which revealed that bundles are relatively rare in blogs (see Section 4), the frequency cut-off point was set to minimum 20 occurrences in the corpus, distributed in at least 5 texts/blogs, which is roughly equivalent to 10 occurrences per million words.[5] In this way, the frequency cut-off point adopted here is in line with that in Biber et al. (1999) (i.e., 10 bundles per million words). This is a far more relaxed cut-off point than the one in Biber et al. (2004), but it was considered appropriate for this study's exploratory goals. With these requirements, after exclusion of bundles automatically generated by blogging software (e.g., *posted at 5 pm by, at am comments Thursday*), 460 different bundles (i.e., bundle types) occurring in at least 5 different blogs were retained for analysis. Most bundles however occurred in many more texts: 79% occurred in at least 20 texts, 19%

[5] Biber (2006: 175, fn.9) adopted a similar approach for the analysis of university textbooks.

occurred in 15-19 texts, while only 2% occurred in 5-14 texts. Thus, dispersion here is rather robust.

To allow comparisons of the distribution of bundles in blogs with other registers (e.g., conversation, academic prose), the frequency of occurrence of bundles in American blogs was normed to one million words. The distribution of types (i.e., the number of different lexical bundles), however, was not normed because vocabulary type distributions are not linear; hence, as Biber and Barbieri (2007) point out, 'it is not possible to directly normalize the number of lexical bundle types to a rate per million words' (268, fn. 5; see also Biber 2006).

3.3 Functional classification of lexical bundles

This exploratory study of bundles in blogs comprises three types of analyses: functional analysis (analysis of discourse function), analysis of person reference, and analysis of grammatical structure. The analysis of discourse function adopts Biber et al.'s (2004) functional taxonomy (see also Conrad and Biber 2005; Biber and Barbieri 2007), which comprises three main categories: stance expressions, discourse organizers, and referential expressions. Stance bundles express feelings, attitudes, or assessments towards the following proposition. Common stance expressions

include: *I don’t want to, I don’t know if, I would like to, I have to say, I want to be, I don’t think I:*

(1) I just don’t want to get hurt again. And, *I don't want to* hurt her either.

(2) So last night we had these perfect moments that *I don't know if* I can describe, but it sums up why I am so crazy about Scott.

(3) And finally, *I would like to* thank all of the MacWorld Boston attendees throughout the years […].

Discourse organizers provide information about the structure of discourse, such as introducing, clarifying, and elaborating on topics, and identifying or focusing on entities, individual, or groups. They include bundles such as: *on the other hand, nothing to do with, as well as the, if you have a, here are a few, take a look at:*

(4) The ICJ, *on the other hand*, lacks an effective enforcement mechanism; in turn, […].

(5) If they don't, FEDERAL LAW prohibits such sales. That's why we have the NICS system *in the first place*.

(6) I recently came across a poet that is worth mentioning if you have not heard of her and am a closet poet. *Take a look at* some of her work.

(7) *Here are a few* things you CAN try to help you lose weight and stay motivated: […]

Referential expressions typically identify an entity or single out some characteristic or feature of an entity (in the extra-linguistic context) as especially important. They include bundles such as: *the end of the, in the middle of, one of the most, at the same time, in front of the, one of my favorite.*

(8) The current dip in the polls is due to the following: Iraq, Katrina, and the Border. But it's not *the end of the* man's Presidency.

(9) We all just basically show up *at the same time* and do our own thing.

(10) Fajitas is *one of my favorite* family meals.

(11) I still remember coming out of the movie theatre crying after seeing this film. Spike Lee is *one of the most* thought provoking writer/directors of his time and this film proves it.

Biber et al.'s (2004) taxonomy also includes sub-functions for the main functional categories. Thus, stance expressions are sub-divided into two main categories: epistemic stance bundles (e.g., *I don't know, I don't think*

so) and attitudinal/modality stance bundles, which in turn are subdivided into desire bundles (e.g., *I don't want to*), obligation/directive bundles (*you have to be*), intention/prediction bundles (*I'm not going to*), and ability (*to be able to*). Discourse organizers are classified as topic introduction/focus bundles (*if you look at*) or topic elaboration/classification bundles (*on the other hand*). Referential expressions are sub-classified as identification/focus (*that's one of the*), specification of attributes (*a little bit of*), imprecision (*or something like that*), and time/place/text/multi-functional reference bundles (*at the same time, in front of me, in the middle of*).

This is a very fine-grained taxonomy; in fact, some of the subcategories are further sub-divided into more specific categories: for example, sub-categories of stance bundles are also classified according to person reference, namely as personal (*you might want to*) or impersonal (*it is important to*); referential expressions of specification of attributes are further classified as quantity specification (*a lot of people*), tangible framing expressions (*in the form of*), and intangible framing attributes (*when it comes to*).

Functionally classifying bundles is challenging because of the inherent multi-functionality of some of them. Like many functional taxonomies, Biber et al.'s (2004) taxonomy has been criticized for lack of reliable and clear-cut criteria (Ädel and Erman 2012). Some of the sub-categories are

arguably appropriate for more than one macro-category. For example, identification/focus bundles (e.g., *that's one of the, one of the things*) – a sub-category of referential expressions – could be regarded as discourse organizers. And indeed, this problem has led to some inconsistency in previous research. This particular category is treated as referential expression in Biber et al. (2004), Chen and Baker (2010), and Ädel and Erman (2012), and as discourse organizer in Cortes (2004) and Biber and Barbieri (2007).

Given the exploratory goals of the present study, bundles here were classified only in the macro-categories (stance expressions, discourse organizers, referential expressions). Identification/focus bundles were classified as referential expressions, following Biber et al. (2004), and not as discourse organizers, contra Biber and Barbieri (2007), in order to maximize comparability with the registers in Biber et al. (2004).

4 Lexical bundles in American blogs

For the purposes of this exploratory study, 460 different four-word bundles (i.e., bundle types) occurring at least 20 times and distributed in at least 5 different blogs were retained for analysis of discourse function, person

reference, and grammatical pattern. Before turning to the functional and grammatical analyses, let's take a look at the frequency of bundles in blog writing. Table 1 shows that there are 6,071 (tokens) and 93 different (types) four-word bundles occurring at least 40 times in the corpus. This is equivalent to a normed rate of occurrence of 2,759 bundles per million words.[6] However, this is probably a somewhat inflated normed rate, because it is based on a raw frequency count obtained with a cut-off point of 40 occurrences in the corpus, not per million words. With a cut-off point of minimum 80 occurrences in the corpus, the corpus yields 2,073 bundles, or 942 bundles per million words – a very low rate of occurrence, with a likely excessively restricted number of types (17 types). Setting the cut-off point between 40 and 80 (i.e., 60) will likely give a closer estimate. With a cut-off point of 60, there are 3,414 four-word bundles (36 types) in the corpus, or 1,551 bundles per million words.

Table 1: Distribution of bundles in American blogs

Minimum # of occurrences per bundle	# of different bundles (types)	# of bundles in the corpus	# of bundles / million words
80	17	2,073	942

[6] These rates of occurrence do not include bundles automatically generated by blogging software (e.g., *posted at pm by*), which were manually removed.

60	36	3,414	1,551
40	93	6,071	2,759
20	460	15,650	7,113

A comparison of this rate of occurrence (1,551/million words) with the frequency of bundles in conversation and academic registers in Biber et al. (2004, Figure 3) reveals that bundles in blogs are rather infrequent compared to spoken registers: they are about one fourth as common as in conversation, and less than one fifth as common as in classroom teaching. More surprising though is that they are also more infrequent than in university textbooks, and only about half as common as in academic prose. Bundles are overall far less common in blogs than in most of the registers in Biber et al. (2004) even considering the possibly 'inflated' normed rate of 2,759 bundles per million words. Specifically, they are less than half as common as in conversation, and about one third as common as in university classroom teaching, but also less frequent than in academic prose and only slightly more frequent than in textbooks.

If comparisons of (normed) rates of occurrence based on different distributional criteria should be taken cautiously, comparing lexical bundle types (i.e., the number of different lexical bundles) is especially problematic because vocabulary type distributions are not linear. This effectively means

that the number of lexical bundle types cannot be converted to a normed rate (Biber 2006; Biber & Barbieri 2007). Accordingly, normed rates of lexical bundle types are not provided here. Sections 4.1., 4.2, and 4.3 report findings for the functional analysis, the analysis of person reference, and the structural analysis of bundles.

4.1 Functional characteristics of lexical bundles

Following Biber et al.'s (2004) taxonomy, lexical bundles were classified into three main categories: stance expressions, discourse organizers, and referential expressions. Biber et al. (2004; see also Conrad & Biber 2005) also include a fourth, minor category, for 'special conversational functions', which comprises politeness formulae (e.g., *thank you very much*), simple inquiry phrases (e.g., *what are you doing*) and reporting phrases (e.g., *I said to him*). This category was not adopted here; however, a fourth category became necessary for misfits, that is bundles that did not fit into the three main categories of stance, discourse, and referential expressions. These were mostly verb phrase-based bundles including a dependent clause fragment (e.g., *to take care of, to get out of*), verb phrases which typically serve a narrative function (e.g., *I was a kid, turned out to be, I was in the, and I went to, I used to be*), and conversational, formulaic noun phrases

(e.g., *a lot of fun*). Because the predominant function of these bundles appears to be narrative, I call them 'narrative expressions'.

The functional analysis revealed that stance expressions are the most common type of bundles in blog writing, accounting for 45% of the 460 bundles analyzed here (Table 2). Referential expressions are the second most common functional type, accounting for 39%, while discourse organizers are surprisingly rare in blogs, representing a negligible 3%. Finally, narrative expressions account for a sizeable 14% (Table 2).

Table 2: Distribution of lexical bundles across functional categories in American blogs

Functional Category	**Number of Bundles**	**Percent**
Stance Expressions	207	45%
Discourse Organizers	13	3%
Referential Expressions	177	38%
Narrative Expressions	63	14%
Total	460	100%

A closer look at stance bundles reveals that they tend to be personal expressions of epistemic stance, desire, and intention/prediction including

first person pronoun *I*, while obligation/directive bundles, which usually tend to include second person pronoun *you* (see Table 3, Biber et al., 2004), are far less frequent. Table 3 lists stance expressions occurring among the top 100 most common bundles. The top 100 bundles occurred at least 39 times. A cursory look at Table 3 reveals the overwhelming presence of first person reference among stance bundles – and, consequently, among bundles overall – a pattern which will be confirmed by the analysis of person reference (Section 4.2).

Table 3: List of stance expressions among top 100 bundles.

*Numbers in parentheses refer to raw frequency of the bundle in the corpus

I am going to (188)*, *I was going to* (122), *I don't want to* (121), *to be able to* (104), *is going to be* (92), *I don't know if* (89), *I would like to* (87), *if you want to* (78), *as much as I* (74), *was going to be* (70), *I don't know what* (68), *I have to say* (68), *I have no idea* (66), *going to be a* (65), *going to have to* (63), *the fact that I* (62), *I don't know how* (56), *I want to be* (56), *I feel like I* (53), *I wish I could* (52), *I don't have to* (51), *I need to get* (50), *will be able to* (50), *I was able to* (49), *I'm not going to* (48), *I don't think I* (46), *I don't know why* (45), *I am not a* (44), *I didn't want to* (44), *I just want to* (43), *I thought it was* (43), *let me tell you* (43), *should be able to* (42), *you don't have to* (42), *it would be a* (41), *and I have to*

(40), *are going to be* (40), *don't get me wrong* (40), *for the sake of* (40), *I have to go* (39)

Lexical bundles tend to be structurally incomplete (e.g., *I was going to, at the end of*), but an interesting feature of stance bundles in blog writing is that they are sometimes – perhaps often – structurally complete. Examples include: *I don't want to, I would like to, I have no idea, I wish I could, I don't have to, I'm not going to, I don't know why, I didn't want to, I just want to, I thought it was, let me tell you, you don't have to, don't get me wrong.* Although of course these bundles are not always used as structurally complete units, the point here is that they can be (see examples below), and often are used as structurally complete units in casual conversation. The fact that they occur so frequently in blogs arguably contributes to the conversational flavor of the language of blogs.

(12) Now that I am older, I could get away with reading them if I wanted to. But *I don't want to*.

(13) Did you know I could sing? Well, I can. *I'm not going to*, though.

(14) People seem to misunderstand this vehicle completely and *I don't know why*.

(15) *Don't get me wrong*, disco bowling can be fun. Very fun.

Referential expressions are the second most common type of bundles in blog writing. It is noteworthy that, with 207 occurrences in 150 different blogs, the top blog bundle (*the rest of the*) is a referential expression, and that referential expressions represent 9 of the 20 top bundles (*the rest of the, the end of the, in the middle of, at the same time, at the end of, for the first time, when it comes to, the middle of the,* and *most of the time*).

Referential expressions appear to be distributed across three major sub-categories, namely attributes specification (*the rest of the, a lot of people*), expressions of time/place/text reference (*the end of the, in the middle of, at the same time, for a long time, from time to time, the top of the,*), and identification/focus expressions (*one of the most, is one of the, one of my favorite*), while imprecision bundles appear to be absent. The three sub-categories represented (examples 16-18) serve the narrative function of blog writing, a distinctive feature of blogs (Biber and Egbert 2016; Grieve et al. 2010; Titak and Roberson 2013). Time and place adverbials (here represented by time/place reference bundles) are indeed stereotypical features of narrative discourse (Biber and Egbert 2016) and situated reference (Biber 1988); likewise, attributes specification and identification/focus expressions support the construction of situated reference in discourse.

(16) Even today work was mobbed or anything. We got *a lot of people* who were desperate to find shortcuts/back roads to try to get there in time.

(17) We officially started moving in today. The former owners cleaned it up real nice for us too. :) Rachel's family, the Nunley's, are in town right now because their oldest son flew in from Hawaii today *for the first time* in over a year (he's in the coast guard).

(18) *For those of you* who despise my driveway, for those of you who despise my kingdom on the hill: Burrillville has declared a state of emergency. The clear river isn't looking very clear right now.

The small proportion of discourse organizing bundles in blogs partly reflects the fact that identification/focus bundles here were classified as referential expressions, following Biber et al. (2004), and not as discourse organizers. This in turn contributes to the higher rate of referential expressions. Nevertheless, the low incidence of discourse bundles is remarkable. (See examples 4-7, and 19-20 below.)

(19) If she wants to flatten a building, she will. It has *nothing to do with* us, our country's foreign policy, or that sinful Superbowl flash of Janet Jackson's 40-year-old boob.

(20) My friend Phil is adamant about watching a movie fullscreen. Now *if you have a* small TV, I can understand that widescreen

won't look good (it's too small), but *if you have a* DVD player, you owe it to yourself to get a movie widescreen.

Narrative expressions are more common than discourse organizers. Narrative expressions among the top 100 bundles include: *I went to the, to go to the, to go back to, to take care of, to get out of, turned out to be.* Looking at the bigger pool of narrative expressions reveals that, structurally, these bundles tend to include a verb (see above), and the verb is often in past tense (e.g., *I went to the, turned out to be, I used to be, we went to the, I got to see, so I decided to*). This suggests that many of these bundles support a narrative function.

(21) On my lunch break today *I went to the* Homewood library to read.

(22) So I started worrying and trying *to take care of* her....and that's where I am at now living with my parents […]

(23) The jack-ass *turned out to be* quite rude and told me that I "deserve to rot in hell" because I refused to let him ruin Tina's life.

Another structural characteristic of these bundles is that they tend to be phrasal or phrasal prepositional verbs: in addition to most bundles listed here above, consider *to get back to, trying to figure out, to come up with, to get rid of, what was going on, to figure out how, come up with a, get out of the, I got to see, to check out the, to keep up with.* Phrasal and phrasal prepositional verbs contribute to the informality of blog writing because

they are conversational features; however, what's perhaps even more interesting is that phrasal prepositional verbs are most common in fiction writing (Biber et al. 1999, Ch. 5).

(24) I didn't even have a chance *to come up with* my story, other than to tell him what Dr. Brenda told me […]

(25) I am happy that *I got to see* him 2 weekends in a row!

Finally, some bundles in this category (*he told me that, asked me if I*) serve a reporting function. Biber et al. (2004) included these bundles in the 'special conversational functions' macro-category, which comprises bundles occurring only in conversation. But since reported speech is a key feature of narratives, reporting bundles can also be viewed as supporting a narrative function:

(26) She *asked me if I* was serious, because it's a drug and you don't want to do drugs.

(27) *He told me that* he hadn't thought of that painful time in years.

4.2 Person reference in lexical bundles

The second type of analysis involved classifying bundles for person reference, that is whether the bundle includes person reference (first, second, or third person) or not. This analysis can shed light on subjectivity

and intersubjectivity in blogs discourse. Findings reveal that nearly half of lexical bundles in blogs do not include person reference (Figure 1): bundles with no reference indeed account for 47% (N = 214) of the 460 bundles included in this analysis. When bundles do include person reference, however, it is usually a first person pronoun: bundles with first person reference account for 40% (N = 184); 8% (N = 39) of bundles include third person reference, while only 5% (N = 23) of bundles include second person reference.

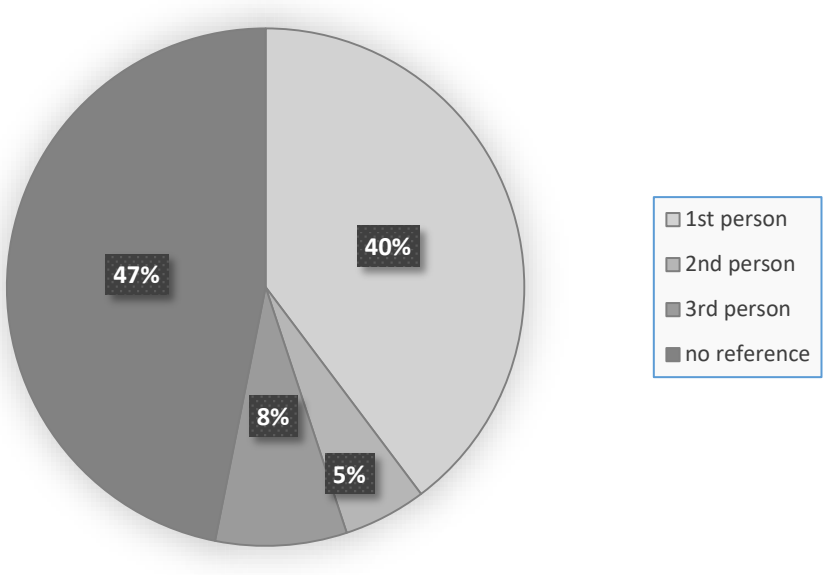

Figure 1: Proportional distribution of person reference in blogs bundles

Bundles with first person reference are overwhelmingly those with a reference to the self (i.e., *I*, more rarely *me*), while only a handful of bundles

include first person plural *we* (*we are going to, we went to the, we were going to, then we went to, we went back to*). As shown above (Table 3), bundles with first person (singular) reference tend to be stance bundles. Bundles including second person reference are far more infrequent, but appear to be more spread across stance and referential expressions. Common stance bundles with second person reference include: *if you want to, you don't have to, as you can see, you know what I, you are going to.* Common referential expressions with second person reference include: *for those of you, those of you who*. A few bundles with second person reference function as discourse organizers: *if you have any, if you are a:*

(28) Furthermore, *if you have any* questions following the reading the report, you will have full access to our customer service who can answer all of your unanswered questions.

(29) Ah, OK, before I forget: *If you are a* New York artist, you should know about these grants.

4.3 Structural characteristics of lexical bundles

Finally, lexical bundles were analyzed for grammatical structure. Biber et al. (1999) and Biber and Conrad (2005) identify 12 structural patterns of lexical bundles, and compare the distribution of lexical bundles in conversation and academic prose across these structural patterns. This is a very fine-grained taxonomy, which however in some cases may hide more basic structural

patterns. For example, patterns like ‘adverbial clause fragment’ or ‘wh-clause fragment’ typically include a verb (e.g., *if you want to*), but not always (e.g., *as soon as I*). This means that by merging a structural type that includes a verb with one that does not include a verb, one loses sight of the more basic distinction between VP-based and non-VP-based structural patterns. Given the exploratory goals of the present study, following Fuster-Márquez (2014), lexical bundles were classified into three main categories: NP-based, PP-based, and VP-based. NP-based bundles are noun phrases with or without a post-modifying fragment, such as: *the back of the, the other side of, a couple of days, a couple of weeks, the New York Times*. PP-based bundles are bundles which start with a preposition followed by a noun phrase with or without a post-modifying fragment (Fuster-Márquez 2014: 96), such as *at the end of, in the first place, for the rest of, for a long time, in front of me*. VP-based bundles are any bundles that contain a verb, such as *I am going to, to be able to, I went to the, when it comes to, if you want to.*

Analyses of grammatical structure revealed that lexical bundles in blogs are overwhelmingly VP-based: bundles including a verb account for 64% (N = 294) of the 460 top bundles, while with 81 and 85 types respectively, both NP-based and PP-based bundles account for 18% of the top 460 bundles (Figure 2).

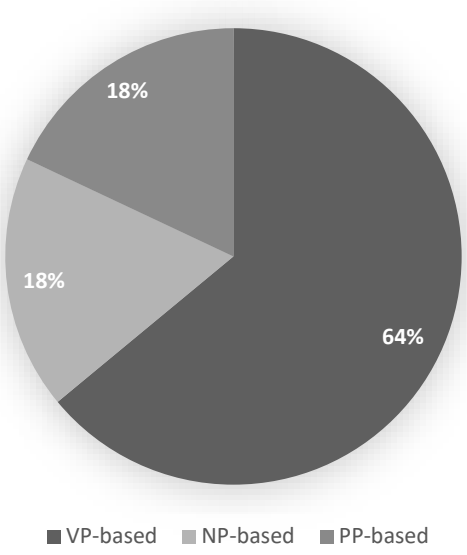

Figure 2: Proportional distribution of structural types in blogs bundles

The finding that bundles are predominantly verb phrase-based, while noun phrases and prepositional phrases are relatively uncommon, suggests that blog writing relies on verbs rather than nouns, nominalizations, and post-nominal modification. This is consistent, of course, with the finding that bundles often include first person reference (first person reference automatically requires a verb). This is also consistent with research showing that blog writing tends to be characterized by a narrative and personal style, resulting from the clustering of first person pronouns, discourse particles, past tense, time and place adverbials, third person pronouns, speech act and communication verbs, and so on. (Biber and Egbert 2016; Grieve et al. 2010).

5 Discussion and conclusion

The present study aimed to examine the linguistic characteristics of lexical bundles in blogs and the extent to which they reflect key communicative purposes of blog writing, such as self-disclosure and proximity, which are best captured by the notions of subjectivity, involvement, and intersubjectivity. Results revealed that blogs rely heavily on stance bundles, especially personal expressions of epistemic stance, desire, and intention, which tend to include explicit reference to the self (i.e., person pronoun *I*). Coupled with the high frequency of bundles including first person reference, this finding reflects the overall style of blog writing in American English, which is strongly characterized by features associated with involved production, personal focus, and interactive discourse (Grieve et al. 2010; Titak and Roberson 2013).

Referential expressions are the second most common bundle type in blogs. This finding is related to the high frequency of bundles with no person reference, and points to a different dimension of blog writing, which has been shown to be characterized by features necessary for thematic development and narrative style (Grieve et al. 2010).

The extremely low frequency of discourse organizers is surprising. While this may partly reflect the fact that identification/focus bundles were classified as referential bundles, it was expected that discourse organizers would be more frequent in blog writing, given that they are fairly common in conversation (and even more frequent in classroom teaching) (Biber et al. 2004; Figure 5), and blogs are strongly characterized by features of oral registers (Biber and Egbert 2016). On the other hand, discourse organizers are very uncommon in textbooks and academic prose, registers which heavily rely on referential expressions (Biber et al. 2004; Biber and Barbieri 2007). Thus, lexical bundles in blog writing are similar to spoken registers (conversation, classroom teaching) in their reliance on stance bundles, but also similar to written academic registers in their use of formulaic referential expressions.

The study also revealed the presence of a small but noteworthy proportion of bundles associated with another distinctive aspect of blog writing: narrative style (Biber and Egbert 2016; Grieve et al. 2010; Titak and Roberson 2013). Narrative style is characterized by past tense verbs, perfect aspect verbs, third person pronouns, reporting/communication verbs (Biber 1988), but of course personal narratives are also characterized by first person pronouns. The narrative dimension of blog writing revealed by narrative expressions and by bundles with third and first person reference is consistent with research showing that narratives are pervasive in blogs, even

in blogs from particular disciplinary communities (Luzón 2013a). The function of narrative bundles in blog writing clearly deserves further investigation.

Comparisons of the frequency of bundles in blogs and other registers revealed that, surprisingly, lexical bundles are very infrequent in blog writing – more infrequent than in conversation, but also than in academic written registers. On one hand, the lower frequency than in conversation might have been predictable, given that bundles have been shown to be far more common in spoken registers than in written (academic) registers (Biber et al. 1999; Biber et al. 2004). Blogs however are strongly characterized by linguistic features typical of personal narratives, and of involved, interactive discourse – characteristics that make them distinctively different from other web-based registers with an informational focus, such as newspaper articles or encyclopedia articles (cfr. Dimensions 1 and 2 in Titak and Roberson 2013, and Dimensions 1-3 in Biber and Egbert 2016), and more similar to conversation and oral web-based registers, such as songs or interviews (Biber and Egbert 2016). Based on this background, the finding here that bundles are more infrequent in blogs than in written registers such as academic prose and university textbooks was unexpected.

At the same time, blogs are very similar to newspaper articles relative to features characteristic of descriptive and opinionated discourse (cfr. Similar scores on Dimension 3 and Dimension 4), such as public and reporting

verbs, *that*-clauses, third person reference, past tense verbs (Titak and Roberson 2013).

Taken together, these findings suggest that lexical bundles in blog writing are characterized by a unique combination of features of different registers, which reflects the influence of two forces: mode and communicative purpose (Biber and Barbieri 2007). Like oral registers, blogs rely on stance bundles to serve one of the main communicative purposes of blogs: the expression of writer's involvement and subjectivity. At the same time, they are similar to literate registers in their reliance on referential expressions. Blogs also use bundles serving a narrative function, a feature that makes blogs reminiscent of fiction. From a structural perspective, lexical bundles in blogs are predominantly verb phrase-based, like in conversation, but rely on nouns and prepositional phrases more than do conversational bundles. In addition, a sizeable proportion of bundles includes phrasal prepositional verbs – verbs especially frequent in fiction and conversation.

The present study has shown, once again, that lexical bundles are an effective tool to uncover the defining, yet sometimes unexpected, characteristics of registers. Previous research has shown that lexical bundles are different from other lexico-grammatical features in that they respond to both physical mode (speech vs. writing) and communicative purpose, while main lexico-grammatical features (e.g., verbal and clausal features, complex noun phrase features) are influenced primarily by mode (Biber and Barbieri

2007; Biber et al. 2004). This point effectively explains the unique combination of functional and structural characteristics of lexical bundles in blog writing in American English. More research is necessary, however, to better understand the communicative purposes of blogs, their linguistic characteristics – including structural and functional characteristics of lexical bundles – and the intersection between communicative purpose and linguistic features.

Finally, the present study focused on lexical bundles in blog writing in American English at the turn of the century. Hardy and Friginal (2012) showed that American blogs differ from Filipino blogs in their linguistic characteristics: American blogs are more extreme than Filipino blogs in personal focus, addressee focus, and narrative orientation. Lexical bundles have been shown to encode speaker or writer stance (Biber et al. 2004), and the present study has shown that blog writing heavily relies on stance expressions. The precise way that stance is expressed, however, might vary across English varieties. Precht (2003) showed that American and British speakers express stance in fundamentally different ways: American speakers favor affect, while British speakers favor evidentiality. Taken together, these findings suggest that the language of blogs might differ across English varieties, and differences might be reflected in lexical bundles. Thus in future research it would be interesting to compare lexical bundles across blogs representing different English varieties.

ACKNOWLEDGMENTS

I am grateful to Eric Friginal and Jack Grieve for granting me access to the American blog corpus.

REFERENCES

Ädel, Annelie, and Britt Erman. 2012. "Recurrent word combinations in academic writing by native and non-native speakers of English. A lexical bundles approach." *English for Specific Purposes* 31:81-92.

Anthony, Lawrence. 2013. AntConc (Version 3.4.1).

Barbieri, Federica. 2015. Involvement in university classroom discourse: Register variation and interactivity. *Applied Linguistics, 36*(2), 151-173.

Baron, Naomi. S. (2008). *Always on*. Oxford: Oxford University Press.

Bednarek, Monika. 2011. The language of fictional television. A case study of the 'dramedy' *Gilmore Girls*. *English Text Construction, 4*(1), 54-84.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge and New York: Cambridge University Press.

Biber, Douglas. 2006. *University language. A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas, & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26,* 263-286.

Biber, Douglas, Conrad, Susan, & Cortes, Viviana. 2004. *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Biber, Douglas, & Egbert, Jesse. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics, 44*(2), 95-137.

Biber, Douglas, Egbert, Jesse, & Davies, Mark. 2015. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora, 10*(1), 11-45.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, & Finegan, Edward. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Blood, Rebecca. 2000. Weblogs: A history and perspective. In *Rebecca's Pocket* (September 2000-September 2013) Retrieved from http://rebeccablood.net/essays/weblog_history.html

Bondi, Marina, & Diani, Giuliana. 2015. *I am wild about cabbage*: evaluative 'semantic sequences' and cross-linguistic (dis)similarities. *Nordic Journal of English Studies, 14*(1), 116-151.

Breeze, Ruth. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics, 18*(2), 229-253.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written*

language: Exploring orality and literacy (pp. 35-53). Norwood, New Jersey: Ablex.

Chen, Yu-Hua, & Baker, Paul. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30-49.

Conrad, Susan & Biber, Douglas. (eds.). 2001. *Variation in English: Multi-dimensional studies*. Harlow, England: Pearson Longman.

Conrad, Susan, & Biber, Douglas. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*(56-71).

Cortes, Viviana. 2004. Lexical bundles in student and published academic writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423.

Crowston, Kevin, & Williams, Marie (1997). Reproduced and emergent genres of communication on the World-Wide Web. *The Information Society, 16*(3), 201-215.

Csomay, Eniko. 2013. Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics, 34*(3), 369-388.

Fitzmaurice, Susan (2004). Subjectivity, intersubjectivity and the historical construction of interlocutor stance: from stance markers to discourse markers. *Discourse Studies, 6*(4), 427-448.

Fuster-Márquez, Miguel. 2014. Lexical bundles and lexical frames in the language of hotel websites. *English Text Construction, 7*(1), 84-121.

Grieve, Jack, Biber, Douglas, Friginal, Eric, & Nekrasova, Tatiana. 2010. Variation among blogs: A multi-dimensional analysis. In A. Mehler, S. Sharoff, & M. Santini (Eds.), *Genres on the web: Computational models and empirical studies: Text, speech and language technology* (pp. 45-71). Dordrecht: Springer.

Hardy, Jack A., and Eric Friginal. 2012. "Filipino and American online communication and linguistic variation." *World Englishes* 31 (2):143-161.

Herring, Susan C., & Paolillo, John C. (2006). Gender and genre variation in blogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Herring, Susan, Scheidt, Lois Ann, Wright, Elijah, & Bonus, Sabrina. 2005. Weblogs as a bridging genre. *Information Technology and People, 18*(2), 142-171.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies, 7*(2), 173-192.

Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Kopaczyk, Joanna. 2013. Formulaic discourse across Early Modern English medical genres: Investigating shared bundles. In A. H. Jucker (Ed.), *Meaning in the history of English: Words and texts in*

context (Vol. 148, pp. 257-300). Amsterdam and New York: John Benjamins.

Krishnamurthy, S. 2002. *The multidimensionality of blog conversations: the virtual enactment of September 11*. Paper presented at the Internet Research 3.0, Maastricht, October.

Luzón, María José. 2011. 'Interesting post, but I disagree': Social presence and antisocial behaviour in academic weblogs. *Applied Linguistics, 32*(5), 517-540.

Luzón, María José. 2013a. Narratives in academic blogs. In M. Gotti & C. S. Guinda (Eds.), *Linguistic Insights. Narratives in Academic and Professional Genres* (Vol. 172, pp. 175-193). Bern: Peter Lang.

Luzón, María José. 2013b. Public communication in science blogs: Recontextualising scientific discourse for a diversified audience. *Written Communication, 0*(4), 428-457.

Mauranen, Anna. 2013. Hybridism, edutainment, and doubt: Science blogging findings its feet. *Nordic Journal of English Studies, 13*(1), 7-36.

Miller, Caroline R., & Shepherd, Dawn. 2009. Questions for genre theory from the blogosphere. In J. Giltrow & D. Stein (Eds.), *Theories of genre and their application to internet communication* (pp. 263-290). Amsterdam: John Benjamins.

Nardi, Bonnie. A., Schianto, Diane. G., Gumbrecht, M., & Swartz, L. 2004. 'I'm bloggin this.' A closer look at why people blog. *Communications of the ACM, 47*(12), 41-46.

Nesi, Hilary, & Basturkmen, Helen. 2006. Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-304.

O'Reilly, Tim. 2005. What is Web 2.0. Design patterns and business models for the next generation of software. Retrieved from http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1

Paquot, Magali, & Granger, Sylviane. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130-149.

Partington, Alan, & Morley, John. 2002. From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk (Ed.), *Practical Applications in Language and Computers--PALC 2013* (pp. 179-192). Frankfurt am Main: Peter Lang.

Precht, Kristen. 2003. Stance moods in spoken English: Evidentiality and affect in British and American conversation. *Text, 23*(2), 239-257.

Puschmann, Cornelius. 2009. *Diary or megaphone? The pragmatic mode of weblogs*. Paper presented at the Language in the (New) Media: Technologies and Ideologies, Seattle, WA.

Puschmann, Cornelius. 2010. *Thank you for thinking we could:* Use and function of interpersonal pronouns in corporate web logs. In H. Dorgeloh & A. Wanner (Eds.), *Approaches to syntactic variation and genre* (pp. 167-194). Berlin and New York: Mouton the Gruyter.

Ruiz-Garrido, Miguel, & Ruiz-Madrid, M. Noelia. 2011. Corporate identity in the blogosphere: The case of executive blogs. In V. K. Bhatia & P. Evangelisti Allori (Eds.), *Discourse and identity in the professions: Legal, corporate, and institutional citizenship* (Vol. 149, pp. 103-125). Bern: Peter Lang.

Titak, Ashley, & Roberson, Audrey. 2013. Dimensions of web registers: an exploratory multi-dimensional comparison. *Corpora, 8*(2), 235-260.