LINGUISTIC VARIATION AND IDENTITY REPRESENTATION IN PERSONAL BLOGS: A CORPUS-LINGUISTIC APPROACH

GONG WENGAO

NATIONAL UNIVERSITY OF SINGAPORE

2009

LINGUISTIC VARIATION AND IDENTITY REPRESENTATION IN PERSONAL BLOGS: A CORPUS-LINGUISTIC APPROACH

GONG WENGAO
(M.A., NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF ENGLISH LANGUAGE & LITERATURE

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

First and foremost, I would like to thank my supervisor A/P Dr Vincent Ooi. It is him who ignited my interest in language practice in internet-based communication settings. It is also through his modules that I have learned how to deal with online discourse data which are quite non-conventional in many ways and how to use them for academic research. As an experienced supervisor, he knows very well when to leave me free exploring for themes of my interest and when to bring my attention back to things with value. He hardly tells me exactly what to do, but offers constructive suggestions and insightful clues for further development. This style suits my personality and age very well. I have genuinely enjoyed the freedom, independence, and trust that my supervisor has given me during my PhD studies.

Secondly, my thanks should go to my committee members: A/P Dr Bao Zhiming and Dr Peter Tan, for their sincere concerns and encouragements. My thanks also go to A/P Dr Lionel Wee, A/P Dr Michelle Lazar, A/P Dr Kay O'Halloran, Dr Mark Donohue, Dr Benny Lee, and A/P Dr Shi Yuzhi. What I have learned from their modules has contributed greatly to the completion of my thesis.

Thirdly, my sincere thanks go to my friends who have given me great moral support and feedback about my research ideas over the past several years. They are: Hong Huaqing, Zhang Ruihua, Paramjit Kaur A/P Karpal Singh, Liu Yu, Li Songqing, and Zhang Yiqiong.

Fourthly, I would like to thank my family, especially my better half, Zhou Hongxia, for their sacrifice, tolerance, and support. I owe them too much. For that it may take the rest of my life to repay. Special thanks go to my son, Zixuan, from whom I have learned quite a bit of the netlingo. His interest in my research and his concerns about what kind of career I could make out of researching online discourses are the two most important driving forces which have sustained me through the past almost five years.

Fifthly, sincere thanks go to the internal and external examiners and the panel members of my oral examination for their valuable feedback about my thesis.

Last but not least, I would like to thank the National University of Singapore for offering me the scholarship, without which my PhD studies would not be possible. Thanks also go to the friendly and hard working staff members of the NUS Central Library. Their service has made my stay in this university a memorable one.

Table of Contents

Summary

Adopting a Wmatrix-based multi-variable approach supplemented with qualitative analysis, I have conducted a comprehensive investigation about how identities are represented or reflected through linguistic variations in personal blogs. The language of personal blogs as revealed by the corpus constructed for this research has displayed certain features which are different from both spoken and written texts. Bloggers in this research have employed seven major strategies to realize orthographic variation. They are: unconventional contracted forms, abbreviations, letter repetition, orthographic representation of paralinguistic features, misspellings, phonetic spellings, and innovative use of special symbols like the asterisk. Apart from orthographic variation, bloggers have also displayed variations in terms of the use of lexicological strategies, slanguage use, preference for semantic domains, and the use of grammatical and pragmatic features. Bloggers' frequent use of non-conventional contracted word forms, unconventional letter repetition, and orthographic representation of paralinguistic features, their use of slanguage words and neologisms related to emergent Internet culture, their use of new or less conventional grammatical features (such as the new usage of the plural marker, the use of *like* as a quotative complementizer, and the use of accusative case of pronouns in subject positions), and their use of new pragmatic markers and vague expressions are found to be closely related to their expression of age-related identities, especially the representation of youth identity. Bloggers' frequent use of initials and acronyms representing laughing and laughter, words with unconventional letter repetition, orthographic representation of paralinguistic features, the asterisk as action markers, and interjections is found to be very closely related to their expression of gender-related identities, especially that of the female gender. The use of neologisms related to IT and video and Internet games, on the other hand, is more closely related to the male gender. The use of slanguage has displayed two opposing patterns in gender representation. For bloggers of the younger generation (those below 25), males outperform the females. For bloggers of the more mature generation (those above 30), females outperform the males. The different preferences for semantic domains displayed by bloggers reveal a close relation between age and gender and the blogging content, reflecting the different social and psychological realities that bloggers are facing. Bloggers' preference for certain slanguage words and grammatical and pragmatic features reflects their regional identity. Apart from reflecting the collective identities of bloggers, linguistic variation is also able to demonstrate bloggers' individual identities, which are more easily observable in their use of new lexical items of nonce formation. This research also shows that deviating from the established writing norms and transplanting oral discourse features into blogging are two major means for bloggers to represent various aspects of their identities. It has also demonstrated the necessity of adopting an eclectic framework in understanding the multi-faceted nature of the concept of identity and an eclectic analysis approach in capturing the various linguistic strategies for identity representation in blogging texts.

List of Tables

List of Figures

List of Abbreviations

BBS: Bulletin board system

BNC: British National Corpus

CIC: Cambridge International Corpus

CLAWS: Constituent Likelihood Automatic Word-tagging System

CMC: computer-mediated communication

CMDA: computer-mediated discourse analysis

COLT: (the Bergen) Corpus of London Teenage Language

EBC: English Blog Corpus

IBC: Internet-based communication

IT: information technology

LDCE: Longman Dictionary of Contemporary English

LL: log likelihood

MWE: multi-word expression

OEFs: orthographically engineered forms

POS: part of speech

UD: Urbandictionary

UK: the United Kingdom

US: the United States

USAS: UCREL Semantic Analysis System

WebCA: Web Content Analysis

Chapter 1 Introduction

This chapter first introduces the research goals of the thesis. Following that, it presents some background information about the development of Internet-based communication and its influence on people's daily language use. After that, a brief discussion about some of the unique features of blogs is presented, followed by a short account of the relationship among linguistic variation, identity, and personal blogs. It concludes with the thesis structure.

1.1 Research goals

The rapid development and popularization of the Internet technology over the past two decades has created many new settings for language use which were simply unimaginable before the advent of the Internet. Among them, personal blogs are a recent example. With the affordances of being both a publishing tool and a social communication platform, personal blogs have rapidly gained enormous popularity among young people worldwide. Despite the multi-semiotic potential of personal blogs, text is still the most important means of expression for blogging. Influenced by the personal nature of the content, the absence of others-imposed editing, and the potential for interpersonal communication, personal blog texts tend to be quite informal in style. It may not be too exaggerated to say that the writing in personal blogs is a type of "written vernacular." The popularity of personal blogs offers a special window for language researchers to observe and investigate how variations are realized through textual means and what functions these written variations are put to perform in representing their identities. This thesis is an attempt in this regard.

The thesis aims to achieve one primary goal and two secondary ones. The primary goal is to describe the strategies employed by bloggers from the United States and the United Kingdom in realizing linguistic variations and explore how these variations are related to bloggers' representation of various aspects of their identities. The two secondary goals are: testing the applicability of the corpus linguistics approach in identity representation research and identifying the challenges that non-conventional written data such as personal blogs could pose on the theory and practice of corpus linguistics. To be more specific, the thesis attempts to answer the following questions:

1. What strategies do bloggers employ to realize linguistic variations in a new written genre of personal blogs?
2. What sorts of social, psychosocial, and cognitive realities are reflected in these linguistic variations? In other words, what information can these variations reveal about bloggers' identities?
3. Methodology-wise, how useful could a corpus-linguistic approach be in revealing bloggers' efforts in identity representation?
4. What challenges could personal blogs pose on mainstream automated language-processing tools and the theory and practice of corpus linguistics?

As linguistic variations can find their expression in almost all aspects of the language system, it is obviously not possible to exhaust all of them within the confines of a single work. Thus, the current study will only focus on variations in the following aspects:

A. **Orthographic representations:** features concerning non-conventional orthographic representations of words and creative exploitation of orthographic symbols.

B. **Lexicological aspects:** features involving the creative use of various word-formation strategies, neologism, and slanguage words.

C. **Preference for semantic domains:** bloggers' preference for semantic domains as defined in Wmatrix (Rayson, 2003, 2008b).

D. **Grammatical features:** new or less conventional morpho-syntactic and syntactic features.

E. **Pragmatic features:** features pertaining to the use of pragmatic markers, interjections, and vague words and expressions**.**

A detailed description of the linguistic variables pertaining to the above-listed aspects will be presented and their relationship with bloggers' identity representation will be explored.

1.2 Research background

1.2.1 Internet and Internet-based communication

According to the latest statistics about global Internet usage published by Miniwatts Marketing Group[1] in August 2008, there are more than 1.45 billion Internet users worldwide, among which 39.3% are from Asia, 26.4% from Europe, and 17% from North America (see Table 1.1 below for details). In other words, the number of Internet-users has already taken up 21.8% of the world's population and this number is still growing rapidly. If we look at the penetration rate of Internet usage for different regions, we will find that in North America region Internet users has taken up 76.3% of the total regional

[1] http://www.miniwatts.com/

population. In Oceania/Australia region, this percentage is 59.5% while in Europe it is 48.1%. In China, the most populous developing country in the world, the number of Internet users has been increasing rapidly over the past several years. According to the *22nd Statistical Survey Report on the Internet Development in China* released by the China Internet Network Information Center[2], Internet users in China reached 253 million by the end of June, 2008, accounting for 19.1% of its whole population.

Table 1.1 World Internet usage and population statistics

World Regions	Population (2008 Est.)	Internet Users 2000	Internet Users 2008	% Population (Penetration)	Usage % of World
Africa	955,206,348	4,514,400	51,065,630	5.3 %	3.5 %
Asia	3,776,181,949	114,304,000	573,538,257	15.2 %	39.3 %
Europe	800,401,065	105,096,093	384,633,765	48.1 %	26.4 %
Middle East	197,090,443	3,284,800	41,939,200	21.3 %	2.9 %
North America	337,167,248	108,096,800	248,241,969	73.6 %	17.0 %
Latin America/Caribbean	576,091,673	18,068,919	139,009,209	24.1 %	9.5 %
Oceania / Australia	33,981,562	7,620,480	20,204,331	59.5 %	1.4 %
WORLD TOTAL	6,676,120,288	360,985,492	1,458,632,361	21.8 %	100.0 %

(Reproduced from the information published by Miniwatts Marketing Group)

The ever-increasing population of Internet users worldwide has contributed considerably to the expansion of the territory which written communication has been claiming ever since the advent of computer-mediated communication (CMC) technologies. According to Herring (1996, p. 1), CMC is “communication that takes place between human beings via the instrumentality of computers.” Ooi (2002, p. 91) redefines it as “a mode of human communication that centrally involves the computer as the medium, and made via a hybrid of speech, writing, graphics and orthography,” giving more prominence to the multi-modal nature of the medium. To better reflect the current status that computer-mediated communication is almost exclusively Internet-based, the term “Internet-based communication” (IBC) will be used thereafter except when literature is being reviewed.

[2] http://www.cnnic.net.cn/

IBC normally covers the following forms: online chat (consisting of Internet Relay Chat and various other real-time chatting platforms such as MSN, Yahoo Messenger, Jabber, Skype, and QQ), emails, Bulletin Boards (or Forums), weblogs (mobile blogging, microblogging, twitter, and plurk), and other social networking websites such as MySpace, Facebook, and Hi5.

1.2.2 Weblog as a unique medium and a new genre

Among the IBC types, the weblog is a rather new member, with only a history of around a decade. It has gained enormous popularity especially over the past few years. Blogs are often defined as "frequently modified web pages in which dated entries are listed in reverse chronological sequence" (Herring, Scheidt, Wright, & Bonus, 2005, p. 1). Weblog can be categorized into three types: blogs, filters, and notebooks (Blood, 2002). Herring et al (2005) change the term "blogs" in this categorization to "personal journals" to avoid confusion. According to Blood (2002, p. 7), blogs (personal journals) are mainly for revealing the blogger's thoughts and internal workings. Filters are characterized by contents such as world events and online happenings and they are hyperlink-heavy. Notebooks tend to be more of a random record of ideas. Despite its short history, the weblog has developed certain conventionalized features in terms of its form, content, and communicative functions thus established itself as a unique genre by absorbing the features of the source genres they adapt and adapting to their distinctive technical affordances (Herring et al., 2005).

With the rapid development of telecommunication technology, the integration of mobile telecommunication with Internet-based communication has become a new trend and thus created several new species in the ecology of weblogs, for instance, mobile blogging, microblogging, twitter, and plurk. Integrating mobile communication with Internet-based communication increases the accessibility of the latter; nevertheless, the inherent constraints of mobile communication (such as length limit, different keypad, and so on) restrict its function in self-expression which relies on extended writing to a large extent. As the technological affordances of moblogging and microblogging place more constraints on the length of the blogging entries, their language may have more resemblance to texts mediated by mobile communication devices. Thus, the current research will focus on blogs in the more conventional sense.

Among the three subcategories of weblog, personal blogs (or "personal journals" in the original term) are arguably the most interesting for linguistic studies due to some of the unique features. First, personal blogs have inherited the personal nature from their offline counterpart - diaries or journals but taken on new features in terms of privacy control. Different from conventional diaries or journals which are normally not meant to be read by people other than the authors themselves, personal blogs are normally meant for others to read, though the authors have full control over the accessibility of their entries. Second, the embedding of commenting technology and other communication components has made personal blogs a social communication tool, which in turn increases the tendency of incorporating more oral features in the language of blogging. Third, the anonymous nature of personal blogs has made blogging a publishing space where authors can enjoy almost the greatest freedom: there is no others-imposed editing of any sort. Last, as an interface bridging the material world with the virtual world, personal blogs offer people a new stage to present (or, more accurately, perform) their identities. All these features of personal blogs will inevitably exert influence on the kind of language that bloggers are going to use.

1.2.3 Linguistic variation, identity, and personal blogs

Linguistic variation has long been a major theme in sociolinguistic studies (especially in studies pertaining to the role of social variables in language change). Such research has almost been exclusively about spoken communication, focusing on the phonological variations across social groups of different age, gender, ethnicity, social classes, and so on. Introducing the concept of identity into sociolinguistic studies was a rather recent venture. In fact, identity is a concept which lends itself to various interpretations, be it in or outside the field of sociolinguistic studies. It can be approached from a variety of perspectives, for instance, philosophy, sociology, psychology, cultural and political studies, to name just a few. Despite the differences in focus, all these perspectives have one thing in common: they are all attempting to answer the fundamental question of "Who we are." No matter how we define identity, we should bear in mind certain basic facts about it. First, identity finds its expressions in almost all aspects of people's daily life, which of course include people's daily linguistic behaviors. Second, understanding the concept of identity will inevitably involve comparison between self and others. To put it in a simplified (maybe a bit oversimplified) way, identity is actually a Me-Us-Them relationship. Or, identity is always a representation of the relationship between self, community, and society. It does not simply imply "sameness" as what the etymological meaning of the term might suggest; instead, it is a fusion of "sameness" and "differences." As social beings, we are trying to identify with other members of the same social group (consciously or subconsciously) while, at the same time, maintain certain level of self independence. Identity features will become more prominent when individuals are pooled together.

The relationship between language and identity is so close that some scholars even define identity as "the linguistic construction of membership in one or more social groups or categories" (Kroskrity, 1999, p. 111). According to Kroskrity, identities may be linguistically constructed through the choice of particular languages, linguistic forms, and communicative practices which are indexical of specific social characters. In a similar vein, linguistic variation and identity are also closely related. To a certain extent, they are inseparable from each other. Linguistic variation is a tool for us to "construct ourselves as social beings, to signal who we are and who we are not and cannot be" (Lippi-Green, 1997, p. 63). It is largely an embodiment of people's identity or at least part of people's identity as there are non-linguistic ways (e.g., dressing, hairstyle, and ways of behaving) for identity presentation as well.

As mentioned earlier, existing research about linguistic variation has mostly taken spoken discourse as the object of investigation. Considering the primary status of speech in the field of linguistic studies, nothing seems to be wrong with that. As Perrera (1984) points out, "[s]peaking is as fundamental a part of being human as walking upright" whereas writing is just an optional extra. In the prototypical setting of human communication (that is, the face-to-face oral communication setting), speakers can make some basic judgment about each other's identity the moment they start the conversation. This is also the case in computer-mediated communication settings such as video conferencing and video chat. Even in mediated oral communication settings such as telephone conversations and voice chat, the conversing parties can still gather some basic information about each other's identity (for instance, gender, age, region, and so on) from the voice quality, the accent, and other features which are embedded in the spoken medium.

Writing, on the other hand, seems to be a rather “lean” resource for mining social identity features. Very often, we may need to read between the lines to find out the age, gender, ethnicity and so on about the author if such information is not explicitly spelled out. There is no readily available information such as voice quality in writing which can help us to identify whether the writer is a male or female. There are no obvious clues like accent in conventional writing which can help us to identify from where and what social background the author is. Moreover, writing in its conventional sense is more closely associated with standardization and prescription which are often imposed and reinforced by government policies, the educational system, and mass media. The standardization process is, to a great extent, a process of trying to eliminate regional and even idiosyncratic features. It is true that spoken language has to go through similar standardization processes as well, but still it seems more easily succumbed to variation. Compared with speech, writing seems to be more stable. Furthermore, unlike speaking which is almost omnipresent in people’s daily interactions, writing in the sense of “composition” used to be of limited relevance to people’s daily life. With the advent of the Internet and Internet-based communication, the relevance of writing to people’s (not everyone’s, of course) daily life has been greatly increased. For instance, millions of people are using emails and instant messaging (IM) tools for communication with other people nowadays and both email and IM are writing-based. Personal blogs are a more recent example for ordinary people to use writing for self expression. Different from the self-presentation in spoken situations, bloggers have no face-to-face confrontation with the audience and they cannot use prosodic and paralinguistic features to help them. The keyboard, the computer screen, the Internet access, and the blogging software are the only instruments available for bloggers regardless of their age, gender, ethnicity, social classes, and origin. In other words, bloggers are presenting themselves mainly through textual means (though they can also use other semiotic means). Trying to express oneself in writing had long been practiced but mostly in private in the pre-Internet days and thus it used to be quite difficult for researchers to obtain sufficient data for identity representation studies of quantitative nature. The popularity of personal blogs has changed this scenario. The relatively easier availability of personal blogs as linguistic data makes it possible for researchers to adopt a corpus-linguistic approach and conduct a more comprehensive and systematic investigation about how ordinary people are actually using variations in written language to represent various aspects of their identities.

1.3 Thesis structure

The whole thesis consists of ten chapters. This chapter (Chapter 1) introduces the main objectives and the background of the current study. Chapter 2 reviews literature related to the current study and discusses the theoretical frameworks that I am drawing on. Chapter 3 introduces the corpus construction and the data analysis methods. Chapter 4 presents an overview of the blogging language as revealed by the English Blog Corpus I have constructed for this study. Chapter 5 describes the strategies that bloggers use in realizing orthographic variations and what functions they are put to perform. Chapter 6 focuses on reporting bloggers' creative exploitation of word-formation strategies, neologism, and slang. Chapter 7 presents the variation brought about by bloggers' preference for semantic domains. Chapter 8 explores linguistic variation resulted from bloggers' use of non-conventional grammatical features and pragmatic features. Chapter 9 explains how linguistic variations in various aspects are related to different aspects of the bloggers' identities. Chapter 10 summarizes the major findings, points out the limitations of the current research, and recommends issues for future research.

Chapter 2 Literature Review

This chapter starts with a review of studies concerning Internet-based communication discourse, focusing on studies about blogs. Then, it introduces the concept of identity and identity-related research, followed by a review of literature on linguistic variation. After that, it discusses the speech-writing relations, followed by some critical comments on the limitations of existing linguistic and identity research. It concludes with a proposal for an eclectic framework for investigating the linguistic variation and identity issue in personal blogs.

2.1 The development of Internet-based communication

The ever-expanding territory new writing has been claiming since the advent of computer and Internet technology has not only become a new sphere for people to represent themselves but also a new test bed for people's linguistic experimentation. This is also why many researchers regard Internet-based communication as a new frontier for linguistic investigation. The following sections present a brief review of literature concerning language use in four major Internet-based communication settings.

2.1.1 Online chat

Online chat is a very special Internet-mediated communication means. For the first time in human history, the written medium has been pushed to the very extreme of functioning as "speech" without the physical co-presence of participants. Online chat's nature of being synchronous, interactive, basically text-based, and anonymous makes itself a social interaction setting which is dynamic, transient, experimental, unpredictable, and predominantly recreational (Crystal, 2001a, 2006). As a medium which invites playful and manipulative behavior due to the fact that users are free to camouflage their real-world personal characteristics (Baron, 2002), online chat offers the opportunity for people to experiment with extended or alternative identities. In an online chat situation, people can try on different names, ages, and genders, different personalities, different attitudes and opinions, different relationships, and so on.

Linguistically, online chat discourse displays many features which are not found in oral conversations, despite its similarity to the latter. For instance, it displays such features like high degree of disrupted adjacency, overlapping exchanges, different repair positions, and topic decay (Garcia & Jacobs, 1999; Herring, 2001; Schönfeldt & Golato, 2003). It is also found to be dense with informal discourse particles, unconventional spellings, and simplified grammatical structures (Werry, 1996). These features can be attributed to the temporal, spatial, contextual, and social constraints and the chatter's efforts to reproduce or simulate the discursive style of face-to-face spoken discourse. Crystal (2001a, 2006) makes a rather comprehensive list of the main features of chatroom English based on his observation. These features include: dominant use of monosyllabic words, highly colloquial constructions and non-standard usage, nonce-formations, heavy use of non-standard formations, jargon, and slang, playing with language, and so on. Al-Sa'di and Hamdan (2005) find that chatroom English is characteristic of short and simple sentences, variously truncated words, intentionally and accidentally misspelled words, and frequent use of taboo words. They conclude that English in online chat shares attributes with both spoken and written English and thus should be viewed as a newly emerging, hybrid form with its own characteristics and uses. Gong and Ooi (2008) offer some explanations of the possible social motivations behind some of the typical lexical and grammatical features of online chat discourse. They attribute chatters' use of non-conventional orthographic, lexical, and grammatical features to the technological affordances of the medium, chatters' efforts to economize on typing, and their intention to appear informal, playful, innovative or impressive by deviating from the established norms.

2.1.2 Emails

Different from online chat which is synchronous, email is basically asynchronous. Research shows that the easy-to-use nature of email system has increased its interactivity, which in turn contributes to the formation of a dialogic character similar to e-messaging (Crystal, 2001a, 2006). The kind of language used in email is often closely related to the social distance between communicators and the purposes of communication. Linguistically, e-mail bears resemblance both to writing and speech: for example, the underlying social dynamics are those of writing, whereas the lexical and stylistic properties more closely resemble speech (Baron, 1998). As far as the discourse features are concerned, the language of email is found to be a mixture of informal letter and essay, of spoken monologue and dialogue. At the same time, it lacks some of the most fundamental properties of conversation, such as turn-taking, floor-taking, and adjacency pair (Crystal, 2001a, p. 148). Emails are also found to display the so-called “e-mailisms” which are characterized by trailing dots, capitalization, excessive use of exclamation and question marks, and the use of emoticons (Colley & Todd, 2002). In this regard, email and online chat are quite similar to each other.

Due to its deeper penetration into people's daily life, email is found to be a good place for gender-related studies. For instance, Colley and Todd (2002, p. 380) find females prefer social and domestic topics such as shopping, night life, and cost whereas males prefer the so-called "impersonal, external" topics of locations, journeys, and local people. They also find that females' emails contain a higher incidence of features associated with the maintenance of rapport and intimacy than those from male participants. Thomson and Murachver (2001) find that females make more references to emotion, provide more personal information, use more modals and use more intensive adverbs in email writing. This finding echoes those of others researchers (e.g., Tannen, 1990) from analyzing non-electronic discourses.

2.1.3 BBS

Bulletin board system (BBS) (currently incarnated as online forums) is an asynchronous situation where interactions are stored in some format and made available to users upon demand, so that they can catch up with the discussion, or add to it, at any time (Crystal, 2001a). Two main features of this kind of communication may have shaped their linguistic features: asynchronicity and interactivity. The former allows participants more time to plan and revise their messages if they like while the latter may contribute to its spoken features. BBS depends heavily on message archival (Taboada, 2004), which has made it a sort of "persistent conversation" (i.e., a conversation-like interaction formed through persistent contributions of posters) (Erickson, 1999) realized through messages posted by different participants (called posters) concerning a thread or a topic over a period of time. Different from online chat which is more oriented towards social interactions, BBS is more oriented towards information seeking. By posting messages, BBS posters can seek information they are interested in, share or impart information or expertise, defend their own stands, or challenge others' knowledge or opinions (in the worst case, verbally attack other people). Due to its nature of being conversation-like, asynchronous, and information-sharing oriented, the language of BBS is found to be characterized by a high degree of involvement (similar to that of spontaneous genres such as interviews, spontaneous speeches, and personal letters) and being non-narrative and highly persuasive (Collot & Belmore, 1996). BBS is also a scenario where gendered differences have been observed and compared with traditional gender role stereotypes. According to Herring (1994), men and women adopt different communication styles in discussion lists (forums). The male style is characterized by adversariality: put-downs, strong, often contentions assertions, lengthy and/or frequent postings, self-promotion, and sarcasm. The female style, in contrast, is featured by two aspects which typically co-occur: supportiveness and attenuation. The former is characterized by expressions of appreciation, thanking, and community-building activities that make other participants feel accepted and welcome. The latter is featured by hedging and expressing doubt, apologizing, asking questions, and contributing ideas in the form of suggestions.

2.1.4 Blogs

The blog is an Internet-based communication type which has gained its popularity over the past few years. Just like all the other IBC types at their emerging days, the blog has also attracted the attention of researchers from various fields. Quite a few studies are devoted to describing the origin of blogs, their technological features, categories, functions, and similarities to and differences from conventional diaries or journals. Compared with emails and online chat (Instant Messaging included), it is easier to obtain blog data for research purposes. The diversity and relatively easier availability of blog data has made blogs a good object for various academic investigations. Researchers interested in sociolinguistic issues investigate gender and age differences in blog discourses. Social constructionist practitioners focus on studying the role that blogs play in people's identity constructions. There are also a few studies which are linguistically oriented. Many studies have been conducted by computational linguists in author gender identification, emotion identification, and automatic text classification as well. The following sections present a review of some of the major issues which have been discussed in existing literature.

2.1.4.1 The evolution of blogs

Blogs are often defined as "frequently modified web pages in which dated entries are listed in reverse chronological sequence" (Herring et al., 2005, p. 142). A more detailed version from Kumar and colleagues defines blogs as "web pages with reverse chronological sequences of dated entries, usually with sidebars of profile information and usually maintained and published with the help of a popular blog authoring tool" (2004, p. 35). Blogs have distinctive technological features that set them apart from other forms of Internet-based communication. First, they are easy to use: no knowledge of web programming languages is needed before they can publish their blogs on the Internet. Second, blogs allow readers to comment on the posted blog entries. Third, bloggers can link to other bloggers through hyperlinks and form online communities known as blogroll or blogosphere (Huffaker & Calvert, 2005). Theoretically speaking, anyone with Internet access can publish blogs, and blogs are written about anything bloggers like and in whatever style they wish, typically with no editorial control (Argamon, Koppel, Pennebaker, & Schler, 2007). Bloggers can make their own decisions concerning publication and distribution at the very moment of writing without intervention from a publishing institution (van Dijck, 2004).

According to Blood (2004), blogs at their early days were all about links. The term "weblog" was first coined by Jorn Barger, editor of one of the original Weblogs, Robot Wisdom[3] in 1997. He defined weblog as "a Web page where a Web logger 'logs' all the other Web pages she finds interesting" (2004, p. 54). Another important development of blogging technology - the trackback technology - was introduced by Movable Type[4] in 2001. Trackback allows bloggers to ping other blogs, placing a reciprocal link (i.e., a trackback) in the entry that they have just referenced. By collating all available responses to an entry and making the formerly invisible connections visible, trackbacks help invite instant responses from other bloggers, thus giving blogs (especially the comment area) a conversational nature (Blood, 2004).

Blood (2002) categorizes weblog into three types: filters, blogs, and notebooks. Herring and colleagues (2005) classify blogs into personal journals, filters, and k-logs (i.e. knowledge logs). Among them, personal journals are the most common. Schaap (2004) makes a distinction between three blog categories: linklogs, lifelogs, and photologs. "Linklogs" is actually another label for the "filters" identified by Herring et al. (2005). They mainly consist of hyperlinks to important events, 'noteworthy' news items and other weblogs or websites. The so-called "lifelogs" are actually what Herring and her colleagues call "personal journals." They are typically created by one author who shares all kinds of personal information with his/her audience on a regular basis. The "photolog," as the name suggests, is a "photo only" weblog usually created by individuals

[3] http:// www.robotwisdom.com/
[4] http:// www.movabletype.org/

interested in photography who want to share their photos (van Doorn et al., 2007). Existing studies concerning blogs are mostly about personal journals or lifelogs.

The integration of mobile telecommunication with Internet-based communication has given birth to several new species in the ecology of weblogs, for instance, mobile blogging, microblogging, twitter, and plurk. According to Wikipedia[5], mobile blogging (or moblogging for short) is a form of blogging where the authors publish blog entries directly to the web from a mobile phone or other handheld device even when they are on the move. Entries posted via moblogging could be text-only or multi-modal, depending on how well-equipped the mobile phone is. Microblogging[6] is a form of multimedia blogging that allows users to send brief textual updates or multimedia entries and publish them. The content of a microblog could consist of a single sentence or fragment or an image or a ten-second video. But, still, its purpose is similar to that of a traditional blog. Twitter[7] is a free social networking and micro-blogging service that enables its users to send and read messages known as *tweets*. Tweets are text-based posts of up to 140 characters displayed on the author’s profile page and delivered to the author's subscribers who are known as *followers*. Twitter is sometimes described as the “SMS of the Internet.” Plurk[8] is another free social networking and micro-blogging service which is very similar to twitter. It allows users to send updates (or plurks) through short messages or links, which can be up to 140 text characters in length. These new developments may contribute to the formation of new linguistic features, but the current research will focus on blogs in the more conventional sense.

[5] http://en.wikipedia.org/wiki/Mobile_blogging
[6] http://en.wikipedia.org/wiki/Microblogging
[7] http://en.wikipedia.org/wiki/Twitter
[8] http://en.wikipedia.org/wiki/Plurk

2.1.4.2 Motivations for blogging

Quite a number of studies about blogs investigate why people keep blogs. According to Nardi, Schiano, Gumbrecht, & Swartz (2004, p. 43), there are five major motivations for blogging: documenting one's life; providing commentary and opinions; expressing deeply felt emotions; articulating ideas through writing; and forming and maintaining community forums. In other words, people keep blogs for both personal and social purposes. In fact, many bloggers take blogging as “a form of social communication in which blogger and audience are intimately related through the writing and reading of blogs” (Nardi, Schiano, & Gumbrecht, p. 224). Using blogs as a platform for building social network can be further evidenced by the findings of Kumar et al. (2004) who studied the blogging behaviors of a group of LiveJournal bloggers. According to them, the average number of blogger friends listed in a blogger's profile is fourteen and in eighty percent of these cases, the expression of friendship is mutual. These friends will form a small community where members might list one another's blogs in a “blogroll” (a sidebar within a particular blog listing the other blogs the blogger frequents) and might read, link to, and respond to content in other community members' blogs (pp. 37-38).

Some people choose to use blogs instead of email or personal web page for very practical reasons such as they do not need to worry about whether the recipients have changed their addresses or whether they can accept large photo files. Blogging is also felt to be less intrusive, because it is the readers who can decide whether and when to read a blog entry (Schiano, Nardi, Gumbrecht, & Swartz, 2004).

2.1.4.3 Features of blogs

Discussions about the features of blogs abound in existing literature, with different researchers emphasizing different aspects of this hybrid medium.

Some researchers find that blogging and radio broadcasting are quite similar: they are both a broadcast medium of limited interactivity and they share many features. According to Nardi, Schiano and Gumbrecht (2004), an early blogging software package was called Radio UserLand. Like in radio broadcasts, bloggers can broadcast messages of their own choice without interruption from the audience. The comments or feedback area of blogs is held to be analogous to listener call-in on a radio station. Just like radio broadcasts, blogs can broadcast anything and everything topically (p. 230). Of course, not everyone thinks that blogs are a kingdom for freedom of expression. Gumbrecht (2004), for instance, finds that in practice bloggers tend to impose constraints on themselves. To avoid repercussions in future interactions, bloggers are found to use ambiguous language and references. Some bloggers will even forewarn their audience about the contents of their blog. These strategies allow them to protect themselves while, at the same time, deliver their message well enough to satisfy themselves and their selected audience. “Bloggers engage their audience but find ways to control interaction so that it is infrequent and less emotional, more reflective, than in other more interactive media or face-to-face communication” (Nardi, Schiano, & Gumbrecht, 2004, p. 228).

Another feature of blogs often mentioned in existing literature is bloggers’ freedom in deciding who can get access to their blog entries. Normally, there are three levels of accessibility (or three levels of privacy): private or password-protected (for oneself), friends–locked (friends only), and public or free access (for anyone) (Kendall, 2007).

Defining one’s readership is to define one’s sense of inclusion in and exclusion from a community (van Dijck, 2004).

Some researchers emphasize the hybrid nature of blogs as a medium. Schiano et al. (2004) refer to blogging as “a surprisingly versatile medium with uses similar to those of an online diary, personal chronicle or newsletter, and much more” (p. 1146). This comment is later echoed by van Dijck (2004). According to her, blogs possess several features that other media do not have, for instance, the ability to combine extensive written comments with links, pictures, music and clips, as well as the possibility to post something online to a large anonymous readership. Blogging may be “a combination of both oral and literate practices, such as diary writing, letter writing, the exchange of cultural objects, printed publications, and even conversation” (van Dijck, 2004, p. 8). Based on two years of ethnographic observation on LiveJournal[9], Kendall (2007) finds that there exist tensions between several models of participation in this medium. As a *diary*, LiveJournal provides a place for bloggers to record their feelings, opinions, daily events and reflections. As a *communication tool*, it provides a forum for connection with others and public expression. As a *performance venue*, it provides a stage for self–presentation and artistic production. Kendall’s research also shows that many bloggers regard blogging not really as writing but as *talking*, as can be demonstrated in one of her informant’s remarks:

> [LiveJournal] really is this huge project in self–expression on the part of people who would not normally get to talk to a sort of wide semi–anonymous public. I think that I get the feeling that a lot of the people who are talking in it, especially people who do most of their posts public, really feel like *they’re talking to the whole universe in a way* (Kendall, 2007).

The hybrid nature of blogs as a medium has triggered much discussion about whether blogs should be taken as a genre. Some scholars (e.g., Herring et al., 2005) hold that blogs

[9] http://www.livejournal.com/

have acquired their genre status while others (e.g., Karlsson, 2006) argue that it might be too early to say whether blogging is a ‘genre’. Among those scholars who acknowledge the genre status of blogs, there are still arguments about whether blogs are an emergent genre or a reproduced one. Karlsson (2006) also discusses the hybrid nature of blogs. According to her, the blog is “a loose baggy monster, content-wise, tool-wise, feature-wise, author-wise, reader-wise,” though its basic format is rather stable (frequently updated date-stamped entries in reverse chronological sequence) (p. 6). Just like personal blog is a hybrid medium, it is also a hybrid genre which draws on an amalgam of online and offline genres (Karlsson, 2006).

2.1.4.4 Age, gender, and blogs

As blogs are widely believed to be the favorite means of Internet-based communication for young people, especially young females, it is no wonder that blogs have attracted the attention of many researchers who are interested in studying age and gender differences. Existing blog-related research covers issues such as preference for blogging subgenres and topics, identity construction, and blogging behaviors.

According to Herring and Paolillo (2006), bloggers can be almost evenly divided between women and men but gender is skewed in relation to blog sub-genre. There seems to be a difference in preference for blog genres. Both males and females write personal journal blogs, but the latter drastically outnumber the former. When it comes to filters and k-logs, it is just the other way round: males are in absolute majority. It is still not well-understood why females prefer personal journal blogs; however, we may get some clues from Hogan’s remarks that “the diary’s valorization of the detail, its perspective of immersion, its mixing of genres, its principle of inclusiveness, and its expression of intimacy and mutuality all seem to qualify it as a form very congenial to women life/writers" (1991, p. 105). Although personal blogs cannot be taken as just diaries or journals transplanted to the Internet, there is still an undeniable link between offline diaries and personal blogs.

Age and gender are found to be closely related to the topics and writing style of blogs in several studies. For instance, Kumar et al. (2004) find high correlation between bloggers' age and their topics of interest. According to them, bloggers' topics of interest demonstrate a steady progression from early high school through college, to 20-something, into a more refined 30s, a somewhat conflicted 40s, and even into later life. Many of the strongly age-correlated interests are completely unfamiliar to most people outside the age group (2004, p. 36). Huffaker and Calvert (2005) find that blogs written by males and females are more *alike* than different. They also find that male teenage bloggers use more emoticons than their female counterparts and that female teenagers "are not using language that is more passive, accommodating, or cooperative" in their blogs (2005, p. 15). According to them, teenage bloggers tend to take blogs as an extension of their real life identities rather than a place to pretend. Van Doorn et al. (2007) also find that blog authors tend to present themselves in almost exclusively 'real life' categories such as hobbies, family, work and place of residence thus "leaving no room for the construction of gender identities that bear no relationship to their offline lives" (p. 156). Nowson, et al. (2005) report that blogs written by female bloggers are more contextualized than those written by male bloggers. Argamon and colleagues (2007) find that older bloggers tend to write about externally-focused topics, while younger bloggers tend to write about more personally-focused topics. They also find that with the increase of age, bloggers' writing styles become more masculine. Pedersen and Macafee (2007) find that, like North American bloggers, British bloggers also demonstrate gender differences in their blogging. Their findings reveal that female British bloggers tend to blog about more personal content and show an orientation towards the social aspects of blogging whereas male bloggers tend to be more information-oriented. Female bloggers have also displayed a preference for lesser technical sophistication and greater anonymity.

2.1.4.5 Blog analysis approaches

Existing studies concerning online discourse in general and blogging discourse in particular have demonstrated the possibility of adopting different analytical approaches. Some of these approaches are worthy of particular mention here. The first one is the five-level computer-mediated discourse analysis (CMDA) approach proposed by Herring (2004a). These levels include: 1) structure, 2) meaning, 3) interaction, 4) social behavior, and 5) participation patterns (p. 341). The structural level focuses on the linguistic aspects such as the use of special typography or orthography, novel word formations, and sentence structure. The meaning level focuses on the meanings of words, utterances, and larger functional units. The interactional level focuses on means of negotiating interactive exchanges such as turn-taking and topic development. The social level (or the sociolinguistic level) focuses on the linguistic expression of social relations. The participation level refers to the extent of involvement as measured by frequency and length of messages posted and responses received. According to Herring, the basic methodological orientation of CMDA is language-focused content analysis which may be purely qualitative or quantitative, though she also mentions that sometimes it is necessary for quantitative CMDA to comprise a qualitative component, especially when the phenomena of interest are semantic in nature. CMDA has been widely applied to the analysis of online discourses which are of conversational nature, for instance, email, discussion forums, chat rooms, and text messaging. Developing from this approach, Herring (2004b, 2008) proposes what she calls an expanded pluralistic paradigm of Web Content Analysis (WebCA) (see Figure 2.1 below).

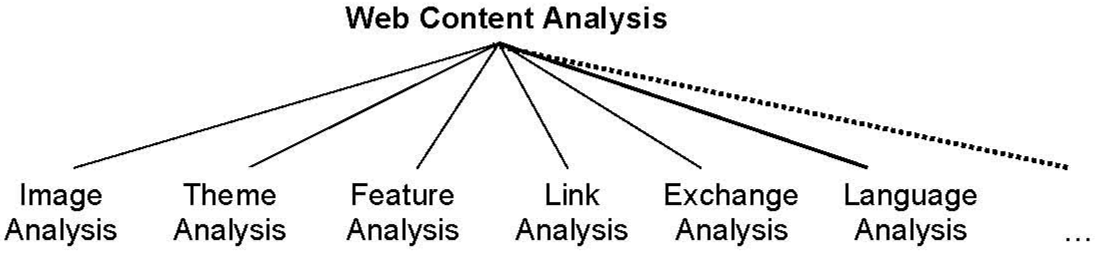


Figure 2.1 Herring's expanded paradigm of Web Content Analysis

In this expanded paradigm, the term "content" has been expanded to cover various types of information contained in new media documents, including themes, features, links, and exchanges, all of which can communicate meaning (Herring, 2008). This approach is applicable to new media such as blogs; nevertheless, no existing studies adopting this approach could be found. In fact, content analysis on blogs can be conducted in other ways as well. For instance, Huffaker and Calvert (2005) have applied DICTION, a content analysis software package, to analyzing the front page of 70 adolescent weblogs to identify gender differences. Nevertheless, their analysis is only restricted to the front page.

Another approach is ethnography, which typically involves participant observation and interviews with small number of informants. The analysis involved is mainly qualitative and non-linguistic. The advantage of this approach is that the researcher is able to experience the dynamics of a particular blogging community. The interviews with bloggers themselves also give the researcher access to what bloggers are actually thinking about when they compose and post a certain entry. The disadvantage of this approach is that it is only suitable for small sample size. For instance, Nardi, Schiano, & Gumbrecht

(2004) have used this approach and conducted audio-taped ethnographic interviews with 23 bloggers and qualitative text analysis of their blog posts. Kendall (2007) conducts a two-year ethnographic study on identity and interactional tensions on LiveJournal. Her analysis is also qualitative and the focus is not on linguistic features.

There are also approaches which are corpus-based and more linguistically oriented. For instance, Nowson, Oberlander, and Gill (2005) have experimented with calculating the F Score of a blog corpus based on the frequency counts of parts of speech to measure the linguistic formality of blogging. Herring and Paolillo (2006) conduct a quantitative analysis using a corpus of 35,721 words to identify gender differences based on their observations about bloggers' use of personal pronouns and some predefined words preferred by males and females. A more recent and more insightful approach which focuses on the association between linguistic features and online identity (or culture) representation is the Wmatrix Approach proposed by Ooi, Tan, and Chiang (2007). Wmatrix is an integrated corpus linguistic tool developed by Paul Rayson (2003, 2008b) from Lancaster University. This system is able to afford word frequency profiles, lexico-grammatical patterning, part-of-speech annotation, and semantic content analysis. By exploiting both the advantages and the limitations of the Wmatrix system, Ooi and colleagues have demonstrated the power of the system in investigating identity representation in unconventional written data such as personal blogs. This approach might have been inspired by the research of Rayson, Leech, and Hodges (1997) which undertakes a comparison of the vocabulary of speakers using a corpus analysis tool and the spoken component of BNC. The Wmatrix Approach has its own limitations, but it does offer some interesting insights about how identity representation in personal blogs can be approached from a corpus-linguistic perspective.

2.1.5 Summary

From what has been presented so far, we can see that the advent of Internet-based communication has started to exert great influence on the linguistic behavior of those people who have Internet access, no matter where they are residing. From the linguistic features displayed in online chat, BBS, and emails we can see the emergence of a range of online discourse features which could be attributed to netizens' creative manipulation of the linguistic forms. We may attribute the prevalence of such features in online chat (instant messaging included) to the chatters' pursuit of speed because this textual conversation is real-time and being quick in response is of vital importance to keep the conversation going. Nevertheless, the pursuit of speed theory cannot explain the presence of such features in asynchronous Internet-based communication settings where the time constraint is no longer an issue. Therefore, we may need to look beyond the field of online discourse analysis and try to obtain insights from studies at a wider context. One potentially useful direction would be identity-related research, which is the focus of the next section.

2.2 Introduction to identity

2.2.1 Defining identity

As Lawler (2008) rightly points out, "identity is a difficult term: more or less everyone knows more or less what it means, and yet its precise definition proves slippery" (p. 1). It is simply not possible to give a single, overarching definition which can fit in all the contexts where the notion of identity is being used. The reason is simple: the same term is used to mean quite different things in different disciplines. In fact, "identity" has become such a buzz word that we can easily find it in almost all social science disciplines. Among the fields where the notion of identity makes its most frequent presence are: psychology, sociology, philosophy, political studies, and sociolinguistics (especially in discourse analysis), to name just a few.

2.2.1.1 Identity as a psychology concept

Identity is, first and foremost, a psychological concept. According to Kroger (2007), Erik Erikson, an American developmental psychologist and psychoanalyst, was the first scholar who had offered detailed academic explanations of this notion in the field of psychology. According to Erikson (1956, 2008), identity has to do with “something in the individual’s core with an essential aspect of a group’s inner coherence” (p. 223). He further explains that the term identity implies “both a persistent sameness within oneself (self-sameness) and a persistent sharing of some kind of essential character with others” (p. 224). Among the key words Erikson uses to talk about the notion of identity are: “self-sameness” “continuity over time,” and “conscious and unconscious process.” To him, identity is “a configuration gradually integrating constitutional givens, idiosyncratic libidinal needs, favored capacities, significant identifications, effective defenses, successful sublimations, and consistent roles” (Erikson, 1959, p. 116). Erikson (1959) holds that a person’s ego identity is shaped by that person’s physiological characteristics, psychological needs, and the social and cultural milieus. For Erikson, identity development is actually a person’s pursuit of proper social roles and niches within a society which can accommodate his or her biological and psychological capacities and interests. This pursuit is normally believed to start during the mid- to late adolescence and will continue and reformulate throughout the life span as one’s biological, psychological, and societal circumstances change (Kroger, 2007). The identity development process, according to Erikson, consists of eight stages, each identifying a different psychological task requiring resolution at different stages of the life span (for details, please refer to Erikson 1963 and Kroger 2007). Following Erikson, Kroger and Adair (2008) define identity as “a configuration, an integration of biological givens, psychological needs, interests and wishes, significant identifications, and meaningful and consistent social roles” (p. 8). There are other lines of thought, of course, for instance, the structural stage approaches to identity represented by scholars such as Jane Loevinger (Loevinger, 1976) and Robert Kegan (Kegan, 1982). According to these approaches, there exist some internal structures in an individual’s ego development. The so-called internal structures, which are believed to follow a predictable, sequential, and increasingly complex pattern of development over the course of childhood, adolescence, and adulthood, are actually psychological filters an individual uses to make sense of his or her life experiences. By focusing on internal identity structures and their functions in facilitating an individual to interpret the content of his or her life experiences, scholars such as Loevinger and Kegan also provide important insights to the understanding of identity in the field of psychology (Kroger, 2007).

It is beyond the scope of the current research to introduce all the different interpretations of the notion of identity in psychological studies. From what has been presented here, we can see some of the core features of identity in psychology. First, identity is a multi-faceted concept which covers biological, psychological, and social aspects. Second, identity is not fixed; instead, it develops with age and is subject to change. Third, identity is a hybrid of intrapersonal sameness (i.e., self-sameness) and partial interpersonal sameness (i.e., partial identification with others in the society). In other words, identity is both individual and social, though what has been emphasized in this field is more of the individual aspect.

2.2.1.2 Identity as a sociology concept

Aside from being a key concept in developmental psychology, identity has also made its way into the field of sociology. As Cerulo (1997) points out, the study of identity forms a critical cornerstone within modern sociological thought. According to him, the notion of identity was first introduced into sociology by two early sociologists Charles Horton Cooley and George Herbert Mead in the first half of the 20th century. Ever since then, identity studies have evolved and grown central to current sociological discourse. In fact, nobody knows when the precise term "identity" was first adopted by sociologists. The original term Cooley and Mead used was "self." Early sociologists primarily focused on exploring the formation of the "me" and the ways in which interpersonal interactions mold an individual's sense of self (Cerulo, 1997). Unlike Erikson and his colleagues in the field of psychology who focus more on the psychological development of an individual's identity, sociologists emphasize the decisive roles that society plays in shaping people's identities. In Cooley's own words, "self and society are twin-born, we know one as immediately as we know the other, and the notion of a separate and independent ego is an illusion" (1909, p.5). Mead (1934) expresses a similar view, saying that the self is not innate but something that "arises in the process of social experience and activity, that is, develops in the given individual as a result of his relations to that process as a whole and to other individuals within that process" (Mead, 1934, p. 135). According to Kroger (2007, p. 20), many different theoretical approaches to identity in the sociological line of thought share a common view that "an individual's identity is the product of the surrounding social context." As Côté (1996) remarks, "for many sociologists there is no identity without society, and society steers identity formation while individuals attempt to navigate the passage" (p. 133 cited in Kroger, 2007, p. 20). Lawler (2008) also contends that "identity, far from being personal and individual, is a deeply social category." According to him, "identities are lived out relationally and collectively. They do not simply belong to the individual; rather, they must be negotiated collectively, and they must conform to social rules" (p. 143). He further points out, "we are engaged social actors, *doing* (rather than having) identities dynamically through time and space, but doing them within the various forms of social constraints" (p. 145).

A review of literature reveals three major lines of thought in identity-related social studies: essentialism, constructionism, and postmodernism. Researchers taking an essentialist stance believe that the attributes and behavior of socially defined groups can be determined and explained by reference to cultural and/or biological characteristics believed to be inherent to the group (Bucholtz, 2003, p. 400). Identity, in this line of thought, is generally held to be associated with pre-defined social groups and considered to be rather fixed. Sociologists taking a social constructionist stance, on the other hand, reject any category that sets forward essential or core features as the unique property of a collective's members. They argue that identity is negotiated and constructed via social interactions. Scholars adopting a postmodernist approach take the variation within identity categories and that across identity categories as equally important. They advocate a shift in analytic focus, deemphasizing observation and deduction and elevating concerns with public discourse (Cerulo, 1997). Cerulo's review of the important studies published since the 1980s shows a diversity of research trends: refocusing attention from the individual to the collective, prioritizing discourse over the systematic scrutiny of behavior, approaching identity as a source of mobilization (rather than a product of it), advocating the concept of identity politics, and discussing new concepts such as "virtual identities" arising out of the advent of Internet-based communication. Despite the shifts and developments in theoretical paradigms and focus, the major topics have remained more or less the same: gender, sexuality, race and ethnicity, and national identity are still on the top of the list. What can be observed about the defining features of identity in sociology include: 1) identity is socially produced; 2) identity is plural in nature; and 3) identity construction implies an agentive role of the individual.

2.2.1.3 Identity as a linguistic concept

Identity is also an important concept in linguistic inquiries, and sociolinguistics in particular. According to Edwards (1985), "sociolinguistics is essentially about identity, its formation, presentation and maintenance" (p. 3). The term "identity" is generally used to mean "social identity" in sociolinguistic studies. According to Ochs (1993, p. 288), social identity is a cover term for "a range of social personae, including social statuses, roles, positions, relationships, and institutional and other relevant community identities one may attempt to claim or assign in the course of social life." Kroskrity (1999) defines identity as "the linguistic construction of membership in one or more social groups or categories" (p. 111). According to him, identities may be linguistically constructed "both through the use of particular languages and linguistic forms and through the use of indexical communicative practices" (1999, p. 111). Although language is not the only means for identity construction, it is generally believed to be the most important means for that purpose. As Ochs (1993) points out, "linguistic constructions at all levels of grammar and discourse are crucial indicators of social identity" and "social identity is a crucial dimension of the social meaning of particular linguistic constructions," though the latter is rarely grammaticized or explicitly encoded in human languages (p. 288). Tabouret-Keller (1997, 2000) holds that our individual identity and social identity are both mediated by language. Language features are the link which binds them together. Such features cover a whole range of language use, from phonetic features to lexical units, syntactic structures, and personal names (p. 317). Social identity in linguistic studies has long been associated with linguistic variation. For instance, Eckert (2000) views identity as “one’s meaning in the world,” which finds its expression in one’s place in relation to other people, one’s perspective on the rest of the world, and one’s understanding of his or her value to others (p. 41). She further points out that the individual’s engagement in the world is a constant process of identity construction and the study of meaning in sociolinguistic variation is a study of the relation between variation and identity. Chambers (2003) gives a more direct explanation of the relationship between linguistic variation and identity, saying that

> the underlying cause of sociolinguistic differences, largely beneath consciousness, is the human instinct to establish and maintain social identity. Linguistic variation shows the profound need for people to show they belong somewhere, and to define themselves, sometimes narrowly and sometimes generally (p. 274).

In fact, all linguistic variation studies involve the issue of identity to a certain extent. Linguistic differences may arise out of age, gender, sexuality, ethnicity, political stance, religion, and many others, all of which could be manifestations of identity.

2.2.2 Creativity, identity, and IBC

Linguistic creativity is generally associated with the notions of novelty, authorship, deviation from norms, and difference. According to Sternberg and Lubart (1999), creativity is “the ability to produce work that is both novel (i.e. original, unexpected) and appropriate (i.e. adaptive concerning task constraints)” (p. 3). Gerrig and Gibbs (1988) define creative language as “any utterance, phrase, or word whose meaning varies with the context in which it is produced in a way that could not be predicted from the lexicalized meanings of its component words” (p. 2). Following Bakhtin’s notion of intertextuality and the dialogic nature of human discourse, Pennycook (2007) holds that creativity rests in the recontextualization of others' expressions rather than new construction, focusing on the Deleuzian philosophical notion that "repetition, which we might have thought to be a matter of the Same, turns out to be a matter of the Different, the obscure" (Bearn, 2000, p. 444). Despite the diversity in definition, many contemporary researchers (e.g., Carter, 1999, 2004, 2007; Carter & McCarthy, 2004; Cook, 1997, 2000; Crystal, 2001b) hold that creativity is a pervasive feature of routine language use rather than a display of special talent in language manipulation only restricted to literary authors and other verbally gifted speakers (Maybin & Swann, 2007). As Kress (2003) points out, creativity is not something rare, special and exceptional which is only allowed to special individuals. Instead, "creativity is normal, ordinary; it is the everyday process of semiotic work as making meaning" (p. 40).

According to Gerrig and Gibbs (1988), linguistic creativity can be conceptually, socially, and pragmatically motivated. Conceptual motivations stem from the necessity of having to express a new concept or idea which is inexpressible within the confines of the standardized repertory of meanings. Social motivations arise out of the need in expressing group solidarity or enhancing one's social status. Echoing this argument, Carter (2001/2002) views creativity in everyday talk as a natural social and interpersonal activity which is more likely to occur "when participants in a speech event feel relaxed and socially at ease with one another" (p. 292). He further points out that creativity is particularly associated with the collaborative sharing of ideas between friends or family members, acknowledging the fact that it may occur across different types of interaction (Carter, 2004). Crystal (2001b) holds that linguistic creativity in the form of language play is often used to establish rapport among interactants. Holding a slightly different view, Cook (2000) argues that language play can be used to perform a range of social functions from creating solidarity or antagonism and competition to preserving or inversing social order. Maybin and Swann (2007) believe that linguistic creativity is particularly suitable for foregrounding an evaluative function. North (2007) finds that linguistic creativity in the form of humor prevails in informal written conversations in online environment and the reason is partly social. The textual cohesion built up through such jointly constructed humor is itself a reflection of the social cohesion of the group, which it also helps to sustain (p. 553). Pragmatic motivations refer to the need in expressing various types of indirect speech acts for reasons such as politeness or avoiding sensitive issues (or taboos). That is also why creativity and novelty are highly valued in persuasive discourse, in which the aim is not just to provide information but to change opinions (Gerrig & Gibbs, 1988).

Linguistic creativity is also connected with the search for and the expression of identities. As Carter (2004) rightly points out, identity is not simply a personal construct nor something pre-existent, singular, fixed and unchanging. Rather, it is multiple and is constructed through language in social, cultural and ethnic contexts of interaction. It is dynamic and mobile and emergent, and is not normally something passively received or assumed (pp. 199-200). Identities can be constructed through creative acts whereas creativity inheres in responsive, dialogic, interpersonal acts of mutuality as well as in individual acts of self-expression (p. 48).

Although linguistic creativity is found to be ubiquitous across a range of text-types, it is especially salient in spoken discourse (Carter, 2004). The rapid development of information and communication technologies over the past two decades has not only triggered a rapid expansion of the lexicon of the English language and created countless examples of lexical innovations, but, more importantly, provided an impulse towards new text types and new forms of creative interaction, in which a new interface has been created between spoken and written language. This new development has created "new spaces for the expression of new identities" (Carter, 2004, p. 190). Due to its special nature of being both public and private, cyberspace "provides new terrain for the playing out of the age-old friction between personal and collective (i.e., social) identity" (Papacharissi, 2002, p. 20). For many people, especially young people in the more industrialized parts of the world, "Internet-based communication media are significant modalities for them to seek answers to identity questions, consciously or unconsciously" (Weber & Mitchell, 2008, p. 26). Many researchers find that Internet-based communication environment is a good place for an individual to play a more agentive role in his or her identity construction. The individual in this context is "the author or playful agent in the production and performance of their own identity" (Merchant, 2005, p. 303). Merchant (2005) explains very clearly why Internet-based communication is so closely related to identity performance, as can be seen from the following quotes:

> Popular electronic communication provides plenty of opportunity for identity work, through multiple and complex interactions with familiar and unfamiliar audiences, and it is in this way that the idea of performing identity becomes salient, not least because acts of performance require an audience. Identity performance becomes important in digital communication when we wish to establish relationships with those whom we have little or no face-to-face contact with, particularly where words on screen are all we have to work with (Merchant, 2005, p. 303).

Observing from these contexts, he further points out, identity is contingent, multiple, and malleable and is quite different from the fixed identities associated with industrial and pre-industrial society. Research findings pertaining to Internet-based communication media such as online chat, discussion forums, and blogs have all demonstrated that these new media have provided rich contexts for users to perform identity with diverse audiences and affinity groups. One important consequence of the popularity of Internet-based communication is that individuals nowadays are able to experience more choice, variety, and idiosyncrasy, which in turn brings about other changes in people’s daily behaviors. As Meyrowitz (1997) remarks, “just as there is now greater sharing of behaviors among people of different ages and different sexes and different levels of authority, there is also greater variation in the behaviors of people of the same age, same sex, and same level of authority” (p. 66).

2.2.3 Pop culture and identity

According to Schwartz and Merten (1967), the special nature of youth as a life stage makes society assert that young people must not prematurely assume adult roles. This ideology actually gives them license to experiment with the possibilities inherent in adult roles and allows them to celebrate the freedom from conventional restraints on social behavior, which, to a considerable extent, helps to form the efflorescence of youth culture. Youth culture is an important constituent of pop culture. Levine (1992) defines pop culture as “culture that is popular; culture that is widely accessible and widely accessed; widely disseminated, and widely viewed or heard or read” (p. 1373). Characterized by its diverse and rapidly changing stylistic practices, youth culture is often taken as a resource for teenagers and young adults to draw on in the construction and display of their identities (Bucholtz, 2000). Language is a flexible and omnipresent set of resources for this culture while at the same time is being shaped by it. The rise of interactive digital media such as the Internet, according to Bucholtz (2000), provides conditions which are more conducive than ever before for the production of innovative styles of youth culture. As a consequence, “language will necessarily take on new forms and uses in a world in which communication has become mediated to a heretofore unprecedented degree" (p. 281). As the youth culture consists of those adolescent norms, standards, and values which are discussed in a language particularly intelligible to members of this age-grade, the data which can best reveal the character of the youth culture are linguistic, and the relevant aspect of adolescent language is obviously semantic (Schwartz & Merten, 1967, pp. 454-457). Bucholtz (2000) has also pointed out the importance of analyzing youth language but she emphasizes that we should approach it as a set of stylistic resources that together produce a multitude of age-based identities rather than just analyzing it at one single linguistic level.

Among the linguistic phenomena pertaining to youth culture which are most widely investigated are slang and sound change. Finegan (2004), defines slang as "a register used in situations of extreme informality, and it may signal rebellious undertones or intentional distancing of its users from certain mainstream values" (p. 335). As Crystal (1995) humorously remarks, "the chief use of slang…is to show that you are one of the gang" (p. 182). Slang can be put into many different uses of which three are believed to be the major ones: expressing informality, identifying group membership, and opposing established authority (Eble, 1996). Slang is especially popular among teenagers and college students, though its use is by no means restricted to such groups (Finegan, 2004). This statement is echoed by Bucholtz's (2000) remarks that slang is the most noticeable linguistic component of youth-based identities.

As Bucholtz (2000) rightly points out, existing literature concerning slang use tends to focus on tracing the origins of particular slang terms and documenting the use and function of slang as an in-group marker. There should also be studies which focus on how slang is being used to differentiate youth identities from one another and the process whereby slang is transmitted and transformed in its movement from group to group are yet to be conducted.

Pop culture is a very important source of entertainment in people's daily life in modern society. People of different age, gender, and even ethnic groups may have different preference for different subcategories of pop culture. These subcategories include pop music, movies, TV series or sitcoms, video or computer games, newly emerged Internet culture like fanfictions, and so on. An individual's preference in this regard is also a very important index of his or her personal and/or group identity (identities). There are a few studies (e.g., Riley, 2007) about pop culture in teenagers' repertoire of daily conversational topics but the contexts are almost exclusively spoken. Personal blogs have provided a good place for researchers to observe how bloggers are using pop culture-related topics to represent their identities. Even for a well-researched theme like the use of slang, personal blogs may be able to give us some new insights. Existing literature has already demonstrated very clearly the indexical function of slang in the construction of age- and gender-based identities mainly in spoken contexts. Whether new slang is emerging as a part of youth culture and how new slang is related to age and gender identities are both topics worthy of systemic investigation.

2.3 Linguistic variation research

2.3.1 An overview

Despite the seemingly close link between social identity and sociolinguistic studies, identity expression was seldom explicitly mentioned as a factor in shaping linguistic variation in the early publications of sociolinguistics. The reason is simple: early sociolinguists, especially variationists represented by William Labov, were primarily interested in unfolding why linguistic variation exists and to what extent it contributes to language change. Researchers' continuous efforts in looking for more satisfactory explanations about style-shifting and the influence from other related disciplines such as sociology and social psychology have steered sociolinguistic studies away from variationism toward social constructionism and gradually brought the notion of identity to the forefront of sociolinguistic studies. Along this process, quite a number of approaches, frameworks or models have been proposed, for instance, Labov's Attention to Speech model, Bell's Audience and Referee Design (1984; 2001), Le Page and Tabouret-Keller's Acts of Identity framework (1985), Eckert and McConnell-Ginet's Community of Practice Model, and Coupland's Speaker Design (Relational Self) Approach (Coupland, 2001), to name just a few. Many of these theories have contributed to the sprout of research on style as a production of identity in which language users creatively draw on available linguistic resources in specific interactional and sociocultural contexts, not without constraints, of course (Bucholtz, 2003, p. 407).

2.3.1.1 Attention to Speech Model

Variationist sociolinguists represented by William Labov and his followers are primarily concerned with establishing a theory to explain the relationship between language variation and language change. According to this school of thought, style-shifts are triggered primarily by the amount of attention people pay to their speech as they converse. In other words, the more attention the speaker pays to his or her speech, the more formal it will become (i.e., closer to the standard variety). Conversely, the less attention the speaker pays to his or her speech, the more casual it will become (i.e., closer to the vernacular variety—the variety the speaker naturally acquires) (Schilling-Estes, 2002, p. 379). Meanwhile, research findings of Labov and others also reveal a strong association between the variants used in more casual styles with lower social class groups and those used in more formal styles with higher social groups.

Despite the insights offered by the Attention to Speech approach, it has been criticized on a number of grounds. For instance, it is very difficult to separate casual speech from careful speech in the conversational portion of the sociolinguistic interview. Moreover, it is also difficult to quantify attention to speech. It has also been criticized for being unidimensional. Some researchers find this approach tends to view speakers as passive respondents who alter their speech only in response to changes in the external situation rather than accrediting them with any agency in their use of stylistic resources (Schilling-Estes, 2002, pp. 382-383). Although stylistic variation is found to be associated with social variables, identity is seldom explicitly mentioned in variationist approaches. In other words, the variationist method is not primarily designed to capture the meaningful social experience or projection of class, race, age or gender, or of situational formality, through language.

2.3.1.2 Audience and Referee Design

The inadequacy of the Attention to Speech approach in accounting for style-shifts has led many researchers to looking for alternative models with greater explanatory power. Bell's Audience and Referee Design model is one of them. This model was initially proposed in 1984 and was originally known as Audience Design. Coupland (2001) calls this model "the first systematic sociolinguistic account of style" since Labov's seminal formulation (p. 185). Bell's model holds that people engage in style-shifting in response to audience members rather than in response to shifts in amount of attention paid to speech. This model has its roots in Speech Accommodation Theory (currently called Communication Accommodation Theory) proposed by Giles and associates (Giles, 2008; Giles & Powesland, 1975). According to Giles (2008), accommodation is a process concerned with how people in interaction are able to reduce or magnify communicative differences between them. The former is known as “convergence” and the latter “divergence.” By enhancing interpersonal similarities, the effect of converging toward or “approximating” another has been shown to win approval. According to Bell (2001), at the heart of audience design is the idea that “speakers design their style primarily for and in response to their audience” (p. 143). Audience design does not refer only to style-shift; it also involves features such as choice of personal pronouns or address terms, politeness strategies, use of pragmatic particles, as well as quantitative style-shift. In other words, audience design is a strategy by which speakers draw on the range of linguistic resources available in their speech community to respond to different kinds of audiences (p. 145). According to this model, there is an association between topic types and audience types, meaning that shifts according to topic echo shifts according to audience (p. 146). The Audience Design model in its latest version consists of two dimensions: the responsive dimension and the initiative dimension. Apart from responding to audience types, the speaker may choose to shift the style so as to initiate a change in the situation. In initiative style-shift, the individual speaker creatively uses language resources often from beyond the immediate speech community, such as distant dialects, or stretches those resources in novel directions (p. 147). According to Bell, initiative style-shifts are in essence “referee design,” by which the linguistic features associated with a reference group can be used to express identification with that group. The so-called “referees” are actually third persons who are not present at an interaction but still possess the power to influence the speaker’s style choices. Initiative style-shift is essentially a redefinition by speakers of their own identity in relation to their audience (p. 147).

The Audience Design approach has been well received since its inception, as Schilling-Estes (2002) remarks, owing to its explanatory power, its greater applicability to speech events besides the sociolinguistic interviews, and its predictive power as well. Nevertheless, this model has also been criticized for its excessive reliance on the responsive dimension of stylistic variation despite its taking on an initiative dimension later on. As Coupland (2007) rightly points out, audience design and accommodation theory "have weighted the scales too heavily in favor of recipiency" (p. 80). The Audience Design model has also been found to be unidimensional because this model implies that all style shifts, even those seemingly related to non-audience effects, are held to be derivative from audience-related concerns (Schilling-Estes, 2002). Some researchers (e.g., Rickford and McNair-Knox, 1994) doubt the link between audience types and topic types. Like the Attention to Speech model, the Audience and Referee Design model does not incorporate the general concept of identity, although it is surely impossible to separate issues of social relationships from issues of self identity (Coupland, 2007).

2.3.1.3 Community of Practice Model

"Community of practice" is a notion initially developed by Lave and Wenger (1991) for explaining the process of learning through engaging in appropriate practice. This notion was soon introduced into language and gender research by Eckert and McConnell-Ginet in 1992 (Holmes & Meyerhoff, 1999), who define it as "an aggregate of people who come together around mutual engagement in an endeavor" (1992). According to Davies (2005), the core of this concept resides in the importance of doing things in a way which reinforces membership in that community of practice. In other words, membership in a particular community of practice is created and maintained through social practices (linguistic or otherwise), rather than global categories being imposed on individuals. Linguistic style shift, according to this model, is neither a result of the amount of attention speakers pay to their speech nor that of audience design. Rather, it is an essential part of speakers' endeavor to construct a social identity (or identities) (Meyerhoff, 2002, p. 534). According to Eckert & McConnell-Ginet (1999), individuals' identity construction is mainly accomplished through their direct engagement with others in common ongoing projects, that is, through jointly developing shared ways of doing and thinking about things and shared ways of understanding. It is the practice component that marks off the Community of Practice model from other frameworks.

Advocates of the Community of Practice model argue that quantitative investigations about stylistic variation characterized by aggregating speakers (particularly according to sex and socioeconomic class) tend to homogenize a broad range of uses, masking the extremes at either end of the variation spectrum (Eckert & McConnell-Ginet, 1999, p. 194). They hold that the search for patterns in language data unconnected to the practices of particular communities may be able to obtain correlational information but can never offer explanatory accounts (Eckert & McConnell-Ginet, 1999, p. 190). Some researchers (e.g., Holmes & Meyerhoff, 1999) hold that the Community of Practice model can be used as a potentially productive means of linking micro-level and macro-level analyses. According to Holmes and Meyerhoff (1999), the community of practice model inevitably involves detailed micro-level ethnographic analysis of discourse in context, which covers identifying significant or representative social interactions, characterizing the processes of negotiating shared goals, and describing the practices that identify the community (p.

181). Meanwhile, a community of practice should also be described within a wider context which gives it meaning and distinctiveness, because “the patterns, generalizations, and norms of speech usage which emerge from quantitative analyses provide a crucial framework which informs and illuminates the ways in which individual speakers use language” (Holmes, 1998, p. 325).

The Community of Practice model is well received among language and gender researchers who take an anti-essentialist perspective, because it allows researchers to focus on the local practices and concrete activities people are mutually engaged in, and thus helps avoid a-priori characterizations of individuals and generalizations about social categories such as sex, class (Freed, 1999). Ehrlich (1999), for instance, uses this model in analyzing the language used by women (a female tribunal member and the complainants or victims) in a sexual assault tribunal. Eckert (2000) has conducted a very influential study of variation in a Detroit suburban high school, Belten High, using the Community of Practice model. She describes how two opposition groups of students have been engaged in their respective communities of practice and constructed their respective group identities as jocks and burnouts. From her studies, Eckert (2000) concludes that the individual’s identity is carved through his or her forms of participation in the group, and the group identity is carved through the interplay of the individual forms of participation that constitute its life. And both individual and group identities are in continual construction, continual change, and continual refinement (p. 43). Eckert views speakers as agents in the continual construction and reproduction of a linguistic system. The social meaning in variation is the result of speakers’ effort in crafting subtly new meaning through the innovative use of linguistic forms. This innovation is no accident but comes in through a process of analysis of the relation between linguistic form and its effect in the world (Eckert, 2000, pp. 215-216). Eckert's emphasis on creative agency does not imply that speakers are constantly looking for new ways to speak or that they are completely free in their adoption of new elements of style; she just wants to counter the prevailing emphasis in the literature on norms and on the constraining effect of social groups.

The Community of Practice model, despite its explanatory power in accounting for local meaning-making, has its own limitations. First of all, the practice component of the model is both its strength and its weakness. Trying to including all social practices within a particular community has the advantage of being able to capture the dynamics involved in identity construction; nevertheless, it may be ill-suited for analyzing communities of practice where language is the most important means for meaning-making, for instance, personal blogs. Second, it is problematic to determine the boundaries of different communities of practice, especially in cases where communities of practice are not maintained through face-to-face interactions. Third, this model lends itself more to micro-level analysis as what it emphasizes is local meaning making. Investigating micro-level meaning making can reveal most of the dynamics this process involves but the findings may not be generalizable due to the limited sample size. Moreover, it is still unknown whether this model can be used to explore communities of practice which do not involve face-to-face interactions such as blogrolls or blogosphere.

2.3.1.4 Social constructionist approaches

The Community of Practice model is just one example of the anti-essentialist approaches to sociolinguistic studies. One perspective worthy of particular mention is the social constructivist approach. Within this approach, language and society are viewed as co-

constitutive. Instead of viewing the linguistic features and patterns speakers use as mere reflections of static identity as defined by one's positions in an existent social order, this approach takes them as resources speakers use to shape and re-shape their social identities (Schilling-Estes, 2002). One example is the acts of identity framework proposed by Le Page and Tabouret-Keller (1985). This framework views linguistic behavior as "a series of acts of identity in which people reveal both their personal identity and their search for social roles." In other words, "language acts are acts of identity" (p. 14). Coupland (2007) views this framework as "an important appeal to a constructivist, process-centered perspective on language and social identity" (p. 108). The underlying hypothesis of this framework is that individual language users *strategically deploy* (my italics) varieties and variation to identify with the social groups they wish to identify, or conversely, to distance themselves from the groups they do not wish to identify (Mendoza-Denton, 2002). According to Le Page and Tabouret-Keller, identity construction is actually a consequence (or maybe a target) of social action. What this framework implies is the agentive role of language users in constructing their identities. As Ochs (1993) points out, "socia1 identities always have a sociohistorical reality independent of language behavior, but in any given actual situation, at any given actual moment, people in those situations are *actively constructing* (my italics) their social identities rather than passively living out some cultural prescription for social identity." In other words, people are still the "*agents* in the production of their own and others' social selves" (p. 296).

The agentive role of language users in constructing their identities has been repeatedly emphasized by many researchers over the past two decades. For instance, Mendoza-Denton (2002, p. 475) views identity as "the active negotiation of an individual's relationship with larger social constructs, in so far as this negotiation is signaled through language and other semiotic means," giving more prominence to language as a means and the individual's conscious efforts in identity construction. Holmes (2006) argues that individuals are "constantly engaged in constructing" aspects of their identities. The words they select, the discourse strategies they adopt, and even the pronunciations they favor may all contribute to the construction of a particular social identity (p. 12). In fact, emphasizing the agentive role can be taken as one of the defining features of the social constructivist approach to language and identity studies.

2.3.2 Gender and linguistic variation

Over the past three decades, there have been a plethora of studies exploring differences in the language behavior of women and men. According to the summary Biber and Burges (2000) made of the existing literature, many studies have focused on aspects of conversational style, including topic choice, topic shifting strategies, the use and function of tag questions, and the use and distribution of overlaps/interruptions and silence. Quite many studies have sought to identify contrasts in the typical linguistic characteristics of female/male language. Among other findings, men are found to be more talkative than women in mixed-gender settings. Many researchers (e.g., Cameron, 1998; Coates, 1993, 2004; Eckert & McConnell-Ginet, 2003; Romaine, 2003) find that women tend to use conversation predominantly as a tool for facilitating social interaction, whereas men tend to use it for conveying information (Baron, 2004). According to Holmes (1995), women use language "to establish, nurture and develop personal relationships" whereas men more typically use conversation as "a means to an end" (p. 2). Even in formal writing, female language is found to exhibit greater usage of features identified by previous researchers as "involved" while male language tends to exhibit greater usage of features which have been identified as “informational” (Argamon, Koppel, Fine, & Shimoni, 2003). In addition, men are found to be more assertive and women more tentative in their language use in both conversation and some forms of writing. Deborah Tannen (1995) finds that males tend to use a direct and forceful style while females use a more indirect and intimate style of interaction. This gendered difference in communication styles and patterns can also be observed in Internet-based communication contexts such as online chat and discussion forums (Herring, 2000).

There are also studies concerning gender-based differences in terms of lexical preference. Certain lexical items are found to be particularly associated with a certain gender. For instance, Lakoff (1973) finds the word *so* to be a noncommittal, characteristically female intensifier. Intensifier use is found to be more often associated with women and some researchers believe that this phenomenon has something to do with women’s inclination for “emotional” topics (Tagliamonte & Roberts, 2005). Based on an examination of 30 existing empirical studies concerning gender and language use, Mulac et al. (2001; Mulac & Lundell, 1994) summarize relatively unambiguous gender effects for 16 language features. According to this summary, typical male language features include references to quantity, judgmental adjectives, elliptical sentences, directives, and first person references. Typical female language features (among others) comprise intensive adverbs, references to emotions, uncertainty verbs, negations, and hedges (2001, p. 125). Argamon et al. (2003) find a strong correlation association between females and the use of pronouns and males with the use of certain intensifiers. Baron (2004) also mentions that females tend to use such features as affective markers, diminutives, hedge words, politeness markers, tag questions and first-person pronouns more often than men whereas men tend to use features such as referential language and profanity more than women. According to the research findings of Mehl and Pennebaker (2003), by and large men use drastically more swear words and considerably more big words in their everyday conversations, more anger words and articles than women while women use more filler words, more discrepancy words, and more references to positive emotions than men. Women tend to use more first person singular references (Pennebaker, Mehl, & Niederhoffer, 2003). Some scholars (e.g., Lakoff, 1975) have attempted to explain why women tend to use a less assertive speech that manifests itself in a higher degree of politeness, less swearing, more frequent tag questions, more intensifiers, and more hedges. They attribute this phenomenon to the general lower social status of women and the lack of power. Whether this is still the case is yet to be found.

The gender-based linguistic variation is not confined to lexical preferences. In fact, existing literature also reveals that men and women show different preferences in syntactic structures in their speech and writing. As early as 86 years ago, prominent linguist (grammarian) Jespersen (1922) remarked that “men are fond of hypotaxis and women of parataxis” (p. 251). Karin Aijmer (1986) discusses adverbial clauses in terms of hedging, a phenomenon that is one of the best-known characteristics differentiating female and male speech. Biber and Burges (2000) find that females favor postposed conditionals, whereas men favor preposed conditionals.

Of course, the gendered differences do not pattern similarly in all age groups. Tagliamonte’s research shows gendered differences seem to be non-existent or at most marginal in the youngest cohort of her informants (the 10- to 12-year olds) whereas these differences become more prominent for the older teen cohorts. This finding seems to suggest that gender differences (at least with respect to the pragmatic features

Tagliamonte has studied) are developmental, and are learned (Tagliamonte, 2005). Whether this trend holds for other age groups is still a topic which needs more investigation.

2.3.3 Age and linguistic variation

Compared with the vast literature on gender-based variation in language use, studies about age-based linguistic variation are quite few, though there emerges a recent interest in investigating the role of age in shaping linguistic variation. Age, according to Peccei (1999), is "an important cultural category, an identity marker, and a factor in producing language variation within a speech community" (p. 114). Age-based differences can be observed in many features of people's speech, for instance, the pitch, pronunciation, vocabulary, and grammar. Certain patterns are appropriate for early and late teenagers but may be less frequent or even absent in the discourse of adults. For example, the use of swear words and slang is very common among teenagers and young adults, but it will be less frequently observed from the discourse of old people (Holmes, 1992, 2001). Based on a large corpus of casual conversation in American English, Barbieri (2008) finds that younger speakers make "outstandingly frequent use of slang and swear words, inserts, attitudinal or personal affect adjectives, intensifiers, discourse markers, first and second person singular reference, and particular quotative verbs" (p. 77). Youth in general often engage in practices that are meant to express rebellion or at least differentiate them in some way from older generations (Brake, 1985). One linguistic manifestation of this rebellion would be the use of slang where "terms become fashionable and serve as markers of in-group membership, and then quickly become outmoded in order to mark their users as outsiders" (Chambers, 2003, p. 187). This rise of nonconformity can be seen in the “ad o les cent peak”- the rise in nonstandard language use by teenagers (see Labov, 2001, pp. 101-120), a peak which flattens out as teenagers become older (Kiesling, 2004, p. 299). As Chambers (2003) has pointed out, the transition from childhood to adulthood is often, almost characteristically, accompanied by extremism. The reason is simple: adolescence requires a purposeful divergence from adult norms in favor of alternative norms. The turbulent and hyper-active nature of adolescence contributes to the linguistic instability of this age group. According to Eckert (1997), adolescence is often seen as the time when linguistic change from below is advanced and adolescents are found to lead the entire age spectrum in sound change and in the general use of vernacular variables. To a considerable extent, this lead can be attributed to their engagement in constructing identities in opposition to – or at least independently of – their elders (p. 163). Sociolinguists have distinguished between “change from above” and “change from below” to refer to the differing points of departure for the diffusion of linguistic innovations through the social hierarchy. Change from above is conscious change originating in more formal styles and in the upper end of the social hierarchy; change from below is below the level of conscious awareness, originating in the lower end of the social hierarchy (Romaine, 2003, p. 103). As people grow older, their attitudes tend to become more conservative. With increasing age, individuals used more positive emotion words, fewer negative emotion words, fewer first person singular self-references, more future tense, and fewer past tense verbs. Age is also found to be positively correlated with an increase in cognitive complexity (e.g., causation words, insight words, long words) (Pennebaker et al., 2003, p. 556). Adults have regularly been shown to be more conservative in their use of variables than younger age groups. This conservatism has been attributed to the pressure for use of standard language at work place (Eckert, 1997, p. 164). Speakers from different age groups may use the same linguistic feature yet for quite different purposes. Erman (2001), for instance, discusses how *you know* is used for different purposes in different age groups. This marker is more text-oriented in adult talk and is typically used in thematic organization of the text and as a cohesive device to bracket utterances. In teenage talk, it is more oriented towards the activity of communicating, ensuring that the channel is open between speaker and hearer, and that messages are understood in accordance with the speaker's intentions (p. 1356).

Age as a contributing factor to linguistic variation can also be observed from language change. As pragmatic markers are vulnerable to change and young people are found to be more active innovators, there will be no wonder for us to find a link between them. As observed by Tagliamonte (2005), the English language has witnessed the emergence of a number of dramatic 'new' discourse/pragmatic markers which have gained considerable high-profile attention in recent years, for instance, *like, just* and *so*. The emergence of these new pragmatic markers can be attributed to the linguistic innovation of the younger generation.

2.3.4 Pragmatic markers and linguistic variation

Despite the numerous studies concerning discourse or pragmatic markers over the past two to three decades, there seems to be no general consensus on what term or label should be used to refer to these markers. A variety of terms or names could be found in existing literature, for instance, "pragmatic marker," "discourse marker," "pragmatic particle," "interactional signal," "small word," to name just a few (for a more detailed list, please refer to Brinton, 1996, p. 33). Andersen (2000) uses the term "pragmatic marker" to describe "a class of short, recurrent linguistic items that generally have little lexical import but serve significant pragmatic functions in conversation" (p. 39). He uses this term as a cover term for what used to be called “pragmatic particles” by European scholars, “discourse markers” by those scholars following the Anglo-American tradition, connectives (such as *so* and *but*), and “pragmatic expressions” such as *you know* and *I mean*.

Existing studies have identified several important features of pragmatic markers, about which Briton (1996) has made a good summary. Pragmatic markers are found to be predominantly a feature of oral discourse. They also appear in written discourse but usually in different forms and for different functions. Tree and Schrock (1999) attribute the different distribution of pragmatic markers in written and oral discourses to different nature of the medium. Written discourse and prepared speech normally allow advance planning and extensive revision time whereas spontaneous talk requires speakers to organize ideas on the fly; thus, they may rely on the use of pragmatic markers to buy more time for planning, organizing, and expressing ideas. Due to its strong association with spontaneous oral discourse, pragmatic markers are stylistically stigmatized and negatively evaluated especially in written or formal discourse. Structurally, pragmatic markers are often found to occur outside the syntactic structure (or just loosely attached to it) and hence have no clear grammatical function (Erman, 2001). Many a time, the absence of pragmatic markers would not affect the grammaticality and intelligibility of an utterance. Semantically, pragmatic markers are often held to contribute very little to the communication of propositional meaning. Instead, they are very closely related to the expression of attitudinal meaning, though some scholars (e.g., Andersen, 2000) argue that it is also possible for pragmatic markers to affect the propositional meaning of an utterance. Being grammatically optional and semantically less relevant does not deny the pragmatic importance of these markers. As Briton (1996) rightly points out, the omission of pragmatic markers will make the discourse appear “unnatural,” “awkward,” “disjointed,” “impolite,” “unfriendly,” or “dogmatic” within the communicative context (pp. 35-36). In fact, pragmatic markers are able to perform multiple pragmatic functions on both local and global levels simultaneously as well as on different planes within the pragmatic component (Andersen, 2000).

Plenty of studies deal with the usage and functions of specific pragmatic markers. Andersen (2000) gives a detailed account of two types of pragmatic markers (i.e., the invariant tags and the pragmatic marker *like*), attaching much importance to the grammaticalization process of these markers and how these markers are used to achieve pragmatic functions from the perspective of relevance theory. He also describes the relationship between the grammaticalization of pragmatic markers and language change and how people from different age groups are using these pragmatic markers. There are also studies which approach the use of pragmatic markers from the perspective of gender-based (the more frequent use of hedges or mitigating phrases by women) or age-specific linguistic variations. One more feature often mentioned in existing literature is that pragmatic markers are found to be more characteristic of women’s speech than of men’s speech.

2.4 Speech-writing relations

2.4.1 The primacy issue

The relationship between spoken language (speech for short) and written language (writing for short) has long been an important research theme in the history of linguistic studies. One issue which had once aroused heated debates and discussions was which one of them should enjoy the primary status in linguistic research. In early modern linguistics, speech was often considered primary and writing secondary and thus the former was regarded as the essential object of study. Writing was once taken as a speech surrogate even by some of the prominent scholars in modern linguistics. According to Saussure (1962), “language (i.e., speech) and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first” (p. 45). Edward Sapir (1921) also regards writing as a representation (or realization) of the primary system - speech. To Bloomfield ([1933] 1984, p. 21), writing is “merely a way of recording language in visible marks,” and it is “merely an external device, like the use of the phonograph, which happens to preserve for our observation some features of the speech of past times” ([1933] 1984, p. 282). For Noam Chomsky, the founder of transformational-generative linguistics, and his followers, who focus on investigating the linguistic competence of the ideal speaker-hearer, written texts are basically irrelevant (Baron, 2002). Of course, not all scholars take the same line of thought. The French theorist Jacques Derrida (1976), for instance, argues that the written word should be seen to have primacy over speech because the former is, by its very nature, more of a permanent record whereas the latter is far more ephemeral (Thompson, 2003, p. 65).

2.4.2 Two different mediums

Instead of arguing about the primacy issue, some scholars hold that speech and writing are actually two sub-systems of language which are used for different tasks. Miller (2001), for instance, contends that speech is not a degenerate form of writing and it is systematically different from the latter in several ways. First, speech (in its prototypical face-to-face context) is produced in real time with almost no opportunity for editing.

Constrained by the capacity of human short-term memory and the demand for quick responses, speakers will usually opt for simpler syntax and simple vocabulary to keep the interaction going. Writing, in contrast, allows pauses and editing; thus enables writers to use more complex syntactic structures and wider range of vocabulary. Second, speech (in its default face-to-face context) is accompanied by non-verbal means such as gestures, eye-contact, facial expressions, and body-postures, all of which complement the spoken word in expressing meanings. Writing in its conventional sense lacks the support of such devices. Third, speech possesses resources of pitch, amplitude, rhythm, and voice quality which speakers can exploit to the full to express various emotions and functions whereas writing in its conventional sense can only turn to words and orthographic conventions to express similar meanings and functions. As can be seen, these differences seem to arise mostly out of the differences in the mediums themselves and their constraints on language production.

2.4.3 Contexts of production

Some scholars believe that speech and writing differ most with respect to the contexts in which each is created and functions. For speakers, language is always produced in the company of a language receiver. For writers, language is produced without the presence of the receiver. Therefore, written texts must function apart from the context of their production. As a consequence, speech is said to be context-bound whereas writing is said to be autonomous and therefore writing must be explicit in order to function acontextually. Nystrand (1983) refutes this notion by saying that context of use in written communication is not concurrent with the production of discourse as with spoken language. A written text does not function communicatively at the time of its creation: it only bears a potential for communication. This potential can only be realized when the text is being read and that is the moment when the writer finally speaks to the reader and the text does its communicative job. According to Nystrand, “speech and writing work differently to maintain reciprocity and the underlying pact of discourse between conversants” (1983, p. 62). Considering the time when Nystrand made such remarks, his view seems to make much sense, although it is possible for a written text to function communicatively if what he means by communication includes intrapersonal ones. Personal diaries, for instance, communicate at the time of their production because the author and the reader are co-present.

2.4.4 Technology and the changing status of writing

The advent of information and communication technologies represented by computers and the Internet over the past two or three decades has considerably changed people’s ways of communication and helped redefine the relationship between speech and writing. One of the changes is the augmented relevance of written communication to people’s daily life. The flourishing of Internet-based communication seems to have created “a type of culture that differs from both oral and literate cultures by changing not only the mode of communication but also the way the writer and reader interact with it” (Shank & Cunningham, 1996, p. 41). Facilitated by the Internet, real-time writing in cyberspace allows people to “communicate rapidly with one another in speeds commensurate with thought and with oral storytelling” (Fernback, 2003, p. 39). Empowered by the Internet, writing has been put into use in conditions which are quite similar to that of prototypical oral communication. This will inevitably lead to writing’s acquisition of spoken features. As mediated human communication becomes more and more non-linear, decentralized, and rooted in multimedia, the distinction between orality and literacy becomes less evident and less important (Fernback, 2003, p. 44). In other words, the gap between speech and writing is narrowing. Despite all that, writing cannot supplant oral communication, although it has replaced it in certain communicative contexts and has even helped to create new ones. Similarly, the electronic media "are only substitutes for oral and written communication in certain contexts and are always dependent on them, just as writing is dependent on the oral use of language, which remains the primary means of human communication" (Goody, 1992, p.12).

2.4.5 Speech and writing as linguistic resources

Due to the constraint of space, micro-level differences between speech and writing will not be presented here. From what has been presented above, we can see the complexity of trying to differentiate speech from writing. They are definitely not the same and they are not completely different. In fact, speech and writing are both alike and different, as Woolbert (1922, p. 271) remarked more than eight decades ago; just how like and how different had never been adequately stated. The advent of Internet-based communication has undoubtedly made the two more alike. Whether the seemingly converging trend between speech and writing will continue and for how long are yet to be known. One thing is for sure: speech and writing as two mediums do have different potentials for manipulation, despite that the language styles could be very similar indeed. Maybe it is high time for researchers to put aside the discussion about whether speech and writing are different or alike and adopt a new perspective by looking at them as two linguistic resources people (who have access to both, at least) can draw on to represent themselves linguistically, especially when we are looking at text-based linguistic data like personal blogs.

2.5 Problems with linguistic variation and identity research

From the review of identity-related studies presented above, we can see the complexity of the notion of identity even within respective disciplines. In addition, the methods and theories employed to approach and account for identity also vary considerably across disciplines. Considering the different orientations and different research objectives, nothing seems to be wrong. It is quite normal for psychologists to focus more on the psychological aspects of identity and sociologists to emphasize its social aspects. Nevertheless, when it comes to the conceptualization of identity in linguistic (especially sociolinguistic) studies, the whole thing becomes very tricky. The reason seems to be quite obvious: identity is so intertwined with language, psychology, and sociology that it is almost impossible to talk about identity without mentioning the other three. Unfortunately, there seems to be a tendency in identity-related linguistic studies to overemphasize the social aspects of identity and overlook its psychological aspects in terms of research scope and a tendency to rely too heavily on social constructionism for explanation in terms of theoretical framework, and a tendency to overemphasize qualitative analysis in terms of methodology. This tendency of overemphasizing certain aspects, despite its strength in revealing less prominent or even hidden features, increases the risk of distorting research findings and may lead to misinterpretations or unconvincing conclusions. The major problems with existing identity-related linguistic studies are summarized as follows:

2.5.1 Conceptual problems

As mentioned earlier, a person's identity is actually a hybrid of personal identity (i.e., being oneself) and social or collective identity (i.e., being a member of a social group or category). Personal identity is derived from personal characteristics and individual relationships whereas social identity is the individual's self-concept derived from perceived membership of social groups (Vaughan & Hogg, 2005). Existing identity-related linguistic studies tend to focus on "social identity," which inevitably involves such issues as social roles, social statuses, social norms, social structures, communities, group memberships, and so on. Once social identity becomes the focus, the personal aspects are often overlooked. It is very true that identity is socially constructed and our behaviors (including our linguistic behaviors) are all constrained by the social and cultural milieus we live in. In fact, many aspects of our identity are imposed on us. We cannot easily change them, for instance, our race and ethnicity, our language, our biological sex, and many others. Nevertheless, we can still choose whether to identify with the social norms (inclusive of some of the imposed identities) depending on the kind of society and culture we are in. For instance, people can choose to identify with their gender preference which may not be in accordance with their biological sexes. This choice may or may not conform to the established norms and it may or may not be suppressed by the communities. Regardless of the approval or disapproval of the community, this kind of choice is personal, private, and most probably psychological. To a very great extent, our identity is actually a reflection of the relationship between the self and various norms. There are group norms which may vary according to the social networks we have, community norms, and societal norms. The so-called norms are simply shared codes for behavior. Norms which affect the whole community are generally institutionalized. Group norms may or may not conform to the norms for a wider community. Norms may not necessarily be good for all social members but they are believed to be so by the majority of the community. Deviations (this is a biased term presupposing a norm) may not necessarily be bad for all social members but they are believed to be not normal by the majority of the community. Yet, trying to be different from the norm, whatever it may be, is an important strategy to construct self-identity, especially for people at a certain age, say, adolescents. A person's identity is a result of choices of identifying with or deviating from various norms. These choices are socially constrained, of course. By identifying with some norms while deviating from some others, an individual formulates his or her unique identity (identities). As language is a key instrument for identity exploration and construction (Huffaker & Calvert, 2005), this principle also applies to people's linguistic behaviors. An individual can choose to identify with or deviate from the linguistic norms within a community, though conforming to an established norm is generally held to be the unmarked option (i.e. the norm). The highest standard regarding the linguistic norms is what is represented by the so-called standard varieties of a certain language (be it spoken or written), however illusionary the uniformity of this standard seems to be. Under that level, there will be various different sub-norms existing in various speech communities. Things will become even more complicated in a multilingual society. Of course, the number of options is never unlimited. In fact, for each individual, the options are always limited and their choices will be affected by social factors, communicative purposes, and so on and the choice will also be constrained by the internal system of the language in discussion. By complying with or deviating from the linguistic norms, a person's identity finds its embodiment in the actual linguistic forms that person uses in a certain context. Of course, language is just one means for identity representation. An individual's identity can find its expressions in other aspects as well, for instance, that person's behavior, style of clothing, and preference in terms of music, games, food, and many others.

Psychological factors can also play a very important role in people's linguistic representation of identity. For instance, adolescents may adopt a very different style of speaking just because they want to distant themselves from adults. At the most extreme, they will diverge if the adults appear to be converging with them. The underlying motive is to extricate themselves progressively from familial dependence in order to take on adult roles (Chambers, 2003, p. 275). To a great extent, a person's pursuit of a personal (individual) identity is a process of psychological maturity, though the nature of the pursued identity is social. During this process we are always involved in resolving certain clashes between what we want to become (the self or personal identity) and what we are allowed to become (the social norms or social identity). As Erikson (1959) points out, identity development is actually a person's pursuit of proper social roles and niches within a society which can accommodate his or her biological and psychological capacities and interests. In other words, a person's identity formation is actually a choice constrained by psychological and social realities. It is both constructive and reflective. This is also true for a person's linguistic representation of identity. Language users are not absolute agents as many researchers taking social constructionism seem to be implying. Their agentive roles may be constrained by factors like their developmental stage, their linguistic competence, the medium they are using (spoken or written), the internal linguistic constraints, their intentionality, and many social factors.

One thing needs to be emphasized here is that, by saying that we should not overlook personal aspects of identity, I am not equating identity and personal identity. There is no such a thing as absolute individual identity. The very nature of human beings as social animals has determined that our identity is always a hybrid of personal and collective (or social) aspects. As Eccles (2009) points out, our collective identities are "those personally valued parts of the self that serve to strengthen one's ties to highly valued social groups and relationships—such as one's gender, race, religion, social class, culture, and family" (p. 78). Having said that, I also agree with Eccles (2009) in that not all aspects of personal identity are grounded in social roles. According to him, "personal identities are those aspects of one's identity that serve the psychological function of making one feel unique" (2009, p. 78). To a considerable extent, our personal identity only becomes conspicuous when we are compared with other members of a group. Identifying oneself with a collective identity is also a part of being oneself.

2.5.2 Methodological problems

Existing identity-related linguistic inquiries have established certain well-accepted tradition in terms of what data to investigate, how to gather such data, what linguistic variables to focus on, and how to account for the findings. For instance, the ideal data for linguistic variation analysis are naturally occurring spoken discourse. The well-researched linguistic variables are phonological variations and the focus is often on variants of certain linguistic forms. The most commonly practiced method of data collection is through sociolinguistic interviews. The analysis methods are either the Labovian quantitative analysis or qualitative analysis methods such as case studies and discourse analysis. Despite that this tradition has given us great insights about why linguistic variation exists and how it is related to language change; it has its own limitations. The major problems are briefly summarized below.

First, existing literature tends to overemphasize spoken discourse. As Bucholtz (2003) rightly points out, most traditions in sociolinguistic studies share a strong preference for spoken over written language, to such a degree that speaker is synonymous with language user. In variationist sociolinguistics, it is the language of "the most vernacular speaker at his most casual and unself-conscious" status that is deemed to be the best data (Bucholtz, 2003, p. 406). This preference for spoken data, however justified for specific cases of research, is likely to make the findings or claims speech-biased and thus less generalizable to other data. The fundamental differences between speech and writing as two mediums may eventually shape the strategies people use in expressing their identities, as discussed in Section 2.4.

Second, existing studies tend to overemphasize phonological variations, though there are a few studies which touch on other aspects such as morpho-syntactic or syntactic variations. Phonological variations may be the most prominent (or the most readily comparable) features where social meanings are expressed but they are not the only features. Linguistic variation should exist in other aspects of the linguistic system as well. Speech is the primary means of human communication but it is not the only one. Variation in new writing should also be included as a part of linguistic variation research.

Third, existing literature tends to under-represent "mainstream" social groups. For a long time, the social groups selected for sociolinguistic studies are often marginalized politically, economically, and socially and hence may not even be recognized by the academy or by dominant society as legitimate subjects of research (Bucholtz, 2003). It is positive for sociolinguists to focus on social groups which are either ignored or suppressed in the mainstream discourse and their research findings are illuminating; nevertheless, the so-called mainstream language users should not be excluded from linguistic variation research.

Fourth, there seems to be a tension between describing the general trends of linguistic variation and understanding individual differences in terms of research focus, which in turn leads to another tension between quantitative analysis and qualitative analysis. As Johnstone (2000) reminds us, intuitive work and quantitative analyses of large corpora of data have important roles to play. But the linguistics of language cannot achieve explanatory adequacy without a linguistics of the individual speaker (p. 420). Many researchers have started to see variability as a resource for the expression of an individual's identity and to see linguistic change, therefore, as potentially originating in expressions of identity. This way of approaching linguistic variation actually implies thinking about how individuals create unique voices by selecting and combining the linguistic resources available to them (Johnstone, 2000, pp. 415-417). Individuals, after all, are the very beginning of a change and they are the ones who produce an innovative form for the first time. In other words, it is very important to record both the general patterns and individual differences. This is only achievable through combining quantitative analysis methods with qualitative ones. That is where corpus linguistics can play an important part. A corpus is always an aggregation of both commonalties and differences. With the assistance of corpus linguistic tools such as the Wmatrix and Wordsmith Tools, researchers can identify both recurrent patterns and individual differences.

In summary, existing identity-related studies have showed a strong preference for using a qualitative analysis method to study the spoken discourse produced by certain special social groups, focusing on the social meaning of phonological variations. Existing online discourse analysis approaches which have been greatly influenced by traditional conversation analysis approaches and qualitative discourse analysis tradition in communication research cannot capture the various aspects of linguistic representation of identity in personal blogs where the major means for meaning making is written language. An approach which combines the strengths of both quantitative and qualitative analysis methods should be called for so that we can obtain a more complete picture of the social meanings of linguistic variation.

2.5.3 Problems with existing frameworks

The strong preference mentioned in the previous section has also limited the applicability of many of the theoretical frameworks emerged in written contexts (new writing contexts included). The Attention to Speech model, for instance, may not hold in analyzing personal blogs. Greater attention to writing may not necessarily lead to more standard forms as there are two possibilities: being closer to the normative standard or more deviant from the norm. The former is self-explanatory whereas the latter could be the result of bloggers' creative manipulation of linguistic forms to achieve certain communicative (pragmatic) functions. The Audience and Referee Design model may make certain sense in analyzing personal blogs because we cannot exclude the possibility that sometimes bloggers choose certain linguistic strategies for the purpose of winning approval from the desired readers. In other words, the style a blogger adopts may well be responsive to his or her audience. Nevertheless, this model cannot explain the fact that personal blogs are also a self-expression means which may have very little to do with how the audience are going to react. The Community of Practice model can be used to explain why certain patterns of behavior (linguistic behavior included) emerge or prevail in a certain community, but the precondition is that the researcher should have access to the community in discussion. In other words, this model is mainly suitable for ethnographic studies. It may be incompatible with studies which intend to describe overall patterns and involve cross-group comparisons. There is a tension between the depth of participation and the sample size. Besides, overemphasizing individual (or individual groups) and contextual differences masks the fact that human behaviors share many commonalities. It is true that each individual possesses a unique combination of identities which arise out of the unique combination of his or her biological characteristics, psychological preference, and sociocultural milieus. In a sense, that is the norm. However, it is the sameness within an individual, a group, and a community and that across individuals, groups, and communities which needs more investigation. As far as the more social constructionist approaches are concerned, the major problem is the overemphasis on the agentive role of language users, overlooking the fact that deliberate identity construction efforts may not happen all the time.

The various limitations outlined above suggest that using one particular framework may not be adequate in interpreting what is actually happening in people’s blogging practices. A possible way out would be to adopt an eclectic framework which takes in the all relevant components from the major existent frameworks and complement it with suitable data collection and data analysis methods. This is the main focus of the following section.

2.6 Towards an eclectic framework

Drawing on the findings of existing studies and the theoretical frameworks (models or approaches) concerning linguistic variation and identity and taking into consideration the unique features of personal blogs, I find it necessary to adopt an eclectic framework in investigating the linguistic representation of identity in personal blogs. This framework consists of the following understanding:

1. Identity is a multi-faceted concept which covers biological, psychological, and social aspects. A person's identity consists of both personal aspects and collective aspects.
2. A person's identity is often manifested in the relationship between self and various social and cultural norms. The identification of a person's identity involves inter-personal comparisons.
3. Language is one of the most important means for identity representation. Speech and writing are both resources language users draw on to represent their identities linguistically.
4. An individual's need in identity representation is one of the most important motivators for linguistic variation.
5. Linguistic variation finds its expression in all aspects of the linguistic system and it is not just a matter of randomly choosing one variant from the linguistic repertoire.
6. A combination of quantitative and qualitative analysis is required for linguistic variation investigation and a corpus-linguistic approach makes this possible.

To be more specific, bloggers' identity representation in personal blogs will find its expression mainly in two places: the relationship between the blogging self and linguistic norms and the biological, psychological, and social realities as revealed by the blogging content. The relationship between the blogging self and the linguistic norms is mainly manifested in bloggers' observation of and/or deviation from norms concerning the following aspects:

1. The orthographic aspect: norms pertaining to spelling, use of symbols and punctuation marks, the use of upper and lower cases, and so on;
2. The lexical aspect: norms regarding word-formation;
3. The grammatical aspect: norms regarding morphological, morpho-syntactic, and syntactic features;
4. The discoursal aspect: norms regulating discoursal organizations;
5. The stylistic aspect: norms concerning conventional speaking and writing in terms of formality (including slanguage use and pragmatic markers).

Any deviation from the norms pertaining to these aspects will result in something non-conventional. The biological, psychological, and social realities are mainly manifested in what bloggers write about. Existing studies such as Ooi, Tan, and Chiang (2007) and Rayson (2003, 2008a, 2008b) have demonstrated the power of the Wmatrix system in identifying non-conventional linguistic features, its capability in conducting inter-corpus comparisons, and its potential in carrying out content analysis based on semantically annotated data. Considering the fact that Wmatrix has its own limitations, I will adopt an analysis approach which fully exploits the strength of Wmatrix and other linguistic tools such as WordSmith Tools while at the same time makes use of qualitative methods when necessary. By adopting a combination of quantitative and qualitative, computer-assisted and manual analysis methods, I intend to identify the potential link between bloggers' realization of linguistic variation and their identity representation.

Chapter 3 Methodology

This chapter first introduces the necessity of adopting a corpus-linguistic approach and the combination of quantitative and qualitative methods in the current research. Then, it gives a detailed description of the important principles and procedures in corpus design, data collection, and the post-processing of the data collected. After that, a description about how quantitative and qualitative analysis methods were actually used in this research is presented.

3.1 Introduction

In Section 2.5.2 of Chapter 2 I have mentioned the necessity of adopting a combination of quantitative and qualitative analysis methods in approaching linguistic variation from an identity representation perspective. Identifying variation and its distribution presupposes an approach which is corpus-linguistic and quantitative in essence, be this variation intrapersonal (intra-speaker), interpersonal (inter-speaker), or both. The reason is that variation identification will inevitably involve comparison which in turn relies on sampling. Sampling is necessary because it is very difficult (if not impossible) to record all the discourses (both spoken and written) produced by an individual in different settings of his or her daily life and use the data for intrapersonal linguistic variation analysis. This is even more the case if the comparison involves many individuals. Therefore, a more practical way would be to base the comparison on samples of an individual’s linguistic repertoire for intrapersonal variation and those of different individuals’ linguistic repertoire for interpersonal variation. This kind of comparison entails corpus-linguistic thinking. Investigating the potential linguistic variation among people from different age, gender, and regional groups requires pooling together a minimum number of texts (discourses) produced by each of these groups in discussion at least. This collection of discourse(s) or texts is actually something which can be called a corpus. According to Sinclair (1991), a corpus is “a collection of naturally occurring language text, chosen to characterize a state or variety of a language.” To be more specific, it is a collection of natural linguistic data, either written texts or transcribed recorded speech, which can be used for linguistic description or verifying hypotheses about a language. The size of the corpus required may vary according to the use it is going to be put into. It can be as big as 450 millions words for a monitor corpus like The Bank of English and its number of words is still increasing. Even that number appears to be tiny if it is compared with the whole Internet which is called a “virtual corpus” by some scholars (e.g., Teubert & Čermáková, 2007). A corpus can be as small as 53,000 words like the Longman/Lancaster Spoken English Corpus, which is regarded as “living proof that small can be beautiful” (Sinclair, 2001, p. ix). In fact, Sinclair almost always advocates the use of corpora which are as big as possible. To him, “there is no virtue in being small” and being small is “simply a limitation” (Sinclair, 2004, p. 189). Again, what Sinclair has in mind is the descriptive analysis of a particular language as a whole. Theoretically speaking, the bigger a corpus is the better. In practice, this may not always be necessary or achievable. The actual size and the type of corpus suitable for a particular study will be affected by the research focus, the availability of data, and other practical constraints. In addition, the concept of identity, as discussed in Chapter 2, implies both intrapersonal and interpersonal sameness and difference, which again presupposes comparison. The sameness (in this case is more of interpersonal type) will mainly find its expression in the recurrent patterns of linguistic features. Identifying this sameness is still a quantitative enterprise. The difference, however, may find its expression in two places:

the recurrent linguistic features which can only (or more frequently) be found in one group and the special set of hapax legomena (words of single occurrences) employed by different groups. Moreover, interpreting the pragmatic functions of these differences requires reference to the specific contexts they are being used. This is where qualitative analysis is of essential importance. In a word, both quantitative and qualitative analysis methods are required in this research.

3.2 Corpus construction

The core of a corpus-linguistic approach to linguistic variation and identity representation study is the construction of the corpus, which consists of such steps as corpus design, data collection, and data processing. The following sections give an introduction of each of these steps.

3.2.1 Corpus design

The targeted corpus consists of two components: the British Blog Component and the American Blog Component. According to Leech (2007), corpus construction should address three important issues properly: representativeness, comparability, and balance. He refers to these three issues (especially the representativeness issue) as “crucial desiderata of corpus design”(p. 144). According to him, though it is extremely difficult to achieve them 100 percent, “we should not abandon the attempt to define them and achieve them.” At least, we should recognize that “there is a scale of representativity, of balancedness, of comparability” and “we should seek to define realistically attainable positions on these scales” (Leech, 2007, p. 144). What Leech has in mind when he emphasizes the importance of these issues (especially the one about representativeness) is the construction of large corpora which are mainly used for general linguistic description purposes. Leech is absolutely right in emphasizing the importance of these issues in constructing such corpora. Nevertheless, there are different typologies of corpora and different researchers use corpora for quite different research purposes. As a result, some scholars (such as Teubert and Čermáková) argue that "it does not make much sense to talk about representativeness" (Teubert & Čermáková, 2007, pp. 64-65), because they find it almost impossible to define the discourse to be represented. Therefore, Teubert (2005) suggests that "it is the linguist's task to define and delimit his or her object of research, to specify which language data he or she wants to analyze. Delimiters include linguistic, spatial, temporal, social, topical and medial parameters" (p. 4). Considering the boundlessness and the potential diversity of the blogging community, it does not seem to make much sense to talk about representativeness, either, as it is almost impossible to define the discourse the corpus is meant to represent. Therefore, it is more advisable to adopt Teubert's (2005) understanding of corpus linguistics as a guiding principle for data collection in this research. According to him, the essence of a corpus-linguistic approach to language study is actually "an insistence on working only with *real language data* taken from the discourse *in a principled way* and compiled into a *corpus*" (2005, p. 4, my italics). In other words, what really matters is the notion of a corpus of "real language data" constructed in a "principled way." In this research, real language data is not an issue, as the blog entries are all naturally-occurring data with no researcher intervention of any sort. Thus, the more important issue is how to construct a corpus in a principled way. Being principled entails the concepts of comparability and balance. Theoretically speaking, this can be achieved by applying a set of pre-defined sampling principles compatible with the research objectives. These principles specify the criteria for what kind of data and how much to be included into the corpus. In practice, these pre-defined principles need to be tested and modified through pilot studies before they are adopted as the final working principles for data collection. The pilot studies should cover all the aspects of data collection and the major aspects of data analysis. Practical constraints such as time, accessibility of data, availability of language processing software, and so on should also be taken into serious consideration.

As the current research is about linguistic variation and identity representation in personal blogs, the data will only consist of blog entries from personal blogs. In other words, other subgenres of blogs such as filters and notebooks (as defined by Blood, 2002) will not be included. The sampling principles for data collection include: (preferably) native speakers, five blog entries (main text, regardless of size) from each blogger which can represent the blogger's writing style and length preference, topics restricted to daily life experiences and reflections, written between 2006 and 2008, and published in mainstream blogging websites. The controlled variables for the corpus include: age group, gender, region, language, and the number of bloggers for each group. According to the original plan, the target corpus would consist of 2,400 blog entries from a total number of 480 bloggers. To be more specific, there would be 240 bloggers from the United Kingdom and the United States respectively. The total number of words of the whole corpus was expected to be around 720k with the British component and the American component taking a half-half proportion. Table 3.1 shows the details of the planned corpus structure.

Table 3.1Planned corpus structure

	UK Component		US Component	
Age Group	Male	Female	Male	Female
15-17	20*5 entries	20*5 entries	20*5 entries	20*5 entries
18-19	20*5 entries	20*5 entries	20*5 entries	20*5 entries
20-24	20*5 entries	20*5 entries	20*5 entries	20*5 entries
25-29	20*5 entries	20*5 entries	20*5 entries	20*5 entries
30-34	20*5 entries	20*5 entries	20*5 entries	20*5 entries
35-40	20*5 entries	20*5 entries	20*5 entries	20*5 entries
Subtotal	600 entries	600 entries	600 entries	600 entries

As mentioned in Chapter 2, personal blogs have been found to be characterized by two features. One is that the blogging community is a sphere for young people, especially teenagers and young adults. According to Huffaker and Calvert (2005), nearly half of all blog entries come from teenage bloggers. McGann (2004) reports that 98.2% of blogs are authored by bloggers below 40 and blogs created by bloggers between the ages of 13 and 29 occupy 91.9%. The other is that female bloggers tend to be the dominant ones in the blogging community. According to Orlowski (2003), most bloggers are teenage girls. Herring and colleagues have also found that personal blogs are dominated by female teenagers and preferred by females in general (Herring, Kouper, Scheidt, & Wright, 2004). That is to say, existing studies have shown that age and gender distribution is not symmetrical in the blogging sphere, with females outnumbering males and teenage bloggers outnumbering older ones. In order to find out whether and to what extent maturity will have influence on bloggers' linguistic representation of identity in personal blogs, I decided to include bloggers from a wider age range from 13 to 40. They will be classified into six age groups: 13-17 for mid teens; 18-19 for late teens; 20-24 for early young adults; 25-29 for adults; 30-34 for mature adults; 35-40 for older adults. The terms used for labeling these groups may not be accurate: the main purpose is for easy presentation later on. Each of these group categories roughly corresponds to a developmental stage from secondary school, high school, college, after college or work starters; career development period; and a period for assuming family responsibility. From the corpus structure of the current research, we can see that the number of bloggers for each gender group within the same age range has all been set to be 20. This design is meant to ensure the comparability between groups. Setting the number of entries to be collected from each blogger to be five is based on two major considerations. First, collecting five entries from each blogger allows for a minimum amount of representativeness in terms of content, style, and length while at the same time makes the data collection more manageable. As the data collection and the subsequent post-processing for this research will be conducted manually, collecting more entries from each blogger means more time and effort, both of which are actually constrained by a rather fixed time frame like PhD research. Second, according to my pilot studies, five entries from each blogger can produce an average of around 1,500-1,800 words of data for linguistic analysis. This is close to the word-limit of 2,000 adopted by the constructors of British National Corpus (BNC) for each author. In order to make the data as comparable as possible, only the main text of the blog entry will be collected. Comments from blog readers will be excluded, as they seldom lead to the direct rewriting or amendments of the entry in discussion, though these comments might have influence on the blogger's future posting and they are of research value in their own right. The topics will be limited to daily life experiences, reflections, and emotional expressions. Fanfictions (fictional writings produced and published by bloggers in their blogging websites) and full-length reviews about literature, movies, music, games, and political issues will be excluded. Entries which have long chunks of quotes from other bloggers will not be included either, just to make sure that linguistic data collected roughly reflect the blogger's own writing style. I did not set a fixed sample size for each blogger except the total number of entries to be collected, which means the five entries collected from each blogger may not be of the same length and even the five entries from the same blogger may vary in length. It does not make sense trying to trim the length of the entries according to a pre-defined word-limit in this research and it is not practical to do that. Of course, the size factor will be taken into consideration when comparative analysis is required. Normalization procedures will be executed before cross-group comparisons are conducted. As far as the regional varieties of the language (English) are concerned, choosing the United Kingdom and the United States is purely a matter of convenience. Moreover, regional differences will be touched upon only when they are closely related to identity representations. It would be an interesting topic for future exploration to include other major regional varieties of English.

3.2.2 Data collection

No one will doubt the abundance of blog data on the Internet. Nevertheless, that does not mean it is easy to obtain the data targeted. In fact, the actual data collection process was long and difficult and it involved considerable decision making. Among the numerous blogging service providers, Blogger or Blogspot (www.blogger.com), WordPress (http://wordpress.org/), Xanga (http://www.xanga.com/), and LiveJournal (http://www.livejournal.com/) have gained international popularity over recent years. These websites are the right places to start data mining. As the current research uses age, gender, and region as three major independent variables for organizing the corpus, the actual data collection ran into trouble from the very beginning, because there was no easy way of locating the desired data. Popular search engines such as Google and Yahoo! could not help much in locating the data. The search engines offered by the blogging service hosts lend themselves better to tag-based search but very few of them offer combined search. After trying out the search engines on the mainstream blogging websites, I find that one of them allows for flexible combined search: the LiveJournal. LiveJournal provides a search function called “Directory Search” which allows the user to search for bloggers by location, frequency of updating, age, interest, friends or any combination of these options. This new function has only become available quite recently. Powerful as the LiveJournal search function is, it can only be used to locate blog entries within its own site. Considering the amount of data to be collected and the popularity of LiveJournal enjoys, I decide to gather all the blog data from this blogging site, despite that including blog entries from other blogging sites would help to reduce bias. This has considerably reduced the difficulties in locating potentially useful blog entries.

Of course, the combined search function of LiveJournal cannot solve all the problems concerning data collection. One reason is that this search engine does not include gender (or sex) as an option in the combined search, which means the job of differentiating male and female bloggers can only be done manually. As Ooi, Tan, and Chiang (2007) point out, a more challenging aspect of compiling a corpus of blogs is to identify the nationality (or native speaker status in the case of the current study), age, and gender of bloggers. Due to the anonymous nature of blogs, not all bloggers put their demographic information explicitly in their profiles. As a consequence, reading through the blog entries is often a prerequisite for determining the gender and native speaker status of the blogger.

There are also cases where even explicit gender information may not be very helpful. For instance, it is not always easy to get equal number of bloggers for a particular age group. My experience shows that it is extremely difficult to find blog entries written by bloggers under the age of 15, especially for male bloggers. This is also why I changed the age range of mid-teens from 13 to 17 into 15-17. This adjustment did not make the data collection for this age group easier. As a result, I failed to collect comparable amount of data for male bloggers from UK despite several attempts with intervals of two months. As far as the US male mid-teens group is concerned, I have eventually managed to locate 20 bloggers but the whole process has turned out to be extremely difficult and time-consuming. Due to limited availability, the final corpus has no data for the UK male mid-teens group. This is quite out of expectation, as many research findings claim that almost half the blog entries are from teenage bloggers. Most probably their findings are only based on the number of accounts existing within a blogging website, without taking the frequency of updating into consideration. In fact, there are a great amount of inactive blog accounts, some of which have not been updated even once in several months. There are also accounts which were friends-locked at the time when they were located and were therefore discarded due to time constraints. This experience of mine seems to echo the findings of previous studies that female teenage bloggers are the dominant ones among the adolescents.

Identifying the nationality or the native speaker status of a blogger could be as problematic as identifying their gender. Although most of the bloggers will specify their location (e.g., the city/state name and the country name) in their profiles, that information can only be taken as reference in determining their nationality or native speaker status. A person who is staying in the United States and blogging in English may not necessarily be a native speaker of English. To make judgments in this regard, reading the education background information from their profiles and browsing through the blog entries become a must. For instance, if the schools (from primary to higher education) a blogger lists in his or her profile are all in UK or US, he or she is more likely to a native speaker of English or a competent speaker of that language. Of course, many bloggers state their nationality in the biography section of their profiles. Trying to identify the nationality or native speaker status of bloggers is meant to reduce as much as possible the potential noise of the final data due to the accidental inclusion of too much data produced by non-native speakers. As a non-native speaker of English, I will have to look for stronger signals or more prominent indicators. Consequently, the data collection becomes more time-consuming and misjudgment is inevitable. Closely related to the native speaker status issue is another one: ethnicity. This issue is even more complicated than that of gender and native speaker status because it is more difficult to get the exact information. It is possible to roughly identify the ethnicity of the blogger from their avatars, their descriptions about themselves in their biographical data, and their blog entries, but again there is no guarantee. In fact, there is no guarantee that the information bloggers put on their profiles is all true. There seems to be no better ways for researchers than choose to believe what they can read from the profiles and the blog entries. There are bloggers from different ethnicity in the EBC, but this inclusion is not meant to be representative of the ethnicity distribution of bloggers in general. In other words, ethnicity is not a controlled variable in this research, although it might be an important factor in linguistic representation of identities.

As far as the text size is concerned, I have adopted two rather pragmatic principles. For each blog entry, the main text is selected as a whole regardless of its actual length. For entry selection, try to be as representative as possible length-wise. The blogger’s entries were first browsed for a general impression about the length patterns and then decisions about the representative length for inclusion were made afterwards. If the blogger’s entries are typically long, then five long ones will be selected. If the entries are all very short, then five short ones will be selected. If there are both long ones and short ones and the long ones are of a greater proportion, then three long ones and two short ones are selected. Otherwise, two longer ones and three shorter ones are selected. One problem with this way of data selection is the total number of words collected from each blogger will be different. As a consequence, nominalization will be necessary when comparisons are being carried out.

After determining the blogger's age, gender, location, native speaker status, the exact blog entries to be included, a metadata file is created which contains such demographic information as nickname, gender, age, education, occupation, location, about me (mini biography), and blogging web address. Each blog entry is saved in three different formats: the original HTML format, the PDF format, and the TXT (plain text) format. The HTML version is reserved for future consultation, as that is the fullest version. The PDF version is also for future consultation in case the HTML version may not be working for whatever reasons. The TXT version is created by copying the textual content from the blogs and pasting it to EditPlus (a text file editor). This version will be used for linguistic analysis after going through the post-processing and annotation procedures. Each blog entry from a particular blogger will be saved as a separate file, following a uniform file labeling code.

Following the principles and procedures described above, I eventually constructed a corpus consisting of 2,300 blog entries from 460 bloggers: 220 British bloggers with 100 males and 120 females and 240 American bloggers with 120 males and females respectively. Altogether, there are 220 male bloggers and 240 female bloggers. The total sample size of the final corpus is 689,437 words. The British component consists of 1,100 blog entries from 220 bloggers, amounting to 334,046 words in total. The average length for each blogger is 1518.39 with a standard deviation of 701.878. The American component consists of 1,200 blog entries from 240 bloggers, amounting to 355,391 words in total. The average length for each blogger is 1480.8 with a standard deviation of 634.859. The final corpus structure is represented below as Table 3.2.

Table 3.2 Final corpus structure

	UK Component		US Component		Total
Age Group	Male	Female	Male	Female	
15-17	NIL	100 entries	100 entries	100 entries	300 entries
18-19	100 entries	100 entries	100 entries	100 entries	400 entries
20-24	100 entries	100 entries	100 entries	100 entries	400 entries
25-29	100 entries	100 entries	100 entries	100 entries	400 entries
30-34	100 entries	100 entries	100 entries	100 entries	400 entries
35-40	100 entries	100 entries	100 entries	100 entries	400 entries
Subtotal	500 entries	600 entries	600 entries	600 entries	2,300 entries
No. of Words	149,255	184,791	167,619	187,772	689,437

3.2.3 Data processing

After a blogger has been selected, he or she will be assigned an ID. This ID is actually a numeric-character string containing such information as the blogger's country of origin, gender, and age group, followed by a serial number. It will also be used as part of the file names for the five blog entries selected from the blogger. A metadata file will be created for the blogger as well. For instance, if the first blogger is a female from UK falling into the age group of 15-17, she will be assigned an ID: uk_f_15-17_01. A metadata file will be created and labeled as uk_f_15-17_01_biodata.txt and the five blog entries will be labeled as uk_f_15-17_01_01.txt to uk_f_15-17_01_05.txt. Labeling the blog entries separately makes it possible for tracking who has used which feature in what context. This is going to be very important to both quantitative analysis and qualitative analysis, as will be discussed later. These five files can be easily merged into one bigger file (using file merging tools) representing one particular blogger's sample. In other words, files uk_f_15-17_01_01.txt to uk_f_15-17_01_05.txt can be merged into one file labeled uk_f_15-17_01.txt, representing the data from the first blogger in this group. This merged file can be further combined with those from the rest 19 bloggers within the same group to generate a dataset which can represent the whole group. This arrangement is necessary both for later interfacing the data with language processing software tools for part-of-speech and semantic tagging, intergroup comparisons, and for other analysis basing on individual files. For data merging across groups, there are two ways of doing it. One is through the same file-merging software I use for merging individual files. The other is through the file-merging tools embedded in the language processing system such as Wmatrix.

Before the individual files are ready for merging, they will go through the following procedures:

1. Standardizing encoding into UTF-8 for all English files;
2. Removing all unnecessary blanks and empty lines;
3. Removing all the symbols and signs which are not allowed by the language processing tools.

After all these procedures, the data will be ready for interfacing with language processing tools for wordlist generating, part of speech tagging, semantic tagging, and various intergroup comparisons.

3.3 Data analysis

The data analysis for this research relies heavily on two language processing software tools: Wmatrix (Rayson, 2003, 2008b) and WordSmith Tools (Scott, 1999). WordSmith is mainly used for its concordance and collocate computation functions. It is also used for triangulation purposes, that is, for checking whether the analysis results obtained from other software tools such as Wmatrix are reliable.

Wmatrix is a software tool which provides a web interface to the USAS and CLAWS corpus annotation tools. The CLAWS (Constituent Likelihood Automatic Word-tagging System) is a system for part-of-speech tagging with 96-97% accuracy based on conventional written English. The USAS (UCREL Semantic Analysis System) (Piao et al., 2005) is a framework for undertaking automatic semantic analysis of text, with a success rate of about 91%, also based on conventional written English. In addition, Wmatrix also provides standard corpus linguistic methodologies such as frequency lists and concordances. As the tagsets Wmatrix employs for both grammatical and semantic tagging are meant for handling conventional linguistic data (i.e., Standard English data) such as the BNC Sampler Corpus, the system still has problems in annotating unconventional word forms. The developers of the Wmatrix system have been adding new features which allow for the creation of personal dictionaries which can extend or override the existing semantic lexicon and multi-word expression (MWE) list used the current system (Rayson, 2008b). This is a rather exciting new development as this new feature makes it possible for Wmatrix to annotate prominent online discourse features. That said, the default setting of Wmatrix (the one meant for processing conventional data) can still be very helpful in spotting new features from the user’s corpus data, especially when it comes to online discourse data (personal blog data inclusive). Wmatrix has no problems in identifying new word forms just like all the other language processing software tools. What is really challenging for the system is the grammatical and semantic annotation of these new word forms, as they are nowhere to be found in the existing lexicon of Wmatrix. For an unknown word form, Wmatrix will make a guess according to its own algorithms and assign it a grammatical category. The success rate of this kind of guessing is not very high but the annotation is largely consistent. For the semantic tagging, the system will assign a label of Z99 to an unknown word form. All the unknown words from a dataset will be pooled together by the system and made available for downloading as a single file. This feature of Wmatrix is very useful for identifying creative linguistic forms which might be important markers of group or individual identities. Another important function of this software is that it allows the user to conduct intergroup comparisons at the word, POS, or semantic level. The original intention of this kind of comparison is to help identify keywords, key grammatical categories, and key semantic domains. This can be extended to identify linguistic differences between two datasets. This is extremely useful for the current research because comparison is an important means for identifying linguistic similarities and differences between bloggers from different age and gender groups, as mentioned at the beginning of this chapter. Moreover, Wmatrix allows the user to perform a comparison of the frequency list generated from their own corpus or corpus components against another larger normative corpus such as the BNC sampler. Again this comparison can be conducted at different levels. Using the log-likelihood statistics, Wmatrix helps identify the overuse of words, grammatical categories, or semantic domains against a reference corpus. This reference corpus can be what is specified by the user and uploaded onto the Wmatrix system or the BNC sampler as specified by the system itself.

The default reference corpus used by the Wmatrix system is the BNC Sampler Corpus. It is a sub-corpus of the 100 million-word British National Corpus, with a roughly equal amount of written and spoken materials of one million words each. The Sampler Corpus is part-of speech tagged, and all the part-of-speech tags assigned to words have been manually checked and corrected, which means the number of errors has been reduced to the minimum. As BNC is a corpus representative of British English, when it is used as a reference corpus, it will help reveal non-British features. BNC is also a corpus representative of conventional spoken and written discourses, when it is used as reference corpus, it can also help reveal non-conventional features. This feature is particularly useful for identifying the major features of personal blog data. As online discourses are generally believed to be having both spoken and written features, I decide to use two corpora offered by the Wmatrix system for actual comparisons: BNC Sampler Corpus Spoken and BNC Sampler Corpus Written. The reason for selecting the BNC Sampler Corpus Spoken is almost self-explanatory. It is used for measuring how similar (or different) the language of personal blogs as revealed by the EBC is to conventional spoken language. The BNC Sampler Corpus Written is used as reference corpus for measuring how similar different the language of the English Blog Corpus (EBC) is from conventional written language.

Apart from allowing for comparisons between the user corpus and the BNC Sampler corpus, Wmatrix also allows for inter-group comparisons within the user corpus. This is one of the major means for identifying the linguistic strategies which bloggers from different groups employ to represent their identities.

Despite its usefulness, comparison-based analysis is not the only method I have adopted. As this research covers a whole range of linguistic variables from lexical features to pragmatic features, it is more advisable to employ an eclectic approach in data analysis. Besides, different linguistic features may lend themselves to different analysis methods. For comparison-based analysis and recurrent patterns or features, a quantitative method has been used whereas for features of more individual nature and explanatory analysis a more qualitative method has been adopted.

3.3.1 Quantitative analysis

By quantitative analysis, it means all analytic methods which involve the counting of frequencies, distributional calculations, or comparisons based on observed frequencies. Two major tools have been used for this kind of analysis. The first one is Wmatrix, which is mainly used for identifying differences across groups in preference for semantic domains. The second one is actually a combination of tools such as WordSmith tools and Excel. The concordance function of WordSmith is used to pool together all the instances of a particular linguistic feature. The concordance file is then saved as a plain text file and copied to an Excel worksheet. As each line of the concordance file contains such information as a concordance line and the file name from which that line is taken, among other things, we can use this feature to obtain the frequencies of each linguistic feature for each blogger. This plain text file is then converted into table format first and then sorted according the column of file source. After that, the Subtotal function offered by Excel is used to automatically count the number of occurrences for each blogger. This process will be repeated until all the desired data are obtained. This method is time-consuming but very useful for getting quantitative data. Wmatrix is mainly used to analyze variation in preference for semantic domains between different age and gender groups. The WordSmith tools are mainly used for analyzing the use of slang words, morpho-syntactic variations, the use of pragmatic features such as pragmatic markers, vague expressions, and so on.

3.3.2 Qualitative analysis

Of course, not all linguistic variables lend themselves to quantitative analysis. Some linguistic variables require what I call "quanti-qualitative analysis," for instance, the processing of the unknown words identified by the Wmatrix system. The so-called unknown words are actually words or word forms which do not match the lexicon of the Wmatrix system. They are either typos, or linguistic forms which have undergone orthographic, morphological, or semantic engineering. They are the most important candidates for linguistic creativity and thus may be important markers of group or individual identities. The first step of the quanti-qualitative analysis is to manually annotate or label the data according to pre-defined criteria. After assigning each word form a label, quantitative analysis procedures will be carried out. This analysis method is applicable to linguistic variables such as orthographic variations, word-formation strategies, use of new or non-standard syntactic features, and use of unconventional contractions.

Qualitative analysis is called upon when local context becomes very important in understanding certain features. For instance, Wmatrix can identity almost all spelling “mistakes” (or orthographically engineered spellings), but it cannot tell the user whether these “mistakes” are accidental or intentional and what pragmatic functions they are employed to perform if they are intentional. To answer such questions, I will have to go back to the specific context where a certain form is used and then decide why it takes that particular form. This can only be done in a qualitative manner.

This chapter describes the procedures I have followed in collecting and analyzing the data. These descriptions will make greater and more concrete sense when the findings and results are reported, which is the focus of the next five chapters.

Chapter 4 The Language of Blogging at the First Sight

This chapter is actually an introductory one for the following four chapters (Chapters 5-9). Following a brief introduction, I present some preliminary observations about blogging language obtained through the comparison of the top 20 words generated from the English Blog Corpus (EBC) with those from the Cambridge International Corpus. After that, an account of how the EBC is different from the two sub-corpora of BNC in terms of top 150 key words is discussed.

4.1 Introduction

Existing studies have demonstrated that linguistic variation can be observed in various aspects of people's language use. Since the majority of these studies are concerned with variations in spoken settings, prominence has naturally been given to phonological variations and their social meanings. As mentioned in Chapter 2, the spoken medium (especially in face-to-face settings or where interactants are co-present) lends itself to a variety of manipulations on the part of the speaker, so to speak. For instance, the speaker can adjust the loudness, tempo, pitch, rhythm, and other prosodic features of his or her speech according to the relationship with the listener, the topic, and the communicative purpose intended. Meanwhile, this kind of manipulation is often accompanied and sometimes reinforced by paralinguistic features. As this manipulation seems to be so effortless and natural and people get so used to it that its existence is seldom fully aware of. To add to its ordinariness, everyone seems to be able to do it and can do it very well, regardless of their level of literacy. Aside from that, people's ways of speaking carry lots of information about themselves, for instance, their age, gender, region, ethnicity, social status, and educational background, among other things. All this information discloses certain aspects of people's identities. To put it simply, we are how we speak.

When we shift our mode of expression into writing, however, the ease of production and the ease of manipulation are nowhere to be found. All those features related to speech sounds become less easily achievable. For expressing simple paralinguistic behaviors such as smiling and laughing, the writer will have to turn to descriptive account or use other strategies to mimic the sounds people produce in spoken settings. In other words, the manipulation of linguistic forms becomes more difficult and less desirable in writing except for achieving special effects, as it runs counter to the affordances of the medium (i.e. writing). Instead of relying on the combination of sounds to express meaning, the writer has only the letters (or orthographic symbols) to manipulate and this manipulation is more strictly constrained, especially in conventional writing contexts. There are a number of reasons for this. First, unlike speech which is more often acquired (almost effortlessly), writing can only be mastered through instructed learning. Second, writing itself is a constraining medium: It requires a writing instrument and a medium to carry it. Third, writing is often associated with standardization, which is mainly meant to eliminate regional and idiosyncratic differences for the purpose of enhancing mutual intelligibility and accuracy of information conveyance. Writing in its conventional sense is both a carrier of and a tool for maintaining and reinforcing the standard variety of a national language. To a great extent, the norm associated with writing and standardization is an imposed collective identity which is supposed to be identified with by all members of a community. Denying such identity is generally not officially encouraged and many a time not easily accepted by both the norm enforcers and the general public. This is also why using colloquial expressions or non-standard variety in writing (especially those of more formal kind) are found to be stigmatized. Besides, writing used to be a medium which can be easily “censored” (so to speak) due to the long process a piece of writing would have to go through from being produced and self-edited by the author to being edited by the publisher. All these procedures are in place to make sure that the published writing is compatible with the established norms. With the advent of the Internet and the popularity of Internet-based communication such as personal blogs, the monopoly which has been enjoying by standardization enforcers and their agents sees signs of breaking. Characterized by being a mainly textual communication tool and a publishing tool allowing for great freedom, personal blogs have great potential for linguistic manipulation. As for which aspects of the written forms tend to be manipulated, how they are manipulated, for what purposes, and whether and to what extent they are associated with the bloggers’ linguistic representation of identities, they are the focus of the following chapters.

4.2 Top 20 words in EBC and their implication

Before reporting the specific findings, I would like to present a rough sketch about the general features of the EBC as a whole by making reference to (or comparison with) some corpus-based findings concerning conventional English spoken and written discourses. Existing literature about personal blogs tends to show that the language of blogging is a hybrid of speech and writing. Evidence can be obtained from comparing the most frequently used words (or word forms) in a blog corpus with those from a spoken corpus and a written corpus respectively. Carter and McCarthy (2006, p. 500) make a list of the top 20 most frequent word-forms used in the Cambridge International Corpus (CIC) respectively for the spoken and the written texts. I have also generated a list of the top 20 word-forms from the English Blog Corpus (EBC) specially constructed for the current research. By putting the three lists side by side, we can observe some interesting differences among them.

Table 4.1 Top 20 word-forms in EBC

Spoken*	EBC	Written*	Rank
THE	*I*	THE	1
I	THE	TO	2
AND	AND	AND	3
YOU	TO	OF	4
IT	A	A	5
TO	IT	IN	6
A	OF	WAS	7
YEAH	*MY*	IT	8
THAT	THAT	*I*	9
OF	WAS	*HE*	10
IN	IN	THAT	11
WAS	IS	*SHE*	12
IT'S	'S	FOR	13
KNOW	FOR	ON	14
IS	BUT	*HER*	15
MM	*ME*	*YOU*	16
ER	N'T	IS	17
BUT	SO	WITH	18
SO	HAVE	*HIS*	19
THEY	DO	HAD	20

*Frequency based on Cambridge International Corpus reported by Carter and McCarthy (2006, p. 12).

As can be seen from Table 4.1, the list for the EBC is different from those for CIC Spoken and CIC Written. Among the top 20 word-forms in the EBC, three are related to self-mention (*I*, *my*, and *me*), and the first person singular pronoun (*I*) ranks the first, revealing the egocentric tendency of blog writing. Blog entries are mainly stories about the bloggers themselves, after all. In the list for CIC Spoken (which is actually the Cambridge and Nottingham Corpus of Discourse in English or CANCODE for short), however, the dominant pronouns are *I* (ranking the second) and *You* (ranking the fourth), reflecting the interactive nature and the informality of daily conversation. The interactive and informal nature is further evidenced by the frequent occurrences of discourse markers (or back-channeling devices) such as *yeah, mm, er* and the frequent use of word forms such as *it's* (an indicator of informality) in the data. In the top 20 wordlist for CIC Written, the prominence of first personal singular pronoun (*I*) has dropped considerably, only ranking the ninth. Nevertheless, the variety of personal pronouns is much greater, covering all the singular personal nouns (*he, she, you, her,* and *his*), reflecting the multiple perspective potential of written discourses.

The most striking difference between the top 20 wordlist for the EBC and those for CIC Spoken and CIC Written is that the definite article (*the*), which is supposed to be the most frequently used word in almost any extended piece of discourse, ranks the second, following the first person singular pronoun (*I*). Despite that the prominence of *I* in the EBC can be partly explained by the inherent nature of personal blogs in being egocentric, I still find it enticing to examine the distribution of *I* and *THE* among texts produced by bloggers from different gender, age, and regional groups. For that purpose, three sets of top 20 wordlists were generated from the EBC. The first set consists of a list for all blog entries produced by American bloggers and one for texts produced by British bloggers. The second set consists of a list for texts produced by all female bloggers and one for those produced by all the male bloggers. The third set consists of six separate lists for texts produced by bloggers from six different age groups. After that, various wordlist comparisons were conducted (using the WordSmith Tools) to determine whether and to what extent bloggers from different groups use the first personal singular pronoun (*I*) and the definite article (*the*) differently. Here are the results:

Gender-wise, female bloggers as a whole have used *I* more frequently than their male counterparts, with a log likelihood value of 34.1(at the P value < 0.0001). Male bloggers as a whole have used the definite article (*the*) more often than their female bloggers, with a log likelihood value of 116.7 (at the P value < 0.0001). This pattern seems to support the finding of Mehl and Pennebaker (2003) that by and large men use more articles than women. It also supports Pennebaker and colleagues' (2003) finding that women tend to use more first person singular references.

If we take a look at the distribution of *I* and *THE* between females and males within the same age groups, the picture becomes more complicated (See Table 4.2 below). The comparison results show that gender differences across groups are not straightforward. Six out of the 11 age groups displayed gender differences which are statistically significant in the use of first person singular pronoun (*I*). These groups include: the British 18-19 group, the British 25-29 group, the British 35-40 group, the American 15-17 group, the American 25-29 group, the American 35-40 group. The rest five age groups did not show gender differences of statistical significance. The results for the use of the definite article (*the*) are even less straightforward. Four age groups (the British 18-19 group, the British 20-24 group, the British 30-34 group, and the American 15-17 group) have demonstrated gender differences of statistical significance, with the males outperforming the females. There are three age groups (the British 35-40 group, the American 25-29 group, and the American 35-40 group) whose gender differences are statistically significant, with the female bloggers outperforming the male counterparts. The rest four age groups have demonstrated no significant gender difference in the use of the definite article (*the*).

Table 4.2 Gender difference in the use of I and THE

Blogger Groups		**Log Likelihood Value**	
Female	Male (Reference list)	I	THE
uk_f_18-19	uk_m_18-19	25.4	-38.9
uk_f_20-24	uk_m_20-24	N/S	-36.88
uk_f_25-29	uk_m_25-29	84.1	N/S
uk_f_30-34	uk_m_30-34	N/S	-41.1
uk_f_35-40	uk_m_35-40	90.1	63.4
us_f_15-17	us_m_15-17	40.3	N/S
us_f_18-19	us_m_18-19	N/S	-33.85
us_f_20-24	us_m_20-24	N/S	N/S
us_f_25-29	us_m_25-29	208.5	73.2
us_f_30-34	us_m_30-34	N/S	N/S
us_f_35-40	us_m_35-40	314.1	129.2
N/S means Not Significant			

Region-wise, no difference was found in the use of the definite article (*the*) between American bloggers and British bloggers, but American bloggers have used a greater number of *I* than their British counterparts, with a log likelihood value of 250.3 (at the P value < 0.0001).

The comparison across age groups, however, reveals certain patterns that invite comments. Table 4.3 (see below) shows the top 20 word lists for all the six age groups. If we put our focus on the top 5 words on each of these six lists, we will find some neat patterns. All age groups share the same words (*I*, *THE*, AND, TO, and A) for the top 5, nevertheless, the rankings of these words in the lists reveal some interesting patterns: the teens groups (the 15-17 group and the 18-19 group) share the order of I←AND←TO←THE←A; the 20-24 group shares the order of I←THE←TO←AND←A with the 25-29 group; and the 30-34 group shares the order of THE←I←TO←AND←A with the 35-40 group. A further examination of the percentage that the frequency of each of these five words accounts for in the total word tokens produced by bloggers from a particular age group reveals certain tendency about the use of two words: *I* and *THE*. The use of first person singular pronoun (*I*) decreases with the increase of blogger age, whereas the use of the definite article (*the*) increases with the increase of blogger age. Statistical comparisons between adjacent pairs of age groups in terms of the use of *I* and *THE* show that all the adjacent pairs except for pair of the 30-34 group and the 35-40 group have displayed differences of statistical significance, though the late teens group outperformed the mid teens group. If we consider the teens (15-19), the early and mid-adults (20-29), and the more mature adults (30-40) as three separate groups, we can see that age is an important factor which affects the use of the first person singular pronoun (*I*) and the definite article (*the*) in the EBC. What we can observe from Table 4.3 is that linguistic variations seem to be related to bloggers' age and gender, two important aspects of their identities.

Table 4.3 Top 20 word forms by age group

15-17	%	18-19	%	20-24	%	25-29	%	30-34	%	35-40	%	Rank
I	4.58	*I*	4.68	*I*	4.20	*I*	3.90	*THE*	4.03	*THE*	4.10	1
AND	3.11	*AND*	3.05	*THE*	3.25	*THE*	3.49	*I*	3.65	*I*	3.56	2
TO	2.76	TO	3.02	TO	3.07	TO	3.13	TO	3.15	TO	3.11	3
THE	2.64	*THE*	2.96	*AND*	2.92	*AND*	2.85	*AND*	2.79	*AND*	2.86	4
A	1.85	A	2.16	A	2.19	A	2.37	A	2.50	A	2.36	5
IT	1.47	OF	1.46	OF	1.69	OF	1.68	OF	1.72	OF	1.71	6
MY	1.37	MY	1.34	MY	1.41	IT	1.34	IT	1.31	IT	1.26	7
OF	1.19	IT	1.33	IT	1.39	THAT	1.31	IN	1.28	IN	1.26	8
WAS	1.12	THAT	1.13	THAT	1.19	MY	1.30	MY	1.23	THAT	1.25	9
THAT	1.11	IN	1.13	IN	1.18	IN	1.25	THAT	1.20	MY	1.17	10
IN	1.03	WAS	1.08	FOR	0.92	FOR	0.95	FOR	0.99	WAS	0.96	11
BUT	0.95	ME	0.92	IS	0.90	IS	0.92	WAS	0.91	IS	0.94	12
SO	0.94	SO	0.92	WAS	0.85	WAS	0.86	IS	0.91	FOR	0.92	13
ME	0.91	IS	0.88	ME	0.83	ON	0.82	ON	0.85	ON	0.87	14
IS	0.89	BUT	0.88	SO	0.79	ME	0.78	HAVE	0.74	HAVE	0.81	15
FOR	0.77	FOR	0.85	BUT	0.75	BUT	0.77	ME	0.72	SO	0.73	16
LIKE	0.74	ON	0.77	HAVE	0.75	SO	0.73	WITH	0.72	BUT	0.72	17
ON	0.71	HAVE	0.71	ON	0.73	HAVE	0.71	BUT	0.71	WITH	0.69	18
I'M	0.62	WITH	0.65	WITH	0.62	WITH	0.70	SO	0.69	ME	0.65	19
HAVE	0.62	BE	0.63	BE	0.62	BE	0.68	AT	0.61	AT	0.65	20

% refers to the proportion of a word's occurrence frequencies in the total word tokens produced by bloggers from an age group.

4.3 Keyword comparison with BNC Sampler Corpus Spoken

Table 4.1 gives us a very rough idea about how blogging language is similar to or different from conventional spoken and written language, but it is difficult for us to make any further claims about the similarities or differences just based on 20 most frequently used word forms. If we want to identify more features which can better reflect the language and content of personal blogs, we may need to expand the scope of inspection. This is where Wmatrix (Rayson, 2003, 2008b) can come into an important play. As spelled out in Chapter 3, Wmatrix allows for comparisons between the user’s own corpus and the BNC Sampler Corpus. The system can generate a list of key words (words which are used statistically more frequently) for the user corpus based on the frequency difference of a particular word (or word form) between the user corpus and the reference corpus (BNC Sampler Corpus in this case). Wmatrix offers a number of sub-corpora for users to choose according to their research needs. I have chosen the BNC Sampler Corpus (Spoken) and BNC Sampler Corpus (Written) as the reference corpora for key words identification. Wmatrix has identified 3,905 overused word-forms (that is, words with log likelihood value greater than 6.63) with reference to the BNC Sampler Corpus Spoken and 3,431 overused word-forms with reference to the BNC Sampler Corpus Written. I will only focus on the top 150 word-forms from the two lists of key words generated. Due to the constraint of space, only those words which are more of blogging nature are presented here. Table 4.4 shows the key words identified by Wmatrix to be overused as compared with the BNC Sampler Corpus Spoken.

A number of features can be observed from Table 4.4. First, seven out of the top 20 keywords as compared with the BNC Sampler Corpus Spoken are related to self-mention: *I*, *me*, *my, myself*, and *im*. They all refer to the blogger. The frequent occurrences of *am* and *'m* also point to the blogger since both word forms presuppose a subject of *I*. This echoes what has been discussed earlier in that the blogger is the central character of blog texts. Second, certain words could not be found (or only appear once or twice) in the BNC Sampler Corpus Spoken but they are high-frequency ones in the EBC. They are: *lol*, *mom*, *thats*, *awesome*, *ive*, *kinda*, *cant*, *haha*, *yay*, *tv*, *blog*, *sucks*, *anyways*, *LJ*, *gay*, *Internet*, *wont*, *random*, *wasnt*, and *gym*. It is not unexpected for *s*ome of these words to be identified as key words when the BNC Sampler Corpus Spoken is used as the reference corpus, for instance, *mom*, *blog*, and *Internet.* This has a great deal to do with the nature of the reference corpus. Despite its popularity among corpus linguistics circle, BNC only consists of texts produced or transcribed speech recorded before the 1990s. At that time, the Internet was still emerging and blog was still unborn. Naturally, these two words would not appear in people's daily conversation as represented by the BNC Sampler Corpus Spoken. In fact, both words (*Internet* and *blog*) are surely to be identified as key words even if the whole BNC is used as the reference corpus. In addition, as BNC is a corpus designed to represent the British English, anything typically American will be easily identified. That is also why the word *mom* is among the key words. Of course, the relatively high frequency of the word form *mom* has also played an important role.

If we take a further look at the words (word forms) with zero or very low frequency in the BNC Sampler Corpus Spoken but with high frequency in the EBC, we will see traces of other features as well. For instance, word-forms such as *thats*, *ive*, *cant*, *tv*, *wont*, and *wasnt* have something in common: all of them do not comply with the spelling regulations in standard written English. In other words, the established spelling norm for written English is not well-observed in personal blogs. From all these word forms, we can see strong influence of spoken language. For instance, the omission of the apostrophe from word-forms like *thats*, *ive*, *cant*, *wasnt*, and *wont* will not affect the proper understanding of them if they are read out. In fact, there is no such a thing as the apostrophe in spoken language; it is just a symbol arbitrarily designated for indicating (representing) contracted forms in writing or transcribed speech. In a similar vein, spoken language does not differentiate upper case letters from lower case ones. That is a non-issue in speech. The difference between upper and lower case letters only makes sense in writing. It does not make any difference in people's understanding of the short form of the word "television" whether it is spelled in all lower case letters (*tv*) or all upper case ones (*TV*). Nevertheless, the relative high frequency of unconventional contracted forms does not suggest that normal contracted forms are never used. In fact, conventional contracted forms such as *that's*, *I've*, *can't*, *wasn't* and *won't* are also widely used in personal blogs. Two types of contracted forms actually co-exist: the ones with the apostrophe and the ones without. Despite their difference in orthographic representation, both of them are associated with informality, with the latter being arguably greater in degree. Other word-forms such as *kinda*, *haha*, and *yay are* typical markers of spoken discourse, with *haha* and *yay* as two newly emerged word-forms for expressing laughter and excitement typically in online discourses. Spelling a word according to how they are actually pronounced is a strategy often adopted by netizens when they are producing online discourses. This phenomenon is often called phonetic spelling or eye dialect. According to Harold Wentworth - editor of *American Dialect Dictionary*, "eye dialect is phonetic respelling of words merely to burlesque the words or their speaker" (Bolinger, 1946, p. 337). From the date of this definition, we can see that eye dialect (or phonetic spelling) is not a new trick. It used to be employed by novelists to represent the non-standard form of English that their characters speak in their literary works. Nowadays, eye dialect or phonetic (re)spelling is often used to represent ordinary standard English said in an informal way rather than non-standard pronunciations (Cook, 2008). More detailed discussion will be presented in the following chapter where orthographic variation is the focus.

Table 4.4 Key words in EBC relative to BNC Sampler Corpus (Spoken)

Word	EBC	BNC SP	LL	Rank	Word	EBC	BNC SP	LL	Rank
my	8279	2354	6505.7	1	*yay*	120	0	222.96	73
i	32677	31907	3253.85	2	*tv*	125	1	221.59	75
me	5194	2861	2044.44	3	*blog*	115	0	213.67	78
am	1536	281	1571.17	4	excited	135	5	212.72	80
im	726	5	1294.14	8	weird	191	32	203.62	84
nt	637	24	1001.41	9	tired	242	63	202.24	86
fun	452	34	627.56	14	sad	170	24	194.77	87
'm	3446	2887	573.83	15	damn	176	29	189.03	91
myself	627	149	555.59	16	crazy	127	7	188.05	93
lol	284	0	527.68	18	*sucks*	104	1	182.94	97
friends	520	111	490.72	21	watched	177	33	179.37	101
mom	213	0	395.76	26	ass	113	5	173.58	103
thats	198	0	367.89	30	*anyways*	93	0	172.8	104
guess	328	58	341.04	32	fuck	225	71	163.26	110
awesome	189	1	339.68	33	crap	147	25	155.65	116
ive	182	0	338.16	34	seriously	153	29	153.78	118
cool	240	18	333.45	37	*LJ*	81	0	150.5	119
kinda	183	1	328.6	38	*gay*	80	0	148.64	120
cant	161	0	299.14	42	movie	104	7	148.03	121
haha	161	0	299.14	43	*Internet*	78	0	144.93	123
shit	319	70	296.37	44	*wont*	78	0	144.93	124
fucking	432	157	277.43	50	*online*	103	8	141.93	126
anymore	156	4	256.46	57	*random*	85	2	140.9	127
post	223	34	247.67	63	*wasnt*	74	0	137.49	133
guy	274	65	243.05	65	amazing	159	41	133.72	138
guys	166	13	228.28	70	*gym*	66	0	122.63	149

LL: Log likelihood

Another feature which can be observed from the list of overused words in personal blogs is that some typical markers of online discourse rank very high. For instance, *lol* (an acronym for *laughing out loud*) ranks the 18th among the top 150 word-forms. Apart from *lol*, the word-form *anyways* is another new word which only appears in online discourses. By deviating from the established norm, bloggers are actually representing themselves in a different way. Looking from this perspective, we can take the employment of markers of online discourse as an identity marker. Of course, the spelling norm is not the only rule under challenge. If we look at word-forms such as *awesome* and *sucks* - two words (or word forms) which have gained currency quite recently in spoken English, we will form an impression that bloggers do not seem to care much about the regulations about lexical choice (which have mainly been prescribed by language experts or representative enforcers of the standard variety) either. What they seem to be doing is use whatever linguistic materials readily available to achieve their intended communicative purposes. One of the most readily available linguistic resources is their spoken discourse repertoire. This is also one of the reasons why slang words such as *awesome* and *suck* are frequently used in the blog corpus.

4.4 Keyword comparison with BNC Sampler Corpus Written

One recurrent claim about blog texts in existing literature is that blogging is a hybrid of speech and writing. If this is the case, comparing personal blog data with formal writing data should be able to reveal more of the spoken features of the former. Based on this consideration, I have carried out a key word comparison between EBC and the BNC Sampler Corpus (Written) with the help of Wmatrix. Table 4.5 below lists some of the top 150 word-forms. This list is slightly different from the one with the BNC Sampler Corpus Spoken as the reference corpus. One striking feature is the huge difference between two corpora in the use of self-mention words as represented by the top three key words: *I, my, and me,* echoing what has been observed from Table 4.4. Another striking feature is the high frequency of private verbs and verbs for emotion expression, for instance, *get, got, getting, know, think, like, want, feel, love, hate, and feel like*. These words are also frequently used in spoken or informal discourses. Again, this feature is a reflection of the generic feature of personal blogs. One of the most important functions of personal blogs is to voice out bloggers' feelings, emotions, thoughts, and reflections. If we look at the word-forms with zero or extremely low occurrences in the BNC Sampler Corpus Written, we will get a more concrete picture about the language of personal blogs.

Table 4.5 Key words in EBC relative to BNC Sampler Corpus Written

Word	EBC	BNC_WR	LL	Rank	Word	EBC	BNC_WR	LL	Rank
i	32672	6904	30507.38	1	myself	627	142	562.47	41
my	8279	1914	7335.45	2	getting	574	111	562.2	42
'm	3446	375	4269.96	3	stuff	433	46	540.62	43
me	5194	1438	4085.05	4	anyway	474	68	532.04	44
so	4825	1503	3469.3	5	shit	319	7	526.55	45
n't	5089	1758	3351.78	6	lol	284	0	522.65	46
just	3191	919	2438.26	7	na	339	14	520.29	47
do	4199	1682	2397.25	8	gon	267	2	469.8	53
it	10796	8226	2208.02	9	guy	274	10	427.84	60
really	1984	296	2191.66	10	mom	213	0	391.99	68
get	2109	457	1941.43	11	yeah	313	33	391.72	69
've	1573	303	1543.78	12	ok	281	20	390.36	71
am	1535	288	1526.85	13	thats	198	0	364.38	73
got	1560	321	1478.24	14	fuck	225	8	352.53	75
2007	799	0	1470.41	15	right_now	224	9	345.16	78
im	726	1	1321.9	16	hate	280	32	341.47	80
mood	781	37	1173.49	17	ive	182	0	334.94	81
nt	637	4	1127.75	18	weird	191	4	316.55	85
like	2218	1052	1043.05	19	awesome	189	5	306.45	86
know	1571	552	1019	20	cant	161	0	296.29	91
think	1372	457	932.98	21	kinda	183	5	295.72	92
going_to	979	201	928.82	22	haha	160	0	294.45	94
2008	487	0	896.23	23	hopefully	184	7	285.7	98
today	1047	273	858.68	24	anymore	156	2	267.67	102
things	1040	295	803.66	27	feel_like	181	10	264.8	104
because	1397	562	794	28	guys	166	6	259.53	106
want	1167	395	782.61	29	cool	240	38	258.51	107
fucking	432	16	673.22	30	damn	176	11	251.41	109
feel	739	168	661.45	32	cos	126	0	231.88	116
go	1142	457	652.65	33	crap	147	5	231.63	117
nice	531	75	599.72	35	yay	119	0	219	124
actually	603	116	592.2	36	blog	115	0	211.64	127
'll	1056	431	591.1	37	ass	113	2	189.82	144
pretty	433	34	587.79	38	2006	103	0	189.55	145
love	735	201	583.18	39	pretty_much	112	2	188.01	146
fun	452	45	575.59	40	blah	101	0	185.87	150

Among the zero or low frequency word-forms in BNC Sampler identified by Wmatrix are: *im*, *pretty, fucking, stuff, thats, na, shit, lol, ok, gon, fuck, right now, weird, awesome, cant, kinda, haha, guys, damn, cos, crap, ass,* and *blah.* Some of them have been identified because of their unconventional spelling, for instance, *im*. Some are colloquial expressions which are seldom used in formal settings, for instance, *pretty*, *fuck*, *fucking, stuff*, *shit*, *crap*, *damn,* and so on. Some are spoken or online discourse features, for example, *gon (gonna), na* (*gonna or wanna*), *lol*, *ok*, *kinda*, *haha*, and *blah*. What can be concluded from the substantial presence of these words or word forms in the EBC is that the language of personal blogs is characteristic of oral discourse features. This echoes the dominant impression in existing literature that personal blogs are a hybrid of speech and writing.

4.5 Chapter summary

This chapter presents a very rough sketch about what the language of personal blogs looks like. By comparing the top 20 most frequently used word forms generated from the EBC with those generated from the CIC Spoken and CIC Written, we see that the language of personal blogs is different from both conventional spoken language and conventional written language in that the first person singular pronoun (*I*) ranks ahead of the definite article (*the*). Although we can take the prominent presence of the first person singular pronoun (*I*) as an indicator of the nature of personal blogs as a genre for self-expression, it is still somewhat out of expectation to see the definite article (the) ranking the second on the top 20 wordlist. Further examination concerning the distribution of these two words across the texts produced by bloggers from different age and gender groups points to the influence of bloggers' effort in expressing their age- and gender-related identity. A further comparison between the wordlist generated from the EBC and those from BNC Sampler Corpus Spoken and BNC Sampler Corpus Written shows that the language of personal blogs is a hybrid of speech and writing, as existing studies have already revealed.

Examining the most frequently used word-forms in the EBC can give us a flavor of the language of personal blogs and offer us some clues about the potential link between linguistic variation and certain aspects of bloggers' identities such as age and gender, but it can only tell us something very general and impressionistic about the aggregated whole. It cannot tell us much about the specific features of blogging language and its respective constituting components. As the focus of the current research is on linguistic variations in personal blogs and their relationship with bloggers' identity representation, examining an aggregated wordlist is and should only be taken as a starting point for more specific and local analysis. In the following four chapters (Chapters 5-8), I will give a more detailed discussion about the linguistic variations in terms of orthographic representations (Chapter 5), lexicological strategies (Chapter 6), preferences for semantic domains (Chapter 7), and grammatical features and pragmatic features (Chapter 8). In Chapter 9, I will explore the links between these variations and their significance in bloggers' representation of identities.

Chapter 5 Orthographic Variation

This chapter presents a detailed description about the six linguistic strategies that bloggers employed in realizing orthographic variation and the functions of the non-conventional orthographic representations of words. A discussion about the use of orthographic symbols as represented by the asterisk in the blogging texts is also presented.

5.1 Categorizing orthographic variation

Compared with the syntactic and semantic aspects of a language system, the lexical aspect seems to be the one over which a language user has a greater control. That may be one of the important reasons why lexis is widely acknowledged as the most active part of language change. Among the various lexical aspects of the English language, the orthographic representation of word forms seems to be quite vulnerable to linguistic manipulation. As mentioned earlier, writing has an established set of norms regarding the orthographic representation of word-forms. Deviation from these norms is normally not encouraged, especially in written publications. One exception would be in literary writing where deliberate deviation is sometimes employed as a technique for achieving special effects. In this case, deviation from the norm is a privilege entitled to professional writers. The flourishing of personal blogs has offered the general public a channel to publish their own writing with no others-imposed censorship and editing. As a consequence, whether to comply with or deviate from the established norms as represented by conventional publications becomes a matter of choice.

If a language user chooses to deviate from the conventions regarding the lexical aspect in writing, there are a number of ways of doing it. However, the easiest way would be to change the outlook of word-forms by engineering on their orthographic representations. An English word can be orthographically engineered in several ways, for instance, abbreviating, lengthening, replacing letters or morphemes, blending two words together, shifting between upper and lower cases, adding other orthographic symbols, and so on. Identifying orthographic variation in the EBC is not a difficult task, as most mainstream linguistic analysis software tools can generate a wordlist out of a corpus in just a few seconds. Nevertheless, trying to pool all these word-forms together for pattern analysis could be very problematic, as no language processing tools can tell whether a word-form is a new one or not. This is where the semantic annotation tool of Wmatrix can contribute a great deal. As introduced in Chapter 3, Wmatrix automatically assigns a semantic domain of Z99 for any word (word-form) which does not belong to its own lexicon and pools all such words or word-forms together under a category labeled "unknown words" for users to download for further analysis. Users can then conduct manual recategorization of these "unknown words" according to their own schemes.

Among the 689,437 word-forms of the EBC, Wmatrix identifies 16,587 unknown items. Among these items, 3,906 are common words in disguise: they are identified as "unknown" due to the encoding problems arising out of the file conversion process. In other words, there are approximately 12,681 actual "unknown" word-forms. These word-forms can be roughly classified into two types: the ones which have undergone orthographic engineering and those which have not. The former mainly consist of proper nouns and new words as compared with the BNC Sampler Corpus whereas the latter comprise word-forms which are different from their conventional forms for whatever reasons. I have adopted a six-category scheme for classifying the orthographically engineered word-forms according to the strategies (reasons) involved. These categories include: 1) unconventional contracted forms (word-forms resulted from omitting the apostrophe, for instance, *dont* for *don't*), 2) abbreviations (word-forms resulting from deliberate shortening of any kind), 3) letter repetition (word-forms involving repetition of one or more letters), 4) e-paralinguistic words (word-forms imitating laughter and other non-verbal behaviors), 5) misspellings (word-forms resulted from slips of the keyboard and intentional erroneous word-forms), and 6) phonetic spellings (word-forms resulted from attempts of mimicking how words are actually pronounced in speech by the blogger or other people). Table 5.1 shows the details. Each of these categories will be discussed in turn and their functions will be explored in the following sections.

Table 5.1 Categorization of unknown word-forms

Total Sample Size		689,437
Total Unknown Word-forms		16,587
Normal Words in Disguise		3,906
Actual Unknown Words		12,681
Orthographically Unengineered	Names	2,839
	New Words	1,704
	Interjections	331
	Vulgar Terms	212
	Words with Asterisks	287
	Others	518
Orthographically Engineered	Non-conventional Contracted Forms	1,839
	Abbreviations	1,757
	Letter Repetition Words	599
	E-Paralinguistic Words	304
	Misspellings	1,004
	Phonetic Spellings	1,287

5.2 Non-conventional representation of word forms

5.2.1 Non-conventional contracted forms

As mentioned earlier, any deviation from the established norm of English spelling (be it intentional or not) will result in a new word-form which will contribute to the realization of orthographic variations. One simple way of violating the spelling norm is to omit the apostrophe in contracted forms, for instance, spelling *I'm* as *im*. There are two possible reasons for this kind of omission in blogging. First, omitting the apostrophe speeds up the typing. If a blogger wants to capture his or her flow of thoughts, he or she may choose to ignore those semantically unimportant orthographic details. Second, the apostrophe is a symbol deliberately designed to mark contracted forms but is unpronounceable itself, therefore, its omission would normally not cause comprehension problems for the readers. This in turn encourages bloggers to omit it. Of course, there are cases where this omission may cause confusion, for instance, spelling *I'll* as *ill*. The reader will have to rely on the context to determine whether the blogger is talking about a future action (or status) or a status of being unwell. Nevertheless, such cases are the absolute minorities. Not many word-forms involving the use of apostrophe are likely to cause such problem when the apostrophe is omitted. The following table (Table 5.2) shows the ten most frequently used non-conventional contracted word-forms in the EBC. These top ten word-forms account for 92% of the total number of occurrences of non-conventional contracted forms.

Table 5.2 Top 10 non-conventional contracted word-forms

Non-conventional Spelling	Conventional Spelling	Frequency
Im	I'm or I am	657
dont	don't or do not	323
thats	that's or that is	173
Ive	I've or I have	136
didnt	didn't or did not	118
Ill/ill	I'll or I will/shall	88
wont	won't or will not	71
wasnt	wasn't or was not	54
haven't	haven't or have not	40
couldn't	couldn't or could not	34
Total		1,694

5.2.2 Abbreviations

The second way of deviating from conventional spelling is through abbreviations. According to Plag (2003), abbreviation is a word-formation strategy which involves the amalgamations of parts of different words. Abbreviations are generally formed by assembling initial letters of multi-word sequences, though in some cases they do incorporate non-initial letters. Here in this research, the term "abbreviation" is used in a broader sense. Following Gong and Ooi (2008, p. 933), abbreviations refer to word forms created by removing one or more components of a word or phrase. They can be further classified into acronyms and initials (word-forms created by combining the first letter of each constituting words, for instance, *lol* for *laughing out loud* and *idk* for *I don't know*), clippings (word-forms created by taking away either the initial or ending part of a word, for example, *pic* for *picture*, and *toon* for *cartoon*), forms with total vowel omission (word-forms created by removing all the vowel letters from a word, e.g., *gd* for *good*, and *lvl* for *level*), and abbreviated compounds (word-forms created from keeping the initial letter of the first word of a compound, for instance *f-list* for *friend list*). Table 5.3 shows the details.

Table 5.3 Types of abbreviations

Type	Frequency	Percentage
Acronyms & Initials	1,019	58%
Clippings	648	37%
Total Vowel Omission	43	2%
Abbreviated compounds	38	2%
Others	9	1%
Total	1,757	100%

5.2.2.1 Acronyms and initials

A closer examination of the acronyms and initials present at the EBC reveals that more than half of them fall into two major categories: markers of online discourse and abbreviated noun phrases. By markers of online discourse, they refer to the initials and acronyms which are often used in online chat (be it public chatting or instant messaging) to express paralinguistic features (e.g., *lol* for laughing), emotions (*wtf* (*what the fuck*) for showing anger and *omg* (*Oh my God!*) for showing surprise), and other shorthands (such as *imo* for *in my opinion*). As Table 5.4 shows, there are 682 occurrences of such markers, accounting for 38.8% of the initials and acronyms. Abbreviated noun phrases account for 15% and they cover a variety of semantic domains such as technology (mainly information and communication technology), education, entertainment, place names, medical care, terms related to daily life experiences, company names, social organizations, and government departments. In fact, initials and acronyms are also commonly used in conventional writing, especially in journalism. Many a time, these abbreviated forms have to do with names of institutions (e.g., MIT), organizations (e.g., NATO), or job positions (CEO). From Table 5.4, we can see that this is also true for the blog data of this research.

Table 5.4 Top 10 subcategories of acronyms and initials

Subcategories	Frequency	Percentage	Rank
Markers of Online Discourse	682	38.8%	1
Technical Terms	62	3.5%	2
Education-related Terms	60	3.4%	3
Entertainment-related Terms	47	2.7%	4
Place Names	30	1.7%	5
Medical Terms	18	1.0%	6
Terms related to Daily Life	17	1.0%	7
Organization Names	12	0.7%	8
Company Names	10	0.6%	9
Government Departments	7	0.4%	10
Subtotal	945	53.8%	

If we take a look at the top ten acronyms and initials identified from the EBC, we will find that word-forms representing laughter or laughing add up to 384 occurrences, accounting for 21.9% (see Table 5.5 below). The word-form *lol* alone occurred 347 times, establishing itself as the most frequently used acronym in the EBC. Why a word-form describing a paralinguistic behavior occurs so frequently in a written genre is something that needs explanation.

Table 5.5 Top 10 acronyms and initials

Initialism	Frequency	Variants	Meaning
LOL	347	lol (296), LOL(38), lolz(2), lolol (2), lololol(2), LOLZ(1), LoLs (1), LOLOL (1), loll (2), lol'ed (2)	laughing out loud
LJ	89	LJ(77), lj(12)	LiveJournal
OMG	71	OMG(34), omg(22), OMFG(3), zmog (4), ZOMG(3)	oh my (fucking) God
WTF	48	WTF(29), wtf(16), Wtfeck(1), wtfed (1)	what the fuck
LMAO	30	lmao(16), LMAO(7), lmfao(6), WLmao(1)	laughing my (fucking) ass off
IDK	30	idk(26), idkk(2), idkkk(2)	I don't know
BTW	20	btw(13), BTW(4), b.t.w (3)	by the way
TBH	8	tbh(7), tbqh(1)	to be (quite) honest
FTW	7	FTW(4), ftw(3)	for the win
ROFL	7	ROFL(5), rofl (2)	rolling over the floor laughing
Total	657	Taking up 37% of the total number of acronyms and initials.	

Before we can explain the presence of paralinguistic features in personal blogs, we may need to understand how acronyms and initials are used in online discourses in general. The proliferation of acronyms and initials in online discourses may have a great deal to do with online chat. As Gong and Ooi (2008) point out, the time constraints chatters are facing may have triggered the extensive use of abbreviated forms in online chat. In other words, the principle of economy has played an important role. Abbreviation cuts down on the number of strokes needed for typing a word. Chatters use these forms to shorten the lapse between utterances so as to keep the communication going. Thus, the abundance of abbreviations in real-time textual chatting situations seems to be quite reasonable. The question is, however, can we still turn to the principle of economy for explanation when it comes to blogging where time constraint is no longer a major issue? The answer is both yes and no. The principle of economy may still have its relevance in certain cases. For instance, when the blogger wants to capture his or her online thoughts before they flash away, he or she will have to increase the typing speed and use abbreviated forms in order to save time. This is just one possibility. In fact, there are other possibilities. For instance, a blogger can use acronyms and initials to screen potential readers, as we will see later in this chapter.

5.2.2.2 Clipping

Clipping is the second most frequently used linguistic strategy for bloggers to shorten word-forms. There are 648 tokens of clipped word-forms, covering a total number of 103 word types. Compared with acronyms and initials, clippings are orthographically more complicated yet semantically less opaque. It is easier to recover the original spelling of a clipped word and this recovery is less dependent on the context and mutual knowledge. The reason is that in clipping, the more important part of a word will normally be kept, which makes the remaining part a better reminder of the original word. Table 5.6 shows 20 recurrent clipped word-forms. Many of the words are about dates, which reflect a very important discourse feature of personal blogs. For each blog entry, the blogging software will automatically add on the date and time. Depending on the region where the blogger is located and the template he or she chooses, the date and time will be displayed in slightly different ways. Certain templates use the clipped forms for names of the week and month. In fact, the most frequently used word-forms in Table 5.6 have other implications, too. The word-form *pic(s)* is an important one in personal blogs, as pictures are a very important part of blog entries. *Fiction* is another important word in blogsphere, as many bloggers write fictions either in their LiveJournal or in their own blogging sites. That is why the word-form *fic(s)* ranks the second. One more difference between initialism and clipping is that the latter seems to work more often with common words.

Table 5.6 Top 20 recurrent clipped word-forms

Clipped forms	Conventional Spelling	Frequency	Clipped forms	Conventional Spelling	Frequency
Oct	October	214	*esp*	especially	7
pic(s)	picture(s)	61	*fest*	festival	6
Sep	September	67	*hol(s)*	holiday(s)	5
fic(s)	fiction(s)	47	*eps*	episode(s)	5
xmas	Christmas	36	*prolly*	probably	5
emo	Emotional	32	*fave(s)*	favorite(s)	5
Fri	Friday	11	*pro*	problem	4
vid(s)	video(s)	10	*cig(s)*	cigarette(s)	4
Thurs	Thursday	7	*chem*	chemistry	4
Tues	Tuesday	7	*appt*	appointment	4

5.2.2.3 Total vowel omission

The third linguistic strategy for shortening word-forms is total vowel omission. This seems to be a rather recent phenomenon, although we may find this on road signs where limited space would lead to the removal of vowel letters. Another setting where abbreviation of this kind is often found is the mailing addresses. Wherever it is used, this is a typically written practice. There is no way to omit all the vowels in speaking because it is extremely difficult if not totally impossible to speak without vowels. In writing, on the other hand, it is possible for us to fully understand the meaning of a sentence if we remove all the vowel letters and keep the consonant ones. It is not that vowels are not important but that we can automatically recover the vowels according to our knowledge of literacy and the local context. As far as the cognitive effort is concerned, total vowel omission is the most demanding on the part of the blogger, because it works against the common spelling habit. This might be a reason why there are only 43 occurrences of total vowel omission occurrences in the whole corpus. Table 5.7 lists all the words which have undergone vowel removal process.

Table 5.7 Word-forms undergone vowel removal

New forms	Conventional spelling	Frequency	New forms	Conventional spelling	Frequency
tht	that	7	*plz*	please	1
yr(s)	years	6	*plzkthx*	please ok thanks	1
lvl	level	2	*plzthx*	please thanks	1
bck	back	1	*Thnx*	thanks	1
bldg	building	1	*rly*	really	1
Blvd	boulevard	1	*Rt*	route	1
bzns	business	1	*SMRT*	smart	1
cmnts	comments	1	*srry*	sorry	1
fwds	forwards	1	*sry*	sorry	1
gd	good	1	*std*	student	1
hve	have	1	*thngs*	things	1
Itr	later	1	*ths*	this	1
kthx	ok thanks	1	*txtd*	texted	1
MdM	madam	1	*utd*	united	1
msg	message	1	*VK*	vodka	1
nght	night	1	*Wht*	what	1

From this table we can see that most of the words are high-frequency everyday words and they resemble texting messages in many ways. Unlike in text messaging where the length of one text is limited to 256 characters, there is no word limit for blog entries. Obviously, the principle of economy cannot account for this phenomenon any more. Besides, the new forms are all semantically less transparent. We may wonder why bloggers use such odd forms. This issue will be picked up later in Section 5.2.7 where the functions of non-conventional word-forms are being discussed.

5.2.2.4 Abbreviated compounds

Abbreviated compounds are also a linguistic strategy for shortening words. These words or word forms are often created by truncating one of the constituent words of a compound.

As this strategy involves two different processes, it subsequently takes longer processing time. Besides, the final word-forms so produced are also semantically less transparent than the original spellings. Possibly because of this, there are only 38 occurrences of such word forms in the blog corpus, representing 15 word-form types. They are: *bday*/*b-day* (birthday)(16), *flist*/*f-list* (friendlist)(7), *ex-gf* (ex-girlfriend)(1), *ex-bf* (ex-boyfriend) (1), *ex-Mr* (ex-husband)(1), *IBM'ers* (users of IBM computer)(1), *jrock* (Japanese-rock)(1), *LJ-cut(1), LJfriends (1), ljsecret(1), RP-er* (player of role-playing game)(1), *sex-ed* (sex education)(2), *t-storm* (thunder-storm)(2), *V-day* (Valentine's Day)(1), and *wkend* (weekend)(1). Among them, *b-day* is the most frequently used, occurring 15 times in the blog data.

As can be observed from the description presented above, abbreviation is an important way for bloggers to create orthographic variations. However, shortening the word-forms is not the only thing which bloggers can do in deviating from the established orthographic norms of conventional writing, as will be demonstrated in the following section.

5.2.3 Letter repetition

Orthographic variation may not necessarily result from bloggers' efforts in shortening word-forms; it can also result from their efforts in lengthening word-forms. One way of lengthening a word form is through letter repetition. Letter repetition is a very basic orthographic strategy in the English spelling system: it simply increases the possibilities of using the same set of alphabets or orthographic symbols to represent more words or word-forms. Among the 26 English letters, more than half can be repeated and used as part of a word. Certain letters are seldom repeated, for instance, letters *A, H, J, K, Q, U, V, W, X,* and *Y*. Letters *I* and *N* also belong to this type but they occasionally appear in forming present participle forms of certain verbs, for instance *skiing* and *beginning*. A typical pattern (or local context) for consonant letter repetition is XX+LE, for instance, -*bble*, -*ddle*, -*ffle*, -*ggle*, -*pple*, -*ssle*, -*ttle*, and -*zzle*. For vowel letters, only two letters (*E* and *O*) appear in repetition to represent a sound which is different from what the letter represents in isolation. One of the reasons why letters *I* and *U* do not appear in repetition could be that repeated *Es* are used to contrast *Is* and repeated *Os* to contrast *U*s. The same may also be true for the two semi-vowel letters *Y* and *W*, which resemble *I* and *U* in many ways. Regardless of the context, there is one rule which applies to all cases which involve letter repetition in the English orthographic system: A letter can only be repeated twice and consonant letter repetition seldom occurs at the initial position of a word. What we can observe from the letter repetition patterns is that letter repetition is not random and arbitrary as it seems. There are actually certain regulations behind it. In other words, there is an established norm for orthographic representation of a standard language and people are expected to identify with this norm. The question is: do bloggers identify themselves with this norm? If no, what norms are they identifying with? Do bloggers from different age and gender groups identify with different norms? To answer these questions, we need to pool those word-forms with letter repetition together so that patterns of repetition can be investigated and association between age and gender and patterns of letter repetition can be studied. One thing needs to be clarified here is that the word-forms involving letter repetition I am going to account for below do not include those following the conventional spelling rules. In other words, only those unconventional letter repetition cases will be discussed. Again, this is where Wmatrix can play an important role.

Just like abbreviated word-forms which are identifiable because of their unconventional orthographic representations, word-forms created out of unconventional letter repetition will inevitably be categorized as unknown words by Wmatrix. Again, this provides a shortcut for me to identify them and pool them together for pattern analysis. As Table 5.1 shows, there are 599 occurrences of word-forms with letter repetition. These word-forms are actually orthographic variants of 183 word types, the top 25 of which are presented in Table 5.8 below. If we read those word types more closely, we will soon find that they roughly fall into three grammatical categories: inserts, intensifiers, and adjectives. According to Biber et al. (1999, p. 56), inserts are a relatively newly recognized category of word. They characteristically carry emotional and interactional meanings and are especially frequent in spoken texts (for information about subcategories of inserts in the English language, please refer to pp 93-94.). Intensifiers are lexical devices which are often used to strengthen or emphasize a comment or statement. Evaluative adjectives are often used to express speaker attitude in conversation.

Table 5.8 Top 25 word-forms created via letter repetition

Lexical Items	Grammatical Category	Frequency	Lexical Items	Grammatical Category	Frequency
so	Intensifier	111	*blah*	Insert	6
haha	Insert	41	*oh*	Insert	6
grr	Insert	38	*way*	Intensifier	6
ah	Insert	19	*eek*	Insert	5
really	Intensifier	19	*huge*	Adjective	5
yay	Insert	16	*yes*	Insert	5
argh	Insert	14	*aw*	Insert	4
ugh	Insert	13	*ow*	Insert	4
yeah	Insert	13	*bye*	Insert	4
aha	Insert	12	*good*	Adjective	4
and	Conjunction	10	*old*	Adjective	4
no	negation	9	*please*	Insert	4
well	Intensifier	7	Total		379

By repeating one or more letters of a particular word, a lengthened word-form will be created. The increased length of the word will naturally remind the reader of the lengthening of sounds in spoken language. In other words, spelling a word with repeated letters is actually reminding the reader that the word-form so produced should be emphasized. Or it is more accurate to say that letter repetition plays a similar role as what stressing or lengthening sounds plays in speech. When an intensifier is spelled with repeated letters, it expresses a stronger emotion or attitude than what the intensifier itself normally conveys. For inserts spelled with letter repetition, they can kill two birds with one stone: mimicking the manner these words may be pronounced (for instance, the lengthening of certain sounds) and expressing a stronger emotion. For evaluative adjectives, letter repetition is just an innovative way of emphasizing the intended meaning while at the same time inviting the reader to say these words in an accentuated way. To a great extent, letter repetition plays both semantic and pragmatic roles. Tables 5.9 to 5.10 give a flavor of how bloggers are manipulating the orthographic representations of words by applying the strategy of letter repetition.

Table 5.9 Orthographic variants of SO

Variants of SO	Frequency
soooo	33
sooo	31
soo	17
sooooo	15
soooooo	4
soooooooooooooo	3
sooooooo	2
sooooooooo	2
soooooooo	1
sooooooooo	1
sososo	1
ssoooooo	1
Total	111

From Table 5.9 we can see that the major pattern of letter repetition for *SO* is just repeating the vowel letter *O*, with only one exception which is actually a repetition of the whole word. This pattern seems to have something to do with the phonological realization in speech. If we want to emphasize the word *so*, normally we will lengthen the vowel sound. In a written situation mediated by the keyboard, repeating the letter *O* is as easy and natural as lengthening the vowel sound in speech. What the blogger needs to do is hold the *O* key and stops whenever he/she wants. As a two-letter (monosyllabic) word with the structure of CV (consonant plus vowel), there does not seem to be much choice in terms of letter repetition. A more natural choice would be repeating the vowel letter. When it comes to disyllabic or multi-syllabic words, the repetition pattern will become less predictable, as can be seen from the orthographic variants of the word *really* (see Table 5.10 below). There are 16 variants for this word: almost all constituting letters have been repeated. Most of these resulted forms are not pronounceable. In fact, they are not meant to be pronounced as what the letter combinations might be suggesting. All these forms are telling the readers one thing: "I'm emphasizing this word." Whether this seemingly random repetition pattern is linked to some other pragmatic purposes will be discussed later.

Table 5.10 Orthographic variants of REALLY

Variants of REALLY	Frequency	Variants of REALLY	Frequency
realllly	2	reallyreally	1
realllyyyy	2	reeaaallllyyyy	1
reeeeeeeally	2	reeaalllly	1
realllllyyyy	1	REEALLLY	1
reallllyy	1	reeeeally	1
reallllyyyy	1	reeeeeallly	1
reallly	1	rreeaallllyy	1
realllyyy	1	rreeaallyy	1
Total: 19			

To summarize, the word-forms presented in these two tables do not seem to follow a fixed pattern. Different bloggers have adopted different standards in terms of how long a word should be and which letters should be repeated. It seems that the actual phonetic realizations of these words have not played a part except the stress triggered by the letter repetition itself. This seemingly random behavior of bloggers in terms of letter repetition actually has something to do with the "writing" instrument: the keyboard. The easy operation of hitting the keyboard makes the typing of one letter and the typing of repeated letters not much different. The only thing the blogger needs to do is hold the same key slightly longer. It is almost as effortless as a speaker lengthens a speech sound. If the bloggers have to literally write all these letters out, say, on a piece of paper, they may be less motivated to do so, or at least the length of the resulted word-forms will be shorter. They may simply work out a different way for achieving similar effects. Similar to orthographic variation resulted from abbreviation, variation caused by letter repetition is normally not intended for attaching a new grammatical or semantic feature to the resultant word-form, but rather for realizing stylistic and the pragmatic functions. This will be discussed in greater detail in Section 5.2.7.

5.2.4 E-paralinguistic words

Another type of non-conventional word forms comes from bloggers' efforts in trying to represent laughter or laughing in orthographic forms in their blog entries. As mentioned in Section 5.2.2, certain acronyms and initials are also results of bloggers' efforts in trying to represent some paralinguistic features in words. One typical example would be *lol* (*laughing out loud*), which is the acronym of a phrase describing a paralinguistic behavior. The more exaggerated form *rofl* (for *rolling over the floor laughing*) presents a more vivid image of the blogger when such forms are being read by the audience. Although, as readers, we can hear the laughter in our minds loud and clear when we come across these acronyms and initials, they are still descriptions of actions. Not every blogger likes this rather indirect way of expressing laughter or the action of laughing (though it adds more flavor of performance to the text), because there is an easier and more direct way of achieving the same effect: onomatopoeia – imitating the laughing sounds directly.

Table 5.11 Word-forms representing laughter

Word-forms	Frequency	Word-forms	Frequency
haha	163	hahahahahahaha+	1
hahaha	22	hahahahahahasj+	1
hah	14	Hahahahahhaha	1
hahahaha	8	Hahahha	1
hahah	6	Heh	40
hahahhaa	2	Hehe	32
hahahaa	1	Hehehe	5
hahahah	1	Heheh	1
hahahahaha	1	Hehehehe	1
hahahahahaha	1	Teehee	1
hahahahahahaha	1	Total	304

*The symbol + indicates that there are more letter following it.

Acronyms and initials such as *lol*, *lmao*, and *rofl* can present a very vivid image about the action of laughing, but they cannot present the quality of the laughter. Word-forms generated from directly mimicking the laughing sounds can show the difference between belly laugh and muffled laughter. According to Urbandictionary, *hehe* is muffled laughter which differs from *lol*, suggesting a sneaky aspect to that being laughed at and *teehee* is laughter gives out by school girls. Sometimes *lol* and *haha* (and its variants) can be used interchangeably but not always. Bloggers of different age and gender groups may have different preference for one of the two forms. This issue will be addressed later in Section 5.2.7.

5.2.5 Misspellings

Spelling a word wrongly either unconsciously or deliberately will produce a new word-form. This is also a major contributor of orthographic variations. Just like there are slips of the tongue in speech and slips of the pen in handwriting, there are slips of the keyboard in keyboard-mediated writing. In fact, handwriting and typing do not work exactly the same way. When writing by hand, the actual process of spelling a word on a piece of paper or a notebook page is strictly linear and is carried out by a hand holding a writing instrument. In other words, the writing hand and the writing instrument are working together as a whole. As the handwriting process can be affected by the writer's mood, status of concentration, language proficiency, and time constraints, spelling mistakes are unavoidable. Nevertheless, misspellings are more likely to be caused by the incomplete recall of the orthographic representation of the language in the writer's mind. Typing, however, involves a more complicated process. It is still linear, which is determined by the basic nature of human languages, yet, the linearity becomes more likely to be affected due to the QWERTY arrangement of the keyboard and involvement of ten fingers in the typing process. With the linearity of conventional handwriting being replaced by a coordinated action of ten fingers, chances for misspellings increase to a considerable extent. The easy operation of typing (pressing a key once will produce a letter on the monitor screen), the speed, and the ease of correction give rise to many misspellings which do not normally occur in hand-writing situations. Meanwhile, the relatively easier operation of typing (as compared with handwriting) also makes it easier for the writer to manipulate the orthographic forms of words if he or she chooses to do so. For instance, in conventional handwriting, very few people will spell the word *the* wrongly. In typing, however, it is highly likely to type *the* as *teh.* In fact, this misspelling form (*teh*) has become so common that many people have started to deliberately spell it this way to achieve stylistic effects in online discourse such as online chat. The auto correction function of mainstream text processing software such as Microsoft Word has contributed to a growth of indifference to common misspellings, as they will be automatically corrected anyway, which in turn increases people's tolerance of misspellings, especially in not-so-formal writing situations. Moreover, there are some inherent problems with the English spelling system, which have actually become the source of the spelling confusions and misspellings.

As a matter of fact, spelling confusion has long been a concern of the British-based Spelling Society, an international organization that has advocated simplified spellings since 1908. According to the survey results of this society in 2007, more than half of their 1,000 British adult informants could not spell *embarrassed* or *millennium* correctly and more than a quarter struggled with *definitely, accidentally* and *separate* (Fitzpatrick, 2008). Misspellings are so common in papers submitted by first year undergraduate students in the United Kingdom that many lecturers find them very annoying. According to a news report written by Fitzpatrick (2008), a British university lecturer Ken Smith suggests that "we'd be better off letting the perpetrators off the hook and doing away with certain spelling rules altogether." Among the ten words most frequently misspelled identified by Ken Smith from his students' papers are *Febuary* (instead of *February*), *twelth* (instead of *twelfth*), and *truely* (instead of *truly*). All these words involve certain confusion over silent letters.

With the help of Wmatrix and manual identification, 933 erroneous word-form types (1,004 word-form tokens) are sorted out from the EBC. These word-forms do not include the ones which are deviated from the conventional spellings yet they seem to comply with certain phonological rules in one way or another. The latter will be discussed in details in the section to come. Table 5.12 lists the top 20 English words which have been most frequently misspelled in the blog data for this research. Strangely enough, the most frequently misspelled words are not difficult words; rather, they are all commonly used ones. Table 5.13 lists six words and their spelling variants. Some are obviously typos – spelling mistakes resulted from haste in typing, for instance, *beause* for *because*, leaving one important letter out whereas the most frequently form *becuase* may have a great deal to do with the writing instrument - the keyboard. The word *tomorrow* seems to have aroused some confusion among certain bloggers in terms of which letter should be repeated: *m* or *r* or both. The word *definitely* seems to be another problem for some bloggers. Many people do not seem to be happy with the conventional spellings of *until* and *argument* thus they have decided to add on one letter to each word to make them look more complete.

Table 5.12 Top 20 English words easily misspelled in blogs

Lexical Items	Frequency	Lexical Items	Frequency
because	12	experience	4
until	9	extremely	4
tomorrow	8	filming	4
argument	7	finished	4
definitely	7	going	4
about	5	just	4
received	5	sandwich	4
something	5	that	4
apparently	4	truly	4
decision	4	usually	4

Table 5.13 Examples of typos

Conventional Spelling	Erroneous Spelling
because	becuase (7), becasue (2), bacause (1), beacuse (1), beause (1)
tomorrow	tommorow(5), tommorrow(3)
until	untill (8), unti (1)
definitely	definately (3), definatelly (1), definetly (1), definiterly (1), defintely (1)
argument	arguement (5), arugment (2)
about	aout (2), abbout (1), abotu(1), nabout(1)

From Tables 5.12 and 5.13 we can see that not all the words which have been misspelled involve confusion over silent letters. According to Barbara Wallraff, a columnist of the *Atlantic* and *King Features Syndicate* who writes about language and writing problems, “people who spell a lot of words incorrectly either aren’t paying attention or don’t care,” therefore, there is no need to accommodate them (Fitzpatrick, 2008). Wallraff’s words make sense to a certain extent but they do not always hold. If many people (literate and highly literate people) are making more or less the same spelling mistakes again and again, the issue is no longer a matter of attention or care. There must be something in the orthographic system which allows such deviations.

A closer examination of the misspellings sorted out from the EBC shows that they can be roughly divided into the following four types:

1. Slips of the keyboard;
2. Misspellings induced by spelling confusion such as silent letters;
3. Misspellings induced by incompetence;
4. Intentional misspellings

Among the four types of misspellings, intentional misspellings are the most difficult to determine. One important principle for determining the intentionality of a misspelled form is to check whether similar forms are spelled in similar ways by the same blogger. In other words, spelling consistency is one of the most important criteria. To obtain this kind of information, both the local context of a misspelled form and a wider context (that is, all the blog entries written by the blogger) need to be consulted. Table 5.14 illustrates my point. From this table we can see that all the words ending with *–ing* are spelled with a new ending of -*ign* (with one exception) and these spellings all come from the same blogger. There is apparent consistency in these misspellings; therefore, we cannot say they are simply slips of the keyboard. As for what this obvious intentionality aims at, it will be discussed later. Of course, this is just one type of intentional misspelling and no obvious phonetic/phonological principles seem to be at play. Similar manipulation of the orthographic forms could be found in misspellings such as *ahd* (for *had)*, *ahve (*for *have)*, *ym* (for *my)*, *crhis* (for *chris)*, and *crhurch* (for *church)*. None of these new forms can be properly pronounced. In misspellings such as *dwunk* (for *drunk)*, *hoinh* (for *going)*, and *engliz* (for *English)*, we can see a deliberate replacement of one or more letters for effects. As for misspellings like *tiem* (for *time)*, *liek* (for *like)*, *langwadge* (for *language)*, and *kicced* (for *kicked)*, we can still feel the influence of phonological factors.

Table 5.14 Examples of intentional misspellings

Misspelling	Conventional Spelling
startign	starting
workign	working
buildign	building
callign	calling
somethign	something
somethign	something
talkig	talking
transferign	transferring

Leaving misspellings uncorrected and deliberately changing the conventional orthographic forms of common words will both run the risk of reducing the intelligibility of the resultant word-forms and thus demand more efforts on the part of the reader (in deciphering the secret codes). There must be something which is even more important than the correctness of the orthographic forms. From the random nature of the variant spellings I have demonstrated above, we may conclude that bloggers are not really intended to alter the shape of the spelling for its own sake. Nor do they seem to be interested in actually decreasing the intelligibility. What they are aiming at seems to have something to do with the pragmatic aspects of language use, as we will see in Section 5.2.7.

5.2.6 Phonetic spellings

Another major contributor of orthographic variations is phonetic spelling. Strictly speaking, phonetic spelling is a subcategory of misspellings discussed in the previous section. One major difference between the misspelling and phonetic spelling is that the latter can be explained from the phonological perspective. All the phonetic spellings in discussion are results of bloggers' endeavor to better approximate the actual pronunciation of the words intended. That is to say, the bloggers have chosen these forms because they think the new forms are better orthographic representations of the words intended. The identification of phonetic spelling is exactly the same as that of identifying misspellings, which involves the use of Wmatrix for initial identification and manual categorization afterwards. The defining principle for phonetic spelling identification is that the misspelled forms must have something to do with the actual pronunciation of the words in discussion.

There are altogether 1,287 occurrences of word-forms which can be called phonetic spelling, representing 723 word types. Table 5.15 lists the top 40 words which have been most frequently spelled according to their actual pronunciations. These words cover 32% of the total number of phonetic spelling tokens.

Table 5.15 Top 40 words undergone phonetic spelling

Lexical Items	Frequency		Lexical Items	Frequency
a lot	50		you know	8
cause	33		tomorrow	8
sort of	20		apparently	8
you all	19		gonna	7
freaking	17		never mind	7
at least	13		surprise	7
out of	12		in fact	7
each other	12		love	6
weird	12		happened	6

something	12	disappointed	5
like	11	ever	5
I must have	10	hello	5
definitely	10	interesting	5
as well	10	let me	5
damn it	10	trying	5
night	10	already	4
absolutely	9	awkward	4
in front	9	basically	4
surprised	8	chilling	4
whatever	8	god damn it	4

The 1,287 phonetic spelling tokens can be categorized into four types according to the strategies involved in orthographic engineering. The first type (Type 1) involves the omission of silent letters. In this type of phonetic spelling, letters which are not pronounced or are not thought to be pronounced are normally removed from the orthographic representation of the word. For instance, one way of spelling the word *whatever* phonetically would be *watever* since the letter *h* is silent in the actual pronunciation. The second type (Type 2) involves the replacement of the original letter or letter combinations by new ones which are considered to be better representation of the actual pronunciation. For instance, many people think *phone* should be spelled as *fone*, as the letter *f* better reflects the actual sound *[f]* than the letter combination of *ph*. This type of phonetic spelling enjoys greater diversity, as different people may have different understanding about how certain words are actually pronounced. This might be affected by the regional variety of English the blogger is speaking. The third type (Type 3) involves the combination of two separate words into a single one. For instance, *a lot* is often spelled as *alot*. There are also cases which involve two processes at the same time. For instance, *sort of* is sometimes spelled as *sorta*. This actually involves the combination of *sort* and *of* and then the replacement of *of* by *a*, thus *sorta*. For cases like this, the actual spelling decides which category a word-form will go. If it is spelled as *sortof*, it goes to third type (combination); if it is spelled as *sorta*, it goes to the second type (letter replacement). The fourth type (Type 4) involves playing with the pronunciation. Usually these word-forms are the orthographic representations of some funny ways of pronouncing certain words. For instance, *absent-minded* is spelled as *apsind-minded* as a representation of a funny way of saying the word.

Among the four types of phonetic spelling, Type 2 is the most frequently used and Type 4 is the least frequently used. In fact, Type 4 can also be considered to be a special type of letter replacement (Type 2). Table 5.16 lists the major strategies.

Table 5.16 Categories of phonetic spellings

Category	Strategies employed	Frequency
Type 1	Omission of silent letters	472
Type 2	Letter replacement	639
Type 3	Word infusion	170
Type 4	Playing with the pronunciation	6
	Total	1,287

Among the 472 tokens of Type 1 phonetic spelling (representing 302 word-form types), 123 (representing 77 word types) involve the omission of the letter *g* from the present participle ending –*ing*, accounting for 26%. This omission has resulted in a new ending –*in*, which is actually a closer approximation of how –*ing* is pronounced by some people or certain ethnic groups, especially in informal situations. Table 5.17 shows the top 15 words bearing phonetic spelling with silent letters omitted. If we compare the conventional spellings with their corresponding phonetic spellings in Table 5.17, we can easily notice that all the missing letters (but one) are silent ones. The only exception is the phonetic spelling (*lemme*) of *let me*. In this case, the letter *t* of the word *let* is not really missing but rather replaced by the letter *m* to better reflect the assimilation in the actual pronunciation.

Table 5.17 Top 15 words spelled with silent letter omission

Conventional Spelling	Phonetic Spelling	Frequency
freaking	*freakin*	17
surprise(d)	*suprise(d)*	15
something	*somthing*	12
absolutely	*absolutly*	9
happened	*happend*	6
disappointed	*disapointed*	5
let me	*lemme*	5
trying	*tryin*	5
already	*alredy*	5
hello	*helo, 'ello*	5
chilling	*chillin*	4
smoking	*smokin*	4
awkward	*akward*	4
basically	*basicaly*	4
whatever	*watever*	4
actually	*actualy*	3

As mentioned earlier, Type 2 is the most commonly occurred among the four types of phonetic spelling. There are 639 occurrences altogether (representing 395 word types), accounting for 49.7% of the total number. Table 5.18 shows the top 15 words which have been spelled with one or more letters being replaced. Again, if we compare the conventional spellings with the corresponding phonetic spellings, we will notice that most of the phonetic spellings appear to be closer to the actual pronunciation of these words. Some of the phonetic spellings even reveal regional differences. For instance, *cuz* [kʌz] has been used by American bloggers only, showing that it is a form mainly used in American English. The word-forms *intresting* and *intrested* have only been used by British bloggers. As British people tend to pronounce the words *interesting* and *interested* [ˈɪntrɪstɪŋ] and [ˈɪntrɪstɪd], they may find the conventional spellings do not match the actual pronunciations very well and thus some people prefer the phonetic spelling forms. The Americans, on the other hand, may not find the conventional orthographic representations problematic because it is quite common to for them to pronounce the words *interesting* as [ˈɪntərestɪŋ] and *interested* as [ˈɪntərestɪd]. One more example which probably shows the influence of American accent in phonetic spelling is the word-form *tomarrow* (for *tomorrow*). American English speakers tend to pronounce the word as [tə ˈmɑrəʊ], with the second vowel sound at more open and back position. That may explain why the word-form *tomarrow* is preferred by some American bloggers.

Table 5.18 Top 15 words spelled with letter replacement

Conventional Spelling	Phonetic Spelling	Frequency
because	*cuz (AmE)*	30
sort of	*sorta*	20
you all	*yall, y'all*	19
(to)night	*(to)nite*	18
out of	*outta*	12
weird	*wierd*	12
like	*liek, lyk*	11
definitely	*definately*	10
I must have	*I'ma, I'mma*	10
interesting (intrested)	*intresting (intrested)(BrE)*	8
gonna	*gunna, goona, gonne*	7
love	*lurve, lave*	6
whatever	*whateva, whatevah*	6
ever	*evar, evah, eh-ver*	5
tomorrow	*tomarrow(AmE)*	5

Unlike the first two types of phonetic spelling which try to orthographically represent as closely as possible how certain words are actually pronounced, Type 3 is rather a roundabout way of indicating the phonetic nature of the new spellings. By infusing two words together, the blogger is actually indicating that these two words should be pronounced as a whole. For example, infusing *damn* and *it* into *damnit* does not really tell the reader how the blogger tends to read it, but the reader can still hear the pronunciation *dammit* in his or her mind based on their knowledge about how it is normally pronounced when these two words are combined together. As mentioned earlier, some word-forms in Type 2 have gone through two processes: infusion and letter replacement, for instance, the word-form *sorta* is actually resulted from the combination of *sort* and *of* and then replacing *of* by the letter *a* to reflect the actual pronunciation. As the categorization is mainly based on the final word-form, *sorta* has been included in Type 2 rather than Type 3. There are 171 occurrences of infused word-forms. Table 5.19 shows the top 10 infused word-forms.

Table 5.19 Top 10 infused word-forms

Conventional Spelling	Phonetic Spelling	Frequency
a lot	alot	50
damn it	damnit	14
at least	atleast	13
each other	eachother	12
as well	aswell	10
a bit	abit	10
in front	infront	9
you know	y'know	8
never mind	nevermind	7
in fact	infact	7

Type 4 is the least frequently occurred one among the four, nevertheless, it is not unimportant. It is actually the only type of phonetic spelling which involves obvious innovation and creativity on the part of the blogger. For instance, the word-forms *heylo* and *sexction* can achieve certain special effects which other forms like *hello* and *section* cannot achieve. The former implies a naughty way of saying *hello* and the latter is a humorous way of spelling *section*, indicating a playful tone.

To summarize, phonetic spelling is an important linguistic strategy which bloggers employ to realize orthographic variations. In fact, it is the second most important strategy only next to abbreviations. From the strategies bloggers have used in creating these unconventional spellings, we can see that users of English are fully aware of the inconsistency between the English spelling system and the sounds of the language. In fact, the academic circle has long noticed the problems with the English spelling system.

According to James B. Carter (2006), the English spelling system is “archaic and dysfunctional” (p. 83) and “the spelling of English is now practically unique in being so far away from a consistent phonetic basis” (p. 90). Based on this judgment, Carter proposes that essential reformation of the English spelling system should be conducted to make it more logical, consistent, and easier to learn. Carter is not the first one to make such proposals. There have long been debates in Britain and the United States about whether and how English spelling reforms should be conducted. Nevertheless, no solutions have yet been found. The existence of various phonetic spellings echoes the concerns of advocators of English spelling reforms. Of course, just like online chatters who tend to use phonetic spellings, bloggers who adopt phonetic spellings may have no intention to demonstrate their support for an English spelling reform; they have other purposes in mind. The following section will focus on discussing these purposes.

5.2.7 Functions of non-conventional word forms

The previous five sections have focused on describing the linguistic strategies bloggers employed in realizing orthographic variations. This section explores their possible functions. Changing the spelling of a word (even a high-frequency one) many have two immediate consequences: making the text more difficult and making the text look different. The former is apparently not what bloggers are intended for as it runs against the very basic principle of human communication. When a blog entry is put online, it is meant to be read, although some bloggers claim that they do not care whether their blogs will be read or not. If the blogger is not intended to make their entries more difficult, they may well want to make their blogs look different. At the surface level, changing the spelling of words will make the resultant text different from conventional writing.

Normally, it will make the text more informal because the defining nature of informality is deviation from the established norm (which is represented by and maintained in formal conventional writing), especially in terms of the orthographic representation of words. To a certain extent, this is a just stylistic choice. At a deeper level, however, we can take this as bloggers' deliberate efforts in representing their identity. Many of the non-conventional forms created by blogs cannot be found in offline writing situations. From these non-conventional forms we can see very clearly the intentionality of the bloggers in this regard. As mentioned earlier, conventional spelling is an essential part of the norm which aims to maintain the orthodox of the standard variety of a language. This norm is maintained and reinforced by professionals, the media, the publishers, and the educational system. It is something imposed on the members of a society and is supposed to be taken as a part of people's collective identity as a native speaker of a language. The omnipresence of this imposed norm makes the choice of being deviated from it more prominent and meaningful. In other words, choosing to deviate from this norm is an identity marker itself, which not only distinguishes bloggers from non-bloggers but also distinguishes themselves from their real life writing styles. At a local or more contextual level, non-conventional word-forms can be used to achieve pragmatic functions. Despite the similarities in being able to alter the outlooks of orthographic forms of words, the five major categories of orthographic engineering strategies seem to display different preference for the realization of pragmatic functions. In order to depict a clearer picture about what each strategy can contribute to the realization of pragmatic functions, I will discuss each of them in turn.

The use of non-conventional contracted forms can be taken as a stylistic marker. As mentioned earlier, using contracted forms in writing is already a marker of informality.

Omitting the apostrophe seems to be able to make the resultant word-form even more informal. Whether the omission of apostrophe is associated with the linguistic representation of identity of bloggers from different age and gender groups is the topic to be discussed in the following section.

Abbreviations, regardless of the cognitive efforts required on the part of the blogger (for instance, acronymy and initialism are cognitively less demanding than total vowel omission), can actually reduce the amount of typing and may thus increase the typing speed. However, the resultant word-forms may increase the readers' decoding efforts unless they are familiar with what these shorthands stand for. This is especially the case for acronyms and initials of noun phrases. As Table 5.4 shows, 27% of the acronyms and initials are related to noun phrases, especially of names of various kinds. Semantically speaking, acronyms and initials are more opaque than their full version, especially to outsiders. A blogger decides to use an acronym or initial instead of the full version normally because he or she knows whether the target audience is not going to have great problems understanding it. Using a clueless initial or acronym does not make any sense as it is the most opaque kind of orthographic representation of words and the only consequence is communication breakdown. A more sensible interpretation would be acronyms and initials (especially those related to proper nouns) are used to screening target audience. As Plag (2003) rightly points out, "within certain groups of speakers, the use of an abbreviation can be taken as a marker of social identity: speaker and listener(s), but not outsiders, know what the speaker is talking about" (p. 129). They can also be taken as markers of ingroup membership. This ingroup may be as big as a country or as small as a group of role playing game participants. To cite just one example, the initial *RFT* (River Front Times) is an ingroup marker for those bloggers who read or at least know this local American newspaper. Blog readers who come across this initial but cannot understand it are apparently not the intended ones. As mentioned earlier, acronyms and initials relating to nouns of various kinds only occupy 27% of the total. In fact, the biggest category among all abbreviations for the current study is the group of word-forms which are often associated with online discourses, for instance, *lol* and *omg*. Apart from working as shorthand for their original semantic meaning and expressing their emotional flavor, these word-forms are also used as identity markers. This time, they mark off netizens from non-netizens and their virtual identities from their real life ones. Of course, bloggers from different age and gender groups may behave differently in their use of abbreviation strategies, as we will see in the next section.

In Section 5.2.3, I have given a rather detailed account of the word-forms created out of unconventional letter repetition. Different from abbreviations, the use of letter repetition will not make the semantic meaning of the resultant word-form more opaque. There are two reasons for this. First, word-forms with unconventional letter repetition are mainly very high-frequency ones (i.e. very common words), as can be seen from Table 5.8. Second, the original word is contained in the new word-form and easy to spot out. As a result, these word-forms are less likely to be used to screen readers as acronyms and initials are. They are definitely a stylistic marker of informal discourse as they do not follow a fixed pattern and are radically different from orthographic principles in conventional writing, often adding a playful tone to the text. To a great extent, using these forms can also be considered an identity marker which separates netizens from non-netizens. Of course, this is not the only function they are intended for. A more important function is to represent prosodic features in an orthographic way. The normal way of representing prosodic features in writing is through punctuations (such as exclamation mark and question mark) and textual description. Influenced by online chatters' practice, many bloggers choose to use letter repetition to represent emphasis. What we can see from this kind of effort is bloggers' intentionality of trying to infusing oral features into written texts to create the impression of informality.

Similar to letter repetition, word-forms representing different kinds of laughter are apparently not used for audience screening. Even their stylistic function may well be a by-product as their main function is to mimic paralinguistic behavior. What is really interesting is the presence of these word-forms in the written text. Laughter is something which often accompanies face-to-face communication: it is a typical oral discourse feature. Deliberately introducing typical oral discourse features into written texts is again a deviation from the conventional writing norm which not only makes the resultant texts more informal but also adds a flavor of performance to the whole act of blogging. This is quite similar to the use of letter repetition. In both cases, what the blogger readers are doing is not actually reading but rather listening to a piece of writing which is being read out aloud by the author.

Phonetic spelling, as another major strategy for realizing orthographic variation, can be used to achieve several purposes. The very first one should be stylistic as phonetic spelling is an obvious deviation from the conventional spelling. Phonetic spelling normally suggests intentionality to a lesser or greater extent on the part of the blogger. The reason is that the blogger has to invent (so to speak) a new orthographic representation for a word if he or she is not happy (or not so sure) about its conventional spelling. If the latter is the case, the blogger will apply his or her folk linguistic knowledge about letter-sound correspondence which is, more often than not, different from the original spelling. As has already discussed in Section 5.7, bloggers tend to apply three major strategies in trying to more closely approximate the actual pronunciation of words they are familiar with: omission of silent letters, letter replacement, and word infusion. As the new word-forms are different from the original ones, they will make greater sense if they are read out. This will also give readers an impression of listening to people reading out their stories. The frequent occurrences of word-forms with phonetic spelling may be another piece of evidence to show bloggers' intentionality in creating a new form of writing by infusing oral features into a written genre. This intentionality should also be considered an identity marker which separates bloggers from non-bloggers. Phonetic spelling is also an aspect of bloggers' language use where linguistic creativity can be observed.

The function of misspellings has a great deal to do with the nature of the misspellings, that is, whether they are intentional or unintentional. For unintentional misspellings such as slips of the keyboard or spelling errors resulted from bloggers' incomplete command of the word-forms, they can also be taken as a marker of informal style or even a marker of online discourse genres. The reason is not how ridiculously some bloggers are spelling the English words but rather their tolerance of these spelling errors. Again, this is a direct violation against the spelling conventions, revealing a rebellion against a collective identity. As for intentional misspellings, their functions are multi-fold. First of all, they are also a stylistic marker like those unintentional spelling errors as their presence gives the text containing them a different outlook from a conventional text and thus makes the text more informal. Second, they are also a marker of bloggers' linguistic creativity. Third, some intentional misspellings are meant for achieving some special effects which might have something to do with the identity of the blogger or someone the blogger is describing. The following two excerpts taken from the blog corpus for this study may help to illustrate the last point more clearly.

> **Excerpt 1**
>
> Jun. 6th, 2008
> Katie asked me to post this.
> hi! i *halp* momy! i *cary* the *luchbox*! its big! i *wak funy* to *cary* it to the room *thet gos* up and *don*! i *cary luchbox* to the car! im a *goo* girl! hi five!
> She is a good girl, that little scamp. She has taken on the morning duty of carrying M's lunch bag to the car for her. …

There are many misspellings in the second paragraph of Excerpt 1 (note the italicized word-forms). These misspellings are obviously intentional. The blogger is trying to mimic her little daughter’s way of speaking through intentional misspellings. If we look at the underlined words closely, we can find that the blogger employed different strategies to represent his daughter’s way of speaking. For instance, there are cases of omission of silent letters (*cary* for *carry, wak* for *walk*, *funy* for *funny*, *gos* for *goes*, *don* for *down,* and *goo* for *good*) and letter replacement (*halp* for *help, thet* for *that*, *luchbox* for *lunchbox*). These misspellings have presented a very vivid image of a lovely little girl who has just started to pick up the language. To a certain extent, the blogger is actually using the misspellings to represent the identity of a little girl. People may argue that this is just an example of using eye dialect to represent a character in literary works. This kind of argument makes much sense in this particular case. If we look at other cases where misspellings are intentionally used, we may find that misspellings can also be used for other purposes, as can be shown by Excerpt 2 below.

> **Excerpt 2**
> 12 Oct 2007
> So, I began the career this week - WP School now has +1 of me. I'm a Learning Support Assistant, and more excitingly one of only two men in the school. I have to set an example. Oh dear. I work with KS1's, split into two classes. I am focused on literacy due to my fantastic grasp of *teh engliz langwadge*. Innit. So that's spelling, letter formation, grammatical correctness and the increasing of vocabulary. There are 52 students, between the ages of early 6 and late 7. Of that, 13 Special Educational Needs students who require intense tutoring during lessons. And one student who arrived in this country a little over seven days

> ago with no english. I like that one best, because they try very hard. Today I taught them numbers 1 - 10, and how to write them.

If we look at the italicized words (*teh engliz langwadge* for *the English language*), we will find that these misspelled words are results of playing with the language. These misspellings can achieve a humorous and playful effect which their conventional spelling counterparts can never achieve. Misspellings can be used to express bloggers' mood as well, as can be shown in Excerpt 3 below.

> **Excerpt 3**
>
> October 19th, 2007
> the keyboard was a great idea...
> In case you missed it, I posted a bulletin on myspace a few nights ago. Intoxicated. Since then, I've read over and thought it was worth re-posting. Here it is:
> titled: is (I thought it was a facebook status)
> *definiterly* not completely sober, and wishes he had a girlfriend to date and kiss and hug.
> and yes, i realize there are typos and *thast* i should correct them, but i figure this way you get a better idea of my state of mind.
> i don't even like vodka. except maybe with orange soda, or *grapge* juice.
> but yes, i need a girlfriend. i'm a *loely* fuck. but i don't just want a *girlfirrnd* for the sake of having a girlfriend. it would just be nice to have someone special there for you.
> fuck girls.
> they're stupid.
> but some give me booze, <u>ad</u> for that i am eternally *greatful*.
> lolz. no one evens says that here. bummer. does anyone even say bummer anymore? what is wrong *witrh* me? huh?
> the sad truth is, i meant every word of it. [From us_m_18-19.txt lines 202-214]

The misspellings in this excerpt involve different strategies such as phonetic spelling (*definiterly* for *definitely*, *greatful* for *grateful*) and intentional misspellings (*thast* for *that*, *grapge* for *grape, loely* for *lonely*, *girlfirrnd* for girlfriend, *ad* for *and*, and *witrh* for *with*). The seemingly random and chaotic misspellings were a reflection of the blogger's status of mind or emotional status (of feeling unhappy) when he was composing this blog entry. Again, all these misspellings have not changed the grammatical or semantic nature of the original words but they have contributed to the informality of the text and they are used to achieve certain pragmatic functions.

5.3 Asterisks matter

Apart from playing with the orthographic representations of words, bloggers tend to make creative use of other linguistic symbols as well. Online discourses are characteristic of innovative use of punctuation marks. For instance, the use of questions marks and exclamation marks is a recurrent topic in existing literature about online chat discourse. What I am going to present here is not about any punctuation marks but about the use of a special symbol – the asterisk (*).

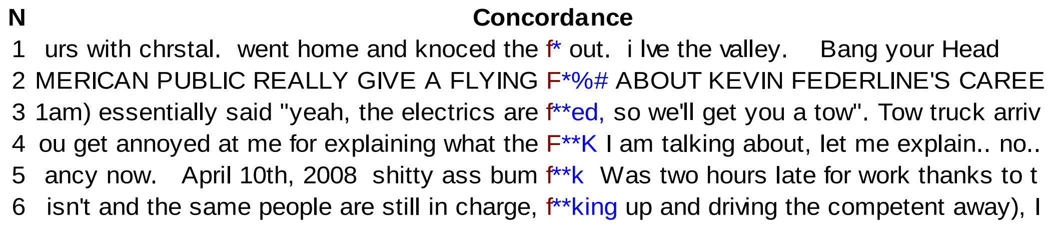
There are four basic uses of the asterisk (*) in personal blogs: 1) as an emphasis marker, 2) as a euphemism marker, 3) as a marker for comments and whole chunks of text, and 4) as a separator marking off blogger behaviors from the main text (or action marker). The first two belong to lexical features whereas the last two belong to discoursal features. In terms of occurrence frequencies, asterisks as emphasis markers and action (or behavior) markers are much more common than the other two. Below is an account of each of them.

The first use of the asterisk, the one as emphasis marker, can be taken as an innovative way of expressing prosodic features in written form. The function of the asterisks is to emphasize the words or expressions enclosed. By putting the asterisks on a word form, the blogger is actually making it more prominent orthographically and thus achieving the effect of emphasizing it. It is very similar to other strategies like spelling the whole word in upper case letters or lengthening a word by letter repetition. There are 96 occurrences of this use. Figure 5.1 shows the concordance lines of one third of the total occurrences. Among the 96 occurrences, 63.5% (61 cases) are from bloggers aged from 25 to 40; only 36.5% (or 35 cases) are from younger bloggers. In fact, only eleven occurrences are from the mid- and late-teens groups (three from the former and eight from the latter), accounting for less than 15% of the total occurrences. It seems that teenage bloggers do not find this way of saying things innovative enough. Gender-wise, female bloggers contribute around 60% of the occurrences and male bloggers around 40%. Region-wise, American bloggers contribute more than their British counter parts, with the occurrences from the former accounting for 54% and those from the latter 46%.

N	Concordance
1	d 60 pounds since the beginning of the year. *35* of those pounds have been packed on in
2	head stopped shouting "THE FUCKER HAS **ABANDONED** YOU" I did what any desper
3	g emotion, you not only have to deal with the *actual* disappointment, but the fact that you
4	*anyone* who isn't family or, recently, work in *ages*. Tuesday, May 13th, 2008 Keep
5	d up in Games Design and Development with *any* company during those two years then I
6	s go completely unchallenged and those that *are* challenged remain regardless of weathe
7	us feel safe in situations where we will never *be* safe. And then once we feel safe, life c
8	ost adorable squee-filled ship LIEK EVAR. :D *can't wait* The Commander is still huggable
9	ome true. I got all dressed up too, I looked so ~*classy*~. And I only lost 銃?0. ;) A fit you
10	econded the opinion. The doc also said that *currently* he's not concerned about her bon
11	ails, even MSN etc. Hence, what little time I *do* have is being spent on the most importa
12	p me from killing everyone, at least i won't be *entirely* alone when i cark it. I shall have 20
13	a bliss from Sonic. And no, I don't drink them *every* day, but I do drink them a lot as well
14	ople I want to marry or have a life with. I don't *fit*. And that's another thing. I don't feel lik
15	running over the weekend because they take *FOREVER*). Had to cut through some pre-
16	g with her now, so it's not me worrying about *her*. I just don't know how I should be towar
17	oneself rather than on the cheater... Ie: Can *I* trust that this person will not ever cheat o
18	to be mounted on the wall. Sometimes rather *intelligent* students manage to get their deg
19	I can take it the second week of July. So we *might* be going down there then. I still don't
20	oblem with my 'real life' fics because they are *mine* and I don't give a flying fuck what peo
21	we hung out with Hannah! Anyway, I'll give a *much* more detailed description of my week
22	r's, and having already spent that and more, I *need* that lump sum to pay off the money I
23	s of the Caribbean: At World's End. She had *no* idea she was coming early to watch the
24	ta drive to force it to only un at sata standard, *or* get a sata II raid pci card. so im in the m
25	e on MONDAY for JB. ugh. i have to find the *perfect* outfit i guess. so...major diet time.
26	bout it, then he said the one thing that is the *real* source of the problem at hand. "It's lik
27	really cool and definetely makes my costume *so* much more complete! I'm so glad FIRES
28	r since 1. we're a college and 2. we're even a *teaching* school. But I guess it's good they
29	a la "Fatal Attraction." As if. So who needed *that* bullshit?? Not I. Off I ran, only this time
30	do that from time to time, again, as a special *treat*. I need to get back into the habit of
31	th. What else...well me and Kel are officially *very* over, if I didn't mention that already. It
32	u're about to be arrested. Still, the coffee is *way* better here in the USA :) September 2

Figure 5.1 Asterisks as emphasis marker

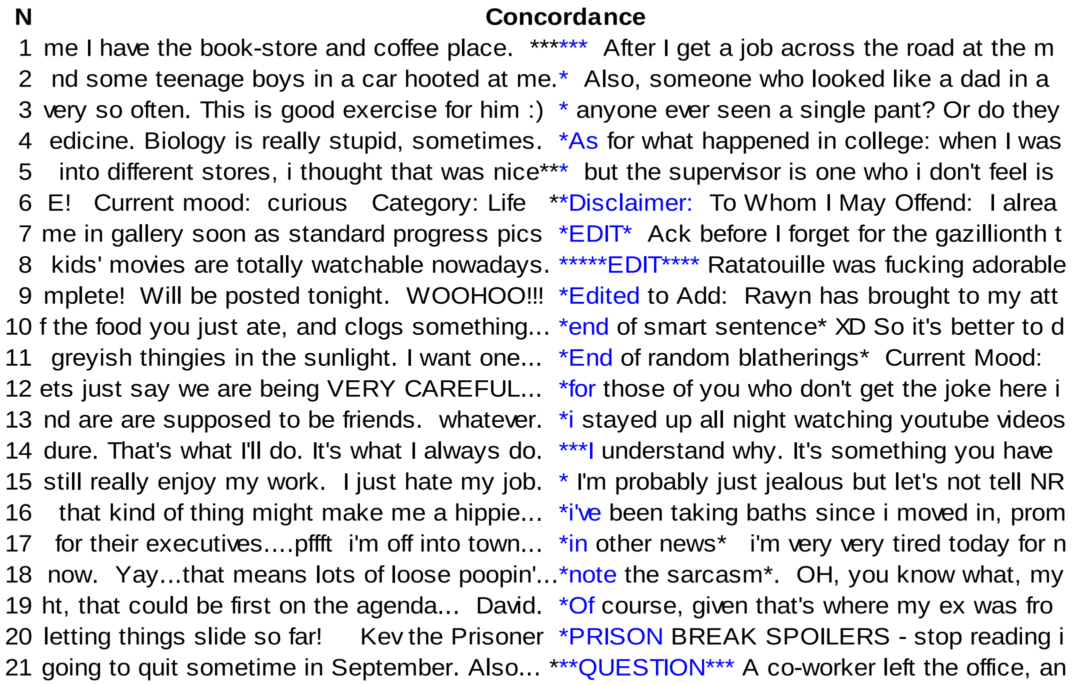
The use of the asterisk as euphemism marker is not very common in the corpus. Only six cases are found in the whole corpus and they exclusively appear in different forms of the word *fuck*, as the following concordance lines (Figure 5.2) demonstrate. The function of this use is to reduce the impact of using vulgar terms. Five out of the six occurrences are from bloggers (both male and female ones) aged from 30 to 40.



```
N                                   Concordance
1  urs with chrstal.  went home and knoced the f* out.  i lve the valley.    Bang your Head
2 MERICAN PUBLIC REALLY GIVE A FLYING F*%# ABOUT KEVIN FEDERLINE'S CAREE
3 1am) essentially said "yeah, the electrics are f**ed, so we'll get you a tow". Tow truck arriv
4  ou get annoyed at me for explaining what the F**K I am talking about, let me explain.. no..
5  ancy now.   April 10th, 2008  shitty ass bum f**k  Was two hours late for work thanks to t
6  isn't and the same people are still in charge, f**king up and driving the competent away), I
```

Figure 5.2 Asterisks used as wildcards

The third use of the asterisk, the one as a device for bloggers to mark off important chunks of text or add comments to their own statements, is more frequently observed in the corpus than that of the second use. This use falls into the category of discoursal features, as it is mainly for highlighting or reminding purposes. Figure 5.3 shows half of the total occurrences of this use.



```
N                                        Concordance
 1  me I have the book-store and coffee place.  ******  After I get a job across the road at the m
 2   nd some teenage boys in a car hooted at me.*  Also, someone who looked like a dad in a
 3  very so often. This is good exercise for him :)  * anyone ever seen a single pant? Or do they
 4   edicine. Biology is really stupid, sometimes.  *As for what happened in college: when I was
 5    into different stores, i thought that was nice***  but the supervisor is one who i don't feel is
 6  E!   Current mood:  curious   Category: Life   **Disclaimer:  To Whom I May Offend:  I alrea
 7  me in gallery soon as standard progress pics  *EDIT*  Ack before I forget for the gazillionth t
 8   kids' movies are totally watchable nowadays. *****EDIT**** Ratatouille was fucking adorable
 9  mplete!  Will be posted tonight.  WOOHOO!!!  *Edited to Add:  Ravyn has brought to my att
10 f the food you just ate, and clogs something... *end of smart sentence* XD So it's better to d
11   greyish thingies in the sunlight. I want one...  *End of random blatherings*  Current Mood:
12 ets just say we are being VERY CAREFUL...  *for those of you who don't get the joke here i
13  nd are are supposed to be friends.  whatever.  *i stayed up all night watching youtube videos
14  dure. That's what I'll do. It's what I always do.  ***I understand why. It's something you have
15  still really enjoy my work.  I just hate my job.  * I'm probably just jealous but let's not tell NR
16    that kind of thing might make me a hippie...  *i've been taking baths since i moved in, prom
17   for their executives....pffft  i'm off into town...  *in other news*   i'm very very tired today for n
18  now.  Yay...that means lots of loose poopin'...*note the sarcasm*.  OH, you know what, my
19 ht, that could be first on the agenda...  David.  *Of course, given that's where my ex was fro
20  letting things slide so far!     Kev the Prisoner  *PRISON BREAK SPOILERS - stop reading i
21 going to quit sometime in September. Also... ***QUESTION*** A co-worker left the office, an
```

Figure 5.3 Asterisks as comment markers

The actual use may vary from blogger to blogger. Some bloggers use the asterisks to mark the beginnings of new paragraphs (see lines 1, 3, 4, and 13). Some use them to highlight certain parts of the text which they hold to be of importance and to which they expect the readers to pay greater attention (see lines 5-9, 12, 17, and 21). Some bloggers use the asterisks to raise the readers' awareness of the comments the blogger is making (see lines 10, 11, and 18). They are actually giving guidance as to how certain parts of the text should be read or interpreted. To a certain extent, this use is similar to that of the first one in that both are intended for attracting more attention from the readers.

The fourth use of the asterisk, which is also the most interesting one among the four, is to mark off blogger actions (or behaviors) from the main text of the entry. There are 266 cases of such use in the whole corpus, a sample of which is shown in the following concordance lines (see Figure 5.4).

```
 N                              Concordance
 1  en damned and their mortal forms sacrificed. *ahem*  bought and played through the penn
 2   sure they have an excellent service from me *beams* plus they actually appreciate you,
 3  12th-Dec-2007  Mock exams... AHHHRGH!   *bites fingernails*  I had my first Mock GCS
 4  ages - runners up matches and all that. Heh.  *blows whistle*  Life is good, and hopefully w
 5   entric because American is not an ethnicity. *boggles*   There are so many comebacks t
 6   ate/Laura Cadman fics have been nommed.  *bounces*    Iz ded    2 Oct 2007    mood: exh
 7  ooked after removing his wig and facial hair..  *cries*    Well recommended :)   and thank y
 8 ' to me.  - Oh, Rhys. You are so so so lovely. *cuddles him*  - THAT FIGHT WAS SO AW
 9    Feb. 16th, 2008 | 06:16 pm   mood:  artistic  *dances*  YAY!!!  We have a week-long brea
10    08 October 2007    Jiggity Jig!      I'm home!   *drops bags on the bed to unpack later*   I n
11   hours. Bed now. *collapses*   3rd-Apr-2008  *dusts off LJ*  So, time for a change. I'm in t
12  ambition has been to be Just Like Hermione. *facepalm* I really hope this doesn't mean s
13   en's problems these days?  26 March 2006  *falls to the floor dead* I'm fucking shite at wr
14     gonna love Christmas when it finally arrives *flops*  Still I got another 3 days of overtime
15 orning? I think we all know the answer to that *goes and watches tv in bed*.    Thursday, D
16  be translating into masses of money just yet *grumble*  A client today was enthusiastic a
17    I'm freaking out right now unnecessarily??!!  *jumps up and down* Well, do ya? Do ya?..
18  his first words you "You've got a cold, then?"  *laughs with embarrassment at the memory*
19  ay. but i've already said that multiple times..  *looks down and thinks* i don't feel like wishi
20   ING YOUTH. damnit... take advantage of it.. *mumbles off*... i feel horrible..... i don't kno
21 , I didn't get straight to sleep when people left *nudge, wink etc* and we had to be up early
22  e Impala = bouncy happy fangirl.  Oh, Dean.  *pats fondly*  ETA (again):  Who couldn't lov
23 e especially.  No school for a week and a bit.  *Punches air*. I'm actually shattered. I alms
24  worst pain i've ever felt. :'(  goodnight all...      *rolls eyes*   come on sam... like your a lost
25  verse for those who'd like such things ...  :F!  *sets the bowl o' of candy by the front door*
26 I like the agree to disagree thing. I don't know *shrug*   I'm going to go maybe do somethin
27   ve to watch it instead ;) what a shame! Still...*sigh* i suddenly feel very depressed :(. Im g
28  OR spoiler. Wow. That sounds interesting...  *slaps self*  Must. Do. Biology. Revision.   *
29  id go I ended up swallowing my lunch twice.  *sticks tongue out in gagging gesture*  We'r
30   qualify and we don't we will never live it down *sulks*  Oh well...I went to my lecture and s
31  t pie...u no what im talking bout megan..lol...**vomitz**....ugh nasty!!!!! yea so we got outta
32 . 15th, 2007  October Minimeet  mood:  tired  *Yawns and blinks* God im tired. Getting up
33  sniffs*.  But it'll be back in a couple of weeks *YAY*.  I'm finally uploading my 67th London
```

Figure 5.4 Concordance lines for asterisks used as action marker

A number of observations can be made from Figure 5.4. First, a great majority of the words and phrases enclosed by asterisks are verbs. Other categories include noun phrases and interjections but they are not many in number. Second, most of the verbs or verb phrases are in the third person singular form, which is quite unusual. Since the logical subject of each of these verbs is exclusively *I* (the blogger), the verb should be in its first person present tense form as in lines 16, 21, 26 and 27. By enclosing the actions with asterisks, the blogger is actually signaling to the reader: "It's show time!" To a great extent, the asterisks are signals which mark the beginning and ending of the time when the blogger exits from the narration and does something here and now. This is quite similar to play scripts where actors' actions are marked off from their lines by bracketed instructions which are featured by the use of simple present tense. The third person singular form is almost the default as the play scripts are often arranged according the chronological order of the actors' utterances, one line after another. By adopting a discourse structure which is characteristic of play scripts, the blogger is actually turning his or her autobiographical account into a sort of narrating plus performing. Consequently, the reading of blogs has also been turned into a sort of watching. By inserting actions, the blogger is also making the entries more appealing to the readers. These actions and interjections help visualize the blogger in the reader's mind and achieve the effect of chatting via webcam. The origin of this use of asterisks may have something to do with a common practice in online chat where the system marks off the chatter's intended actions with an asterisk on the initial position (to compensate for the lack of paralinguistic features due to affordances of the early text-based chatting tools). Third, a great majority of the asterisked words or phrases are related to bloggers' emotional statuses while blogging. Table 5.20 lists the top 30 asterisked words functioning as actions in the corpus. From this list we can see that many of these words are verbs or expressions related to body language or paralinguistic features. Through these words and expressions the bloggers are actually trying to create a sense of presence, thus shortening the social distance between the blogger and the intended readers.

Table 5.20 Top 30 asterisked words and phrases

Item	FRQ	Item	FRQ	Item	FRQ
sigh(s)	52	ahem	4	shudder	2
bounce(s)	7	fingers crossed	4	smile(s)	2
cough(s)	7	crosses fingers	3	sniffle(s)	2
shrug(s)	6	hugs	3	sniffs	2
yawn(s)	6	rolls eyes	3	ugh	2
lol	5	yay	3	beams	2
cries	5	blink	3	breathes	2
facepalm	4	growl	2	cry	2
snigger(s)	4	grumble	2	dies	2
squee	4	rejoices	2	grin(s)	2

5.4 Chapter summary

From what has been presented in this chapter we can see that bloggers have employed a variety of strategies to realize orthographic variations. Some of the commonly used strategies include: apostrophe omission, abbreviation, letter repetition, orthographic representation of paralinguistic features, spelling words according to how they are pronounced, and even misspellings. In addition, special symbols such as the asterisk are also used to perform new functions and thus giving the blogging texts a different outlook. Almost all of these strategies are intended for achieving certain stylistic and pragmatic functions. All these strategies have contributed to the informality of the blogging discourse. By adding informal and oral discourse features to a written genre, the bloggers have actually turned blogging into talking and static silent letters into dynamic, audible sounds accompanied with paralinguistic features. By deviating from the established norm of conventional writing, bloggers have created a new writing style which is undoubtedly more suitable for the purpose of communicating with people via information sharing. Of course, these are not the only functions of orthographic variation. In fact, orthographic variation is also a good place for observing bloggers' identity representation, which is going to be discussed in Chapter 9.

Chapter 6 Lexicological Variation

This chapter first presents a discussion about two commonly used word-formation strategies bloggers employed to create new words: compounding and derivation. Then, it describes some of the minor word-formation strategies. After that, a description about neologisms related to IT and emergent Internet culture is presented. Following that, a detailed discussion about the use of slanguage is presented.

6.1 Introduction

The previous chapter has offered a detailed description about the various strategies that bloggers employ to realize orthographic variation, but orthographic variation can only reveal what bloggers are actually doing with the outlook of the word-forms. Variation at this level has not yet touched upon other aspects (such as the grammatical and semantic aspects) of the linguistic system. In fact, bloggers' innovative manipulation of the linguistic system does not stop here. If we shift our focus from the surface level of orthographic representation (or variation) into a deeper level of lexicology, we may be able to obtain more insightful observations about how bloggers are linguistically representing themselves.

If we say that orthographic representation is an established norm and a part of the imposed collective identity of the language users of a particular speech community, word-formation seems to be even more deeply rooted in the social, cultural, and linguistic development history of that speech community. If we say conventional orthographic representation is arbitrary, then word-formation is more rule-governed, although it is closely related to the orthographic system. There are a number of ways of forming new words in the English language, among which the most common ones include: compounding, derivation, abbreviation, coinage, conversion, reanalysis, and backformation. As the current research focuses on how bloggers are representing their identities in linguistic ways, a detailed account of all the lexicological strategies that personal bloggers have adopted is not intended. I will only discuss compounding, derivation, and some special ways of word-formation. Apart from word-formation strategies, I will also discuss bloggers' use of neologism and their use of slang words.

6.2 Compounding

According to Plag (2003), compounding is the most productive type of word-formation process in English, yet it is perhaps also the most controversial one in terms of its linguistic analysis (p. 132). To a great extent, compounding is also the most convenient way of forming new words because what it involves is mainly the combination of two or more words. Of course, this is not to say that there are no restrictions about this kind of combination. Nevertheless, the restrictions are looser than other word-formation processes, if not the loosest. As a matter of fact, the term "word-formation" is somewhat tricky, as different scholars have different understanding about what a "word" actually refers to. *Word* can be used in the sense of *word-form*; it can also be used in the sense of *lexeme*. These two terms refer to quite different things. A lexeme subsumes the different inflected forms of a word base. For instance, the lexeme *BE* subsumes all its inflected forms: *am, is, are, was, were, being,* and *been* (that is, seven words). This may be an extreme example. A more typical example would be the lexeme *UNDERSTAND* which covers word-forms like *understands*, *understood*, and *understanding*. There is another term “lexical item,” which refers to anything that can be listed in a speaker’s mental dictionary (Bauer, 2006). It is also called *listeme*. A lexical item (or listeme) can be as simple as a suffix (e.g., *-ly*) or as complicated as an idiomatic phrase (e.g., *bark up the wrong tree*). According to Bauer (2006), “word-formation is about the formation of lexemes rather than about the formation of word-forms” (p. 484). This definition of word-formation applies very well to processes other than compounding. When it comes to compounding, however, the definition becomes somewhat problematic. Compounding is often defined as “the combination of two words or word-forms to form a new word” (Plag, 2003, p.133). Bauer (2006, p. 485) contends that compounds are words which are made up of two lexemes. She emphasizes three criteria for identifying compounds. First, a compound must contain bases of two independent lexemes. Second, it should have the ability and requirement to inflect just like other lexemes which do not have a complex internal structure. Third, it should not be resulted from the lexicalization of syntactic structure. By these criteria, forms like *forget-me-not*, *love-in-a-mist*, and *shilly-shally* are all not compounds but lexical items. Probably influenced by the generative tradition of morphology which tends to focus on explaining the rules governing the so-called pure word-formations, this definition of compounding is so narrow that it actually excludes the possibility of accounting for words which consist of three or more elements. In fact, if we look at the naturally-occurring data from ordinary language users, we may find it necessary to expand the definition of compounding so that we can cover a wider range of word-forms created out of the process of combination and find out how language users are actually using this strategy to achieve their communication purposes. As Bauer and Renouf (2001) point out after examining the patterns of new compound formations in a large corpus of British newspaper English, considering real data can cause problems for the theoretician of word-formation and for the descriptive grammarian alike. They find that some patterns used productively in the English of the early 1990s break principles that are laid down as absolute in some of the theoretical works (p. 101). As a result, we need a broader definition to cover lexical items which are compound-like but do not fully meet the requirements specified by the traditional definition.

Current literature has already started to pay attention to new compounds which are beyond the scope of explanation of the narrow definition. Plag (2003), for instance, defines a compound as "a word that consists of two elements, the first of which is either a root, a word or a phrase, the second of which is either a root or a word" (p. 135). This definition makes it possible to talk about compounds which consist of three or more elements but it is only applicable to right-headed long compounds. In fact, Bauer (2006) mentions forms like *a don't-mess-with-me look* and *give-me-the-money-or-I'll-blow-your-brains-out scenarios* when she is discussing one of the important features of compound nouns, that is, allowing whole phrase/clause/sentence in the pre-modifying position (pp. 489, 493). However, she does not elaborate on this kind of nominal compounds due to lack of examples. What can be inferred from her description about the two long compound nouns cited above is that the components prior to the head nouns are a clause and a sentence respectively (or maybe two sentences). By referring to multi-word sequences of this sort as clausal or sentential pre-modifiers does not contribute much to answering the question why a clause or even sentence can appear in the pre-modifying position of a noun phrase (or rather compound noun). This is against the basic principle of the English language which tends to put clausal or sentential modifiers at the post-modifying positions. If we insist on calling them phrases, clauses, or even sentences, we may need to explain why this is syntactically possible and whether this phenomenon signifies a new direction of syntactic change for the English language. A simpler or arguably more reasonable way of looking at this phenomenon is to expand the definition of compounding and take these forms simply as cases of compound (or compound-like) words, side by side with the more conventional (or orthodox) categories of compounds. The advantage of doing so is that it can avoid the whole trouble of having to re-examine the syntactic rules governing the English language while at the same time makes it possible to talk about this often-overlooked phenomenon. To follow the terminology in existing literature (e.g., Meibauer, 2007; Wiese, 1996), I will refer to them as phrasal compounds, though this terminology is not an ideal one. A separate section will be devoted to the phrasal compounds occurred in the blog corpus and their roles in helping bloggers to achieve their communicative purposes. Prior to that, an account of conventional compounds and their common internal structures will be presented first.

The highly productive nature of compounding as a word-formation process implies enormous number of possible combinations, which in turn makes it very difficult to automatically (and accurately) retrieve the compound words (even in the strictest sense of the term) from a corpus of even a moderate size. Manual classification could be a more accurate option but it is labor-intensive and extremely time-consuming, thus it may not be very practical to use it as a major means for identifying the distribution patterns of compounds and their internal structures in a corpus. Nevertheless, two defining features of compounds could be exploited in their identification, with the help of the Wmatrix system. One is that the high unpredictability resulted from the high productivity of compounding as a process will produce nonce lexical items which will normally not be included in dictionaries or lexicon of natural language processing tools. The other has something to do with the English spelling principles. According to Bauer (2006), “there is a principle of English spelling whereby any item consisting of more than one orthographic word is hyphenated when it occurs in an attributive position” (p. 485). Although reasonable doubt about the regulative power of this principle exists, we cannot rule out the possibility that this principle is being generally observed. Even if it is not well observed, there are basically two ways of violating it: spelling all the constituent words separately and forcing the readers to do the guess job or infusing the constituent words into an orthographic whole without hyphenation. If it is the former, the identification will be problematic to both human analysts and language processing software tools. If it is the latter, the resultant forms can be easily captured by language processing tools as new (or unknown) lexical items. Taking both features into consideration, I believe that it is possible to obtain many compound words, especially those nonce formations and less established ones, from the unknown word lists generated by the Wmatrix system. As pointed out earlier, the current research is not intended to present an exhaustive description about word-formation strategies employed by personal bloggers but rather to explore how certain word-formation strategies are being exploited for identity representation purposes. Therefore, what is going to be presented below is only based on the unknown words identified by the Wmatrix system. These compounds suffice to demonstrate bloggers’ observation of and deviation from the established word-formation rules despite that the actual number of compounds is definitely bigger than what is presented here.

Among the 16,587 tokens of unknown word-forms, 1,135 have been manually categorized as compound (and compound-like) words. Echoing the major findings in related literature, the compounds created by personal bloggers also fall into three major categories: nominal compounds, adjectival compounds, and verbal compounds. The nominal compounds are most frequently used, taking up 63% of the total, followed by adjectival compounds, 23%, and verbal compounds only 4%. There are also a few cases of adverbial compounds, occupying 1%. Table 6.1 summarizes the overall distributions.

Table 6.1 Types of compounds identified

Category	Tokens	Percentage
Nominal Compounds	717	63%
Adjectival Compounds	265	23%
Phrasal Compounds	100	9%
Verbal Compounds	42	4%
Adverbial Compounds	11	1%
Total	1,135	100%

A closer examination of each category of the identified compound words reveals some patterns of internal structures. Many of these structural patterns are similar to those described in works about English word-formation (e.g., Adams, 2001). Below is an account of the internal structures for each category.

6.2.1 Nominal compounds

As far as the nominal compounds are concerned, 33 different internal structures are identified, among which the pattern Noun + Noun ranks the first, followed by the pattern Adjective + Noun. Both structures are proved to be the most typical ones for forming nominal compounds. Table 6.2 lists the top ten internal structures of the nominal compounds.

Table 6.2 Top 10 internal structures of nominal compounds

Internal Structure	Examples	No. of Tokens
Noun + Noun	workcrush, frog-man, metalhead, meatspace	385
Adjective + Noun	livejournal, nastygram, popart,	125
Determiner + Noun	MySpace	75
Verb + Particle	hangout, lie-in, shout-outs, meetup	42
Noun + Verb + ING	breast-feeding, wine-tasting, screen-writing, mapquesting	30
Pronoun + Noun	YouTube	28
Verb + Noun	touchscreen, kickball, blowjob, jumpsuits	24
Noun + Verb	head-start, powercut, fingersave, shot-put	14
Particle + Noun	up-side, in-breath	8
Noun + Verb + ER	train-goer, money-saver	4

Apart from the two most commonly observed patterns, two other patterns deserve more comments here. They are: Verb + Particle and Determiner + Noun. The Verb + Particle pattern is interesting for two reasons. First, nominal compounds of this kind have actually gone through two word-formation processes: compounding and conversion. The phrasal verbs (or verbal phrases) from which these nominal compounds are formed are actually results of combining verbs and adverbial particles or prepositions. They become nominal compounds through the conversion of part of speech from verbs into nouns. Second, nominal compounds formed this way are usually informal and thus have stylistic implications. The Determiner + Noun pattern is seldom used in new word formations in daily language use. One of the reasons might be that determiners are a closed grammatical category and the members are fixed and very small. Even if such words exist, they will not be counted as orthodox compounds according to the criteria identified by scholars such as Bauer (2006; 2001). Most probably they will be described as syntactic compounds because the pattern itself shows an apparent syntactic relation and the lexical items resulted will be naturally regarded as the lexicalization of syntactic structure. Unconventional as it may sound, it still makes sense, at least syntactically.

6.2.2 Adjectival compounds

Compared with nominal compounds, adjectival compounds are much smaller in number. Nevertheless, they have displayed a similar level of variety in internal structures. The 262 tokens of adjectival compounds fall into 35 different patterns. Table 6.3 lists the top 12 patterns. Half of the patterns listed in this table are listed as common patterns in Adams (2001). They are: Noun + Verb + ED, Noun + Adjective, Adjective + Noun + ED, Noun + Verb+ ING, Particle + Noun, and Noun + Noun + ED. The other six patterns are not listed as common ones, as they may not fully satisfy the orthodox definition of compounds. The pattern Adverb + Adjective, for instance, may well be categorized as lexicalized phrases or syntactic compounds, for example, the items '*nearly-new*', '*too-serious'*, and '*politically-correct*' in the following sentences:

(1) My grandparents are strange; they're very well-to-do and are perfectly comfortable giving away a *nearly-new* car, but they're the sort of people who wrap up random stuff they find lying around the house to give away as Christmas presents (us_f_18-19.txt).

(2) She's young, she's cute, she's got all the fun traits of Maddie and none of the *too-serious* parts (us_f_25-29.txt).

(3) And on that day I was feeling GOOD and PROUD and *POLITICALLY-CORRECT* like any decent, self-righteous vegan would, and breezed right on through (us_f_35-40.txt).

All these items have specific meanings and they are used the same way as other simple words. If we really want to impose an explanation on the syntactic relationship between the items and the modified nouns, we can say that they are results of the lexicalization of syntactic structures. But that does not clash with labeling them as adjectival compounds. The interesting thing is why English does not allow full-clause or sentence to appear at the pre-modifying position. This is a topic which will be picked up in the next section. The pattern Adverb + Verb +ED is of similar nature; therefore no further comments are needed.

Table 6.3 Top 12 internal structures of adjectival compounds

Internal Structure	Examples	No. of Tokens
Noun + Verb + ED	copy-protected, stress-induced, crack-addicted, job-related	44
Noun + Adjective	caffeine-high, baby-proof, blog-worthy, Gwen-heavy	36
Adjective + Noun + ED	light-headed, curly-haired, feathery-leafed, fuzzy-eyed	20
Adverb + Adjective	lesser-smelling, nearly-new, too-serious, politically-correct	18
Noun + Verb + ING	gut-wrenching, gas-guzzling, eye-opening, miracle-performing	18
Noun + Noun	million-dollar, one-page, wedding-type, baby-type	18
Particle + Noun	in-game, on-call, after-lunch, between-act	17
Adverb + Verb + ED	judiciously-chosen, hard-boiled, practically-pissed, well-maintained	16
Adjective + Noun	low-budget, longterm, realtime, low-pay	14
Noun + Noun + ED	family-sized, baby-pitched, finger-looped, steam-powered	12
Adjective + Verb + ING	heavy-going, cool-looking, professional-sounding, sick-making	9
Adjective + Adjective	passive-aggressive, fecal-oral, giddy-like, luke-warm	9

For patterns like Noun + Noun and Adjective + Noun, it may sound rather controversial to call them adjectival compound patterns because the lexical items resulted are very much like nominal compounds. Nevertheless, if we take a closer look at the grammatical functions of such items, we will find that they are almost exclusively used in attributive positions. They may have more or less the same semantic meaning as their noun phrase counterparts but grammatically they are no longer the same. For instance, the noun phrase for the item '*million-dollar*' should be '*a million dollars*'. '*Wedding-type*' means 'something similar to wedding' not 'a type of wedding' as the noun phrase '*wedding type*' may mean, as is shown in the example below:

(4) The first of four *wedding-type* events of this year so far, Sean and I are going to be all weddinged-out by the end of the year I think (uk_m_30-34.txt).

Similarly, for items like '*low-budget*' and '*longterm*', they are no longer nominal compounds as they have become items describing the quality of the items they are modifying. They are different from the noun phrases '*low budget*' and '*long term*', though the semantic tie is still there. The difference between items like '*low-budget*' as an adjectival compound and '*low budget*' as a nominal compound can be observed from their difference in orthographic representations. When two lexical items are spelled as one orthographic word (with or without the hyphen) and placed in an attributive position, they will normally lose some of the features (for instance, plural inflection for countable nouns as the item '*million-dollar*' in '*a million-dollar deal*') when they are used as separate items.

The pattern Adjective + Verb + ING is actually a very common one in adjectival compound formation but the verbal component in this pattern is held to be restricted to a special type. According to Adams (2001, p. 92), "adjectives can be compounded only with present-participial adjectives corresponding to verbs of perception." Seven out of the nine adjectival compounds identified from the unknown word lists are exactly like what Adams claims. Six of them take '*-looking*' and one takes '*-sounding*' as their ending parts. Both '*look*' and '*sound*' are verbs of perception. Nevertheless, the other two items do not seem to follow this principle: one is '*heavy-going*' and the other is '*sick-making*'. Neither '*go*' nor '*make*' is a verb of perception. The following two examples show how these two items are actually used by bloggers:

(5) Other times I'll become bored, disillusioned or just plain confused by a book that is particularly *heavy-going* (uk_f_20-24.txt).

(6) I mean no offense by this - frankly I adore the little buggers - but they carry horrible, *sick-making* germs (uk_f_20-24.txt).

These two examples show that in natural language using situations language users do not always follow the so-called rules or principles spelled out by linguists. What really matters to them is whether their intended meaning can be effectively conveyed.

6.2.3 Verbal compounds

Verbal compounds are much less frequently observed than nominal and adjectival ones. As Adams (2001) points out, genuine verb compounding is typologically a rare phenomenon. The small number of tokens of verbal compounds seems to be echoing this remark. Altogether there are 42 tokens of verbal compounds and they fall into seven different internal structural patterns. Two of the more frequently used patterns are Noun + Verb and Noun + Noun, as Table 6.4 shows.

Table 6.4 Internal structures of verbal compounds

Internal Structure	Examples	No. of Tokens
Noun + Verb	bus-knit, self-mediate, packet-sniff, hug-rape	16
Noun + Noun	rearend, flowerbud, tailgate, , paintball	11
Verb + Verb	jabberjaw, playtest, kickstart, spell-check	5
Adverb + Verb	autocross, almost-fail	4
Adjective + Noun	super-glue, hot-wire	2
Particle + Verb	oversleep, outpee	2
Verb + Particle	sleepover	2
Total		42

For verbal compounds formed through the combination of Noun + Verb, the semantic relationship between the head (i.e. the verbal component) and the modifier does not follow a fixed pattern. There are cases where the noun component works as the object/complement of the verbal head, which is a very commonly observed relationship, for instance, ‘*packet-sniff*’. There are also rare cases like ‘*bus-knit*’ which actually means ‘*to knit on the bus*’ and ‘*hug-rape*’ which means ‘*to hug somebody against their will*’. For verbal compounds formed through the combination of two nouns or an adjective plus a noun, they will normally undergo a process called conversion or transposition before they are used as verbs. For example, ‘*to rearend*’ consists of two nouns ‘rear’ and ‘end’, when they are compounded and undergo conversion they will produce a new verb with inflected forms like ‘*rearended*’ and ‘*rearending*’. Another example is ‘*to super-glue*’. Probably originated from a brand name called ‘super glue’ and originally a compound noun (or a noun phrase), the new item ‘*to super-glue*’ is easily transposed or converted into a verbal compound.

From what has been presented so far, we can see that on the whole personal bloggers are identifying themselves with the mainstream word-formation strategies. Maybe we should take this as an indication of the constraining power of the linguistic system itself. Nevertheless, they do not always follow the so-called rules or principles: they will create new lexical items which suit their own communicative needs as has been demonstrated in some of the less common patterns of internal structures of nominal and adjectival compounds. What is going to be presented in the next section is somewhat different from the neat picture depicted so far.

6.3 Phrasal compounds

Within compounding, phrasal compounds are arguably the most problematic. Their very existence poses some challenge to orthodox morphological theories. As Meibauer (2007) points out, phrasal compounds violate the No Phrase Constraint and the Principle of Lexical Integrity and they display expressivity typical of marginal morphology. What puzzles theoreticians of word-formation is that the non-heads of phrasal compounds can be filled up by phrases (e.g., a collection of *never-to-be-opened* notes), clauses (e.g., various *keep-clothing-off-the-floor* devices), or even sentences (e.g., the *hastily-put-together-and-we're-totally-not-following-it* syllabus). Bauer (2006) mentions that compound nouns allow whole phrase/clause/sentence in the pre-modifying position but she does not explain why. As the English language tends to postpose long modifiers, the presence of extended and maximal projections in the non-heads of phrasal compounds requires an explanation. There are a number of insightful attempts in existing literature. Wiese (1996) proposes the quotation hypothesis, claiming that the non-heads of the phrasal compounds are all quotations. This hypothesis could explain all the irregularities displayed by the non-heads, but it seems to be too perfect to be true. In languages such as Mandarin Chinese, it is quite common to use clauses or even sentences to modify a noun (or noun phrase) and putting these clausal or sentential modifiers on the left-hand position is the only option possible. Therefore, it does not make sense to refer to them as merely quotations. As an alternative explanation, Ackema and Neeleman (2004) have proposed the Generalized Insertion approach, arguing that phrasal syntax can be inserted into word syntax or vice versa. This approach makes greater sense than the quotation hypothesis, but it cannot explain why speakers choose to use these phrasal compounds in the first place. To solve that problem, Meibauer (2007) proposes that a pragmatic module should be added to the General Insertion approach to account for the expressivity displayed by phrasal compounds. Insightful as these models or approaches are, they all focus too much on the formal aspect of phrasal compounds while neglecting the semantic aspect. In order to understand the presence of phrasal compounds in the EBC, we should take into consideration the formal aspect (which includes both the structural aspect and the orthographic aspect), the semantic aspect, and the pragmatic aspect.

Altogether 98 tokens of phrasal compounds have been identified from the EBC, almost all of which are nonce formations. They fall roughly into four categories according to the grammatical functions they are performing in the texts: pre-modifiers, nouns, inserts, and adverbs. Table 6.5 presents the details. As can be seen from Table 6.5, there are only four inserts and two adverbs. Three out of the inserts are formed through the combination of two or more online discourse elements. Linguistically, they are not of much interest. Stylistically, they can be used as indicator of informality. For instance, '*thankskbai*' is a combination of three words ('*thanks*', '*ok*', and '*bye*') and is often used as formulaic language in online discourse to say goodbye. '*Wootroflyeahhh*' is a term for expressing great excitement. It is actually a mixture of three quite different elements with no internal relations: the leetspeak word '*woot*' (expressing excitement), the initialism '*rofl*' (rolling over the floor laughing), and the variant of an insert word '*yeah*' (often used to expressed excitement). '*Omfgsgcaaiatpo*' is another item containing a very popular abbreviation in online discourse: *omfg* (oh my fucking god). As for the item '*yadda-yadda-yadda*', it is used in speaking as a filler word for unstated material or to indicate boredom or distaste for things others are saying or have said. Its function in a written discourse is to add a flavor of colloquialism to the blog entry while at the same time expressing the emotional status of feeling bored. The only two adverbs also seem to have been used as stylistic markers, with '*oh-so-much*' sounding colloquial and full of emotions and 'zero-to-sex' (meaning *doing something directly*) sounding blunt and metaphorical. Interesting as these six terms seem to be, they are not the mainstream ones. The other two types of phrasal compounds are of greater interest here: pre-modifiers (or adjectival phrasal compounds) and nominal phrasal compounds.

Table 6.5 Grammatical functions of quasi-compounds

Grammatical Function	Examples	No. of Tokens
pre-modifier	(a few) less-than-ideal (moments), (a) cop-arrests-hot-woman (romance), the almost-certainly-not-going-to-happen (box), (the) born-live-procreate-die (type of life), (in that) it's-so-stupid-you-can't-help-but-laugh (way)	55
Noun	stupid-frickin-chapter-fourteen, the 10MB-shared-between-one-thousand-plus-students, (the biannual) let's-clean-out-the-kids'-books-so-we-have-room-for-other-crap	37
Insert	thankskbai, wootroflyehhhh, yadda-yadda-yadda, OMFGSGCAAIATPO	4
Adverb	(from) zero-to-sex, oh-so-much	2
Total		98

Among the 55 adjectival phrasal compounds, 32 of them are of phrasal structures (meaning that the items are phrases of different sort), 20 are of clausal structures (that is, these items are actually lexicalization of clauses), and 3 are of sentential structures (i.e. they are lexicalization of full sentences). Quite a number of the items with phrasal structure are actually shortened version of relative clauses. Two questions are of interest here. First, why should a post-modifying clause become a hyphenated phrase when it is shifted into a pre-modifying position? Second, why do personal bloggers choose to use phrasal compounds instead of post-modifying clauses? The former is basically a linguistic issue whereas the latter is more of pragmatic nature. Generally speaking, if the pre-modifier is an adjectival phrase, it can be put directly before the modified, as a noun phrase has the structure of [determiner] + [adjective/noun] + [noun]. In other words, the slot between the determiner and the head noun can only be filled up by adjectival or nominal items if it really needs to. As English does not allow full sentences, clauses, or phrases which are not of adjectival or nominal nature to fill up that slot, they will have to undergo the process of nominalization for qualifying themselves to fill up that slot. There are a number of ways for nominalizing phrases and clauses. For instance, it can be done by putting a complementizer at the very beginning of a clause or changing the clause into an infinitive or gerundial phrase if it is going to be placed at the subject position. However, these strategies are not quite relevant to what is being discussed here. For the pre-head slot of a nominal group (or noun phrase) the strategies for nominalization will be much more restricted, as this slot is more suitable for adjectives. Infinitives and gerunds are possible but not good candidates either. A simpler way would be to hyphenate all the constituting elements together and create a word-like entry out of them. By stringing together the constituting components, the resultant phrasal compound has been made into a single item which is meant to be understood as a whole concept regardless of its length and structural complexity. Of course, the original complementizers or relative pronouns will have to be taken out first. This seems to be what Bauer (2006) implies when she says that compound nouns allow whole phrase/clause/sentence in the pre-modifying position (p. 493). The following examples offer a flavor of what I call adjectival phrasal compounds.

(7) It's like one of my *less-than-favorite* evangelists said on TV yesterday, it's not good to keep putting yourself in a line of unnecessary hurt, especially if the source of hurt isn't genuinely remorseful for it (us_f_20-24.txt).

(8) Worldcon looks a bit *touch-and-go* at the moment, so maybe I'll shoot for World Fantasy instead (uk_m_25-29.txt).

(9) It should, I must point out, be meant in the smutty sense and not in the *get-a-friggin-wash* sense (uk_f_30-34.txt).

(10) So my plan, which was originally going to be to get *stupid-frickin-chapter-fourteen*, which I am now on my *god-knows-how-many* draft of, finished over the holidays, has gone pretty much to pot (uk_f_15-17.txt).

(11) I've never read that kind of thing before in my life. An[d] this isn't even a *cop-arrests-hot-woman* romance novel (us_f_18-19.txt).

(12) Then, 11-year-old Semi-Charmed Life came on, and I embraced the *my-best-years-have-passed* dorkiness to sing along (uk_f_25-29.txt).

(13) However the top is a tad too tight which might result in a *boob-popping-out-incident*. Oh vell, I shall have to deal with that as it comes. The major problem is the *walking-around-with-flab-on-show* problem (uk_f_18-19.txt).

(14) After that we stopped by the mall and had some *oh-so-bad-for-my-diet* orange chicken and did a bit of shopping (us_m_30-34.txt).

(15) Drake and Josh - I actually like this show (in that *it's-so-stupid-you-can't-help-but-laugh* way), but not enough that I would pay for cable (us_f_30-34.txt).

(16) Just wrote two papers that are due tomorrow. There's a third that was supposedly due two or three weeks ago, but my teacher never said anything about it, it's just in the *hastily-put-together-and-we're-totally-not-following-it* syllabus (us_m_20-24.txt).

(17) Along the lines of people are so desensitised by the proliferation of media they suffer badly from the *you're-not-at-home-in-your-sitting-room-so-DON'T-TALK-ALL-THE-WAY-THROUGH-THE-GODDAMN-FILM* (and in particular DON'T answer your mobile) syndrome, you can't believe it really happens.

All the italicized parts but two (the '*stupid-frickin-chapter-fourteen*' in Example 10 and the '*boob-popping-out-incident*' in Example 13) in the above cited examples are functioning as pre-modifiers despite the discrepancies in internal structures. The two exceptions belong to another category - the nominal phrasal compounds, which will be discussed below.

Similar to (and related to) the adjectival phrasal compounds discussed earlier, another group of phrasal compounds also display certain unconventional features: the nominal phrasal compounds. Thirty-nine nominal phrasal compounds have been identified from the EBC. The biggest difference between adjectival ones and the nominal ones is that the latter are either functioning as heads of nominal groups or as objects or complements of verbs (or prepositions). Compared with the former, the latter enjoy lesser degree of syntactic restriction. Here are some examples to show how they might be different from the adjectival phrasal compounds.

(18) I will be happy if I never see *boys-pretending-to-have-vaginas* ever again (uk_f_18-19.txt).

(19) Last week, I found the cutest bookmarks. They're little magnetic clips and they've turned into my best *at-the-counter-impulse-buy* in months (us_f_35-40.txt).

(20) Which reminds me, contrary to what many people have commented about it, I gave into the cold Yorkshire air, and bought that *£200-duffel-coat-reduced-to-£30* from Primark (uk_m_18-19.txt).

(21) but I'm pretty sure the people up in Technical Support switch the ResNet portal servers off at the wall and sit back smugly laughing at how their internal 20MB connection for three Computer Labs is far greater than the *10MB-shared-between-one-thousand-plus-students* on campus whilst downloading HD movies at 2,000kB/s (uk_m_18-19.txt).

(22) Went out for a late lunch with *faux-crush-who-is-rapidly-approaching-full-crush-status* and Daniel (us_m_30-34.txt).

(23) While doing the biannual *let's-clean-out-the-kids'-books-so-we-have-room-for-other-crap*, I stumbled upon a few books the kids own that I hate (us_f_30-34.txt).

From the examples cited above, we can see that adjectival phrasal compounds are not the same as non-modifying ones. One major difference is that nominal phrasal compounds tend to have a "noun plus post-modifier" structure. In fact, one third of the nominal phrasal compounds are of such structure. This structure is less likely to appear in attributive position because it would be syntactically awkward and semantically confusing to have a noun with a post-modifier (especially a relative clause) working as the pre-modifier of another noun. Of course, there are other types of nominal phrasal compounds. One more type which is also of interest here is nominalization through hyphenation, as can be seen from the following examples:

(24) We met her in town to give her money to get them all home, and hung out for an hour or so, by which time I'd had my fix of *must-do-something* and was happy to go home and lounge again (uk_f_20-24.txt).

(25) I sometimes am a *know-it-all*, it is annoying I know, it is a bad habit to just pop in, and break the train of collective thought in a conversation with facts and such showing off (us_m_20-24.txt).

(26) She needs to be held accountable to the laws she has broken, and shown that all her daddy's money is not the *end-all-be-all* to life (us_m_35-40.txt).

(27) I didn't handle it well when my sister got upset about it - I got all *defend-the-minorities* (uk_f_30-34.txt).

It is beyond the scope of this research to explain why exactly these expressions are linguistically possible. What is more relevant here is why personal bloggers actually use these seemingly odd ways of expression. If we take a closer look at all the examples cited in this section, we will find that they are all very clear in meaning despite their unusual length. In many of the cases, the bloggers have employed a semantically direct yet lexically round-about way of saying things. These unusual ways of saying things are attention-catching, vivid, and easy to understand. Their unusualness is actually an indication of bloggers' intentionality of exploiting (or rather playing with) the linguistic structure. The presence of these unusual expressions may well be a cursor pointing to the uniqueness of personal blogs as a platform for language use. These multi-word expressions, especially the very long ones, might be less likely to be found in other online discourses such as online chat and instant messaging. Therefore, their presence may well be a feature of personal blogs as a genre. By deviating from the principles in orthodox word-formation, personal bloggers are not only displaying their innovative power as language users but also creating a new style of using the language to achieve their communicative purposes. Bloggers' innovative way of using the language has posed certain challenges on some of the linguistic theories and some practices in corpus linguistic studies, a topic which will be dealt with in greater detail later in Chapter 10.

6.4 New derivations

The above two sections have demonstrated the productivity of compounding as a means of creating new words and compound-like complex words, and how personal bloggers are exploiting this productivity to achieve their special communicative purposes. However, this is not the only strategy they have employed to achieve that purpose. The reason is simple: compounding is just one of the word-formation processes in English. There are other ways of forming new words or expressions, of which affixation or derivation is an important one. Affixation is a process of forming new words by attaching a bound morpheme to a base. Affixation can be classified into three types according to the positions occurred: prefixation (adding a bound morpheme at the initial position), suffixation (adding a bound morpheme at the final position), and infixation (inserting a morpheme in the middle of a word). Both affixation and suffixation are very common in the English language whereas infixation is rare. Some scholars even argue that infixation is not actually a word-formation process in English, as it does not result in new words. Whether to include infixation as a word-formation process in English is not an important issue here. What is really relevant is whether personal bloggers are using this linguistic strategy to achieve their communicative purposes.

Compared with compounding, affixational or derivational word-formation is more rule-governed. To a great extent, what we call affixational rules is actually a shared norm the current generation of a speech community inherited from the preceding generations. This shared norm is more established and restrictive than the one which is governing the formation of compounds. One piece of evidence would be the relative stable inventory of affixes in the English language over the past century. Having said that, it does not follow that there is no possibility for creating new affixes or changing the meaning of existing affixes. Established norms and internal restrictions are powerful but they are not always unbreakable. Language is a social product of human activities, after all. Language users are the ones who will ultimately push the process of language change. One way of capturing the traces of language change is to look at the new words or word-forms which language users are using (in a corpus). This is where Wmatrix can play an important role.

As mentioned elsewhere, Wmatrix is able to list all the word-forms which are not currently included in its lexicon as unknown words for its clients to download for further analysis. This unknown word list can be used as a starting point for tracing new lexical items. As the English language has a limited inventory of affixes, it is not too difficult to manually identify new words which have undergone affixational processes. Due to the constraint of space, I will only present some of the more commonly used affixes by personal bloggers here. Hopefully this can offer a snapshot of personal bloggers' creativity in language using.

Among the more commonly used affixes identified from the unknown word list and further testified by running the concordance function of WordSmith Tools on the original texts, the following are worthy of describing here: *-y*, *-ish*, *-ness,* and *semi-*.

6.4.1 Suffix *-y*

The suffix *-y* is a useful and commonly used bound morpheme for forming adjectives in online discourses. Most of the time, it is attached to a noun (or a verb) to form an adjective. For instance, the word *achey* is formed by attaching the suffix *–y* to the noun *ache*. The meaning of this suffix may vary slightly according to the specific word base it attaches to, but basically it is used to express the meaning of 'full of X' or 'having the quality of X'. Other examples include: *angsty* (full of angst or angry), *fumey* (full of fume), *geeky* (having the quality of a geek), *sweary* (swearing) and *sucky* (really sucks). Sometimes, it can also be attached to an adjective or a verb to form a new adjective or verb but this time its meaning changes into 'very X (the adjective)' or 'X (the verb) very much'. For instance, '*cheapy*' means '*very cheap'*, '*smarty*' means '*very smart*', and '*likey*' means '*like very much*'. It is used as a diminutive suffix as well to show intimacy, for instance, '*wifey*' is just a more intimate way of saying '*wife*'. The following examples show how words formed this way are used in actual blog entries:

(28) The trip involved two tubes, a bus, a bit of a walk, a brief stop into McDonalds for something *chickeny* (*made of chicken*), a phone call to Ade to tell him to fix it, and a final tube ride home (uk_f_20-24.txt).

(29) They had you waiting in this big, over-heated room, with rows and rows of fuscia, *modern-y (very modern)* chairs that look like a bulk buy from Ikea (uk_f_15-17.txt).

(30) And the lazy slob got *pissy (pissed off or angry)* with me for not "letting him do it" (us_f_35-40.txt).

(31) I'm really sad because on my *cheapy (very cheap)* ipod whatchamacallit thinger the sound keeps cutting out and it's not the headphones because I tried different headphones, but I do have to fiddle and twist the headphone cord to get it to work again (us_f_20-24.txt).

(32) I feel really, really, REALLY happy with life in general, as cliché and *n00by* (*someone new to college life*) as it is, everyone I live with are awesome (uk_m_18-19.txt).

Altogether, there are 282 word tokens (139 word types) which involve the attachment of the suffix –*y*, 193 (75 types) of which have already made it into the lexical repertoire of daily English, as they can be found in one or more authoritative English dictionaries (such as *Oxford English Dictionary* and *American Heritage Dictionary*). Nevertheless, a great majority of these words already recognized by dictionary makers are labeled spoken, informal, or slang. Thus the presence of such words in written discourses reveals the stylistic feature of being informal. Table 6.6 lists words with suffix –*y* with at least two occurrences. The rest 89 word tokens with the suffix –*y* (68 word types) are created out of similar strategies but they are nowhere to be found in authoritative dictionaries. That is to say, they are new words in the exact sense. Another feature of these words is that some of them have rather strange internal structures and their meanings are less transparent than those listed in Table 6.6.

Table 6.6 Words with suffix –y which have made into lexical repertoire

Lexical Item	Frequency	Lexical Item	Frequency	Lexical Item	Frequency
bouncy	13	crunchy	4	bossy	2
shitty	13	giddy	4	bubbly	2
geeky	11	icky	4	edgy	2
creepy	8	freaky	3	hefty	2
buddy (n)	7	girly	3	hubby (n)	2
crappy	7	hippy (n)	3	pinky	2
kitty (n)	7	pissy	3	potty	2
puppy (n)	7	pussy (n)	3	shaggy	2
cranky	6	snarky	3	soggy	2
grumpy	5	tipsy	3	spiffy	2
shady	5	bitchy	2	wacky	2
yummy	5	bobby (n)	2	wifey (n)	2

Table 6.7 lists all these words and their internal structures. Two things are of interest here. One is that all these words follow certain rules of word-formation; the other is that they are created intentionally to express certain emotions. In other words, bloggers are well aware of the basic rules of adjective formation while at the same time they are using these rules creatively to make their feelings and emotions expressed. They occasionally choose to deviate from the norm in tactical ways so as to add a playful, non-serious, and humorous tone to their blog entries.

Table 6.7 List of new words with the suffix -y

New Word	Internal Structure	Word Class
achey (3), angsty(5), arsey, boshty, chickeny, chocolate-y, Christmasy, coughy, crouchy, cuddley, flakey (2), flouncy, fumey, gazey, geography-y, glitchy, grindy, grudgy, grumbly, headachey(2), lawyery, lordy(2), dyke-y, mopey, moppy, mule-y, n00by, nature-y, nighty(3), old-timey, pervy, purdy(4), queeny, ranty(2), romancey, ropey, school-y, sciency, shrinky, slashy, smushy, stomachy, sunshiney, zippy	Noun + -y	Adjective
cry-y, explodey, hacky, screwy, skeevy, sucky(4), sweary, updatey	Verb+-y	Adjective
cheapy, drunky, modern-y, nakey, plasticy, ceramicy, smarty	Adjective + -y	Adjective
requesty, worky, addy(4)	Noun + -y	Noun (Diminutive)
whiffy, likey	Verb + -y	Verb (Diminutive)

Apart from attaching to adjectives and sometimes verbs to form new adjectives, the suffix *–y* is also used as a diminutive morpheme attached to nouns (especially personal names) to form new words which can express a sense of intimacy, casualness, and other emotions. Words like *buddy*, *puppy*, *hubby*, *wifey* in Table 6.6 and *worky, addy,* and *requesty* in Table 6.7 are all examples of *–y* used as a diminutive morpheme. In fact, the suffix *–y* has a variant *–ie* which is also very commonly used as a diminutive morpheme. This variant (-*ie*) is only a noun suffix. It is often used with names to express intimacy. Some common names carrying the suffix *–ie* include: *Abbie*, *Allie*, *Annie*, *Angie*, *Barbie*, *Bennie*, *Billie*, *Bennie*, *Carrie*, *Charlie*, *Davie*, *Debbie*, *Eddie*, *Ellie*, *Frankie*, *Georgie*, *Jackie*, *Jessie*, *Jodie*, *Julie*, *Katie*, *Lizzie*, *Maddie*, *Maggie, Stephanie*, *Susie*, and many others. There are 302 tokens of personal names ending with *–ie* in the blog corpus, covering 64 different names. As personal bloggers tend to talk about their daily experience which inevitably involves their family members and friends, it is quite natural for them to use addressing terms with a diminutive morpheme to express intimacy. Of course, the function of the suffix *–ie* as a diminutive morpheme is not restricted to personal (or pet) names. It is also attached to other words (mainly nouns) to express intimacy (e.g., *roomie*), casualness (e.g., *piccie*), or even to express contempt (e.g., *junkie*), depending on the base words it attaches to. There are another 108 occurrences of non-personal-name words which carry the diminutive suffix *–ie*, covering 64 word types. Table 6.8 gives a full list of them.

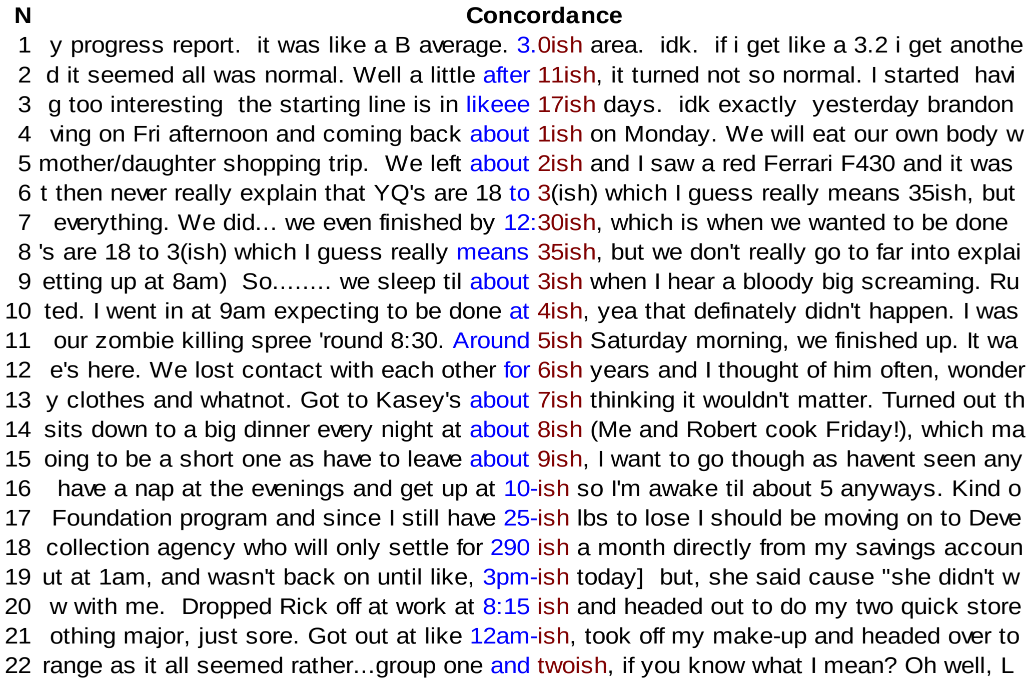
Table 6.8 Words with suffix -ie

Item	FRQ	Item	FRQ	Item	FRQ	Item	FRQ
hippie	8	kiddie	2	ciggie	1	wedgie	1
twinkie	7	newbie	2	commie	1	piccie	1
auntie	4	okie dokie	2	eerie	1	plushie	1
calorie	3	ouchie	2	fittie	1	rookie	1
footie	3	pressie	2	freshie	1	runnie	1
freebie	3	sarnie	2	goalie	1	sharpie	1
homie	3	thingie	2	goodie	1	specie	1
roomie	3	veggie	2	halloweennie	1	sweetie	1
smoothie	3	antoqie	1	hoagie	1	teenie	1
yuppie	3	archie	1	hottie	1	toastie	1
aussie	2	blankie	1	junkie	1	toothie	1
biggie	2	brekkie	1	kaputzkie	1	tortie	1
boogie	2	brownie	1	kookie	1	wheelie	1
doggie	2	budgie	1	lollie	1	woopie	1
girlie	2	cabbie	1	lookie	1	woopsie	1
hoodie	2	chickie	1	lottie	1	worrie	1
Subtotal	52		24		16		16

6.4.2 Suffix *-ish*

Another suffix which is also quite popular among personal bloggers is *–ish*. In fact, this suffix seems to be more established than *–y* (which is arguably an emerging suffix, so to speak). It is one of the common adjectival suffixes listed in Plag (2003). According to Plag, this suffix can attach to a variety of word classes, for instance, adjectives, numerals, adverbs, and even phrases to express the meaning of 'somewhat X, or vaguely X'. When attached to nouns referring to human beings, the new derivatives can be understood as 'of the character of X, like X' (2003, p. 96). Here are some examples from the EBC:

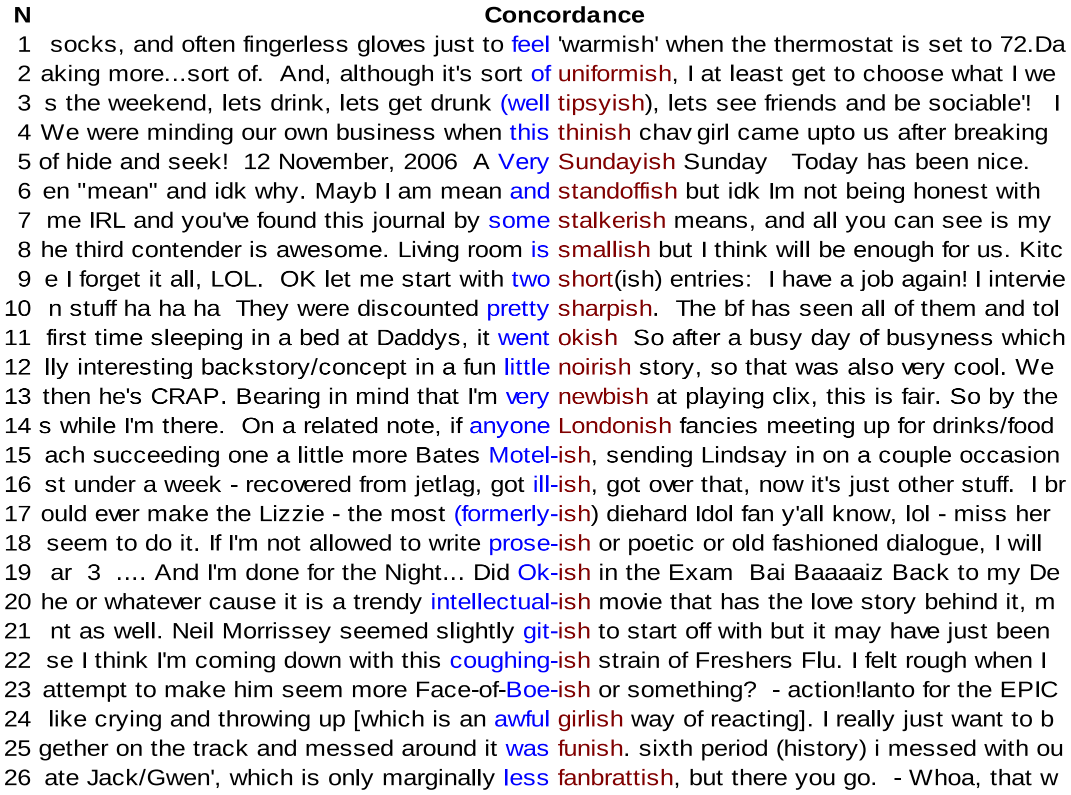
Adjectives: *illish, awake-ish, warmish, easy-ish*
Numerals: *twoish, 290ish*
Adverbs: *soonish, formerly-ish*
Nouns: *boyish, stalkerish, prose-ish, Londonish*
Phrases: *Bates Motel-ish, Face-of-Boe-ish*



N	Concordance
1	y progress report. it was like a B average. 3.0ish area. idk. if i get like a 3.2 i get anothe
2	d it seemed all was normal. Well a little after 11ish, it turned not so normal. I started havi
3	g too interesting the starting line is in likeee 17ish days. idk exactly yesterday brandon
4	ving on Fri afternoon and coming back about 1ish on Monday. We will eat our own body w
5	mother/daughter shopping trip. We left about 2ish and I saw a red Ferrari F430 and it was
6	t then never really explain that YQ's are 18 to 3(ish) which I guess really means 35ish, but
7	everything. We did... we even finished by 12:30ish, which is when we wanted to be done
8	's are 18 to 3(ish) which I guess really means 35ish, but we don't really go to far into explai
9	etting up at 8am) So........ we sleep til about 3ish when I hear a bloody big screaming. Ru
10	ted. I went in at 9am expecting to be done at 4ish, yea that definately didn't happen. I was
11	our zombie killing spree 'round 8:30. Around 5ish Saturday morning, we finished up. It wa
12	e's here. We lost contact with each other for 6ish years and I thought of him often, wonder
13	y clothes and whatnot. Got to Kasey's about 7ish thinking it wouldn't matter. Turned out th
14	sits down to a big dinner every night at about 8ish (Me and Robert cook Friday!), which ma
15	oing to be a short one as have to leave about 9ish, I want to go though as havent seen any
16	have a nap at the evenings and get up at 10-ish so I'm awake til about 5 anyways. Kind o
17	Foundation program and since I still have 25-ish lbs to lose I should be moving on to Deve
18	collection agency who will only settle for 290 ish a month directly from my savings accoun
19	ut at 1am, and wasn't back on until like, 3pm-ish today] but, she said cause "she didn't w
20	w with me. Dropped Rick off at work at 8:15 ish and headed out to do my two quick store
21	othing major, just sore. Got out at like 12am-ish, took off my make-up and headed over to
22	range as it all seemed rather...group one and twoish, if you know what I mean? Oh well, L

Figure 6.1 Concordance lines of -ish with numerals

The concordance lines of *-ish* show that this suffix go more often with numerals, especially numerals expressing time, as can be shown in Figure 6.1. When attached to non-numerals, it displays greater diversity, as Figure 6.2 demonstrates.



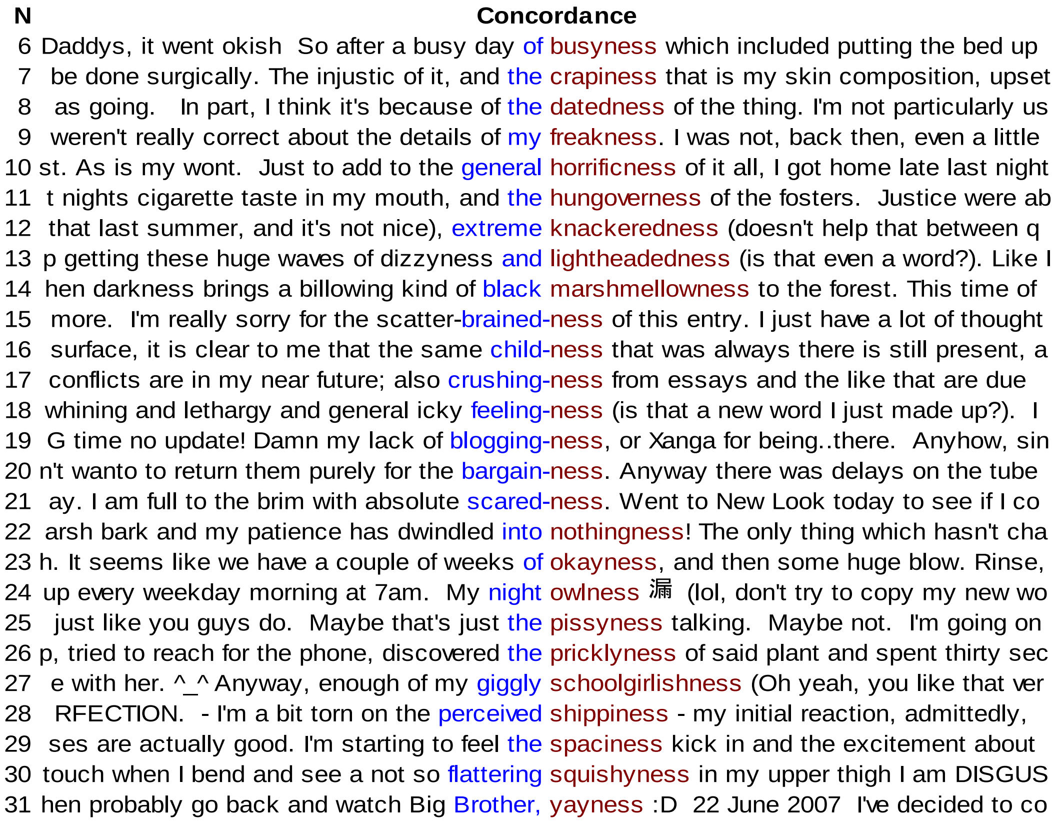
N	Concordance
1	socks, and often fingerless gloves just to feel 'warmish' when the thermostat is set to 72.Da
2	aking more...sort of. And, although it's sort of uniformish, I at least get to choose what I we
3	s the weekend, lets drink, lets get drunk (well tipsyish), lets see friends and be sociable'! I
4	We were minding our own business when this thinish chav girl came upto us after breaking
5	of hide and seek! 12 November, 2006 A Very Sundayish Sunday Today has been nice.
6	en "mean" and idk why. Mayb I am mean and standoffish but idk Im not being honest with
7	me IRL and you've found this journal by some stalkerish means, and all you can see is my
8	he third contender is awesome. Living room is smallish but I think will be enough for us. Kitc
9	e I forget it all, LOL. OK let me start with two short(ish) entries: I have a job again! I intervie
10	n stuff ha ha ha They were discounted pretty sharpish. The bf has seen all of them and tol
11	first time sleeping in a bed at Daddys, it went okish So after a busy day of busyness which
12	lly interesting backstory/concept in a fun little noirish story, so that was also very cool. We
13	then he's CRAP. Bearing in mind that I'm very newbish at playing clix, this is fair. So by the
14	s while I'm there. On a related note, if anyone Londonish fancies meeting up for drinks/food
15	ach succeeding one a little more Bates Motel-ish, sending Lindsay in on a couple occasion
16	st under a week - recovered from jetlag, got ill-ish, got over that, now it's just other stuff. I br
17	ould ever make the Lizzie - the most (formerly-ish) diehard Idol fan y'all know, lol - miss her
18	seem to do it. If I'm not allowed to write prose-ish or poetic or old fashioned dialogue, I will
19	ar 3 .... And I'm done for the Night... Did Ok-ish in the Exam Bai Baaaaiz Back to my De
20	he or whatever cause it is a trendy intellectual-ish movie that has the love story behind it, m
21	nt as well. Neil Morrissey seemed slightly git-ish to start off with but it may have just been
22	se I think I'm coming down with this coughing-ish strain of Freshers Flu. I felt rough when I
23	attempt to make him seem more Face-of-Boe-ish or something? - action!lanto for the EPIC
24	like crying and throwing up [which is an awful girlish way of reacting]. I really just want to b
25	gether on the track and messed around it was funish. sixth period (history) i messed with ou
26	ate Jack/Gwen', which is only marginally less fanbrattish, but there you go. - Whoa, that w

Figure 6.2 Concordance lines for -ish with non-numerals

Adding the suffix *–ish* to a word or phrase gives an impression of imprecision, which functions quite similarly to that of vague language in spoken language. It is used as a means of displaying casualness and thus shortening the distance between the blogger and the reader rather than evidence of showing the blogger's uncertainty or poor memory. The use of vague expressions will be discussed in Chapter 8.

6.4.3 Suffix *-ness*

From what has been presented above, we can see that bloggers are actually exploiting the word-formation strategies to make their messages across in a more interesting way while at the same time trying to maintain a certain rapport with the readers. People may argue that the suffixes *–y* and *–ish* are just special cases because both of them display a tendency of being informal and that may well be the reason why bloggers choose to use them. This kind of argument may make sense to some extent. Nevertheless, bloggers' creative exploitation is not restricted to these two suffixes. In fact, they are also exploiting those very commonly used suffixes in creative ways for communicative purposes. One example would be the noun formation suffix *–ness*. According to Plag (2003), the suffix *–ness* is perhaps the most productive in the English language and it can attach to practically all adjectives. Apart from that, it can also attach to nouns, pronouns, and phrases. There are 31 new lexical items in the corpus which are created through affixation of *–ness*. Most of the base words are adjectives or words of adjective nature, for instance, *angstiness* (cf. angst), *awesomeness*, *annoyingness* (cf. annoyance), *busyness*, *knackeredness,* and *okayness*. There are also examples for base words of other classes: *assness, bargain-ness, childness, blogging-ness, hungoverness, icky-feeling-ness, nothingness, night owlness*, and *yayness*. Some of these new formations may sound a bit strange but they are all formed following the basic principle of word-formation and they are used to express very specific meanings in their own contexts. The following concordance lines (see Figure 6.3 below) give a flavor of how they are actually used in blog entries.



```
 N                              Concordance
 6 Daddys, it went okish  So after a busy day of busyness which included putting the bed up
 7  be done surgically. The injustic of it, and the crapiness that is my skin composition, upset
 8  as going.   In part, I think it's because of the datedness of the thing. I'm not particularly us
 9  weren't really correct about the details of my freakness. I was not, back then, even a little
10 st. As is my wont.  Just to add to the general horrificness of it all, I got home late last night
11  t nights cigarette taste in my mouth, and the hungoverness of the fosters.  Justice were ab
12  that last summer, and it's not nice), extreme knackeredness (doesn't help that between q
13 p getting these huge waves of dizzyness and lightheadedness (is that even a word?). Like I
14  hen darkness brings a billowing kind of black marshmellowness to the forest. This time of
15  more.  I'm really sorry for the scatter-brained-ness of this entry. I just have a lot of thought
16  surface, it is clear to me that the same child-ness that was always there is still present, a
17  conflicts are in my near future; also crushing-ness from essays and the like that are due
18  whining and lethargy and general icky feeling-ness (is that a new word I just made up?).  I
19  G time no update! Damn my lack of blogging-ness, or Xanga for being..there.  Anyhow, sin
20 n't wanto to return them purely for the bargain-ness. Anyway there was delays on the tube
21  ay. I am full to the brim with absolute scared-ness. Went to New Look today to see if I co
22  arsh bark and my patience has dwindled into nothingness! The only thing which hasn't cha
23 h. It seems like we have a couple of weeks of okayness, and then some huge blow. Rinse,
24  up every weekday morning at 7am.  My night owlness 漏 (lol, don't try to copy my new wo
25  just like you guys do.  Maybe that's just the pissyness talking.  Maybe not.  I'm going on
26 p, tried to reach for the phone, discovered the pricklyness of said plant and spent thirty sec
27  e with her. ^_^ Anyway, enough of my giggly schoolgirlishness (Oh yeah, you like that ver
28  RFECTION.  - I'm a bit torn on the perceived shippiness - my initial reaction, admittedly,
29  ses are actually good. I'm starting to feel the spaciness kick in and the excitement about
30 touch when I bend and see a not so flattering squishyness in my upper thigh I am DISGUS
31 hen probably go back and watch Big Brother, yayness :D  22 June 2007  I've decided to co
```

Figure 6.3 Concordance lines for suffix -ness

6.4.4 Prefix *semi-*

Another affix which is also worthy of a note here is the prefix *semi*-. This prefix is often used to quantify the base words, meaning ‘half X or moderately X’, for instance, *semi-challenging* (meaning *moderately challenging*). Although words containing this prefix are not very many in the corpus, most of them are used to display the blogger’s sense of humor or express the blogger’s emotions. For instance, *semi-decide*, *semi-evilness*, *semi-happy*, *semi-naughty*, *semi-challenging*, *semi-drunk*, *semi-impressive*, and *semi-panic attack* all suggest a sense of playfulness. These seemingly vague terms are actually semantically specific. Of course, the prefix *semi*- can also be used to form words with negative sense which will add some emotional force to the statements the bloggers are making. For instance, *semi-fucked, semi-coherent(ly), semi-literate,* and *semi-proper* all suggest a minimum expectation of something the blogger is concerned about and very often this expectation is not met. Again, we see bloggers' effort in taking the advantage of word-formation rules for more accurate conveyance of message and expression of emotions. Figure 6.4 shows how they are used by bloggers.

N	Concordance
2	exhausted Okay, am back from hospital and semi conscious. Had an ODD day all around.
3	s ache but we have a tree with some leaves, a semi filled in dragonfly, the fairy has half her wi
4	ng to kass' and explaining calmly that we were semi fucked It's good, everything will work its
5	hy - its a silly thing). In more constructive and semi impressive (though probably scary for so
6	hing but it was hard. A few weeks later in the semi quiet of our room he is snoring away and
7	r sub-sub-sub culture as boing-boing are. on a semi related note. im thinking of buying some
8	great Mulholland Drive is. (The rest is sort of semi-academic, so here, have a cut.) And ju
9	nothing but window shop, talk, laugh, etc..The semi-anonymous sex just isn't working for me
10	after) I experienced a couple of enjoyable and semi-challenging events. On the 28th he had
11	no Gene Hunt materialised). Then, 11-year-old Semi-Charmed Life came on, and I embraced t
12	s hoping that my offering description is at least semi-coherent. I've been up for waaaaaaay too
13	ing a recovery, to be cognizant enough to write semi-coherently; and worse yet I wish to write
14	ant them all to just leave me the frick alone. I semi-decided to die today. It was funny. OHH
15	love Jun. 3rd, 2007 Like many Brits living in a semi-detached house we share our driveway w
16	t emotional i went to the pub with matt and got semi-drunk and had a nice evening. i didn't sle
17	way and let me brood in peace. Ha, I like my semi-evilness. And alsoooo, I decided that I do
18	t person, I'd get all superstitious and blame the semi-feral group of white-touched black cats th
19	of ourselves on/in it...fun stuff.. i came home..semi-finished my laundry..ate left over chinese
20	t him to change. ever. at least i ended it on a semi-happy note, eh? 10/9/07 So, I'm Eig
21	thening weird fucked up underground tunnel to semi-independency. I lose. 2007.05.19 jes
22	hey call it free. And it is 'free'; free to allow any semi-literate yob to make his case, no matter
23	laundry, starting putting a bag together for my semi-naughty escape tomorrow, etc. it was a
24	ig lesbo. When I first saw her I had an internal semi-panic attack, as something about her re
25	site of a woman who creates rather disturbing semi-pornographic body art). It was nice to swi
26	mb asses. 6. People who cannot use at least semi-proper English when speaking. Using w
27	omeone for saying something because he was semi-raging-drunk. So we had to leave. No big

Figure 6.4 Concordance lines for prefix semi-

6.4.5 Infixation

Compared with prefixation and affixation which are default means of word-formation in English, infixation is barely considered a word-formation process. Morphologists usually agree that English has no infixes (Plag, 2003, p. 101). This statement is quite true in that there are no bound morphemes that qualify for infix status. Nevertheless, having no bound infix morphemes does not imply there is no process of infixation in English. In fact, "there is the possibility of inserting expletives in the middle of words to create new words expressing the strongly negative attitude of the speaker” (p. 101). The process of infixation in English has very strict restrictions on where the expletives can be inserted. They can only be inserted between two feet. According to Plag, a foot is “a metrical unit consisting of either one stressed syllable, or one stressed syllable and one or more unstressed syllables” (p. 102). It is not allowed to interrupt a foot, nor may it appear between an unstressed syllable not belonging to a foot and a foot (pp. 102-103). There are thirteen occurrences of lexical items involving infixation in the blog corpus. For a better understanding of whether bloggers are following the principle of infixation and how exactly they are using this process, all the thirteen occurrences are presented below.

(33) So I went to Asda and fancied some of those mini-eggs. Knew where to find them since they've got an Easter section already. Didn't buy them. *£3-bloody-18p* (uk_f_15-17.txt).

(34) *Fan Bloody tastic* news (title of a blog entry, uk_f_18-19.txt)

(35) The only *fan-fucking-tastic* thing is that they have the German version of Frozen To Loose It All (uk_m_18-19.txt).

(36) I've had fuck all sleep this week as it is with the Expo and editing videos, and now the fucking neighbours are knocking about all SUNDAY *after-fucking-noon* (uk_m_30-34.txt).

(37) oh, check out unarmed for victory on myspace. *infuckingsane*. i cant fake an interests in this (us_m_15-17.txt).

(38) You're *unfuckingbelievable* (us_f_18-19.txt).

(39) yea it's been a *fan-fucking-tastic* day (us_f_20-24).

(40) *Unbefuckinglievable* (title of a blog entry, us_f_25-29.txt).

(41) Tracey got the mula for the mortgage and the car payment. YAY! *Woo-freaking-hoo* (us_f_20-24.txt)!

(42) So, suddenly and out of nowhere, Bowyer has the lead, heading into a green-white-checkered, with *Kyle-fucking-Busch* right on his ass (us_m_20-24.txt).

(43) …the guy did not speak english and had the *absofrickinlutly* brilliant idea of using a socket wrench on my positive battery terminal and having me turn over the engine (us_m_30-34.txt).

(44) *SONOFAFUCKINGBITCH*! i only had my car six months and some dumb, careless bitch rear ends the guy behind me at a stoplight then sends him into me (us_f_30-34.txt).

(45) I then returned to the kitchen for a snack of *Buckwheat-Motherfucking-Crunch* and WTF?!? Right there, in my freezer, sat a pint of Cherry Garcia Ice Cream (us_f_35-40.txt)!

From these sentences we can see that all these new items created by bloggers comply with the rules of infixation. Nevertheless, the resultant words may not necessarily carry negative meaning as Plag suggests. Some items are actually used to express positive emotions, for instance examples 34, 35, 39, and 41. In other words, the expletives inserted are actually amplifying the effect of the original expressions. The use of infixation is not restricted to any age or gender groups. Rare as infixation is in the English language, its presence in personal blogs suggests something about the nature of blogging as a medium for expressing personal emotions.

6.5 Minor word-formation strategies

Apart from the three major word-formation processes discussed in the previous sections, there are actually some minor word-formation strategies which are of interest in this research. There are not many occurrences for each word-formation strategy; however, the very presence of these strategies stands as an indicator of bloggers' creative effort in representing themselves in linguistically interesting ways. By minor word-formation strategies, I am actually referring to those processes which do not involve affixations, be it prefixation, infixation, or suffixation. These strategies include conversion, clipping, clipping compounding, blending, word-manufacture, and using initials as new words. Some of them have already been mentioned in Chapter 5 as strategies for creating non-conventional orthographic representations of words for stylistic or pragmatic purposes. Nothing about the morphological or semantic aspect has yet been discussed. The minor strategies to be discussed below include: blending, clipping, using initials and acronyms as verbs, creative spelling, and leetspeak.

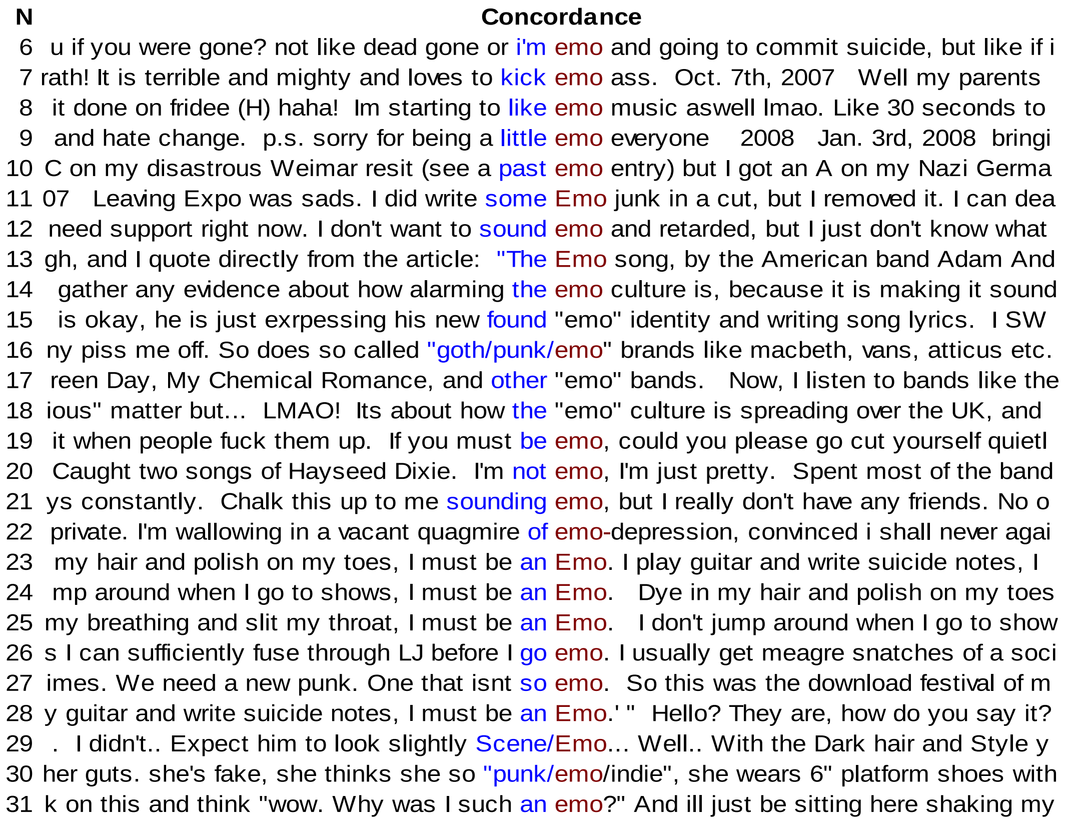
6.5.1 Blending

According to Plag (2003), blending refers to the word-formation process of combining two (rarely three or more) words into one by deleting material from one or both of the source words. There two major types of blending: blending compounds and blends. The former refers to the shortening of existing compounds into single words by taking the initial part of the first words and the last part of the second word. For instance, *sitcom* is a blending compound formed out of *situational comedy*. The latter, however, is a word-formation process which combines the first part of the first word and the last part of the second word to form a new word. For example, *smog* is formed on the basis of *smoke* and *fog*. Blends are semantically different from the clipping compounds in that they share properties of the referents of both elements whereas the clipping compounds do not. For instance, a *motel* (*motor hotel*) is a hotel whereas a *boatel* (*boat hotel*) is both a boat and a hotel (Plag, 2003, p. 122).

As mentioned earlier, words created out of blending are not very many in the EBC, but almost all of them carry colloquial, technical, or even slangy flavor. Here are some examples: *twunt* (*twat cunt*) (a vulgar term), *photo-op* (*photograph opportunity*), *chillax* (*chill relax*), *gianormous* (*giant enormous*), *Spanglish* (*Spanish English*), *sucktastic* (suck fantastic), *Laban-type* (*Labanotation type*, a kind of dance), *fugly* (*fucking ugly*), *humongous* (*huge tremendous*), *snark* (*snide remark*), *craputacular* (*crap spectacular*), *recon* (*retroactive continuity*), *huggle* (*hug cuddle*), *misper* (*missing person*), *mo-fos* (*mother fuckers*), *sci-fi* (*science fiction*), *compsci* (*computer science*), *concall* (*conference* *call*), *sysprog* (*systems programmer*), *winsock* (*windows socket*), *chmod* (*change mode*), *satnav* (*satellite navigation*), *e-zine* (*electronic magazine*), and *pod-cast* (*iPod broadcast*). Among these words, some are of strong slangy flavor, for instance, *crapultacular*, *chillax*, and *sucktastic*. The employment of such words is a rather obvious marker of informality aside from the creativity embodied in the new words. As for those blends which are of more technical flavor, they are more often used as markers of ingroupness. When a shortened form of a technical word is used, the blogger takes it for granted that the intended readers are able to understand it. If the reader cannot understand it, it only suggests that he or she is not a member of the community. A rough examination of the users of such new words reveals no particular pattern. No association between the use of such words with age or gender can be established.

6.5.2 Clipping

Apart from blending, bloggers also use clipping for creating new words. According to Bauer (2006), clipping refers to the shortening of words while retaining the original meaning. Clipping does not create lexemes with new meanings, but lexemes with a new stylistic value (p. 498). Bauer is right in pointing out the stylistic value of words created out of clipping, but her claim that clipping does not create lexemes with new meanings may not always hold, especially in online discourse such as personal blogs where people tend to be very creative sometimes. One word (word-form) which is obviously the result of clipping is *emo* (from *emotive* or *emotional*). If Bauer is correct, then this new word or lexeme *emo* is just a stylistic variant of the original word *emotive* or *emotional,* but this is only half true, as can be observed from the following concordance lines (see Figure 6.5).



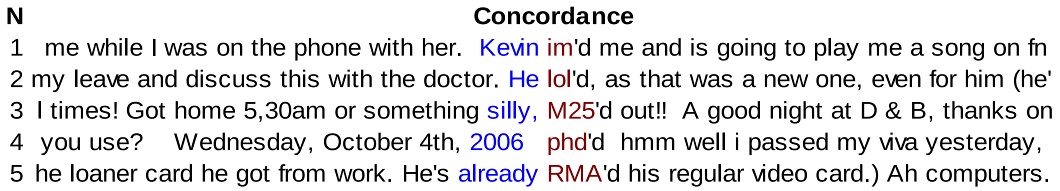
N	Concordance
6	u if you were gone? not like dead gone or i'm emo and going to commit suicide, but like if i
7	rath! It is terrible and mighty and loves to kick emo ass. Oct. 7th, 2007 Well my parents
8	it done on fridee (H) haha! Im starting to like emo music aswell lmao. Like 30 seconds to
9	and hate change. p.s. sorry for being a little emo everyone 2008 Jan. 3rd, 2008 bringi
10	C on my disastrous Weimar resit (see a past emo entry) but I got an A on my Nazi Germa
11	07 Leaving Expo was sads. I did write some Emo junk in a cut, but I removed it. I can dea
12	need support right now. I don't want to sound emo and retarded, but I just don't know what
13	gh, and I quote directly from the article: "The Emo song, by the American band Adam And
14	gather any evidence about how alarming the emo culture is, because it is making it sound
15	is okay, he is just exrpessing his new found "emo" identity and writing song lyrics. I SW
16	ny piss me off. So does so called "goth/punk/emo" brands like macbeth, vans, atticus etc.
17	reen Day, My Chemical Romance, and other "emo" bands. Now, I listen to bands like the
18	ious" matter but... LMAO! Its about how the "emo" culture is spreading over the UK, and
19	it when people fuck them up. If you must be emo, could you please go cut yourself quietl
20	Caught two songs of Hayseed Dixie. I'm not emo, I'm just pretty. Spent most of the band
21	ys constantly. Chalk this up to me sounding emo, but I really don't have any friends. No o
22	private. I'm wallowing in a vacant quagmire of emo-depression, convinced i shall never agai
23	my hair and polish on my toes, I must be an Emo. I play guitar and write suicide notes, I
24	mp around when I go to shows, I must be an Emo. Dye in my hair and polish on my toes
25	my breathing and slit my throat, I must be an Emo. I don't jump around when I go to show
26	s I can sufficiently fuse through LJ before I go emo. I usually get meagre snatches of a soci
27	imes. We need a new punk. One that isnt so emo. So this was the download festival of m
28	y guitar and write suicide notes, I must be an Emo.' " Hello? They are, how do you say it?
29	. I didn't.. Expect him to look slightly Scene/Emo... Well.. With the Dark hair and Style y
30	her guts. she's fake, she thinks she so "punk/emo/indie", she wears 6" platform shoes with
31	k on this and think "wow. Why was I such an emo?" And ill just be sitting here shaking my

Figure 6.5 Concordance lines for EMO

From Figure 6.5 we can see that the new word *emo* has at least three different senses: a subgenre of music or subculture, a type of person who behaves in a particular way, and a quality which is related to the subgenre of punk music or lifestyle. There are several other words which are created out of clipping, apart from the ones already mentioned in Chapter 5. For instance, *combo* for combination, *pedo* for *pedophile, tomoz* for *tomorrow, manips* for *manipulations*, *perv* for *pervert*, and *perving* for *perverting*, *tard* for *retarder*, *convos* for *conversations*, and *crasher* for *gatecrasher*. Small in numbers as they are, they are evidence of personal bloggers' efforts in exploiting word-formation strategies to fulfill their communicative purposes and distinguish themselves from other people.

6.5.3 Using initials and acronyms as verbs

Apart from blending and clipping, bloggers sometimes use certain word-formation strategies which are generally held to be impossible for morphologists. One such strategy is using initials and acronyms as verbs. According to Bauer, initials and acronyms do not appear to be used as verbs although she does not rule out the possibility of the conversion of initialism (2006, p. 500). A closer examination of the unknown word list reveals that bloggers do not seem to take this principle too seriously, as there are instances of initials and acronyms being used as verbs. Let us start with the following concordance lines (Figure 6.6):



```
N                                   Concordance
1   me while I was on the phone with her.  Kevin im'd me and is going to play me a song on fn
2 my leave and discuss this with the doctor. He lol'd, as that was a new one, even for him (he'
3  I times! Got home 5,30am or something silly, M25'd out!!  A good night at D & B, thanks on
4  you use?    Wednesday, October 4th, 2006   phd'd  hmm well i passed my viva yesterday,
5  he loaner card he got from work. He's already RMA'd his regular video card.) Ah computers.
```

Figure 6.6 Concordance lines for initials and acronyms used as verbs

We can make three observations from these five examples. First, all the node words of the concordance lines are abbreviations of some sort. *IM* is the short form for instant messaging. *Lol* is the acronym of a verbal phrase ‘laughing out loud’ and some people used it as a new verb which has its own past form inflection. *PhD* is an abbreviation for ‘Doctor of Philosophy’. *M25* is an abbreviated term for the M25 motorway (also known as the M25 corridor) which is an orbital motorway encircling the Greater London area. As one of the busiest stretches of the British motorway network, M25 is renowned for its traffic congestion. *RMA* is actually the initialism of “Return Merchandise Agreement.” *To RMA* is to return a product, for whatever reason, to the seller. Second, they have all undergone the process of conversion (or transposition) from abbreviated nouns and been used as verbs. Third, their past tense inflections have all adopted the archaic spelling practice called syncope. Syncope refers to the shortening of a word by omitting one or more letters or syllables in the middle. According to Barber (1997), in the middle of the Early Modern period, syncope is often indicated by the spelling which removes the vowel letter of the inflectional morphemes such as *-id* or *-yd*. During the seventeenth century there was a tendency to standardize the spelling *-ed*, but in the later part of the century the spelling *-'d* was often used to indicate syncope, especially for poetry. For instance, John Dryden, one of the most influential poets in the late seventeenth century, regularly used spellings like *chang'd*, *confess'd,* and *disdain'd,* to show that the ending was not syllabic, and this practice was continued in the eighteenth century (pp. 174-175). Of course, the bloggers who have created those lines cited in the above concordance lines may not do it for metrical purposes as the famous poet intended to achieve. Nevertheless, being able to use an archaic inflectional form (such as the syncope in this case) reveals these bloggers' knowledge of the history of the English language. Meanwhile, they are actually exploiting the striking difference between an archaic spelling practice and the very modern terms to create a special effect, although the possibility for these forms to be mere results of phonetic spelling does exist.

Of course, not every blogger uses the syncope as the past tense form and past tense form is not the only verbal form which verbs originated from initials and acronyms take. Here are some examples:

(46) I rather a republican, but if I had to choose I would choose Barack Obama over Hillary Clinton because she would start *PMS-ing* and take it out on China (us_f_15-17.txt).

(47) I was pretty *wtfed* at that point (uk_m_18-19.txt).

(48) A fit young guy at the sauna got half-hard just from looking at my naked body. You best believe I *lol'ed* hard, inwardly (uk_f_25-29.txt).

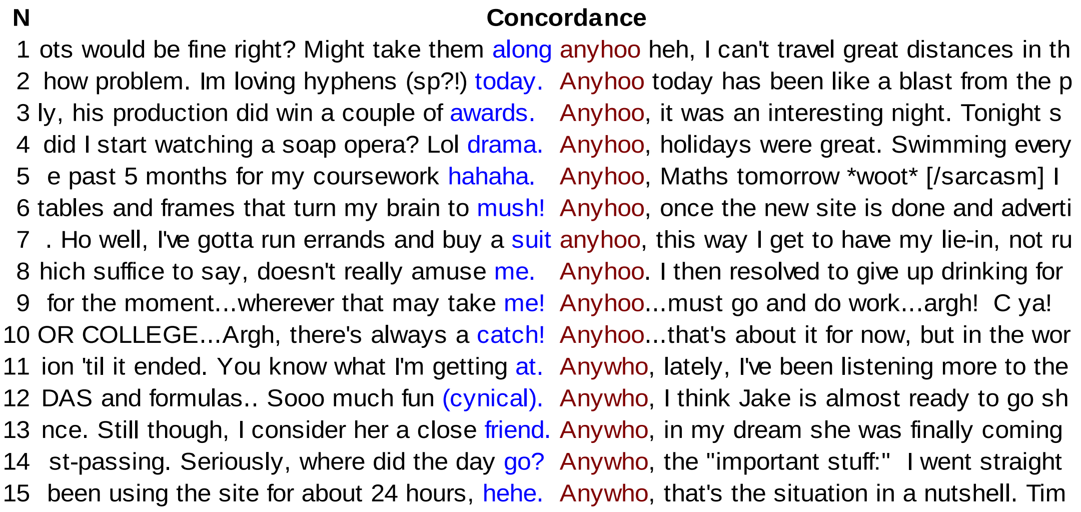
(49) but as G said recently 'RL comes first' and even thoug i feel really abd for kippering their income for this year i really cant justify the cost, primialy monitary, but also mental and physical that *larping* these days intales (uk_f_18-19.txt).

Examples 48 and 49 may not be very convincing to some people, as the acronym *lol* is formed out of a verbal phrase (*laughing out loud*). Considering that original phrase is in present participial form, it makes more sense to say that the bloggers are actually using the acronym *lol* as a new verb.

The minor types of word-formation presented above may not be linguistically very important because very often they arise at the point where system gives way to random creativity (Bauer, 2006). Nevertheless, they are of increasing importance in the lexicon of modern English. Many of these new forms "may appear ephemeral, extremely localized or rather slangy in tone, but so are many words formed by more established word-formation processes" (2006, p. 503). From this random creativity we can actually get a feel of the individual identity of some bloggers.

6.5.4 Creative spelling

Another group of words are actually results of the bloggers' efforts in playing with the English language, especially the pronunciation. For instance, *applorling* for *appalling*, *hyoooge* for *huge*, *naw* for *no*, *reet* for *right*, *fings* for *things*, *alrighty* for *alright*, *barfday* for *birthday*, *exadurate* for *exaggerate*, *froonch* for *French*, *sammiches* for *sandwiches*, *smex* and *smexy* for *sex* and *sexy*, *nekkid* for *naked*, *vell* for *well*, *rocktober* for *October*, *anyhoo* and *anywho* for *anyhow*. Almost all these word-formations involve some play with the sound of the original words and have acquired a sense of playfulness, humor, and informality. Some of these words are quite commonly used in personal blogs, for instance, *anyhoo/anywho*, as can be seen from the following concordance lines (Figure 6.7).



```
 N                                    Concordance
 1 ots would be fine right? Might take them along anyhoo heh, I can't travel great distances in th
 2  how problem. Im loving hyphens (sp?!) today.  Anyhoo today has been like a blast from the p
 3 ly, his production did win a couple of awards.  Anyhoo, it was an interesting night. Tonight s
 4  did I start watching a soap opera? Lol drama.  Anyhoo, holidays were great. Swimming every
 5  e past 5 months for my coursework hahaha.  Anyhoo, Maths tomorrow *woot* [/sarcasm] I
 6 tables and frames that turn my brain to mush!  Anyhoo, once the new site is done and adverti
 7  . Ho well, I've gotta run errands and buy a suit anyhoo, this way I get to have my lie-in, not ru
 8 hich suffice to say, doesn't really amuse me.  Anyhoo. I then resolved to give up drinking for
 9  for the moment...wherever that may take me!  Anyhoo...must go and do work...argh!  C ya!
10 OR COLLEGE...Argh, there's always a catch!  Anyhoo...that's about it for now, but in the wor
11  ion 'til it ended. You know what I'm getting at.  Anywho, lately, I've been listening more to the
12  DAS and formulas.. Sooo much fun (cynical).  Anywho, I think Jake is almost ready to go sh
13  nce. Still though, I consider her a close friend. Anywho, in my dream she was finally coming
14   st-passing. Seriously, where did the day go?  Anywho, the "important stuff:"  I went straight
15   been using the site for about 24 hours, hehe.  Anywho, that's the situation in a nutshell. Tim
```

Figure 6.7 Concordance lines for ANYHOO/ANYWHO

6.5.5 Leetspeak

Different from all the word-formation strategies we have mentioned so far, leetspeak is the only one which is not rooted in natural human languages; rather, it has plenty to do with computing language. Because of this, leetspeak is normally associated with online discourses. According to Wikipedia[10], a free online encyclopedia complied by netizens, leet or leetspeak is an alphabet used primarily on the Internet, which uses various combinations of ASCII characters to replace Latinate letters. Derived from the word “elite,” the term leet is often used to describe a specialized form of symbolic writing. A typical leetspeak would look like the following, which is an example reproduced from the Wikipedia, just to give a flavor of it:

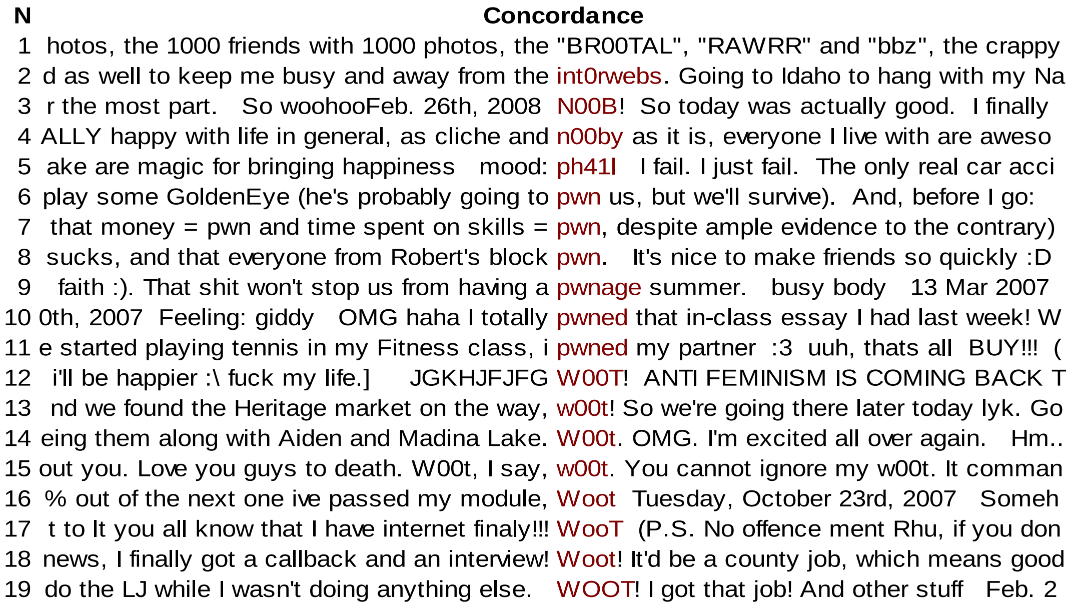
Leetspeak: L337 15 n07 4 c0mm0n 1n73rn37 5p34k 4m0n9 r34l h4x0r

English translation: Leet is not a common Internet speak among real hackers

If we take a closer look at the leetspeak example and its English translation we will immediately find some correspondence between the numbers used in the leetspeak and the English letters they are intended to represent. For instance, 3 for e, 7 for t, 1 for I, 5

[10] http://www.wikipedia.org/

for *s*, 4 for *A*, 0 for *o*, 9 for *g*, and so on. Using numbers and ASCII symbols to replace ordinary letters obviously increases the difficulty in deciphering the message and makes leetspeak a sort of argot among special groups of people. And that is exactly what the inventors of leetspeak intended for. There are not many cases of leetspeak in the EBC, only 30 occurrences, but they suffice to show its influence on netizens' language use. Figure 6.8 gives a flavor of how leetspeak is actually used by bloggers.



```
N                              Concordance
 1  hotos, the 1000 friends with 1000 photos, the "BR00TAL", "RAWRR" and "bbz", the crappy
 2  d as well to keep me busy and away from the int0rwebs. Going to Idaho to hang with my Na
 3  r the most part.   So woohooFeb. 26th, 2008  N00B!  So today was actually good.  I finally
 4  ALLY happy with life in general, as cliche and n00by as it is, everyone I live with are aweso
 5  ake are magic for bringing happiness   mood: ph41l   I fail. I just fail.  The only real car acci
 6  play some GoldenEye (he's probably going to pwn us, but we'll survive).  And, before I go:
 7  that money = pwn and time spent on skills = pwn, despite ample evidence to the contrary)
 8  sucks, and that everyone from Robert's block pwn.   It's nice to make friends so quickly :D
 9  faith :). That shit won't stop us from having a pwnage summer.   busy body   13 Mar 2007
10  0th, 2007  Feeling: giddy   OMG haha I totally pwned that in-class essay I had last week! W
11  e started playing tennis in my Fitness class, i pwned my partner  :3  uuh, thats all  BUY!!!  (
12  i'll be happier :\ fuck my life.]     JGKHJFJFG W00T!  ANTI FEMINISM IS COMING BACK T
13  nd we found the Heritage market on the way, w00t! So we're going there later today lyk. Go
14  eing them along with Aiden and Madina Lake. W00t. OMG. I'm excited all over again.   Hm..
15  out you. Love you guys to death. W00t, I say, w00t. You cannot ignore my w00t. It comman
16  % out of the next one ive passed my module, Woot  Tuesday, October 23rd, 2007   Someh
17  t to It you all know that I have internet finaly!!! WooT  (P.S. No offence ment Rhu, if you don
18  news, I finally got a callback and an interview! Woot! It'd be a county job, which means good
19  do the LJ while I wasn't doing anything else.   WOOT! I got that job! And other stuff   Feb. 2
```

Figure 6.8 Concordance lines for leetspeak

As can be observed from Figure 6.8, two leetspeak words are more commonly used: *pwn* and *woot* (including their variants). In fact, these two words have already become well-accepted slang words among young netizens. According to Urbandictionary, '*pwn*' refers to an act of dominating an opponent. '*Pwn*' is actually the misspelling of '*own*'. It dates back to the days of WarCraft (a very popular online game) when a map designer mispelled 'own' as 'pwn'. As a consequence, what was supposed to be 'the player has been owned' became 'the player has been pwned'. Growing from there, the misspelling '*pwn*' spread over the online world and has eventually become a new word which has gained great currency among young netizens, especially online game players. Aside from inheriting the word class of the original word ‘own’, it has also got its new derivative noun ‘*pwnage*’. The origin of the leetspeak word ‘*woot*’ is also quite interesting. Urbandictionary has it that ‘*w00t*’ was originally a blend for “Wow, loot!,” an expression very common among players of Dungeons and Dragons tabletop role-playing game. This term later entered the Internet subculture of video game communities and lost its original meaning and is now used simply as a term of excitement. This is also why the word and its variants are almost exclusively used as inserts in the concordance lines cited above (Figure 6.9). As for the rest few terms, ‘*n00b*’is also related to online computer games, but the word itself may have something to do with another word “newbie.” ‘*N00b*’ (or ‘*noob*’, ‘*n00by*’) is just a more innovative or playful way of spelling the word ‘*newbie*’. ‘*Br00tal*’, on the other hand, has nothing to do with online gaming. It is associated with heavy metal, a musical genre which is held to be unconventional and rebellious and which is believed to be responsible for several suicide cases of young people in the western world, thus being ‘brutal’. The term ‘ph41l’ is a standard leet form for ‘*fail*’. Neither associated with online games nor lethal music, the meaning of the word itself could be “devastative,” so to speak. Despite the relative low frequency of these leetspeak words, their playful nature and their function as identity markers are undeniable.

What we have presented so far are basically words or word forms created by personal bloggers through the application (sometimes creative application) of existing word-formation strategies. What we can observe from these neologisms is bloggers’ creativity and their knowledge of folk morphology. Some of these neologisms may have something to do with bloggers’ identity representation. This issue will be addressed in greater detail in Chapter 10. Of course, as language users, our main job is not to create new words, but rather use whatever linguistic means which is within our reach for achieving the communicative purposes we are intended for. For certain words, ordinary language users may not be at the position to create them, for instance, neologisms concerning information technology in general and the Internet in particular. By examining the use of neologisms pertaining to IT and the emergent culture on the Internet, we can also obtain certain insights about bloggers' identity representation.

6.6 Neologisms related to IT and Internet culture

With the deeper penetration of computers and the Internet into ordinary people's daily life, neologisms related to information and communication technology and Internet-based new culture have accelerated their pace in making into the daily linguistic repertoire of ordinary language users, and netizens in particular. There are approximately 1,238 tokens of new lexical items (or items which have acquired new semantic meanings) which are related to various information technologies (chiefly computer and Internet technology and the products or services derived from these developments) and emerging Internet culture. These new words are not created by personal bloggers via complying with or deviating from the word-formation principles or strategies of the English language (although most of these items do follow the mainstream word-formation processes of English) per se, but they are frequently used by personal bloggers. They can be roughly divided into two major categories: neologisms about IT hardware and software (e.g., *Wi-Fi, iPod, Google,* and *Wiki*) and neologisms related to emergent Internet culture (e.g., *LiveJournal, MySpace, Facebook, fanfiction, anime,* and *Wii).*

The IT-related new words are mostly technical terms concerning new computer hardware and software and the actual operation of the devices and functions. There are 426 word tokens, covering 78 word types. Words with two occurrences and above in the corpus include: *online* (104), *Internet* (90), *link* (inflectional forms inclusive) (41), *website(s)* (39), *Google* (16), *upload*(*ed*) (15), *wiki*(*pedia*) (13), *XP* (11), *USB* (10), *gb* (gigabyte) (9), mb (megabyte) (6), *Firefox* (5), *LCD* (5), *O2* (4), *wi-fi* (3), *bluetooth* (3), *firewire* (3), *NaNo* (3), *url* (3), *dongle* (2), *drm-free* (2), *rss* (2), *fedora* (2), *reconnect*(2), sata (2), and *sat-nav* (2).

The rapid popularization of the Internet has created many new platforms for people to communicate with one another. These new communication means have contributed considerably to formation and spread of many neologisms. Altogether, there are 1,195 tokens of words referring to social networking websites and related terms. Table 6.9 lists some of the more frequently used neologisms related to emergent Internet culture. The very existence of these words or word forms reveals how human life and human language is being influenced by the development of information technology and the subsequent social changes. In fact, Internet-mediated communication has kept refreshing our daily linguistic repertoire ever since its very existence. Some words have successfully sneaked into daily vocabulary without our awareness, for instance, *email*, *message* and *forum*. Apart from these terms, there are also quite a number of new words which have already made it into many people's daily linguistic repertoire but may not have won the recognition of authoritative dictionaries. Many of these words are names of newly emerged social network websites and related terms. To a considerable extent, being familiar with these terms and being able to use them has also become an identity marker.

Table 6.9 Neologisms concerning Internet-based communication

Neologism (lexeme)	Tokens	Word forms
post	351	post, posts, posted, posting
blog/blogger	180	blogs, blogged, blogging, blogger
update	137	update, updates, updated
email	91	email, emails, emailed, emailing
comment	90	comment, comments, commented, commenting
message	83	message, messages, messaging
MySpace	76	MySpace
Facebook	41	Facebook
YouTube	28	Youtube, YouTube
forum	19	forum, forums
flickr	7	flickr

The less frequently occurred terms include: *Skype* (3), *avatar* (3), *moniker* (2), *spammer(s)* (4), *webcam* (2), *wordpress* (2), *gmail* (2), *irc* (4), *cosplay island* (3), *bbs* (2), *bacn*, *bebo*, *bitchx*, *concalls*, and *renderosity*.

Some neologisms are closely related to newly emerged Internet culture. There are 76 tokens of such words, covering 18 word types. Table 6.10 lists them all. Despite their less frequent occurrences in the corpus, these words represent a very important part of the Internet-mediated non-mainstream cultures.

Table 6.10 Terms related to newly emerged Internet culture

Lexical items	Frequency	Lexical items	Frequency
fanime	11	deviantart	2
fandom(s)	10	fanfiction	2
yaoi	10	cthulhu	1
meme	8	fanart	1
fanfic(s)	7	fanbook	1
retcon(ned)	6	fangasm	1
fanboy (ing)	5	fantascicon	1
fangirl (ing)	4	otaku	1
webcomic(s)	4	xkcd	1

An important example of the Internet culture is the emergence of fantasy fictions (fanfictions for short). The influence of best-selling fantasy fictions represented by *Harry*

Potter and *The Lord of Rings* and the ever-increasing easier accessibility of the Internet has fanned the flames of many people's zest in writing their own fantasy fictions. The popularity of blogging websites has also inspired many bloggers to explore their writing talents and quite a few of them have dedicated much of their enthusiasm and time to publishing short stories or novels which are better known as fanfictions to citizens of the virtual world. Unlike personal blogs which are of autobiographical nature, fanfictions are actually creative (imaginary) writing which requires good skills in plotting, storytelling, and language. Being able to write fanfictions and attract a greater number of readers carries lots of currency among bloggers with this hobby. Another interesting new cultural phenomenon is something called fandom. According to Wikipedia, fandom is a subculture composed of fans characterized by a feeling of sympathy and camaraderie with others who share a common interest. For instance, people who like one particular celebrity or those who share the same hobbies may form an online community, called fandom. Fans typically are interested in even minor details of the object(s) of their fandom and spend a significant portion of their time and energy on their interest. This is what differentiates them from those with only a casual interest. Closely related to this fandom subculture are two other terms: *fanboy* and *fangirl*. *Fanboy* is a term used to describe an individual who is devoted to a single subject in an emotional or fanatical manner, often to the point where it is considered an obsession. According to the *Merriam-Webster Collegiate Dictionary,* the earliest known use of the term 'fanboy' can be traced back to an English-language publication in 1919. By 1990 the term was being used in popular music and science fiction circles. Later, it became increasingly applied to computers and video game consoles. Current subjects of such obsessive loyalty include almost everything from TV shows, movies, and music to video games, computer hardware, and software tools. *Fanboy* was added to the *Merriam-Webster Collegiate*

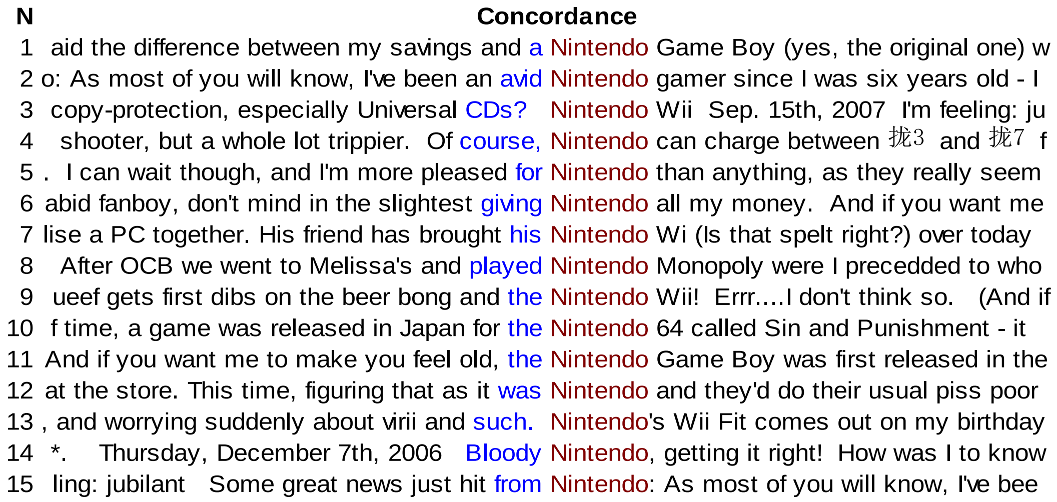
Dictionary in 2008. The term ‘fangirl’ carries slightly different connotations. It is often used to refer to an enthusiastic female fan (regardless of obsessive qualities) and is often used with overtones of ‘teenybopper’. It can also be used or perceived as a derogatory label, depending on the context of use. Mainly used as nouns, the terms *fanboy* and *fangirl* can also be used as verbs and have inflectional forms such as *fanboying* or *fangirling*. Apart from these two fan-related Internet subcultures, there is another type called *meme* or *Internet meme* (some people call it a form of art) which enjoys a great popularity among netizens, especially bloggers. According to Wikipedia, meme at its most basic form is simply the spread of a digital file or hyperlink from one person to another via Internet-based communication forms. The content often consists of a saying or joke, a rumor, an altered or original image, a complete website, a video clip, or animation, among many other possibilities. An Internet meme may stay the same or may evolve over time, by chance or through commentary, imitations, and parody versions, or even by collecting news accounts about itself. Internet memes have a tendency to evolve and spread extremely quickly, sometimes going in and out of popularity in a matter of days.

Due to the great influence of the Japanese culture in the form of computer games, animation movies, and cartoons, some words originated from the Japanese language have become very popular in English, for instance, *anime, manga, and Nintendo*. *Anime* is actually the Japanese version of the English word ‘animation’. In other words, Japanese borrowed this English word and transformed it according to its own pronunciation system and then exported the new word ‘*anime*’ to the rest of the world. *Manga* is a Japanese word for comic cartoons and it has been accepted by western youth as a new English word. One more term which also falls into the category of Internet subculture is *Yaoi*. It is a popular term for fictional media that focuses on homosexual male relationships, yet is generally created by and for females. Originally referring to a type of self-published parody of mainstream anime and manga works, it is currently being used as a generic term for female-oriented manga, anime, or novels about homosexual male relationships. The following concordance lines (see Figure 6.9 below) show how *anime* and *manga* are used in the EBC. A possible explanation for the presence of Japanese words in the blog corpus is that some bloggers may have cultural bonds with Japan. The heterogeneity of the British and American population suggests that there may well be Asian British or American citizens among the bloggers included in the EBC. By talking about the culture of their home country, bloggers are actually revealing a part of their cultural identity.

N	Concordance
1	, and um...Gareth, Amii and I decided to write an Anime Soc fanfiction, which consisted of Alex an
2	d a drink in the SU to while away the time before Anime Soc. Anime Soc was a riot also. Much s
3	long to ensure I never get invited again. My entire anime collection should do nicely. On that note, I
4	rossed for next week's replies. Came home from anime night on the train from Basingstoke. Tom
5	s! :S OH OH! I have started getting interested in Anime again! D: I was wondering if anybody out t
6	S in the world, too!!! And you DONT see them in anime too goddamn often, DO YOU?!? 14th-Jan-
7	play on it. Now I have a bazillion songs, (mostly anime OSTs, although a few albums) on it! And I
8	ust did some work for a few hours. Ive got a new anime im watching atm called Eureka Seven, its
9	here and I can't download anything...so DVDS of Anime are VERY welcome :D I keep watching t
10	HEAP INNO WANTS!!! It also has other Random Anime Cosplay JEwelery that I might buy and we
11	the SU to while away the time before Anime Soc. Anime Soc was a riot also. Much sweets, much
12	found ourselves spending most of our time at the Anime counter. I found like 6 different things i wa
13	soundtrack. Noein: Katie + Spring. Watch this anime it's short 24 eps you will love it. Internet:
14	CKING TIRED OF GIRLS WITH BIGS BOOBS IN ANIME!!! I understand that there are women in t
15	LEDORE DIES ZOMG). It ends different than the Anime, and Shana also told me it ends different t
16	to have a Tim Burton & possibly Chobits day. & anime//manga in general......... Yeah I have a q
17	er happened i bought their new cd, and a manga/anime/jrock mag that had tomo in it and i had ry
18	il is blocked Myspace is blocked. Anything with anime/manga references are blocked. All image
19	s card, Sycorax Warrior, 12 inch Sec, Kerrang, a manga magazine (causa Death Note, which we s
20	sh on Elizabeth!!! . This morning we went to a manga cafe for coffee (which I'm not allowed to h
21	~*happy dance*~ Ok darling, I'm taking you to a manga shop, where you can pick out what you w
22	e a Tim Burton & possibly Chobits day. & anime//manga in general......... Yeah I have a question
23	cked Myspace is blocked. Anything with anime/manga references are blocked. All images inclu
24	r needed to see again (apart from a Demon Diary manga which I'm glad I got back). The thing I'm g
25	Anime Soc was a riot also. Much sweets, much manga (though I read none, as usual), and um...
26	o busy, reading the truckton of Naruto/Bleach/DN manga my parents bought me! As well as playin
27	here. After I get a job I'm going to start ordering Manga and novels from Book World, then cross t
28	ED to go xmas shopping!! D: Ok, so Jess, what manga book would you like? Or what else do yo
29	ot volumes 1 -3 of M盲rchen Awakens Romance manga. It almost looks like something that was
30	at's never happened i bought their new cd, and a manga/anime/jrock mag that had tomo in it and

Figure 6.9 Concordance line for ANIME and MANGA

The influence of the Japanese culture is also great in the gaming industry. As a result, another Japanese word is also well-known among game players: *Nintendo*. *Nintendo* is the name of a Japan-based multinational corporation which is very famous worldwide for its video games. The following concordance lines (Figure 6.10) show this trend.



```
N                                    Concordance
1   aid the difference between my savings and a Nintendo Game Boy (yes, the original one) w
2 o: As most of you will know, I've been an avid Nintendo gamer since I was six years old - I
3   copy-protection, especially Universal CDs?   Nintendo Wii  Sep. 15th, 2007  I'm feeling: ju
4     shooter, but a whole lot trippier.  Of course, Nintendo can charge between 拢3  and 拢7  f
5 .   I can wait though, and I'm more pleased for Nintendo than anything, as they really seem
6  abid fanboy, don't mind in the slightest giving Nintendo all my money.  And if you want me
7  lise a PC together. His friend has brought his Nintendo Wi (Is that spelt right?) over today
8    After OCB we went to Melissa's and played Nintendo Monopoly were I precedded to who
9   ueef gets first dibs on the beer bong and the Nintendo Wii!  Errr....I don't think so.   (And if
10  f time, a game was released in Japan for the Nintendo 64 called Sin and Punishment - it
11  And if you want me to make you feel old, the Nintendo Game Boy was first released in the
12  at the store. This time, figuring that as it was Nintendo and they'd do their usual piss poor
13 , and worrying suddenly about virii and such.  Nintendo's Wii Fit comes out on my birthday
14  *.    Thursday, December 7th, 2006   Bloody Nintendo, getting it right!  How was I to know
15  ling: jubilant   Some great news just hit from Nintendo: As most of you will know, I've bee
```

Figure 6.10 Concordance lines for NINTENDO

Keeping up with the latest fashion is an important part of youth culture. Owning fashionable/trendy IT gadgets or popular video or online games is also an important part of young people’s identity representation. Thus, examining the neologisms related to IT gadgets or computer games can help reveal certain aspects of bloggers’ identities. *iPod* and *iTunes* are two good examples. As iPod and its relating software iTunes have both been widely accepted by the consumer market, it will be quite natural to find this word in the daily dictionary of all blogger age groups. Owning an iPod basically implies the owner has Internet access and knows how to use it. Again, this is a part of people’s lifestyle and a marker of identity. Of course, this does not follow that those bloggers who have not used these terms are not iPod owners and iTunes users. What we can only conclude is that some bloggers explicitly mention that IT gadgets of such kind are part of their daily life. Apart from iPod and iTunes, there are a number of new lexical items which are related to computer games, for instance, *Wii*, *xboxing*, *alpha-complex*, *FFXII*

(*Final Fantasy*), *Torchwood*, *Warcraft*, *CC3 (Command Conqueror 3)*, and *PS2*, *PS3 (playstation)*.

As mentioned earlier, bloggers' preference for certain IT-related terms can tell us something about their lifestyles and their pastimes or hobbies. This kind of information also contributes to an individual's identity pool. Nevertheless, it cannot tell us much about how bloggers are actually constructing their identities linguistically. At most it is just a reflection of certain aspects of their identities. Unlike the new words or word-forms which bloggers have created by either identifying with the established word-formation processes or deviating from them, IT-related new terms are normally not invented (or created) by bloggers themselves. What they normally do is accept them and use them. They may invent new ways of addressing certain terms only when a community is formed and the community members find it necessary to do so. For instance, '*PS'* is an abbreviated form of the game called 'PlayStation'. The use of this abbreviation is only comprehensible to the general public when an adequate population has become familiar with this game either through the producer's advertising or through the word of mouth of the players of this game. It will be quite natural for players of this game to use *PS2* and *PS3* to refer to the newer versions. To a great extent, technical terms are not normally the ones which will undergo major linguistic engineering without causing difficulties in people's communication. Nevertheless, bloggers' preference for IT-related neologisms conveys information about their identities.

6.7 The use of slanguage

6.7.1 Defining slanguage

Trying to define the term ‘slang’ is no different from trying to catch a slimy fish with bare hands. As Crain (2008) remarks rightly and humorously, “like poetry and pornography, slang is easier to recognize than to define.” She also holds that slang is virtually infinite. Despite that there are many slang dictionaries available, what they can capture is just the tip of the iceberg of language users’ slang repertoire. The arrival of the Internet age provides another rich soil for this repertoire to grow. Of course, the sliminess has never stopped lexicographers and researchers from trying to define it, catch it, and analyze it. A good starting point for looking for a proper definition would be authoritative dictionaries. Therefore, let us first take a look at what lexicographers have to say about this slimy fish of slang.

The *Longman Dictionary of Contemporary English* (LDCE) defines slang as “very informal, sometimes offensive, language that is used especially by people who belong to a particular group, such as young people or criminals.” The key words are: *informal*, (possibly) *offensive*, and *group*. The *Oxford Advanced Learners’ Dictionary* defines slang as “very informal words and expressions that are more common in spoken language, especially used by a particular group of people, for example, children, criminals, soldiers, etc.” This definition is similar to the Longman version except for not mentioning the potential offensiveness of slang words. The *Collins COBUILD English Dictionary for Advanced Learners* emphasizes the social nature more by defining slang as “words, expressions, and meanings that are informal and are used by people who know each other very well or who have the same interests.” Lexicographers on the other side of the Atlantic Ocean tend to emphasize such qualities as novelty, playfulness, intentionality, and deviation from the standard variety in their definitions. According to *American Heritage Dictionary*, slang is “a kind of language occurring chiefly in casual and playful speech, made up typically of short-lived coinages and figures of speech that are deliberately used in place of standard terms for added raciness, humor, irreverence, or other effect.” Following more or less the same line, the *Merriam-Webster's Collegiate Dictionary* defines slang as “an informal nonstandard vocabulary composed typically of coinages, arbitrarily changed words, and extravagant, forced, or facetious figures of speech.”

Researchers investigating the use of slang also face the tricky problem of defining the term. In fact, definitions of slang abound in existing literature concerning slang studies. De Klerk (1990) presents a comprehensive overview of the existing definitions of slang prior to 1990 and finds that earlier definitions represented two opposing camps in terms of attitudes towards slang and its users: one negative and the other positive. More recent studies concerning slang take a more neutral stance. Of course, different researchers focus on different aspects of slang and its functions. Here are three relatively newer definitions of slang in existing literature, from which we can get a rough idea about why researchers find slang interesting and important as a linguistic phenomenon. Eble (1996) defines slang as “an ever changing set of colloquial words and phrases that the speakers use to *establish or reinforce social identity or cohesiveness within a group or with a trend or fashion in society at large*” (p. 11; my italics). What can be inferred from this definition is that slang can work as both an identity marker of a particular group and a collective identity marker of a larger community at a particular period of time. Allan and Burridge (2006), on the other hand, define slang from a more linguistic perspective, as can be seen from the following quote:

> Slang is language of *a highly colloquial* and contemporary type, considered *stylistically inferior to standard formal, and even polite informal, speech*. It often uses metaphor and/or ellipsis, and often manifests *verbal play* in which current language is employed in some *special sense and denotation*; otherwise the vocabulary, and sometimes the grammar, is *novel* or *only recently coined* (p. 69; my italics).

This definition focuses more on the strategies employed for slang creation and the subsequent stylistic effects which the use of slang is able to achieve. According to Grossman and Tucker (1997), "slang is a nonstandard vocabulary belonging to a particular culture or subculture. It consists of raw and unrefined expressions, many of which are considered taboo, vulgar, and derogatory" (p. 101). This definition reminds of another term "dirty words," which is closely related to slang yet much easier to recognize and define. Dirty words can be taken as an umbrella term for vocabulary used for swearing and verbal insults. Swearing includes religion-based profanity and blasphemy, as well as a wealth of obscenities which are characterized by language referring to sex, gender, sexuality, sexual behaviors, tabooed bodily functions and effluvia from the organs of sex, micturition and defecation. Verbal insults include epithets derived from tabooed bodily organs (e.g. *asshole),* bodily effluvia (e.g. *shit)* and sexual behaviors (e.g. *fucker, wanker),* epithets that typically pick on and debase a person's physical appearance, mental ability, character, behavior, beliefs and/or familial and social relations (for more details about these sources, please refer to Allan & Burridge, 2006, p. 79). Dirty words can be used to fulfill quite similar functions to those performed by slang words. In some situations the boundary between slang and dirty words is not very clear-cut. As a result, whether to include dirty words into the category of slang becomes an issue open to dispute. In fact, many dirty words are labeled "vulgar slang" in some dictionaries. Slang and dirty words share a number of features. First, both of them are highly colloquial. Because of this, they can both move the style of discourse towards the most informal end of the formal-informal continuum. Second, they both can be used as in-group solidarity markers. Third, both of them can be used to show disrespect for established social conventions. That said, slang and dirty words are not the same. As Crain (2008) points out, “dirty words suggest that the audience is no better than the speaker, and vice versa. Slang, on the other hand, usually suggests that speaker and audience share membership in a group.” Within the category of dirty words, there are functional differences among different subcategories. For instance, verbal insults are normally intended to wound the addressee or bring a third party into disrepute, or both (Allan & Burridge, 2006). In other words, verbal insults are more likely to be used for expressing strong emotions rather than for in-group solidarity building. Considering the similarities between slang and dirty words and the complexity in distinguishing one from another in certain circumstances, I decide to use the term “slanguage” as the superordinate term for both.

6.7.2 Identification of slanguage in the corpus

Although all the definitions of slang cited above have told us something about what slang looks like, none of them can work as an easy-to-operate working criterion for identifying slanguage words. The identification of slanguage words and expressions depends greatly on the native-speaker intuition and the context where a special sense of a particular word is used. As a non-native speaker of English, a more practical way would be to follow the lists of slang words adopted by other researchers while at the same time exploit the unknown word list generated by the Wmatrix system, with the help of dictionaries, of course. Ideally speaking, it is good to identify each and every slanguage word in the corpus and observe their distributions among different blogger groups. In practice, this is hardly possible even for a native English speaker. There are two major difficulties here. One is that slanguage words tend to be in-group markers which can only be understood by the group members unless these words have spread to a wider community and started to gain currency there. That is to say, one may well be familiar with the slang words used by one particular group but know nothing about those used by another group. The other difficulty is that many slanguage words wear a camouflage of ordinary word forms. It is the special semantic senses and the context where they are being used that can determine their slanguage status. As de Klerk (1990) points out, context can play a very important role in deciding whether a term is intended to shock, show disrespect for authority, be witty or humorous, show solidarity among insiders, or exclude outsiders. Without a prior knowledge of these special meanings, certain cases are very likely to be overlooked. For instance, the word *wasted* means *drunk* in its slang sense. I may not list it as a candidate for slanguage use due to the lack of prior knowledge of this special sense. The identification of dirty words is more straightforward and thus less difficult.

Two books are taken as the major sources of reference for identifying slanguage words in this study. One is *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings* written by Stenström, Anderson, and Hasund (2002); the other is *Forbidden Words: Taboo and the Censoring of Language* written by Allan and Burridge (2006). The former has devoted a whole chapter to discussing the London teenagers' use of slanguage words based on the Bergen Corpus of London Teenage Language (COLT). They adopt the term "slanguage" to cover a much wider variety of words than what the current researcher intended to include. The top lists of slanguage words are especially insightful. The latter, as its name suggests, is all about forbidden words which naturally cover slang and dirty words. There are many examples of very recent slanguage words which can be taken for reference. One thing worthy of particular mention here is the identification of newly emerged slanguage words, especially the ones which have often made their appearances in online discourses. For the confirmation of new slang words, an online slang dictionary (the Urbandictionary[11]) is consulted. If a certain word or word-form could not be found in ordinary dictionaries, it is consulted in Urbandictionary. According to Damaso and Cotter (2007), UrbanDictionary.com is an online dictionary of contemporary English slang usage created by Aaron Peckham in 2000. There are over a million definitions for over 400,000 unique headwords and the number of headwords and definitions is still increasing. This online dictionary may not have won official recognition of lexicographers but it has gained great currency among netizens over the past several years. As a complementary source of information, I find it very helpful in making sense out of new words or word-forms. Despite the various measures taken, it is still not possible for me to make an exhaustive list of all the slanguage words used in the EBC. What is going to be presented below is at best an approximation to bloggers' actual use of slanguage.

6.7.3 Distribution of slanguage in the corpus

Following the methods described in the previous section, I have identified a total number of 5,009 slanguage tokens from the blog corpus. These tokens represent 207 word types (or lexemes). As mentioned earlier, slanguage in this research is an umbrella term for slang and dirty words. (In fact, dirty words may not be a good term as it cannot really reflect the real functions of these words. It is just a term of convenience.) The 207 slanguage word types can be roughly classified into five categories. The first category (General Slang) includes highly colloquial words which are used as substitutions for ordinary words, e.g. *ace* (for *excellent, wonderful* or *getting A grade*), *awesome* (for *great* or *excellent*), and *cool*. The second category (SMD Terms) includes words which are related to (or have their origin in) sexual activities, sexuality, SMD organs (organs for sex,

[11] http://www.urbandictionary.com/

micturition, and defecation), and bodily effluvia. The third category (Insulting Terms) is related to offensive terms which are intended to debase other people’s appearance, personality, belief, and especially their intelligence. The fourth category (Drug Terms) refers to slang about alcohol or drug abusing. The fifth category (Profane Terms) includes words which are related to disrespect for religion or talking about religious taboos. For this research, General Slang and Drug Terms belong to slang and the rest fall into the category of dirty words. Table 6.11 lists the details of their distributions. In terms of word types, slang (including General Slang and Drug Terms) and dirty words are more or less the same. In terms of token numbers, dirty words outnumber slang by a wide margin, with the former accounting for 67.7% and the latter 32.3%. I will try to explain why dirty words are more commonly used than general slang later in this section.

Table 6.11 Slanguage category and distribution

Category	Type No.	%_Type	Token No.	%_Token
General Slang	87	42%	1,558	31.1%
SMD Terms	71	34.3%	2,732	54.5%
Insulting Terms	34	16.4%	172	3.4%
Drug Terms	13	6.3%	59	1.2%
Profane Terms	2	1%	488	9.7%
Total	207	100%	5,009	100%

The British bloggers and American bloggers included in this research have displayed very similar overall patterns in the use of slang and dirty words, as can be observed from Table 6.12 below. The slang word types identified from the British blog entries account for 45.3% of the total but their total number of occurrences only takes up 34.8% of the total tokens. The dirty words account for 54.7% of the total types and 65.2% of the total tokens. For American bloggers, their slang words account for 46.3% of the total types but only 29.9% of the tokens. The dirty words take up 53.7% of the types and around 70% of the tokens.

Table 6.12 Distribution of slanguage by blogger region

	British Bloggers			American Bloggers		
Category	Type	Token	%_TK	Type	Token	%_TK
General Slang	60	755	33.9%	51	793	28.5%
SMD Terms	53	1147	51.4%	50	1585	57%
Insulting Terms	26	89	4%	21	93	3.3%
Drug Terms	7	20	0.9%	12	39	1.4%
Profane Terms	2	219	9.8%	2	269	9.7%
Total	148	2230	100%	136	2779	100%

From Table 6.12, we cannot observe the difference in preference for slanguage categories identified by Crain (2008) through the comparison of British and American slang dictionaries. According to Crain, slang can reflect the collective identity of a speech community. She finds that "the American id, viewed through the lens of slang, dwells much on human worthlessness, failure, drug addiction, homosexuality, oral sex, penises and breasts" whereas the collective id of the Commonwealth nations dwells on the fact that they value intoxication, foolishness, money and cheating. Nevertheless, the idea that slanguage as an identity marker seems to make much sense, as we will see later.

Table 6.13 Top 15 dirty words and general slang

Dirty Words		General Slang	
fuck*	892	guy	454
shit*	407	cool	240
hell	247	awesome	192
damn	241	uni	70
suck*	220	man	50
crap*	186	dude	47
piss*	176	chill	43
ass*	172	gig	34
bitch*	138	rock	33
freak	109	emo	32
bloody	93	geek	24
screw	57	quid	23
bastard	56	bloke	22
asshole	31	buck	18
bullshit	31	fit	18
Subtotal	3,056(61%)	Subtotal	1,300 (26%)

If we do a frequency count on the slanguage words identified from the blog corpus, we will soon find that the occurrences flock around a rather limited number of lexemes. Table 6.13 (see above) lists the top 15 slang words and the top 15 dirty words in the EBC. These thirty lexemes account for 87% of the whole slanguage word tokens identified from the corpus, with dirty words accounting for 61% and general slang adding up to 26%. A general list like Table 6.13 is only helpful in obtaining a rough picture of which slang terms or dirty words are more commonly used in blogging. In order to understand why these dirty words and slang are able to sneak into bloggers' entries, we need to take a closer look at how and by whom they have been used.

As it is generally held in existing studies, age and gender are more often associated with slanguage use. To what extent this is also the case in blogging is yet to be seen. As can be observed from Table 6.14, the top 15 slanguage word list for each of the six age groups looks quite similar. If we compare the mid-teens list with the list of the rest age groups one by one, we will find that the difference varies from a minimum of one to the maximum of three (words which are different from the mid-teens group are italicized on the table). The only conclusion we can make based on this comparison would be that these words are the most likely candidates if a blogger chooses to use slanguage in his or her entries. If we take a closer look at the relative frequencies for each word in different age groups, we will see some insightful differences. The mid-teens group outperforms all the rest age groups in 13 out of the 15 words on the list. The only two exceptions are *cool* and *hell*. In fact, the relative frequency for *cool* is the second highest among all the groups and that for *hell* is the third highest. In other words, even though bloggers from different ages share a very similar inventory of commonly used slanguage words, younger age groups, especially the mid-teens, seem to display a greater slanguage density than the older ones. Aside from that, we can also see that bloggers from younger age groups seem to prefer certain slanguage words, for instance, *awesome* and *suck*.

Table 6.14 Top 15 slanguage words across age groups

15-17		18-19		20-24		25-29		30-34		35-40	
Item	/10K	Item	/10K	Item	/10K	Item	/10K	Item	/10K	Item	/10K
fuck	19.3	fuck	12.6	fuck	13.5	fuck	15.8	fuck	10.6	fuck	6.2
shit	10.5	guy	7.4	guy	6.8	guy	6.5	guy	4.7	guy	4.7
guy	8.0	shit	7.1	shit	4.5	shit	5.1	shit	4.2	damn	3.6
suck	5.7	cool	5.6	hell	4.1	damn	4.5	hell	3.3	shit	3.1
awesome	4.9	suck	4.5	awesome	4.1	cool	3.9	piss	3.1	crap	2.8
cool	4.5	hell	4.3	cool	3.4	hell	3.5	crap	2.8	hell	2.7
freak	4.5	awesome	3.9	damn	3.3	awesome	2.5	cool	2.7	ass	2.3
crap	4.1	*uni*	3.8	suck	3.0	ass	2.3	damn	2.6	suck	1.6
damn	3.8	damn	3.3	ass	2.4	suck	2.3	suck	2.5	piss	1.5
piss	3.7	piss	2.5	piss	2.3	piss	2.3	ass	1.8	cool	1.3
hell	3.5	bitch	2.2	bitch	1.8	bitch	1.8	*bloody*	1.7	awesome	1.2
bitch	3.1	freak	1.8	crap	1.4	crap	1.7	*bastard*	1.6	*bloody*	1.2
ass	2.4	crap	1.5	*bloody*	1.4	dude	1.6	bitch	1.3	bitch	1.0
dude	2.0	*bloody*	1.4	*uni*	1.1	freak	1.5	freak	1.3	*screw*	0.9
man	2.0	ass	1.2	freak	1.0	*bastard*	1.3	*screw*	1.2	freak	0.7

As mentioned earlier, slanguage use is also associated with gender in existing literature. In order to explore whether slanguage words have been employed as a marker of gendered identity, I present the top 20 slanguage words from male and female bloggers respectively in Table 6.15. Again, these two lists do not differ much in terms of the words included. The words *emo* and *dude* on the list of female bloggers failed to make it onto the top 20 list for male bloggers whereas the words *chill* and *gig* did not appear on the top 20 list for the females. In terms of relative frequency (tokens per ten thousand words), female bloggers use the following words more frequently than their male counterparts: *shit*, *hell*, *crap*, *freak*, *dude*, and *emo*. The male bloggers use words such as *cool*, *awesome*, *suck*, *damn*, and *chill* more frequently.

Table 6.15 Top 20 slanguage words and gender (total)

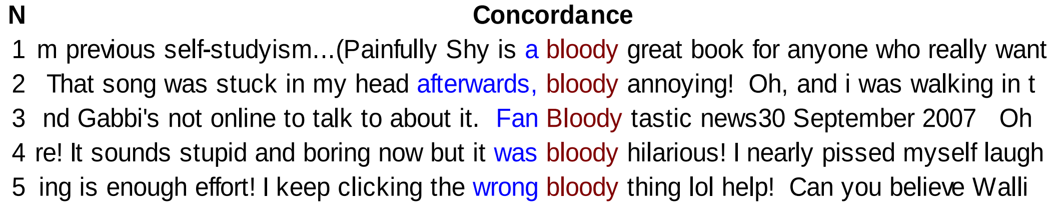
Female Bloggers			Male Bloggers		
Item	Tokens	Per_10k	Item	Tokens	Per_10k
fuck	449	12.1	fuck	416	13.1
shit	218	5.9	guy	240	7.6
guy	214	5.8	shit	151	4.8
hell	153	4.1	cool	138	4.4
damn	119	3.2	damn	122	3.9
suck	105	2.8	suck	107	3.4
cool	102	2.7	awesome	96	3.0
crap	100	2.7	hell	94	3.0
awesome	96	2.6	piss	85	2.7
piss	85	2.3	ass	63	2.0
ass	78	2.1	crap	54	1.7
freak	77	2.1	bitch	49	1.5
bitch	74	2.0	bloody	46	1.5
bloody	47	1.3	bastard	32	1.0
uni	40	1.1	freak	32	1.0
screw	32	0.9	uni	30	0.9
dude	28	0.8	man	27	0.9
bastard	24	0.6	chill	25	0.8
emo	24	0.6	screw	25	0.8
man	23	0.6	gig	23	0.7

Although we can obtain a rough picture of the potential difference between males and females in their used of slanguage words from the top 20 lists presented in Table 6.15, we are not able to observe the more salient differences which may exist in different age and gender groups. As an attempt to look for these potential differences, I list the top ten slanguage words used by all age and gender groups respectively in the following two tables (Tables 6.16 & 6.17), with the data for British bloggers and American bloggers in two separate tables. For easy reference, all the word lists are sorted according to the alphabetic order of the words included.

Table 6.16 Top 10 slanguage words across age and gender groups (UK)

Female Bloggers											
15-17		18-19		20-24		25-29		30-34		35-40	
Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k
bloody	4.1	awesome	3.3	awesome	3.0	*bastard*	2.0	*arse*	2.2	bloody	2.6
cool	5.2	cool	2.4	bloody	3.0	*bitch*	2.0	bastard	2.2	*bugger*	1.6
crap	5.2	*crap*	3.0	*crap*	3.0	*bloke*	1.4	bloody	2.5	*cool*	1.9
damn	4.8	damn	2.1	*damn*	3.9	bloody	1.4	crap	2.9	crap	3.8
emo	4.4	fuck	5.9	fuck	9.4	cool	2.7	damn	1.8	damn	3.5
fuck	20.0	guy	6.2	guy	3.0	damn	4.1	fuck	5.4	fuck	3.5
guy	3.7	hell	3.9	hell	3.9	fuck	24.2	*hell*	4.7	guy	2.9
hell	4.1	*shit*	4.7	piss	2.7	guy	4.8	piss	2.5	hell	3.5
piss	3.7	suck	3.3	shit	3.6	*hell*	4.4	shit	2.9	*piss*	1.6
shit	6.7	uni	7.7	uni	2.4	*shit*	4.1	*suck*	2.2	*shit*	2.9
Male Bloggers											
Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k
n/a	n/a	awesome	4.2	awesome	3.4	bloody	3.1	bastard	3.7	*arse*	1.2
n/a	n/a	*bloody*	3.3	bloody	2.2	*bugger*	2.3	*bloke*	2.0	bloody	1.6
n/a	n/a	cool	7.6	*cool*	2.8	*chill*	2.0	bloody	4.4	crap	2.4
n/a	n/a	damn	3.3	fuck	19.0	cool	5.9	crap	2.4	damn	3.1
n/a	n/a	fuck	11.5	guy	5.9	*crap*	4.7	damn	2.0	*fab*	1.2
n/a	n/a	guy	14.2	hell	5.0	damn	4.3	fuck	16.3	fuck	3.1
n/a	n/a	hell	3.6	piss	3.9	fuck	10.5	*geek*	2.4	*geeky*	1.6
n/a	n/a	*piss*	3.9	shit	4.5	*gig*	2.7	guy	4.1	*gig*	1.6
n/a	n/a	suck	3.6	*suck*	2.0	guy	3.9	piss	3.7	guy	4.7
n/a	n/a	uni	5.1	uni	2.2	*piss*	3.5	shit	4.8	hell	1.6

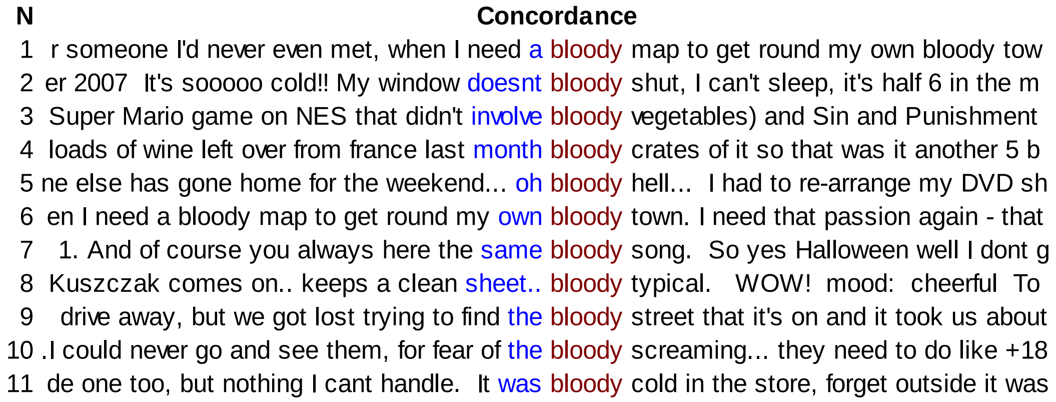
If we compare the list for male bloggers and that for the female bloggers within the same age group, we will find some subtle differences. The mid-teens group will be skipped as there is no male blogger data for comparison. Compared with the rest four age groups, males and females in the late-teens group share the greatest number of words in their top ten slanguage lists: only four words are different. The words *crap* and *shit* did not make it to the male top list. In fact, there are only two occurrences of the word *crap* in the male blog entries for this age group. *Shit* ranks the eleventh on the male list, so to a certain extent it is just a matter of token difference for this word. The two words which make it to the male list but are absent from the female list due to slightly fewer tokens are *bloody* and *piss*. Both words are related to emotion expression. If we take a closer look at how these two words are used by male and female bloggers, we will find that the usage of these two words may not be the same, especially the word *bloody*. The following concordance lines (Figure 6.11) show how female late-teens bloggers use the word *bloody*.



N	Concordance
1	m previous self-studyism...(Painfully Shy is a bloody great book for anyone who really want
2	That song was stuck in my head afterwards, bloody annoying! Oh, and i was walking in t
3	nd Gabbi's not online to talk to about it. Fan Bloody tastic news30 September 2007 Oh
4	re! It sounds stupid and boring now but it was bloody hilarious! I nearly pissed myself laugh
5	ing is enough effort! I keep clicking the wrong bloody thing lol help! Can you believe Walli

Figure 6.11Concordance lines for BLOODY(1)

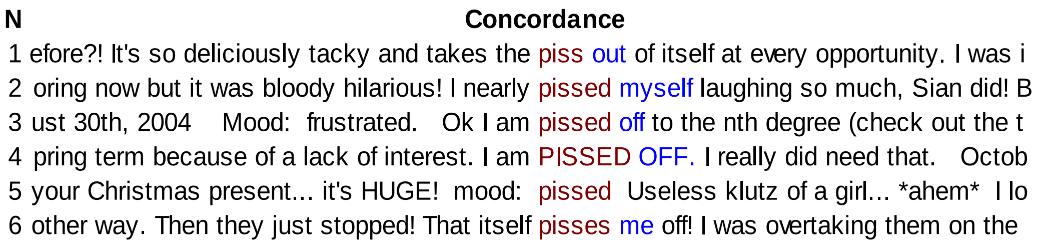
In only one out of the five cases (line 2), *bloody* is used as an intensifier to modify a term with a negative sense. There are three cases where the word is used as an intensifier (or infix) to modify words with positive senses. For the one on line 5, it is used more in a joking manner rather than having any negative connotation. The male bloggers, on the other hand, use this word in a quite different manner, as the following concordance lines (see Figure 6.12) can show.



N	Concordance
1	r someone I'd never even met, when I need a bloody map to get round my own bloody tow
2	er 2007 It's sooooo cold!! My window doesnt bloody shut, I can't sleep, it's half 6 in the m
3	Super Mario game on NES that didn't involve bloody vegetables) and Sin and Punishment
4	loads of wine left over from france last month bloody crates of it so that was it another 5 b
5	ne else has gone home for the weekend... oh bloody hell... I had to re-arrange my DVD sh
6	en I need a bloody map to get round my own bloody town. I need that passion again - that
7	1. And of course you always here the same bloody song. So yes Halloween well I dont g
8	Kuszczak comes on.. keeps a clean sheet.. bloody typical. WOW! mood: cheerful To
9	drive away, but we got lost trying to find the bloody street that it's on and it took us about
10	.I could never go and see them, for fear of the bloody screaming... they need to do like +18
11	de one too, but nothing I cant handle. It was bloody cold in the store, forget outside it was

Figure 6.12 Concordance lines for BLOODY (2)

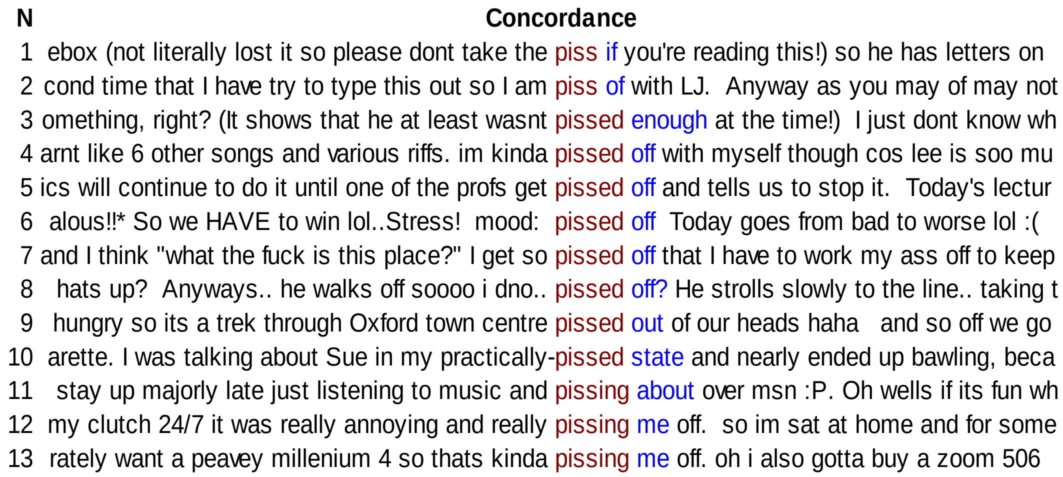
From Figure 6.12, we can see that male bloggers of this age group tend to use *bloody* to express annoyance, though there are two cases (lines 8 & 11) where it is also used as an intensifier. This practice of tending to use the same word for different senses does not show in the case of *piss*. Both male and female bloggers use it to express the feeling of anger or annoyance. Figure 6.13 shows how female bloggers use it.



```
N                                   Concordance
1 efore?! It's so deliciously tacky and takes the piss out of itself at every opportunity. I was i
2  oring now but it was bloody hilarious! I nearly pissed myself laughing so much, Sian did! B
3  ust 30th, 2004    Mood:  frustrated.   Ok I am pissed off to the nth degree (check out the t
4  pring term because of a lack of interest. I am PISSED OFF. I really did need that.   Octob
5  your Christmas present... it's HUGE!  mood:  pissed  Useless klutz of a girl... *ahem*  I lo
6  other way. Then they just stopped! That itself pisses me off! I was overtaking them on the
```

Figure 6.13 Concordance lines for PISS (1)

The male bloggers tend to use the same word in a slightly wider context yet with more or less the same semantic sense of expressing anger or annoyance as the following concordance lines show (see Figure 6.14).



```
N                                   Concordance
 1  ebox (not literally lost it so please dont take the piss if you're reading this!) so he has letters on
 2  cond time that I have try to type this out so I am piss of with LJ.  Anyway as you may of may not
 3  omething, right? (It shows that he at least wasnt pissed enough at the time!)  I just dont know wh
 4  arnt like 6 other songs and various riffs. im kinda pissed off with myself though cos lee is soo mu
 5  ics will continue to do it until one of the profs get pissed off and tells us to stop it.  Today's lectur
 6  alous!!* So we HAVE to win lol..Stress!  mood:  pissed off  Today goes from bad to worse lol :(
 7  and I think "what the fuck is this place?" I get so pissed off that I have to work my ass off to keep
 8   hats up?  Anyways.. he walks off soooo i dno.. pissed off? He strolls slowly to the line.. taking t
 9   hungry so its a trek through Oxford town centre pissed out of our heads haha   and so off we go
10  arette. I was talking about Sue in my practically-pissed state and nearly ended up bawling, beca
11   stay up majorly late just listening to music and pissing about over msn :P. Oh wells if its fun wh
12  my clutch 24/7 it was really annoying and really pissing me off.  so im sat at home and for some
13  rately want a peavey millenium 4 so thats kinda pissing me off. oh i also gotta buy a zoom 506
```

Figure 6.14 Concordance lines for PISS (2)

So, for the word *piss*, the difference between male and female bloggers within the late-teens group is basically that of token numbers. Speaking of relative token frequencies, the male groups use the word *fuck* with greater density than their female counterparts.

For the 20-24 blogger group, four words are not shared by both lists. They are *crap* and *damn* from the female list and *cool* and *suck* from the male list. The 25-29 group has the greatest difference in word types between the male and the female list. Only half of the words on the two lists overlap. The five words which only appeared on the male top list include: *bugger*, *chill*, *crap*, *gig*, and *piss*. The five words which only appeared on the female top list are: *bastard*, *bitch*, *bloke, hell*, and *shit*. Another rather striking feature of the female bloggers within this group is the high relative frequency of the word *fuck*, which is actually the highest among all 23 blogger groups. This is a bit unusual as the use of *fuck* is often associated with male speakers and authors. According to McEnery and Xiao (2003), the use of *fuck* is "a marker of male readership/authorship as it is a marker of male speakers" (p. 511). A consultation of the original blog entries shows that this unusual high frequency has something to do with the intensive use of this word in one particular entry when the blogger is complaining about the Bush Administration and the Iraq War. There are 23 occurrences of the word *fuck* and its variants. The frequent occurrences of the word *fuck* (and its variants) have brought the anger and the anti-war attitude of the blogger vividly onto the screen. Even if we exclude this blogger's use of the word from the calculation, the relative frequency of the word *fuck* is still higher than that of their male counterparts.

For the 30-34 group, there are five words which are not shared by the two lists. They are *arse, hell*, *suck* from the female list and *bloke* and *geek* from the male list. A more noteworthy feature for this group is that male bloggers' use of the word *fuck* is three times that of the female bloggers. For the oldest group, the one aged from 35 to 40, the overlapping of the two lists is the second smallest. Eight words only appeared in one list. They are *bugger, cool*, *piss*, and *shit* from the female list and *arse*, *fab, geeky*, and *gig* from the male list. Female bloggers in this group use more slanguage words and dirty words in particular.

The American bloggers, on the other hand, do not show great difference in top slanguage lists between males and females from the same age groups. The number of different words between gender with the same age group has been maintained at three for four out of the total six groups, exclusive of the youngest (the mid-teens) and the oldest groups (the 35-40 group). The mid-teens group displays the greatest difference in terms of the top ten slanguage words between male and female bloggers: seven out of the twenty words are not shared by both lists. Four of them come from female bloggers. They are: *bitch*, *crap*, *freak*, and *hell*. The word *freak* (and its variants) is the third most frequently used slang on the top list of female mid-teens American bloggers. In fact, this group is also the one which has the highest relative frequency for the word *freak*, almost three times that of the late-teens female bloggers and more than twice the number of female bloggers from the 30-34 group. To a great extent, the word *freak* is a marker of Americanism, female, and youth. There are 23 occurrences of this word in the entries from British bloggers and this may well be an indicator of the influence of American English on the British English and a symbol of younger British bloggers' intention or efforts in identifying with American identity. The three words which only appeared on the male top list are: *ass*, *dude*, and *man*. Among them, *dude* and *man* are typically used by adult males as informal terms of address in informal conversation. This might be an indicator that the potential audience of mid-teens male bloggers is male. The biggest similarity between the male and female bloggers within the mid-teens group is the high frequency of such words as *fuck*, *shit*, and *suck*. As a matter of fact, the mid-teens group hosts the top two highest relative frequencies for the word *fuck* among all the 12 American blogger groups, with the male bloggers ranking the first and the female bloggers ranking the second. For the word *suck*, it is the female mid-teens bloggers who rank the first with their male counterparts ranking the second. Again, this high relative frequency ranks the top among all 23 blogger groups (British and American bloggers put together). For the word *shit*, the female mid-teens rank the top whereas their male counterparts are the fifth among all the 12 American blogger groups. In fact, the female mid-teens' use of the word *shit* is the highest among all the 23 blogger groups in this research.

Table 6.17 Top 10 slanguage words across age and gender groups (US)

Female Bloggers											
15-17		18-19		20-24		25-29		30-34		35-40	
Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k
awesome	4.5	awesome	2.7	ass	3.7	ass	3.4	cool	3.5	ass	4.9
bitch	4.5	*bitch*	4.4	awesome	5.0	awesome	5.1	*crap*	4.1	crap	3.0
cool	4.5	cool	5.3	*buck*	2.4	damn	3.1	damn	3.2	damn	4.0
crap	4.9	damn	3.1	cool	2.9	*dude*	4.2	*freak*	3.8	fuck	13.0
freak	8.9	*freak*	2.7	damn	2.1	fuck	13.5	fuck	11.2	guy	7.3
fuck	17.4	fuck	16.0	fuck	9.8	guy	7.3	guy	4.8	hell	3.8
guy	8.5	guy	7.1	guy	10.9	hell	4.5	hell	4.1	*piss*	2.4
hell	4.1	hell	7.1	hell	2.7	*piss*	3.7	piss	3.5	*screw*	2.4
shit	14.6	shit	11.5	shit	2.9	shit	9.0	shit	5.4	shit	5.7
suck	8.5	suck	5.8	suck	3.2	suck	3.1	suck	2.9	*suck*	3.0
Male Bloggers											
Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k	Item	/10k
ass	4.1	*ass*	2.9	ass	4.1	ass	3.1	*ass*	4.2	ass	2.2
awesome	8.7	awesome	5.4	awesome	4.7	awesome	2.1	*awesome*	1.6	*asshole*	1.1
cool	3.7	cool	7.9	*bitch*	3.8	*bitch*	2.4	cool	4.2	*awesome*	2.2
damn	3.7	damn	5.0	cool	6.7	*cool*	4.8	damn	3.2	*cool*	1.8
dude	3.7	fuck	20.4	damn	5.8	damn	6.9	fuck	9.4	crap	1.8
fuck	20.7	guy	10.4	fuck	15.8	fuck	14.4	guy	8.1	damn	3.7
guy	12.9	hell	3.3	guy	6.7	guy	9.6	hell	2.6	fuck	2.9
man	4.1	*piss*	3.3	hell	5.0	hell	3.4	piss	2.6	guy	3.3
shit	10.6	shit	12.1	shit	7.0	shit	4.5	shit	3.5	hell	1.5
suck	8.3	suck	6.3	suck	5.6	suck	4.1	suck	4.2	shit	2.2

The top slanguage lists for male and females in the late-teens group display a smaller difference both in terms of the words included and their relative frequencies. The male bloggers use the three key dirty words *fuck*, *shit*, and *suck* more often than their female counterparts whereas the female bloggers use *bitch*, *hell*, and *freak* more frequently. This finding seems to fit very well with the general statement of existing literature regarding gender differences in slanguage use. A similar pattern is found within the young adult group (the 20-24 group). Again, male bloggers use words like *fuck*, *shit*, and *suck* more often than their female counterparts. For the 25-29 group, the difference becomes less obvious in the use of *fuck* and *suck* whereas the female bloggers' use of the word *shit* doubles that of the male bloggers. In terms of proper slang words, the female bloggers in this group seem to prefer *awesome* while their male counterparts prefer *cool*. For the 30-34 group, the picture looks a bit different from that of the younger groups. Females surpass the males in their use of dirty words such as *fuck*, *shit*, *crap*, *piss*, and *freak*. It is also different from their British counterparts. This does not seem to consonant with the dominant impression in existing literature that males tend to use more slanguage words than females. The gender difference seems to be even greater in the oldest group of 35-40. The female bloggers outnumber the male ones in almost each and every word on the list. The American female bloggers display a very similar tendency to the British female bloggers of the same age group: female bloggers tend to use more dirty words than males.

6.8 Chapter summary

From what has been presented in this chapter, we can see that compounding and derivation are two major means for bloggers to create new lexical items. From the internal structural patterns of the new compounds created by bloggers and the strategies used in forming new derivatives, we can see that bloggers are mostly following the conventional rules of word formation but they seldom bypass the opportunity of creatively exploiting the linguistic rules for realizing special communication effects. A good example would be the presence of the not-so-conventional phrasal compounds. Bloggers' use of minor word-formation strategies such as blending, clipping, verbalizing initials and acronyms, and leetspeak is another piece of evidence to demonstrate their good sense of creativity in language use and their readiness to experiment with unconventional ways of saying things. Bloggers' intention of trying to be non-conventional is also evidenced by their use of neologisms related to IT and Internet culture and their use of slanguage words. The lexicological variation presented in this chapter offers another window for observing bloggers' identity representation, as will be seen in Chapter 9.

Chapter 7 Variation in Semantic Domains

This chapter focuses on describing whether and to what extent bloggers from different age and gender groups differ in their preferences for semantic domains.

7.1 Introduction

The previous chapters have offered some interesting insights about how bloggers are trying to represent themselves linguistically by deviating from the orthographic norms, exploiting the word-formation processes, and taking advantage of slanguage words and expressions. In fact, we can also observe how bloggers are representing themselves from what they write about in their blog entries. As one of the major functions of blogging is to record bloggers' daily life experiences, different bloggers may choose to disclose different details. Even if people happen to share similar experiences, they may not necessarily feel the same about these experiences. It is sensible to believe that there is a link between what the bloggers write about and some aspects of their identities. If we can find a way to compare the contents of bloggers from different groups, we may be able to identify that link.

This is where Wmatrix comes to play a very important role. As mentioned in Chapter 3, apart from being able to add part-of-speech (POS) tags to linguistic texts, Wmatrix can also perform semantic annotations and comparisons between different sets of data. The system is able to generate a report which contains lists of semantic tags overused by one dataset against the dataset the client designated as reference dataset (or reference corpus). From this list of overused semantic tags, we can get a rough idea about what content has been more frequently mentioned in a particular dataset and from there we expect to obtain some insights about certain aspects of bloggers' identities. We must admit that this job is not easy as it looks. There are at least two problems. One is that semantic tagging itself is difficult due to the fact that words are used in context and no computer software is intelligent enough to be able to tell exactly which semantic domain a particular word should fall into, considering that English words are notoriously polysemous. The other is that online discourse is also notoriously non-conventional, which will inevitably affect the accuracy rate of the semantic tagging results. The less pessimistic side of the story is that the semantic tagger of Wmatrix tends to be rather consistent in making judgments (wrong judgments inclusive). As for the unconventionality issue, we should not exaggerate its negative effects, either. The reasons are twofold: first, compared with other online discourse data, blogging texts are closer to conventional written texts; second, as mentioned in Chapter 5, there are around 16,587 unknown words, only accounting for about 2.4% of the total number of words in the whole corpus. As the Wmatrix system generates lists of words which are found to be overused or underused as against the reference corpus (data) designated by the client, I can always check the lists for tagging errors and decide whether the errors are likely to lead to distorted interpretations. In fact, Rayson (2008a) has warned users of Wmatrix about the possible tagging errors and urged them to take care in interpreting the results. He also asks users to be aware of the fact that "the sense distinctions marked by USAS are coarse-grained and may not match those required in specific studies" (p. 529). Despite the potential uncertainties, it is still worth trying. In fact, Ooi, Tan, and Chiang (2007) have explored using Wmatrix for blog content analysis and they find this method insightful. The following section presents the distribution of preferred semantic domains among different blogger groups.

As there are 23 different blogger groups in this research, theoretically speaking there are more than 250 combinations of group comparisons. Obviously this is not what I can handle within a constrained time frame. In order to avoid getting lost in trivial comparisons, I decide to focus on two major variables: age group and gender. In other words, the regional factor will not be considered, as it may be less relevant to the semantic domains which bloggers write about. Even after the scaling down of the scope, there are still 28 pairs of comparisons: one overall comparison between male and female bloggers, 15 inter-group comparisons (between age groups), and 12 intra-group comparisons (between males and females within the same age groups). As Wmatrix is able to generate result lists which contain both overused and underused semantic domains at one go, that has saved half of the time and trouble. Otherwise, it will take 56 rounds of comparisons to get the same results.

One thing worthy of particular mention here is the identification of preferred semantic domains for each age group. As there are six age groups altogether, for each age group there will be five sets of overused semantic domains relative to respective age groups. By pooling all the overused semantic domains of one particular group together, we can see how many times each overused domain has appeared. Based on the number of appearance, each of these categories is assigned a degree of prominence value. The minimum value would be one and the maximum five. The greater the value, the more prominent the domain is. If a domain's degree of prominence is greater than two, it will be taken as one of the preferred semantic domains of that age group. An intra-group comparison will be conducted between the data of male and female bloggers for potential gender differences within that age group.

7.2 Males and females overall

With the help of the file merging function of Wmatrix, I regrouped the EBC data into two datasets according to gender, without considering their age groups. After that, a comparison between the male and female datasets was conducted with both datasets as mutual reference datasets. Table 7.1 lists the top 20 overused categories for both gender groups.

Table 7.1 Top 20 preferred semantic domains across gender

Female Bloggers			Male Bloggers		
Semtag	**Semantic domain**	**LL**	**Semtag**	**Semantic domain**	**LL**
Z8	Pronouns	170.8	K2	Music and related activities	95.86
S4	Kin	60.07	O2	Objects generally	65.16
Z6	Negative	50.61	I3.1	Work and employment: Generally	43.34
E4.1-	Sad	43.88	K5.1	Sports	41.56
B1	Anatomy and physiology	38.46	Z5	Grammatical bin	35.79
B2-	Disease	36.16	Y2	Information technology and computing	28.01
S2.1	People: Female	32.59	N5	Quantities	27.71
B5	Clothes and personal belongings	29.72	T1.3	Time: Period	24.43
S1.2.6-	Foolish	28.91	Q1.2	Paper documents and writing	22.87
T3-	Time: New and young (babies)	28.17	K5	Sports and games generally	22.25
P1	Education in general	25.98	O3	Electricity and electrical equipment	20.78
E2+	Like	25.53	Q4.3	The Media: TV, Radio and Cinema	20.36
B3	Medicines and medical treatment	23.76	N5-	Quantities: little	20.36
Q2.2	Speech acts	18.33	G3	Warfare, defense and the army; weapons	20.31
E1	Emotional Actions, States And Processes General	16.84	S7.1+	In power	19.15
A5.3-	Evaluation: Inaccurate	16.8	M1	Moving, coming and going	19.07
O4.2-	Judgment of appearance: Ugly	16.04	S1.1.3+	Participating	16.38
N3.6-	Measurement: Area (weight)	16.01	A12+	Easy	15.12
X2.1	Thought, belief	15	T1.1.3	Time: Future	13.75
Z99	Unmatched	12.65	A5.1+	Evaluation: Good	12.91

From this table we can see some interesting differences between male bloggers and female ones. Female bloggers write more about people (revealed by Z8, S2.1, S4, and T3-), body (revealed by B1 and N3.6-), sickness (revealed by B2- and B3), clothes and personal belongings (B5), emotions (revealed by E1, E2+, and E4.1-), education in general (P1), and evaluation of people and things (revealed by S1.2.6-, and O4.2-). They also mention more about oral communication with other people, as can be revealed by the overuse of Q2.2. They write more about their thoughts, their belief, and feelings than male bloggers (X2.1). There are three major tagging errors on the female bloggers' overuse list: N3.6-, O4.2-, and T3-. The words categorized by the system as N3.6- are actually words talking about weight control, not about area measurement. T3- does not really refer to time, but rather words related to little babies. O4.2- is only half-correct as many words are really related to judgment of appearance but it has also included other words related to judgment about personal traits. The male bloggers, on the other hand, talk more about entertainment (music, games, TV) and sports (revealed by K2, K5, K5.1, and Q4.3), work and employment (I3.1), general objects (O2), electronic gadgets (O3), and computers and the Internet (Y2). They mention more about moving around (M1) and their participation in social activities (S1.1.3+). Unlike the female bloggers who seem to 'talk' or 'chat' more with other people, male bloggers appear to be involved in written communication more (Q1.2). They also write more about power relations, especially in workplace (S7.1+). In addition, their language involves more grammatical words such as articles and prepositions (Z5).

7.3 The 15-17 age group

However impressionistic the observation we can obtain from Table 7.1 may appear to be, it encourages me to take a closer look at the potential differences across age groups and the potential gender differences within each age group. Table 7.2 lists the typical overused semantic domains for the mid-teens group (i.e. the 15-17 group).

Table 7.2 Preferred semantic domains for the 15-17 group

Semtag	Semantic domain	Degree of Prominence
A5.1-	Evaluation: Bad (worse)	5
E1	Emotional Actions, States And Processes General	5
Z4	Discourse Bin	5
Z99	Unmatched	5
E4.1-	Sad	4
N1	Numbers	4
P1	Education in general	4
Q2.1	Speech: Communicative	4
S3.2	Relationship: Intimacy and sex	4
Z1	Personal names	4
Z8	Pronouns	4
A14	Exclusivizers/particularizers	3
A5.4-	Evaluation: Unauthentic	3
B1	Anatomy and physiology	3
E2+	Like	3
E4.1+	Happy	3
K2	Music and related activities	3
K5.1	Sports	3
K5.2	Games	3
L1-	Dead	3
S9	Religion and the supernatural	3
Y1	Science and technology in general	3

Compared with the lists generated from the datasets of other blogger groups, the mid-teens list is the longest, suggesting that this group is more different from the rest age groups. The mid-teens write a great deal about their education (P1 and Y1), as school life is an important part of young people at this age period. They are more concerned about body (B1), people (Z8 and Z1) and relationships (S3.2). Music (K2), sports (K5.1), and games (K5.2) are also important part of their daily life. Feelings and emotions (E1), be it sadness (E4.1-) or happiness (E4.1+), are also important topics for them. They seem to be frequently involved in oral communication (Q2.1). They tend to use plenty of interjections or colloquial discourse markers in their blog writing (Z4) and their language tends to be more unconventional (Z99). They are not really eager to die or they are interested in talking about death as the overuse of L1- seems to be suggesting. In fact, they are just being humorous or exaggerating when they use words such as *die* and *kill* and their inflectional forms. Figures 7.1 and 7.2 show the concordance lines for *die* and *kill* (and its inflectional forms) generated from the dataset of the mid-teens bloggers respectively.

N	Concordance
1	that :[pfft. lol seriously. and then i want to die 4 years later ahah nah im jk.. but i DO
2	ttle monkey that's all 'Teehee, you're going to die a horrible death Zack~' and I'm like 'ARG
3	ase you were having trouble sleeping, I didn't die from the flu. I am alright now! Oh right, I
4	od is all-powerful, then why did Jesus have to die on the Cross for God to forgive our sins?
5	hare and be healthy and stay young until we die since there would be no stress. What a
6	t leave me the frick alone. I semi-decided to die today. It was funny. OHHHH. Last night
7	oo late, I will know about it >=D And you will die you know. Ugh... Eyes... getting sleepy.
8	stupid. seriously. my cheeks hurt i hope i die. Feb. 23rd, 2008 i dont know what to
9	ing is off of my forehead. This bitch is gonna die. I want it out of my sight. Kendell just ca
10	s scared to death. I thought I was gonna like die. Then he sped up and made a turn. I st
11	t the dead people ahahhaha dear jesus imma die;; and were also going to get a quija boar
12	llow in my room wandering when I'm going to die? Er, no. I know, I know, before anyone s
13	n, to me, when you "die" your physical body dies on this plane of life. Not passed on from
14	sn't have to go back with Gabby when Ginny dies) finally progressing somewhat. She will
15	gernails* I had my first Mock GCSEs today *dies* English in the morning - went rather be
16	*slaps self* Must. Do. Biology. Revision. *dies*Dec. 13th, 2007 Dear Shakespeare, I

Figure 7.1 Concordance lines for DIE

N	Concordance
1	VER brings bacardi to the creek again i will kill a nigguhh and i hope next weekend is be
2	t other one everyone seems to have read..To Kill A Mockingbird. Done well this xmas have
3	irst two weeks of school because she tried to kill herself. 7. My friend is juggling three boy
4	st, but theres a serial killer loose and tries to kill him and stuff, it's just dead good haha! FI
5	go to school, I can't miss school. Shawn will kill me. And my math and chemistry homew
6	ouglass Academy High School. Yea I know, kill me. Summer is over and it鈥 檚 not comi
7	really that's not exactly polite and Coey may kill me. So I'll probably go anyway... sigh. No
8	eral dossing. Quote: 'they use the money to kill people' haha =D Saturday I went to a lat
9	each time I go to town that 'its ok, I wont be killed by an axe murderer.' I kinda lost it a bi
10	teresting SORRY. That wasp sting FUCKING KILLED... but thank the Lordy for VINEGAR!
11	pressed because then people may have been killed...but it was only 5.2 on the Richter sca
12	official. HERE IS THE LIST OF THINGS I'VE KILLED: * 3+ computers * 2 TVs * 4 sets o
13	that Joseph and I can hurt each other without killing each other. I believe that one the bigg
14	layer of acid over itself? Seriously, why? It's killing me, it's hard to swallow, let alone talk.
15	ad grades on my brain and lack of sleep. It's killing me. I guess I could just try talking to
16	nd fainting!!! no one cares anymore.. and its killing me.. :(its making me feel dead inside,
17	me it was my fault. I was silently and slowly killing myself on the inside. I was doing it to
18	-Got STUNG BY A FUCKING WASP! fucking KILLS and I swore way too much. and cried
19	g on. my feet hurt too. im so stupid. my head kills. i just stared at the screen for like 5 min

Figure 7.2 Concordance lines for KILL

We can obtain a rough idea about what mid-teens bloggers tend to write about in their blogs from Table 7.2, but it cannot tells us whether there is any gender difference within this group. In order to get that information, an intra-group comparison in terms of semantic domain overuse is conducted and the results are presented in Table 7.3 below.

Table 7.3 Gender differences within the 15-17 blogger group

Female			Male		
Semtag	**Semantic domain**	**LL**	**Semtag**	**Semantic domain**	**LL**
E2-	Dislike	22.35	K2	Music and related activities	23.64
X2.1	Thought, belief	15.84	K5.1	Sports	15.67
A14	Exclusivizers/particularizers	13.56	G2.1-	Crime	15.01
Z6	Negative	13.07	W4	Weather	14.02
L1-	Dead	12.18	T1.3	Time: Period	13.34
Z99	Unmatched	12	Q1.2	Paper documents and writing	13.06
X2.5+	Understanding	7.9	A5.1+	Evaluation: Good	12.32
E4.1-	Sad	7.16	M1	Moving, coming and going	10.06
A11.2+	Noticeable	7.02	S4	Kin	9.9
A5.1-	Evaluation: Bad	7.02	S1.2.4+	Polite	9.35

From this table we can see that female mid-teens bloggers appear to disclose more about their thoughts, feelings (X2.1), and the things or people they dislike (E2-). They try to make sense out of things (X2.5+). They tend to feel sad, depressed, and frustrated (E4.1-). They prefer words like *die* and *kill* than their male counterparts and they tend to use these words in a joking or exaggerating manner. In fact, the greatest majority (32 out of the 35) of concordance lines in Figures 7.1 and 7.2 are from female bloggers. In other words, only three lines are from male bloggers. The language of female bloggers in this age group seems to be more unconventional than that of the male bloggers as can be seen from the overuse of the domain Z99. The male bloggers, on the other hand, talk more about music and bands (K2), sports (K5.1), crime-related topics (G2.1-), Internet-based written communication (Q1.2), and family members (S4). They pay greater attention to the changing of weather conditions (W4). They appear to be more dynamic as well, as revealed by the overuse of domain M1.

7.4 The 18-19 age group

The list of preferred semantic domains for the late teens group (i.e. the 18-19 group) is much shorter than that of the mid-teens group, as can be seen from Table 7.4 below. Nevertheless, ten out of the thirteen categories in this list are also included in the list for the 15-17 age group, suggesting that both groups may have many things in common. One of the more salient features of this group is their preferred use of boosters (e.g. *really*, *so*, *seriously*) (A13.3) and the preference for words like *just* and *only* (A14). These two categories do not tell us much about what they write about but rather about how they write about things. They tend to talk more about people: themselves and people around them (Z1 and Z8). They are also interested in talking about relationships (S3.2). They like to write about what they love and like (E2+). Many of their blog entries mention topics related to religion and the supernatural (S9). Some of these topics are really about religion while others are remotely related to religion in one way or another. Like the mid-teens group, the late teens also appear to favor a more unconventional writing style, as can be revealed by their overuse of categories Z4 and Z99. They are not really interested in talking about coldness as the overuse of the domain O4.6- seems to be suggesting. In fact, this has something to do with the wrong semantic annotation of two very frequently used words in the dataset of this group: *cool* and *chill*. Both words can be used to refer to temperature but they are more frequently used as slang words to mean *good* and *relax* respectively. The overuse of this domain actually points to the late-teens' preference for more fashionable language.

Table 7.4 Preferred semantic domains for the 18-19 group

Semtag	Semantic domain	Degree of Prominence
A13.3	Degree: Boosters	5
N1	Numbers	4
P1	Education in general	4
S3.2	Relationship: Intimacy and sex	4
Z1	Personal names	4
Z99	Unmatched	4
A14	Exclusivizers/particularizers	3
E2+	Like	3
O4.6-	Temperature: Cold	3
S1.2	Personality traits	3
S9	Religion and the supernatural	3
Z4	Discourse Bin	3
Z8	Pronouns	3

Within the 18-19 age group, male and female bloggers have displayed certain differences in preferred semantic domains. Table 7.5 lists the details. From this table we can see that late-teens female bloggers talk more about arts and crafts (C1), and photographs in particular. Like the mid-teens female bloggers, they also tend to disclose their negative feelings and emotions such as sadness, depression, and frustration (E4.1-). They talk about education in general (P1) and their school life (Q4.2) more often than the male counterparts. The domain Q4.2 is a bit misleading because it has included words like *paper* and *papers* which almost exclusively (30 out of the 32 occurrences) mean *term paper* or *research paper* rather than *newspaper*. In other words, this domain reflects that school work rather than newspapers is one of the important topics for female late-teens bloggers. Female bloggers in this age group seem to be involved more in communication with other people via cell phones and Internet-based communication tools like email (Q1.3). They also mention plants, trees, and flowers (L3) more in their blog entries. In addition, their mention of people is slightly more frequent than that of the male bloggers, as can be seen from the overuse of pronouns, mainly personal pronouns (Z8).

Table 7.5 Gender differences within the 18-19 blogger group

	Female			Male	
Semtag	Semantic domain	LL	Semtag	Semantic domain	LL
C1	Arts and crafts	35.42	S9	Religion and the supernatural	18.13
E4.1-	Sad	27.64	T3+	Time: Old; grown-up	14.99
P1	Education in general	21.6	K5.1	Sports	11.35
Q4.2	The Media: Newspapers etc.	19.16	O2	Objects generally	10.34
S6+	Strong obligation or necessity	15.08	K2	Music and related activities	8.56
Q1.3	Telecommunications	13.02	A8	Seem	7.81
L3	Plants	9.23	O4.6-	Temperature: Cold	7.75
Z8	Pronouns	8.26	S3.2	Relationship: Intimacy and sex	7.27
A5.4+	Evaluation: Authentic	7.87	F2	Drinks and alcohol	7.25

The male bloggers in this group, on the other hand, seem to be more interested in talking about religion and the supernatural. One possible reason is that some of the blog entries are written during the second half of the year when several religious festivals and holidays are celebrated. Another possible reason would be male bloggers talk more about electronic games where many fictional characters are remotely related to religion. Like the mid-teens male bloggers, the late-teens male bloggers have shown greater interest in talking about sports (K5.1) and music and bands (K2). What is different from the mid-teens is that the late teens talk more about drinks and alcohol (F2). Other than that, they are also more interested in talking about general objects (O2) and relationships (S3.2). Like the mid-teens, they also like to use words like *cool* and *chill*, which are actually slang words for *good* and *relax* respectively as mentioned earlier. Issues regarding annotation inaccuracy due to the slangy use of ordinary words will be discussed in greater detail in Chapter 10.

7.5 The 20-24 age group

The 20-24 age group shares some preferred semantic domains with the late-teens group. For instance, they also tend to write about their likings (E2+), education in general (P1), and religion and the supernatural (S9). Their language is also full of unconventional elements (Z99). Apart from these topics, they write about unexpected things in their daily life (X2.6-), entertainment in general (K1), and music and related activities (K2). Compared with many other groups, this group mentions more about death of friends or relatives (L1-), though some of the words in this domain are used for exaggerating purposes (for instance, the word *kill*). They also talk more about their thoughts, ideas, and opinions and their efforts in making sense out of what other people said, as revealed by the overuse of categories X4.1 and X2.5+ against three other groups. This group seems to be involved in plenty of reading activities (Q3). A closer examination of the words included in this domain shows nearly half of the word tokens can be attributed to the word *read* and its inflectional forms. Table 7.6 lists all the overused categories for this group.

Table 7.6 Preferred semantic domains for the 20-24 group

Semtag	Semantic domain	Degree of Prominence
T1	Time	5
A5.2+	Evaluation: True	3
E2+	Like	3
K1	Entertainment generally	3
K2	Music and related activities	3
L1-	Dead	3
P1	Education in general	3
Q3	Language, speech and grammar	3
S9	Religion and the supernatural	3
X2.5+	Understand	3
X2.6-	Unexpected	3
X4.1	Mental object: Conceptual object	3
Z6	Negative	3
Z99	Unmatched	3

Table 7.6 shows the common features of the whole group. If we compare the dataset of the males with that of the females, we will see some interesting differences. Table 7.7 lists the preferred semantic domains for male and female bloggers in this group respectively. The female bloggers appear to be more interested in topics related to people and personal relationship (S2.1, S3.1, S4, and Z8). They talk more about health, disease, and medical treatment (B2, B2-, and B3). Food seems to be one of favorite topics, as can be seen from the overuse of categories F1 and F1-. The overuse of domain E1 has something to do with the frequent mention of the word *mood* in the female bloggers' data. This is actually resulted from one of the discourse features of blogging, which forces the blogger to choose a word to represent their emotional mood at the time of blogging. The male bloggers, on the other hand, seem to be more interested in topics related to television programs and movies (Q4.3), computer and Internet (Y2), music and bands (K2), sports (K5.1), objects in general (O2), and brand names (Z3). They often find things weird, strange, odd, or incredible (A6.2-). The overuse of this domain may have

something to do with the frequent occurrences of the word *weird*, which might be a fashionable word for male bloggers from this age group. They also tend to show a rather positive attitude towards things as the domain A5.1+ reveals.

Table 7.7 Gender differences within the 20-24 blogger group

	Female			Male	
Semtag	Semantic domain	LL	Semtag	Semantic domain	LL
S4	Kin	37.93	Q4.3	The Media: TV, Radio and Cinema	30.06
S3.1	Personal relationship: General	26.31	Y2	Information technology and computing	16.49
Z8	Pronouns	23.55	O2	Objects generally	14.38
B2-	Disease	19.23	A6.2-	Comparing: Unusual	14.12
B3	Medicines and medical treatment	18.4	K5.1	Sports	13.57
B2	Health and disease	16.54	K2	Music and related activities	13.01
S2.1	People: Female	13.59	A5.1+	Evaluation: Good	12.04
E1	Emotional Actions, States And Processes General	12.3	I1.3	Money: Cost and price	11.4
F1	Food	11.33	Z3	Other proper names	10.75
F1-	Lack of food	11.03	N2	Mathematics	10.74

7.6 The 25-29 age group

The list of preferred semantic domains for the 25-29 age group shows a rather different picture from the three younger age groups I have already described, as Table 7.8 shows. From this table we can see that bloggers from this age group seem to be quite concerned about health problems and medical treatment (B3). Houses, buildings, and architecture have become more prominent in their life (H1). This might be a reflection of the fact that people at this particular age group have started to own their own houses, apartments, or flats. Consequently, they will care more about their backyards, their neighbors, and the neighborhood (H3). It is also natural for them to mention the furniture and household fittings (H5). They are more interested in talking about other living creatures (L2). Work and employment have become an integral part of this group's daily life (I3.1). So is driving or travelling on public transportation means (M3). It is quite difficult to figure out why they talk about the smell of things (X3.5). They tend to use negation very frequently (Z6).

Table 7.8 Preferred semantic domains for the 25-29 group

Semtag	Semantic domain	Degree of Prominence
A7+	Probability	5
B3	Medicines and medical treatment	3
H1	Architecture, houses and buildings	3
H3	Areas around or near houses	3
H5	Furniture and household fittings	3
I3.1	Work and employment: Generally	3
L2	Living creatures: animals, birds, etc.	3
M3	Vehicles and transport on land	3
X3.5	Sensory: Smell	3
Z6	Negative	3

Table 7.9 Gender differences within the 25-29 blogger group

	Female			Male	
Semtag	**Semantic domain**	**LL**	**Semtag**	**Semantic domain**	**LL**
Z8	Pronouns	37.43	I3.1	Work and employment: Generally	39.62
S3.2	Relationship: Intimacy and sex	32.46	K2	Music and related activities	28.45
L2	Living creatures: animals, birds, etc.	28.86	K5.1	Sports	18.74
Z6	Negative	19.2	I1.1	Money and pay	12.77
F3	Smoking and non-medical drugs	19.09	S7.1+	In power	12.55
B5	Clothes and personal belongings	18.16	T2-	Time: Ending	12.26
N3.2-	Size: Small	13.61	Q4	The Media	11.36
X2.5-	Not understanding	12.2	N5++	Quantities: many/much	10.76
S2.1	People: Female	11.89	N3.8-	Speed: Slow	9.59
F4	Farming & Horticulture	11.83	K6	Children's games and toys	9.59

Table 7.9 shows the gender difference within the 25-29 group. The female bloggers write more about people (Z8, S2.1), relationship (S3.2), and other living creatures (L2). They care more about clothes and personal belongings (B5). Many of them mention cigarette

smoking (F3). The male bloggers, on the other hand, write more about work and employment (I3.1), music and bands (K2), sports (K5.1), money matter (I1.1), and power relations in workplace (S7.1+).

7.7 The 30-34 age group

The 30-34 age group is what a blogger calls “not young, not old” group. Its overall list of overused semantic domains is also the shortest and the least exciting. What this list seems to suggest is that bloggers of this age group are less sure about their statements about many things, as revealed by the overuse of the word *seem* (A8). They mention moving around and transportation very often (M2 and M3). This might have something to do with the fact that this group is among the major work force. The O4.4 domain cannot tell us much about this group of bloggers, as some of the high-frequency words included in this domain are not really about geometrical shapes. For instance, the word *line* is used to mean quite different things such as underground line, a queue, and telephone line. This group seems to be interested in writing about new things (for instance, new cars, new hard disks, new jobs, new clothes, and new albums) (T3-). Unlike the language of the younger groups which is full of unmatched categories, bloggers from this age group tend to be more conventional in their use of language, as the overuse of domain Z5 seems to be suggesting. Table 7.10 lists the preferred categories for this group.

Table 7.10 Preferred semantic domains for the 30-34 group

Semtag	Semantic domain	Degree of Prominence
A8	Seem	4
M2	Putting, pulling, pushing, transporting	3
M3	Vehicles and transport on land	3
O4.4	Shape	3
T3-	Time: New and young	3
Z5	Grammatical bin	3

An intra-group comparison between male and female bloggers shows that they are different in a number of preferred semantic domains (see Table 7.11 below). Quite similar to female bloggers in younger groups, female bloggers in this group also appear to be more interested in writing about themselves and the people around them (Z8 and S4). This has a great deal to do with the nature of the blogging genre. They are more concerned about their body (B1), their weight (N3.5), and their physical wellness (B3) than their male counterparts. They mention education in general (P1) more. They are more willing to talk about their thoughts and feelings (S2.1) and they also mention their efforts in trying to make sense out of the things around them. The domain A5.3- cannot tell much about this group, as it includes some high-frequency words like *miss* which can be interpreted in different ways, depending on the context. Mainly it is not about inaccurate evaluation but about thinking of someone. The male bloggers in this group have displayed their interest in writing about their more dynamic life style. They move around more frequently (M1, M3, M4, and M7) and they participate more in social activities (S1.1.3+). They are more interested in things related to entertainment (K3).

Table 7.11 Gender differences within the 30-34 group

	Female			Male	
Semtag	Semantic domain	LL	Semtag	Semantic domain	LL
Z8	Pronouns	41.96	M4	Sailing, swimming, etc.	19.04
S4	Kin	23.94	Z5	Grammatical bin	13.86
B1	Anatomy and physiology	20.59	M3	Vehicles and transport on land	13.54
B3	Medicines and medical treatment	15.39	M1	Moving, coming and going	12.41
A13.3	Degree: Boosters	15.22	K3	Recorded sound	10.96
N3.5	Measurement: Weight	13.93	N5	Quantities	10.74
X2.5+	Understanding	13.86	S8-	Hindering	10.7
S2.1	Thought, belief	12.25	M7	Places	10.41
P1	Education in general	11.61	X3.2	Sensory: Sound	10.31
A5.3-	Evaluation: Inaccurate	9.85	S1.1.3+	Participating	8.94

7.8 The 35-40 age group

The list of preferred semantic domains for the 35-40 age group, the oldest among the whole six groups, shares two categories with that of the younger group next to it (the 30-34 group): M3 (driving and travelling in vehicles) and Z5 (grammatical bin, an indicator of conventionality of language). Different from the 30-34 group, bloggers from the 35-40 group talk more about what usually or normally happens in their life (A6.2+), what they have or possess (A9+), topics related to health care (B3), trees, plants, and flowers (L3), social interactions realized via phone (Q1.3) and visiting (S1.1.1), and activities related to shopping and selling (I2.2). They are more aware of the weather conditions (W4). This time the overuse of this domain has nothing to do with slangy words such as *cool* and *chill*. They also mention water quite frequently in their blogs, drinking water or otherwise (O1.2).

Table 7.12 Preferred semantic domains for the 35-40 group

Semtag	Semantic domain	Degree of Prominence
B3	Medicines and medical treatment	4
W4	Weather	4
A6.2+	Comparing: Usual	3
A9+	Getting and possession	3
I2.2	Business: Selling	3
L3	Plants	3
M3	Vehicles and transport on land	3
N3.8	Measurement: Speed	3
O1.2	Substances and materials: Liquid	3
Q1.3	Telecommunications	3
S1.1.1	Social Actions, States And Processes	3
Z5	Grammatical bin	3

If we take a closer look at exact words included in some of the semantic domains, we will find some extra information about this blogger group. For instance, the domain B3 includes around 17 occurrences of the word *doctor* which is not used to refer to any medical doctor in real life but rather a TV serial entitled *Doctor Who*. This may well be an indicator that this TV serial is very popular among bloggers from this age group. Apart from that, there are around 12 cases for the abbreviated form of this word (i.e. *dr* or *Dr*) as well, which, again, is not really related to medical treatment. See Table 7.12 above for full list of the preferred categories for this group.

Table 7.13 lists the preferred semantic domains for the male and female bloggers in this group. The female bloggers from this group write more about people (Z8) and family members (S4) than their male counterparts. They have also mentioned more of education in general (P1) and driving and travelling via transportation vehicles (M3). They seem to be more involved in spoken communications with other people (Q2.1 and Q2.2). There are quite some negative comments about people and things (S1.2.6-). They mention the word *life* (L1+) more often, although their blog entries are actually records of their real life. They use negation more frequently than their male counterparts.

Table 7.13 Gender differences within the 35-40 group

Female (the 35-40 group)			Male (the 35-40 group)		
Semtag	Semantic domain	LL	Semtag	Semantic domain	LL
Z8	Pronouns	102.3	K2	Music and related activities	55.87
S4	Kin	40.65	Q4.3	The Media: TV, Radio and Cinema	30.58
S1.2.6-	Foolish	28.32	T1.1.3	Time: Future	30.08
Q2.1	Speech: Communicative	27.04	K4	Drama, the theatre and show business	26.32
P1	Education in general	23.84	Z5	Grammatical bin	24.83
M3	Vehicles and transport on land	18.65	O2	Objects generally	23.8
S1.1.2+	Reciprocal	13.73	O3	Electricity and electrical equipment	23.79
Z6	Negative	13.37	K5.2	Games	20.2
Q2.2	Speech acts	13.1	I3.1	Work and employment: Generally	19.65
L1+	Alive	13	Y2	Information technology and computing	17.97

The male bloggers from this group, however, seem to be more concerned about music and related activities (K2), television programs and movies (Q4.3), drama and performances

(K4), future activities or events (T1.1.3), electronic devices (O3), computers and the Internet (Y2), games (K5.2). They also talk more about work and employment than their female counterparts (I3.1). Their language contains more grammatical words, which might be an indicator of more conventional use of language (Z5). The overuse of the domain O2 might be a bit misleading as it includes one high-frequency word *thing* which is often used as a vague term instead of objects in general. Nevertheless, the overuse of this domain tells us indirectly about the male bloggers' use of vague terms.

7.9 Chapter summary

From what has been presented in this chapter, we can see that bloggers from different age groups have displayed different preferences for semantic domains. Examining bloggers' preferences for semantic domains is actually a way of exploring variations in blogging content. The preferred semantic domains of bloggers from each age group can give us clues about at what developmental stages of human life they are and what social roles they are currently assuming. Cross-gender comparison within each age group has demonstrated consistent differences between male and female bloggers in terms of blogging content. As to how variation in semantic domains is related to bloggers' representation of age- and gender-related identity, it will be discussed in Chapter 9.

Chapter 8 Variation in Grammatical and Pragmatic Features

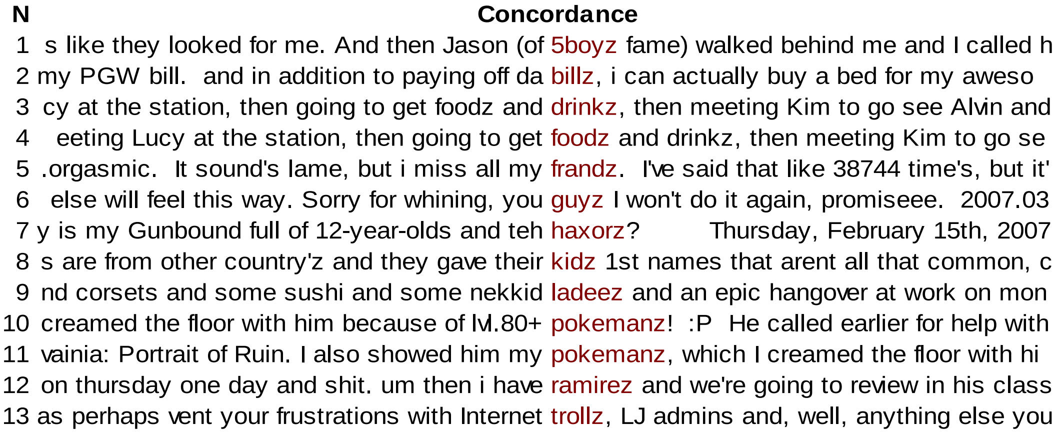
This chapter has two themes: variation in grammatical features and that in pragmatic features. The first half of this chapter focuses on the less conventional (and archaic) morpho-syntactic and syntactic features. The second half discusses bloggers' use of three pragmatic features: discourse markers, interjections, and vague words.

8.1 Variation in grammatical features

Compared with orthographic variations, lexical variations, and semantic variations which are determined by bloggers' intentionality to a considerable extent, variations in morpho-syntactic and syntactic features are more likely to be markers of group identity. Two speakers from the same speech community may vary greatly in their choice of words, but they may vary very little in their use of morpho-syntactic or syntactic structures. To a great extent, morpho-syntactic and syntactic rules are more established and are less likely to be manipulated without creating a sense of oddness, though this oddness may well be exploited for special purposes. I am not suggesting that these less-prone-to-change norms can never be changed. Quite to the contrary, I will present some evidence to show that even in such an aspect of the language system which is generally held to be rather stable, change is also taking place. Meanwhile, I also want to show that bloggers' preference for certain morpho-syntactic and syntactic features can be either results of choice or simply reflection of part of their collective identity. I have no intention to exhaust all the unusual morpho-syntactic and syntactic features employed by bloggers in their blog entries. Instead, I will only focus on five features and try to explore to what extent they are related to the linguistic representation of bloggers' identities. These features include: new meaning of plural forms, the case issue, the pattern of *go/come* plus bare infinitives, *like* as quotative complementizer, and the recycling of archaic morpho-syntactic features. The first two features concern the core grammatical categories in the English language: number and case. The features concerning *go/come* plus infinitives and *like* used as quotative complementizer are two new (or newer) features which are related to grammaticalization. The archaic morpho-syntactic features are the ones which are not supposed to be active in contemporary language use. Examining bloggers' use of these features can give us some snapshots about how variation in grammatical features looks like.

8.1.1 New meaning of plural forms

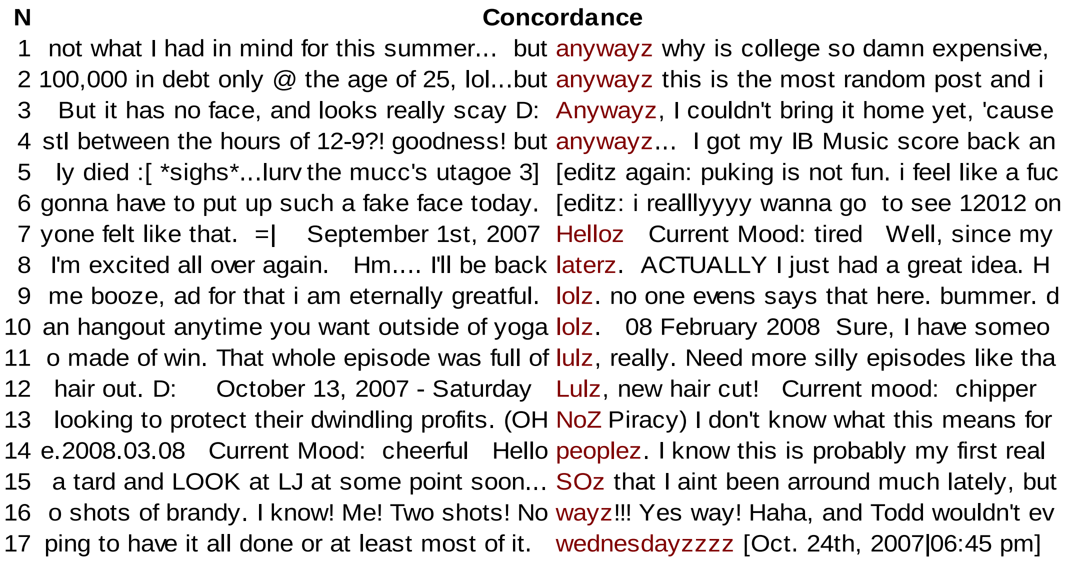
To start with, let us take a look at the change which is taking place with the form and meaning of the plural inflectional morpheme. As mentioned above, certain domains of the linguistic system do not lend themselves for immediate change. Inflectional morphemes are one of them. Modern English still maintains certain morpho-syntactic regulations on some word categories: verbs (for marking person, tense, aspect, and mood), nouns (for marking singular and plural forms), adverbs and adjectives (for marking comparative and superlative degrees), and pronouns (for nominative and accusative cases). These regulations are generally well observed in conventional writing. Even when they are not well observed it is a matter of change of spelling of the inflectional morphemes in most of the cases. For instance, many people will use the form *–in* to replace *–ing* for inflections regarding present participles or gerund forms of regular verbs. In fact, we have already touched on this issue in Chapter 5 when orthographic variations are being discussed. The newer form –*in* (as in the word form *tryin*) is just a phonetic spelling of the conventional form (-*ing*) as many people find that the former is closer to the actual pronunciation of the inflectional morpheme. The grammatical meaning of the new form is no different from the conventional one. The difference lies in the stylistic aspect and the attitude towards established linguistic norm. The same thing happens with the plural inflectional morpheme –s (as in *boys*), which is often spelled as –z (as in *boyz*), again because many people feel that the latter better represents the actual pronunciation of that morpheme. Figure 8.1 shows how bloggers use this new morpheme in their entries.



```
 N                                 Concordance
 1  s like they looked for me. And then Jason (of 5boyz fame) walked behind me and I called h
 2 my PGW bill.  and in addition to paying off da billz, i can actually buy a bed for my aweso
 3  cy at the station, then going to get foodz and drinkz, then meeting Kim to go see Alvin and
 4   eeting Lucy at the station, then going to get foodz and drinkz, then meeting Kim to go se
 5 .orgasmic.  It sound's lame, but i miss all my frandz.  I've said that like 38744 time's, but it'
 6   else will feel this way. Sorry for whining, you guyz I won't do it again, promiseee.  2007.03
 7 y is my Gunbound full of 12-year-olds and teh haxorz?        Thursday, February 15th, 2007
 8  s are from other country'z and they gave their kidz 1st names that arent all that common, c
 9  nd corsets and some sushi and some nekkid ladeez and an epic hangover at work on mon
10  creamed the floor with him because of lvl.80+ pokemanz!  :P  He called earlier for help with
11  vainia: Portrait of Ruin. I also showed him my pokemanz, which I creamed the floor with hi
12  on thursday one day and shit. um then i have ramirez and we're going to review in his class
13 as perhaps vent your frustrations with Internet trollz, LJ admins and, well, anything else you
```

Figure 8.1 Concordance lines for plural morpheme –z

As can be observed from these lines, the grammatical meaning of the new form –*z* is still the same, but the stylistic and pragmatic functions are different. However, bloggers' innovation does not stop here. For this new plural morpheme –*z* bloggers have expanded its scope of application and attached some new functions and new meanings to it, as the following concordance lines may demonstrate (see Figure 8.2 below). In the English language the plural marker is only attached to nouns and countable nouns in particular, but if we take a closer look at the words taking the morpheme –*z*, we will soon find that many of them are not nouns at all, for instance, *anyway* (adverb), *edit* (verb), *hello* (interjection), *later* (adverb), *lol* and its variant *lul* (acronym of a verbal phrase *laughing out loud*), *no* (adjective or adverb), and *so* (conjunction). Apparently, this deviates tremendously from the established norm for inflectional morpheme use in the English language and it is actually not allowed in the English grammar.



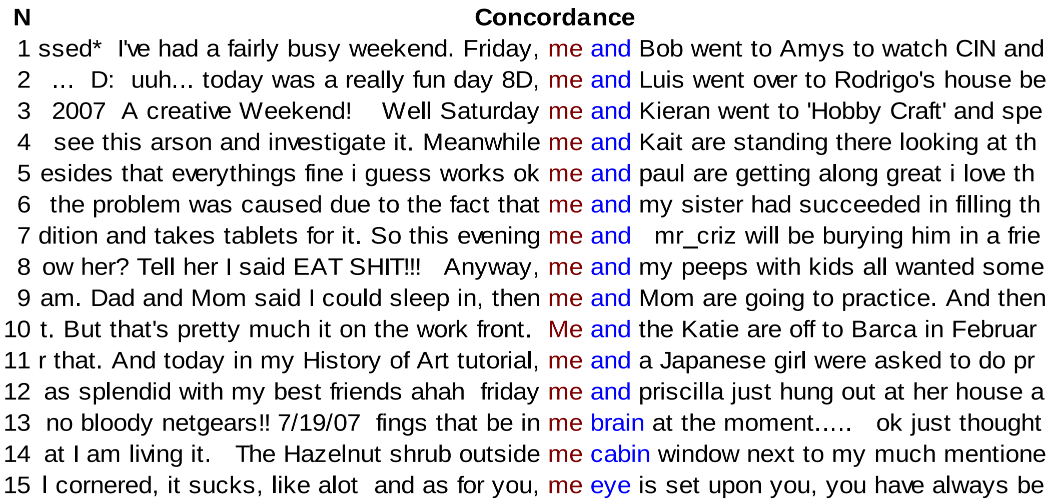
```
 N                                  Concordance
 1  not what I had in mind for this summer...  but anywayz why is college so damn expensive,
 2 100,000 in debt only @ the age of 25, lol...but anywayz this is the most random post and i
 3   But it has no face, and looks really scay D:  Anywayz, I couldn't bring it home yet, 'cause
 4 stl between the hours of 12-9?! goodness! but anywayz...  I got my IB Music score back an
 5   ly died :[ *sighs*...lurv the mucc's utagoe 3]  [editz again: puking is not fun. i feel like a fuc
 6 gonna have to put up such a fake face today.  [editz: i realllyyyy wanna go  to see 12012 on
 7 yone felt like that.  =|    September 1st, 2007  Helloz   Current Mood: tired   Well, since my
 8  I'm excited all over again.   Hm.... I'll be back laterz.  ACTUALLY I just had a great idea. H
 9  me booze, ad for that i am eternally greatful.  lolz. no one evens says that here. bummer. d
10  an hangout anytime you want outside of yoga lolz.   08 February 2008  Sure, I have someo
11   o made of win. That whole episode was full of lulz, really. Need more silly episodes like tha
12    hair out. D:      October 13, 2007 - Saturday   Lulz, new hair cut!   Current mood:  chipper
13    looking to protect their dwindling profits. (OH NoZ Piracy) I don't know what this means for
14 e.2008.03.08   Current Mood:  cheerful   Hello peoplez. I know this is probably my first real
15   a tard and LOOK at LJ at some point soon... SOz that I aint been arround much lately, but
16   o shots of brandy. I know! Me! Two shots! No wayz!!! Yes way! Haha, and Todd wouldn't ev
17  ping to have it all done or at least most of it.   wednesdayzzzz [Oct. 24th, 2007|06:45 pm]
```

Figure 8.2 Concordance lines for -z with new meaning

If we stick to the conventional concept of plural, we may find it very difficult to explain why these new forms are possible. From the word forms cited above, we can see that the conventional meaning of "more than one" is still there, but that is no longer the chief meaning intended. In most of the cases listed here, the new plural morpheme is actually playing the function which letter repetition is playing, that is, emphasizing the word (to which the morpheme is attached) by lengthening it. In other words, *anyways* or *anywayz* equals to something like '*aannnyyywaaay*'; *helloz* equals to *helloooo*; *laterz* equals to *laaaaater*; *lolz* or *lulz* meaning something like *loool*; *noz* equals to *noooooo*; *no wayz* equals to *no waaaaay*, *peoplez* meaning something like *peopleeee*; and *soz that* means something like *soooooo that*. The new morpheme is not intended for conveying the sense of "more than one" but rather used as creative means for marking emphasis. By deviating radically from the established norm, the bloggers manage to achieve several effects at the same time. First, due to the novelty of the word forms, they are better at attracting readers' attention and getting them more engaged in the reading. Second, by using deviated forms, the bloggers show their preference for an informal style of writing. Third, they display their linguistic wisdom by playing with the language system though bricolage (i.e. constructing new things with existing materials). Last, through the use of such unusual morpho-syntactic features, these bloggers manage to mark themselves off from other bloggers or people who stick to the more established ways of writing.

8.1.2 The case issue

Apart from the more radically innovative morpho-syntactic features described in the previous section, bloggers have also attempted to challenge some other aspects of the language system for achieving informality and other purposes. One example is the case issue. In English, only pronouns are formally marked for cases: the nominative case, the accusative case, and the possessive case. These are actually the relics of old English. Each case of a pronoun can only be used in a particular slot of the syntactic structure which matches the case. In written English, violation of this rule will be either considered ungrammatical or uneducated. Thus, the nominative case of *I*, for instance, is only expected to appear at the subject position of a clause or sentence. It is also possible for it to appear in structure such as *It is I*, which is often considered extremely formal and unnatural. The accusative case *me* is generally expected to appear in object or complement positions like *It's me* in speech. Other usages are often considered either grammatically wrong or as evidence of being uneducated. In informal speech, these regulations are not well observed. An examination of the blog corpus shows that many bloggers prefer to use *me* at the subject position. Occasionally, they will use it to replace the possessive case *my*. The following concordance lines (see Figure 8.3) give a flavor of how *me* is used by some bloggers.



```
 N                                     Concordance
 1 ssed*  I've had a fairly busy weekend. Friday, me and Bob went to Amys to watch CIN and
 2  ...  D:  uuh... today was a really fun day 8D, me and Luis went over to Rodrigo's house be
 3  2007  A creative Weekend!    Well Saturday me and Kieran went to 'Hobby Craft' and spe
 4   see this arson and investigate it. Meanwhile me and Kait are standing there looking at th
 5 esides that everythings fine i guess works ok me and paul are getting along great i love th
 6  the problem was caused due to the fact that me and my sister had succeeded in filling th
 7 dition and takes tablets for it. So this evening me and   mr_criz will be burying him in a frie
 8 ow her? Tell her I said EAT SHIT!!!   Anyway, me and my peeps with kids all wanted some
 9 am. Dad and Mom said I could sleep in, then me and Mom are going to practice. And then
10 t. But that's pretty much it on the work front.  Me and the Katie are off to Barca in Februar
11 r that. And today in my History of Art tutorial, me and a Japanese girl were asked to do pr
12  as splendid with my best friends ahah  friday me and priscilla just hung out at her house a
13  no bloody netgears!! 7/19/07  fings that be in me brain at the moment.....   ok just thought
14  at I am living it.   The Hazelnut shrub outside me cabin window next to my much mentione
15  I cornered, it sucks, like alot  and as for you, me eye is set upon you, you have always be
```

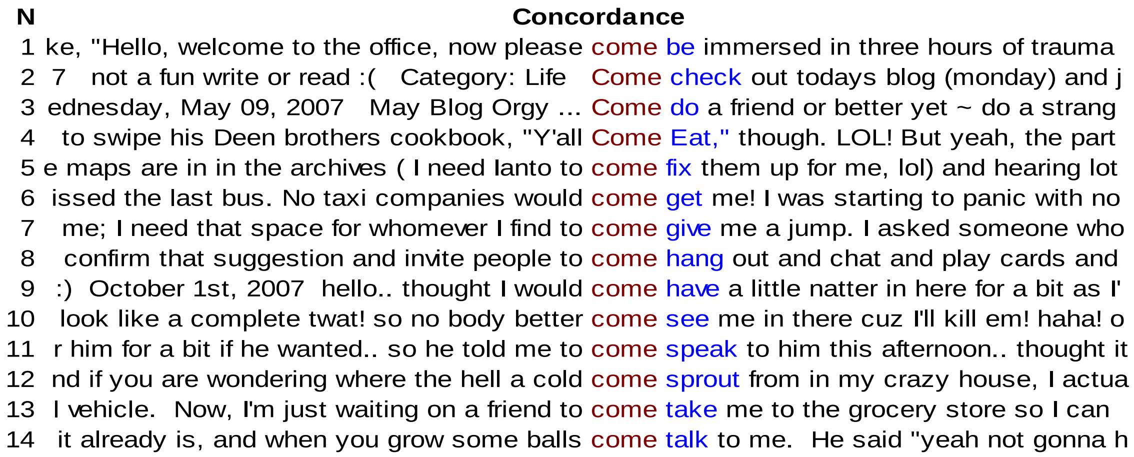
Figure 8.3 Concordance lines for ME

8.1.3 *Go/come* plus bare infinitives

One more syntactic feature which may reflect part of the users' identities would be the absence of syntactic elements, for instance, the absence of infinitive marker *to* in the V+V (verb plus verb) pattern. This pattern is very common in Mandarin Chinese, but it is rare in the English language. Probably this is also the reason why such patterns have seldom been described in English grammar books. The common pattern in English is modal auxiliaries (or supportive do) plus bare infinitives. Very few other verbs can fill the initial position of this pattern. *The Longman Grammar of Spoken and Written English* (Biber et al., 1999) only mentions two such verbs: *dare* and *help*. *The Cambridge Grammar of the English Language* (Huddleston & Pullum, 2002) mentions three, adding the word *know* and pointing out that bare infinitives only appear after the present aspect of the verb *know*. In fact, two more verbs often appear in this pattern but they have seldom been mentioned in any leading English grammar books: They are *go* and *come*. When *go* is followed by a verb, there are several possible patterns. The first one is "*go* plus infinitives" as in *go to* *check my mailbox*. The second one is “*go and* plus bare infinitives” as in *go and check my mailbox*. The third one is “*go* plus bare infinitives” as in *go check my mailbox*. The same patterns also hold for *come*. The second and third patterns are very close in meaning but the meaning of the first pattern could be quite different sometimes. Generally speaking, the “*go* plus infinitives” pattern emphasizes the purpose more. There are 158 occurrences of the pattern “*go/come* plus bare infinitives” in the blog corpus, with 144 occurrences for “*go* plus bare infinitives” and only 14 for “*come* plus bare infinitives.” Here are the sampled concordance lines for “*go* plus bare infinitives” (Figure 8.4) and the complete concordance lines for “*come* plus bare infinitives” (Figure 8.5).

```
 N                                      Concordance
 1 omeday soon she'll be in love with me, too.  I gotta go bribe Dad to take me out to do parallel parking
 2 s trying to tell me that it is actually a check and to go cash it. I am not going to embarrass myself by
 3 smilies here)  Incoming storm now. I think I should go climb into my non-organic bed and put on my t
 4 ndows open, but prepared on a moment's notice to go close them all, as the weather people kept thre
 5 -chatting / debating will be sparse.  Now I've got to go continue catching up with stuff.  Aug. 8th, 2007
 6  ek Night (and a Zipcar at the end of the evening to go deliver someone's ebay winnings).  Friday mor
 7 re. haha. I love you chris, but get your ass up, lets go do something!   Writing - I've been doing two ty
 8 . I'll take pics of that later. Though we now need to go find some nice plates and such to make everyt
 9 leep. oo did i also mention i had to wake up at 7 to go get blood work done? yea it's been a fan-fuckin
10 n elise jarrod and i. i dont even want to get into it) i go hangout with luke and then her and i go bar ho
11 re room in her dental practise to therapists. Gonna go have a chat with her and maybe do a couple of
12  ation that stops when I work at my desk and don't go hide in the lab.   Away from bloody Cardiff and
13 ey'd get out of my crappy normal people dorm and go live in their palaces, kthx.  back to band, i love
14  an't spend time with him. It's upsetting.  I'm gonna go make some tea me thinks.  Sunday, Decembe
15 s more willing to teach me. ; ] After that, I'm gonna go pick out a gift for myself and a certain someon
16 l hours later, we still weren't finished and he had to go pick up his 2 year old daughter from daycare.
17 Funny-Face  Feb. 23rd, 2008 at 8:29 PM    I had to go prove my existence to the government today. T
18  T thing I want and I told him so. Then he offered to go stay with his parents awhile to give me a break
19   days? I just don't understand.  Yeeaahh, I should go study or something. And I'll try to update more
20   ING MONTH TO GO! I'm so PUMPED!  I'm gonna go take a shower now! BYE.  Mood:  exhausted
21  cool. I'm not sure what I'll be doing, but I expect to go visit the guys at work.   As another side note, I
22  ion before.  It's the one with Jim Carrey.  Guess I'll go watch it then...  Now all I need is a Yuletide mi
23  y gets kicked out of a black business, he can just go work somewhere else. It doesn't work the other
```

Figure 8.4 Concordance lines for GO plus bare infinitives



```
 N                                 Concordance
 1 ke, "Hello, welcome to the office, now please come be immersed in three hours of trauma
 2  7   not a fun write or read :(   Category: Life   Come check out todays blog (monday) and j
 3 ednesday, May 09, 2007   May Blog Orgy ... Come do a friend or better yet ~ do a strang
 4   to swipe his Deen brothers cookbook, "Y'all Come Eat," though. LOL! But yeah, the part
 5 e maps are in in the archives ( I need Ianto to come fix them up for me, lol) and hearing lot
 6  issed the last bus. No taxi companies would come get me! I was starting to panic with no
 7   me; I need that space for whomever I find to come give me a jump. I asked someone who
 8   confirm that suggestion and invite people to come hang out and chat and play cards and
 9  :)  October 1st, 2007  hello.. thought I would come have a little natter in here for a bit as I'
10   look like a complete twat! so no body better come see me in there cuz I'll kill em! haha! o
11  r him for a bit if he wanted.. so he told me to come speak to him this afternoon.. thought it
12  nd if you are wondering where the hell a cold come sprout from in my crazy house, I actua
13  I vehicle.  Now, I'm just waiting on a friend to come take me to the grocery store so I can
14   it already is, and when you grow some balls come talk to me.  He said "yeah not gonna h
```

Figure 8.5 Concordance lines for COME plus bare infinitives

From these two sets of concordance lines we can see that both *go* and *come* have lost much of their original semantic sense of moving from one place to another and picked up a more grammatical meaning instead. *Go* seems to be more often associated with future events whereas *come* seems to be associated with events happening at the current moment. *Go*'s association with future events has plenty to do with its original meaning of leaving one's present place for another place, therefore it is pointing forward in terms of time reference. *Come*, on the other hand, refers to moving from another place to the speaker's place thus it is pointing to "now," the time of speaking or writing.

This kind of usage might be indicating that both *go* and *come* are in the process of grammaticalization. One defining feature of words undergoing grammaticalization is their loss of original lexical meaning and their gaining of grammatical meaning. Of course, their grammatical meaning is often remotely related to their original lexical meaning. When the pattern "*go/come* plus *to* infinitive" is used, the infinitive complement tends to be more like a clause of purpose, as can be seen Figures 8.6 and 8.7. Moreover, there are much fewer cases of this pattern in the blog corpus. Altogether, there are only 23 cases (the pattern *go to work* has been excluded in the concordance lines as it is hard to tell whether work is used as a verb or noun; there are about twenty cases of *go to work* in the corpus) .

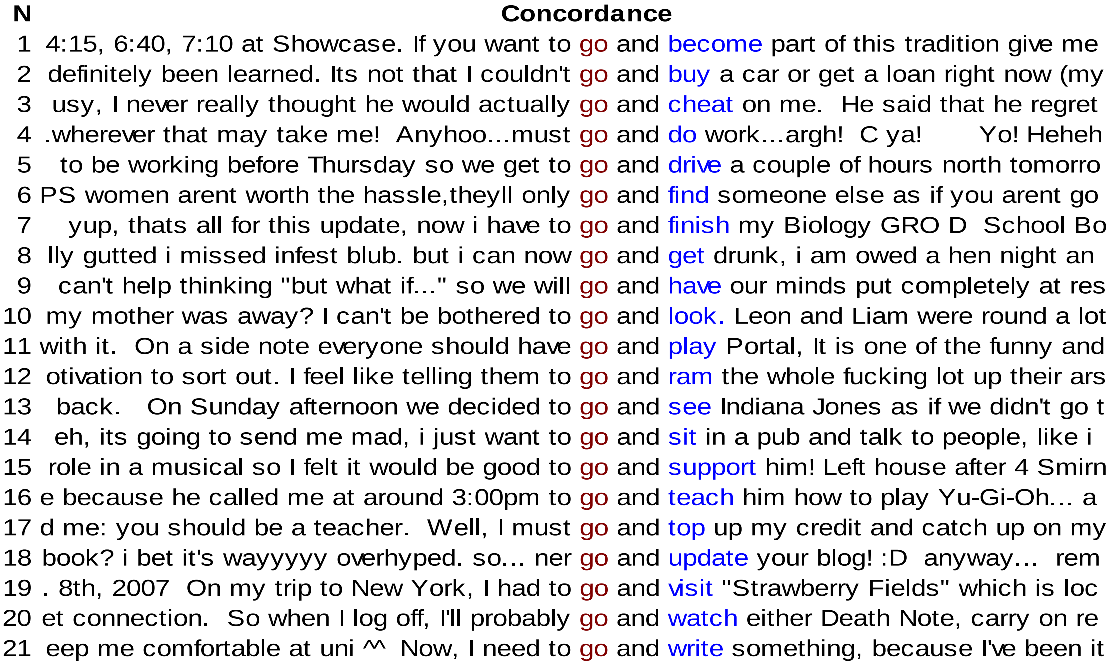
```
N                                  Concordance
 1  I keep expecting to find her dead each time I go to check on her, but she's bouncing arou
 2  t will start being filled up with cars as people go to enjoy an evening meal in one of the re
 3  I'm too young, yada yada yada bullshit.  So I go to get the mammogram Tuesday, wasn't
 4  nt, the lady was gone.   I went back inside to go to get ready for bed. I told my mom and
 5  oblem is that its always too damn crowded. I go to get away, take in some endorphins, an
 6   how I do not know what's wrong with it and I go to return to MSN as my Boyfriend is leavi
 7   nd I forgot to ask my parents if they want to go to see Elizabeth - The Golddn Age at the
 8   a fake face today.  [editz: i realllyyyy wanna go  to see 12012 on april 6 :[ but i have a girl
 9  idn't quite make it in time to see the test, but go to see video and then went out to lunch w
```

Figure 8.6Concordance lines for GO plus infinitives

```
N                                  Concordance
 1  oom with the door shut. Every weekday they come to empty the trash too, but I started p
 2   out casually, "Yeah."   Basically what I had come to expect from him. The bartender bro
 3  nt, and certainly smaller, than you may have come to expect (and dread). This is simply
 4  er half brother as a result. She was a Jones. Come to find out her brother is a Smith. Heh
 5  after she's on the bus, and won't know we've come to get her, if we don't catch her at the
 6   e of the many snopesters whose names I've come to know and admire. I've never met hi
 7   f course seeing all my nice fur friends I have come to know: Southpaw (naturally), Tungro
 8    pretty daunting. Now they're the ones we've come to love and have said goodbye to, mini
 9  as prior to the events of October 21st 2007.  Come to think of it, bad things generally hap
10   rote when I was younger and stupid (though come to think of it, one of them is missing, t
11  o vent my tension. I need to relax. Do I ever? Come to think of it... no, I never do.
12  her days of the week. That would actually fit, come to think of it.   Anyway, that means th
13   communication problems? Is it that we have come to think of marriage as a disposable si
14  now but it is my hopes that, in time, you will come to understand my feelings and know t
```

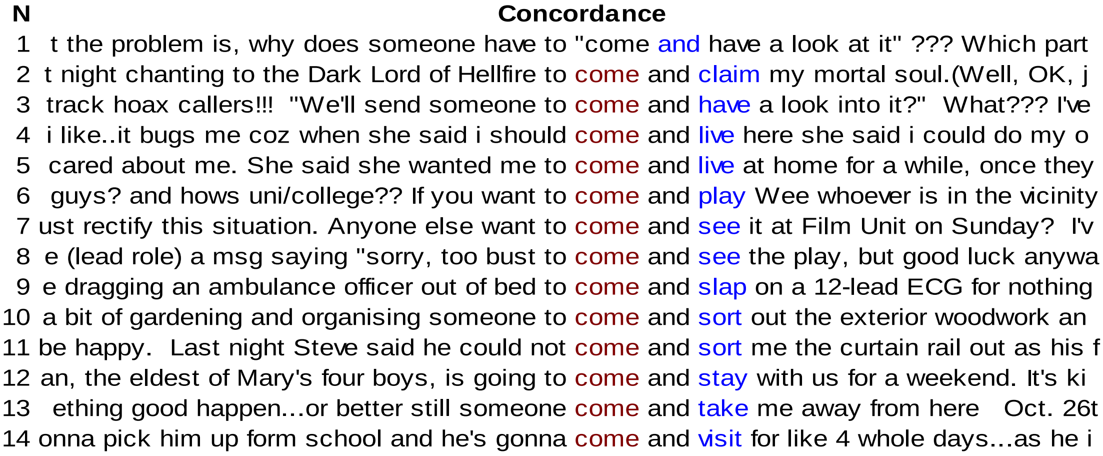
Figure 8.7Concordance lines for COME plus infinitives

There are 48 occurrences of the pattern "*go and* plus bare infinitives" and 14 occurrences of "*come and* plus bare infinitives." Figures 8.8 and 8.9 show how these two patterns have been used by bloggers included in this research. The relationship between the use of these patterns and the representation of blogger identity will be discussed in Chapter 9.



```
 N                                        Concordance
 1  4:15, 6:40, 7:10 at Showcase. If you want to go and become part of this tradition give me
 2  definitely been learned. Its not that I couldn't go and buy a car or get a loan right now (my
 3  usy, I never really thought he would actually go and cheat on me.  He said that he regret
 4 .wherever that may take me!  Anyhoo...must go and do work...argh!  C ya!       Yo! Heheh
 5   to be working before Thursday so we get to go and drive a couple of hours north tomorro
 6 PS women arent worth the hassle,theyll only go and find someone else as if you arent go
 7    yup, thats all for this update, now i have to go and finish my Biology GRO D  School Bo
 8  lly gutted i missed infest blub. but i can now go and get drunk, i am owed a hen night an
 9   can't help thinking "but what if..." so we will go and have our minds put completely at res
10  my mother was away? I can't be bothered to go and look. Leon and Liam were round a lot
11 with it.  On a side note everyone should have go and play Portal, It is one of the funny and
12  otivation to sort out. I feel like telling them to go and ram the whole fucking lot up their ars
13   back.   On Sunday afternoon we decided to go and see Indiana Jones as if we didn't go t
14   eh, its going to send me mad, i just want to go and sit in a pub and talk to people, like i
15  role in a musical so I felt it would be good to go and support him! Left house after 4 Smirn
16 e because he called me at around 3:00pm to go and teach him how to play Yu-Gi-Oh... a
17 d me: you should be a teacher.  Well, I must go and top up my credit and catch up on my
18 book? i bet it's wayyyyy overhyped. so... ner go and update your blog! :D  anyway...  rem
19 . 8th, 2007  On my trip to New York, I had to go and visit "Strawberry Fields" which is loc
20 et connection.  So when I log off, I'll probably go and watch either Death Note, carry on re
21  eep me comfortable at uni ^^  Now, I need to go and write something, because I've been it
```

Figure 8.8 Concordance line for GO AND plus bare infinitives

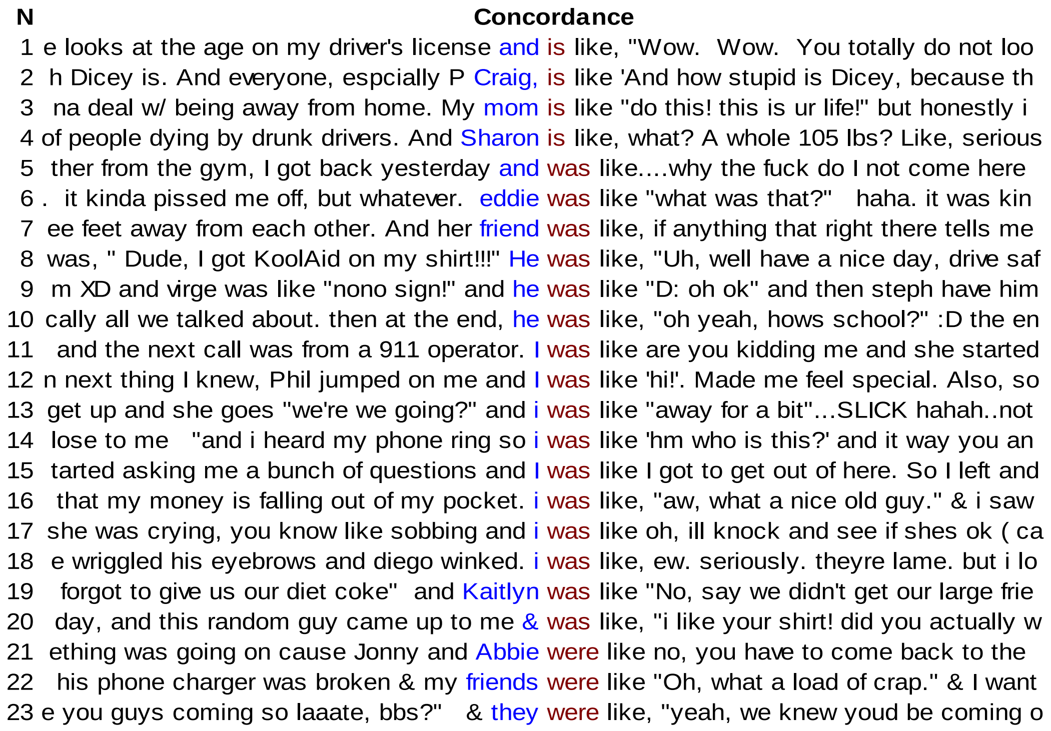


```
 N                                        Concordance
 1  t the problem is, why does someone have to "come and have a look at it" ??? Which part
 2  t night chanting to the Dark Lord of Hellfire to come and claim my mortal soul.(Well, OK, j
 3  track hoax callers!!!  "We'll send someone to come and have a look into it?"  What??? I've
 4  i like..it bugs me coz when she said i should come and live here she said i could do my o
 5  cared about me. She said she wanted me to come and live at home for a while, once they
 6  guys? and hows uni/college?? If you want to come and play Wee whoever is in the vicinity
 7 ust rectify this situation. Anyone else want to come and see it at Film Unit on Sunday?  I'v
 8  e (lead role) a msg saying "sorry, too bust to come and see the play, but good luck anywa
 9  e dragging an ambulance officer out of bed to come and slap on a 12-lead ECG for nothing
10  a bit of gardening and organising someone to come and sort out the exterior woodwork an
11 be happy.  Last night Steve said he could not come and sort me the curtain rail out as his f
12 an, the eldest of Mary's four boys, is going to come and stay with us for a weekend. It's ki
13   ething good happen...or better still someone come and take me away from here   Oct. 26t
14 onna pick him up form school and he's gonna come and visit for like 4 whole days...as he i
```

Figure 8.9 Concordance lines for COME AND plus bare infinitives

8.1.4 *Like* as a quotative complementizer

Similar to *go* and *come* to some extent, the word *like* is also undergoing grammaticalization. It has, according to Stenström et al. (2002), developed a range of new uses over its process of grammaticalization: approximative, exemplifactory, metalinguistic, hesitational or linking, and quotative or interpretive. I will only focus on its syntactic function as a quotative complementizer here. According to Stenström et al., (2002), *like* can be used as an obligatory component of the grammaticalized quotative complementizer and the typical construction for this usage is “copula plus like” (or *be like*). The major function of the word *like* in this construction is to “mark off the following linguistic material as a thought, attitude or feeling which is meta-represented, but which has not necessarily been explicitly uttered” (p. 116). They have also noted that in British English the expression “be like” has not been grammaticalized to the same extent as in American English. In other words, this expression may be more strongly associated with Americanism. There are 65 cases of the expression “be like” being used for its quotative function. Figure 8.10 shows one third of the total number of occurrences. From these examples we can see that “be like” mainly appear in past tense. This may have something to do with the nature of blogging which tends to record what has already happened in the blogger’s real life. In other words, they are actually reporting what they or their friends probably said at a particular context of their social interaction. As Stenström and colleagues (2002) point out, the quotative function of this construction may not be necessarily used for direct quotations of people’s words, though this function is more prominent. There are also cases where the blogger is actually expressing a thought or feeling, as can be seen from lines 5, 11, 15, and 17 in the following concordance (Figure 8.10).



N	Concordance
1	e looks at the age on my driver's license and is like, "Wow. Wow. You totally do not loo
2	h Dicey is. And everyone, espcially P Craig, is like 'And how stupid is Dicey, because th
3	na deal w/ being away from home. My mom is like "do this! this is ur life!" but honestly i
4	of people dying by drunk drivers. And Sharon is like, what? A whole 105 lbs? Like, serious
5	ther from the gym, I got back yesterday and was like....why the fuck do I not come here
6	. it kinda pissed me off, but whatever. eddie was like "what was that?" haha. it was kin
7	ee feet away from each other. And her friend was like, if anything that right there tells me
8	was, " Dude, I got KoolAid on my shirt!!!" He was like, "Uh, well have a nice day, drive saf
9	m XD and virge was like "nono sign!" and he was like "D: oh ok" and then steph have him
10	cally all we talked about. then at the end, he was like, "oh yeah, hows school?" :D the en
11	and the next call was from a 911 operator. I was like are you kidding me and she started
12	n next thing I knew, Phil jumped on me and I was like 'hi!'. Made me feel special. Also, so
13	get up and she goes "we're we going?" and i was like "away for a bit"...SLICK hahah..not
14	lose to me "and i heard my phone ring so i was like 'hm who is this?' and it way you an
15	tarted asking me a bunch of questions and I was like I got to get out of here. So I left and
16	that my money is falling out of my pocket. i was like, "aw, what a nice old guy." & i saw
17	she was crying, you know like sobbing and i was like oh, ill knock and see if shes ok (ca
18	e wriggled his eyebrows and diego winked. i was like, ew. seriously. theyre lame. but i lo
19	forgot to give us our diet coke" and Kaitlyn was like "No, say we didn't get our large frie
20	day, and this random guy came up to me & was like, "i like your shirt! did you actually w
21	ething was going on cause Jonny and Abbie were like no, you have to come back to the
22	his phone charger was broken & my friends were like "Oh, what a load of crap." & I want
23	e you guys coming so laaate, bbs?" & they were like, "yeah, we knew youd be coming o

Figure 8.10 Concordance lines for LIKE as quotative complementizer

There are 68 occurrences of other collocations such as *it's like* and *it was like*, but none of them are used in the quotative sense as described above. No instances of *like* following reporting verbs *say* or *go* have been identified from the English blog corpus for this research.

8.1.5 Use of archaic morpho-syntactic features

Another phenomenon related to grammatical features is also of interest in this research: the use of archaic morpho-syntactic features. When certain features became archaic, the chance for them to be recycled would be reduced to near zero. The process of language development or language change is actually a process of eliminating obsolete words or usages and replacing them with new words or new meanings. Of course, chances for using obsolete words or even obsolete pronunciation cannot be ruled out in modern society. One typical occasion where obsolete words or usages could be found is in fictions, movies, television plays (or series) with historical themes. The main purpose is to add a flavor of historic authenticity to these literary works. Generally, it is the professional writers who are entitled to do this. With the development of free publishing platforms like blogging, this situation may have started to change. Ordinary people are also using certain obsolete words or word-forms to achieve special effects in their own writing. Many personal bloggers, for instance, intentionally use certain archaic words or inflectional suffixes to achieve special purposes. These words and inflectional suffixes are mainly from the Early Modern English period, as will be discussed in a greater detail later. Using ancient suffixes or ancient ways of saying things in a platform equipped with the most modern technologies is no different from putting these old forms under strong spotlight, which will inevitably create effects which are unachievable otherwise. Of course, not every archaic word is naturally qualified for this purpose. Using a lexical item which has long walked into the remote history of the English language is obviously a not very wise thing to do on a platform where daily vocabulary is the norm and few people will be able to understand it as the case may be. A more reasonable strategy would be to apply some grammatical principles which are characteristic of a particular period of history and yet different enough from the present-day English. This will achieve an effect of being playful or humorous. There are around 56 instances of use of archaic inflectional suffixes or words in the EBC. Most of them involve the concept of inflection - the change of word-forms due to requirements in person, number, time, mood, and voice. They neatly fall into two categories: verbal inflection and the use of personal pronouns.

8.1.5.1 Archaic inflectional forms of verbs

Two kinds of archaic inflectional forms of verbs are present in the EBC: the third person singular suffix *–(e)th* and irregular past tense forms. Figure 8.11 below shows how the archaic third person singular inflectional suffix is used by bloggers.

N	**Concordance**
1	USA :) September 24th, 2007 the weekend endeth. So, I spent most of the weekend wit
2	o be doing! 24 August 2007 I have returneth-eth-ed... I just got back from London last
3	rkness round me close, Songs in the night it giveth. No strom can shake my inmost calm
4	ing FORGIVING and HUMANE today鈥 ell hath no fury like a neurotic, attention-whoring
5	d talking, but the teacher snapped at us. Hell hath no fury like an old conutry woman. We
6	st have forgiven them from the start? Sir Kay hath somethinge to saye Mar. 2nd, 2008 S
7	t tempest round me rears, I know the truth, it liveth. What though the darkness round me
8	. 13th September 2007 The Central Line, It Sucketh Muchly. Took over two hours to get

Figure 8.11 Concordance lines for verbs ending with (E)TH

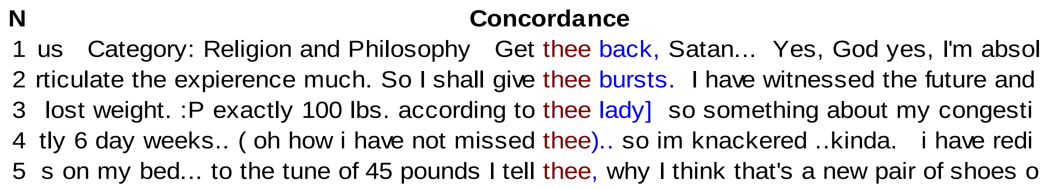
According to Nevalainen (2006, p. 89), Early Modern English verbs typically mark person and number contrast in the second *(-(e)st*) and the third person singular (-*(e)th*/-*s*) as opposed to zero marking in the first person singular and the whole of the plural. At the beginning of the Early Modern period, the verb has seven forms: the base form of the verb, completely unmarked; the second-person singular (-*(e)st*), for concord with thou, the third-person singular (-*(e)th*), progressive form (-*ing*), past form or subjunctive form, past form marking for concord with thou, and past participle (Barber, 1997, p. 164). By the early seventeenth century, the suffix -*(e)th* which was of southern origin had largely been replaced by the northern suffix -*(e)s* in the General dialect although it prevailed in some regional dialects and formal genres much longer. Nevertheless, three verbs *do*, *have*, and *say* were slow to acquire the northern suffix -*(e)s* in the General dialect. *Hath* and *doth* persisted well into the second half of the seventeenth century when -*(e)s* was the regular ending with other verbs (Nevalainen, 2006, pp. 90-91). In fact, during Shakespeare's time, the -*(e)s* ending had started to replace the -*(e)th* ending, but the latter did not die out. According to Crystal (2008), this phenomenon has often been attributed to metrical constraints. The *-eth* ending normally adds an extra unstressed syllable to a word, thus was often exploited by poets for metrical purposes. Barber (1997) holds that the continued use of the suffix *-eth* after about 1590 was actually an example of the conservatism of the written language. He believes that this suffix was probably used in highly formal and solemn speech. Moreover, poets continued to use it for rhythmical purposes. In the writings of the first half of the seventeenth century, it continued to occur quite frequently, mainly in formal styles. The King James Bible invariably used *-eth*, partly because of its dependence on earlier translations, but partly too, no doubt, because *-eth* was more formal and dignified (Barber, 1997, p. 167). It is beyond the scope of this research to present a more detailed description of the emergence and dying out of the suffix *-eth*. The interesting question here is: why do people still use this suffix which had largely left the linguistic stage over 500 years ago? The words taking this suffix are just simple ones like *end*, *return*, *give*, *have*, *live,* and *suck.* Semantically speaking, adding an archaic inflectional suffix would not change the basic meaning of these words except for making them look a bit more special. Morphologically speaking, they are in contrast with the current inflectional suffix *(-s/-es*) for third person singular present tense. Stylistically, they do not really make the blog entries more formal. It is just like someone walks on the street of a metropolitan wearing an ancient costume. However inharmonious with the modern surroundings as it may look, it captures people's attention. For people who understand the history or meaning of that costume, they may marvel at that. This may well be the effect those bloggers want to achieve when they chose to use that special suffix. When a blogger says something like '*the weekend endeth' or 'it sucketh muchly', or 'Sir Kay hath somethinge to saye'* (also note the other two ancient forms *somethinge and saye*), we can see the intentional playfulness behind that special mask. "Being formal in order to be informal" would be a reasonable summary of the strategy the bloggers are actually using. This strategy is unconventional and its major function is pragmatic, though the means is morphological. Of course, it may sound rather absolute to say using the suffix *-eth* in modern text is just for being playful or non-serious. In fact, even in present-day English, there are formulaic phrases or sayings which bear the linguistic fossils and very often these fossils have something to do with religion and music. Recall that the King James Bible uses invariably the suffix *-eth*. For instance, lines 3, 4, 5, and 7 are examples of this kind. The clause '*songs in the night it giveth*' in Line 3 is actually a part of the lyrics of a song entitled "How Can I Keep from Singing." The writer of this song might have borrowed this from the Biblical sentence "But none saith, Where is God, my maker, who giveth songs in the night? (Job 35:10)." It is also likely that both forms of *giveth* and *liveth* have been deliberately used by the song writer for metrical purposes. The phrase in lines 4 and 5 "*hell hath no fury like…*" is a formulaic one with religious origin, nevertheless, the actual usage has no religious meaning as can be told from the words following *like*. In both cases, the bloggers were actually exploiting the part "*hell hath no fury*." Apart from words with suffix *-eth*, there are also a few other irregular verb inflections, as can be seen from *greatened* and *sware* (the past form of swear) in the following two examples:

> (1) George decided to pish off to Campus this morning. Hence Fear has *greatened* :XD (uk_f_15-17.txt).
>
> (2) I *sware*. I don't get Guitar Hero for the wii (us_f_20-24.txt).

Whatever reason it might be, using and understanding of words or phrases with archaic grammatical features presupposes a reasonable amount of knowledge about the history of the language and its cultural heritage. This can also be exploited for identity representation purposes as will be discussed later in Chapter 9.

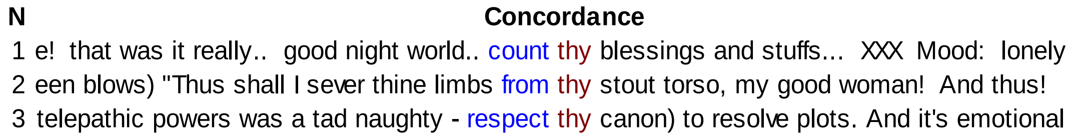
8.1.5.2 Archaic personal pronouns

Another major category which involves the use of archaic forms is personal pronouns. As pointed out by Görlach (1991), present-day English personal nouns are marked for number, case, and in the third person singular for gender. The exception is *you*, which is not even marked for number. The Early Modern English personal pronouns, however, had four forms, each marking a different grammatical aspect. The second person singular pronoun has the following forms: *thou* for the nominative case, *thee* for the accusative case, *thy/thine* for possessive forms. The second person plural forms included: *ye* for the nominative, *you* for the accusative, and *your* for the possessive. In the course of the Early Modern period, *you* became the normal form for both nominative and accusative, and *ye* became just a minor variant. By the end of the seventeenth century *ye* in stressed position had fallen out of use except as a literary archaism (Barber, 1997, p. 149). Figures 8.12 to 8.13 show how these archaic forms of personal pronouns have been used by bloggers. Basically, the respective meaning of these forms is not different from what they were around 500 years ago. Again, the question is: why are they here in the personal blogs?



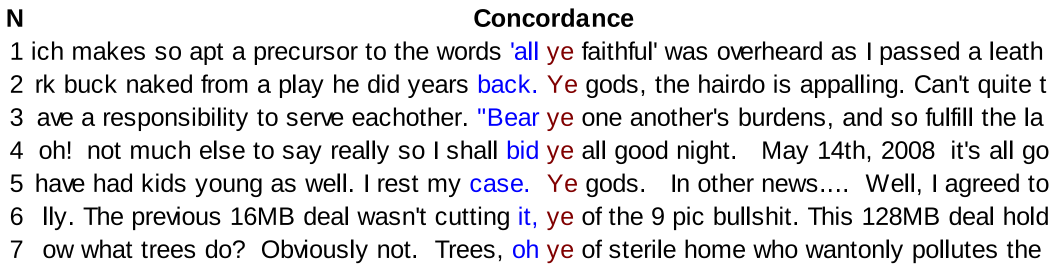
```
N                                   Concordance
1  us   Category: Religion and Philosophy   Get thee back, Satan...  Yes, God yes, I'm absol
2 rticulate the expierence much. So I shall give thee bursts.  I have witnessed the future and
3  lost weight. :P exactly 100 lbs. according to thee lady]  so something about my congesti
4  tly 6 day weeks.. ( oh how i have not missed thee).. so im knackered ..kinda.   i have redi
5  s on my bed... to the tune of 45 pounds I tell thee, why I think that's a new pair of shoes o
```

Figure 8.12 Concordance lines for THEE



```
N                                   Concordance
1 e!  that was it really..  good night world.. count thy blessings and stuffs...  XXX  Mood:  lonely
2 een blows) "Thus shall I sever thine limbs from thy stout torso, my good woman!  And thus!
3 telepathic powers was a tad naughty - respect thy canon) to resolve plots. And it's emotional
```

Figure 8.13 Concordance lines for THY



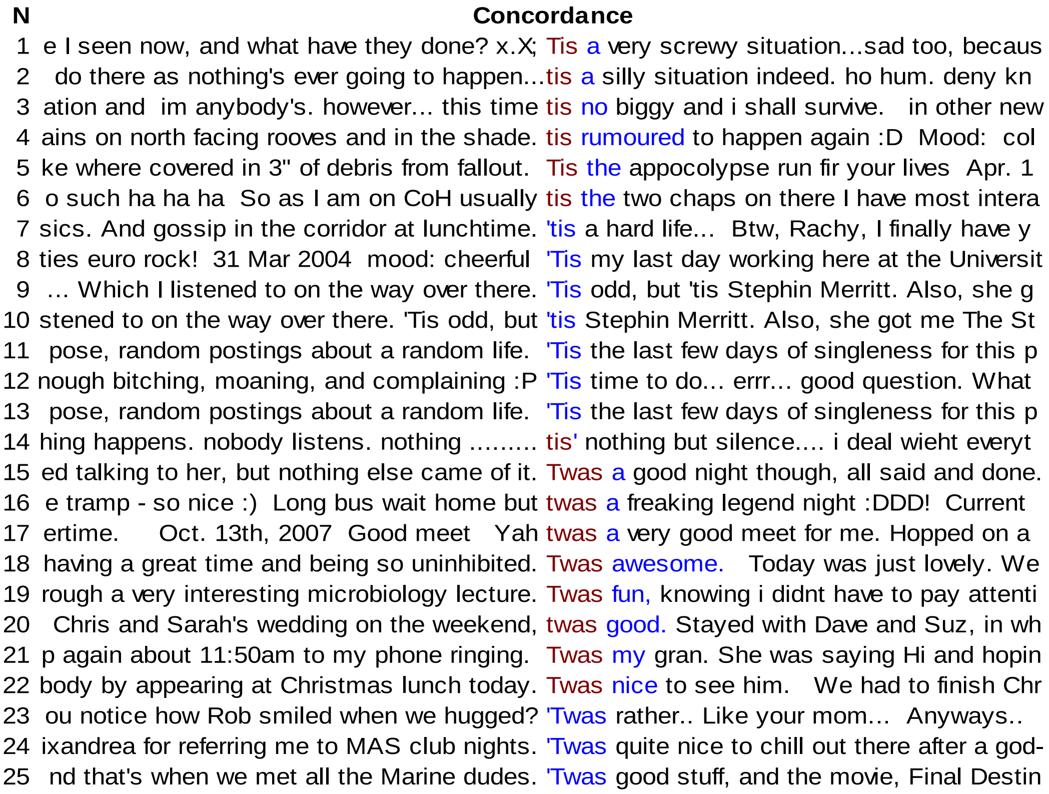
```
N                             Concordance
1 ich makes so apt a precursor to the words 'all ye faithful' was overheard as I passed a leath
2 rk buck naked from a play he did years back. Ye gods, the hairdo is appalling. Can't quite t
3 ave a responsibility to serve eachother. "Bear ye one another's burdens, and so fulfill the la
4 oh!  not much else to say really so I shall bid ye all good night.   May 14th, 2008  it's all go
5 have had kids young as well. I rest my case.  Ye gods.   In other news....  Well, I agreed to
6 lly. The previous 16MB deal wasn't cutting it, ye of the 9 pic bullshit. This 128MB deal hold
7 ow what trees do?  Obviously not.  Trees, oh ye of sterile home who wantonly pollutes the
```

Figure 8.14 Concordance lines for YE

As we can see from the concordance lines, most of the archaic forms have been used as variants of the modern forms *you* or *your*, singular or plural, without particular semantic difference from the modern forms. There are some special ones which may have their roots in the Bible. One example is the line 1 in Figure 8.12 which contains a sentence "Get thee back, Satan." This may well come from a version of Bible. According to the King James Bible, the original saying is "Then saith Jesus unto him, *Get thee hence, Satan*: for it is written, Thou shalt worship the Lord thy God, and him only shalt thou serve" (Matthew 4:10). The exact phrase the blogger used here might come from Shakespeare's famous tragedy Macbeth. There is one line in this play which says "But get thee back; my soul is too much charged with blood of thine already." Another example is line 2 of Figure 8.13 which carries the sentence "Thus shall I sever thine limbs from thy stout torso, my good woman! And thus! And thus!" This was actually the blogger's quote of what has been engraved on a castle since the seventeenth century, a true reflection of the language of that time. Two more examples are lines 2 and 5 in Figure 8.14 which carry the phrase "ye gods." Unlike the *ye* at the rest of the lines which simply means *you*, '*ye gods*' is a seventeenth-century version of 'oh my god', an expression used to show exclamation at that time.

Of course, the second person personal pronouns are not the only relics which are still being kept in the linguistic repertoire of present-day English speakers. Two other forms have appeared more often than the archaic forms of *you* (i.e., *thee*, *thy,* and *ye*). They are *‘tis* and *‘twas*, both of which have a great deal to do with the impersonal pronoun *it*. According to Barber (1997), the original form of the nominative and accusative of *it* was actually *hit,* which was still in use in the sixteenth century. Starting from Middle English (11th to 15th century), the initial sound */h-/* was regularly omitted in unstressed syllables just like what modern speakers do with *her*, *him,* or *his* in unstressed positions. The loss of the initial sound gave birth to a new form *it* for both stressed and unstressed positions. The disappearance of *hit* as a pronoun took place during the sixteenth century, and by 1600 *it* became the normal form. A further weak form */t/* arisen from *it* in unstressed position (which was often represented orthographically as *‘t*) became very common in the late sixteenth century, which resulted in such forms as *‘tis* (an abbreviated form of *it is*), and *‘twas* (a shortened form of *it was*). The apostrophe indicates that a sound has been omitted from the position it marks. This kind of shortened forms often appeared in literary language (Barber, 1997, p. 150).

There are 25 cases of the use of *‘tis* and *‘twas* in the EBC as Figure 8.15 demonstrates. Of course, there are variant forms for these two terms: some bloggers spelled them with the apostrophe omitted; some put the apostrophe at the final position of the word-form; whereas more than half of the cases were spelled as they were 500 years ago. How these two forms emerged is an interesting topic in historical linguistics but what is more interesting here in this research is why some bloggers chose to use them in their blogs. A closer reading of the concordance lines shows both forms have been used mainly for expressing informality. This can be observed from the colloquial or slangy words and expressions following them, for instance, line 1 (*a very screwy situation*), line 2 (*a silly situation*), line 9 (*odd*), line 16 (*a freaking legend*), line 18 (*awesome*), line 19 (*fun*), lines 22 and 24 (*nice, quite nice*), and line 25 (*good stuff*). Obviously, the formalness which used to be carried by both terms (*'tis and 'twas*) was no longer there. It is quite likely that bloggers who employ these terms are actually trying to achieve a sort of spokenness by mimicking the actual pronunciation of the two phrases *it is* and *it was* in daily speech, just like what the playwrights had been doing to "give an illusion of everyday speech" at the late sixteenth century (Barber, 1997, p. 150).



N	Concordance
1	e I seen now, and what have they done? x.X; Tis a very screwy situation...sad too, becaus
2	do there as nothing's ever going to happen...tis a silly situation indeed. ho hum. deny kn
3	ation and im anybody's. however... this time tis no biggy and i shall survive. in other new
4	ains on north facing rooves and in the shade. tis rumoured to happen again :D Mood: col
5	ke where covered in 3" of debris from fallout. Tis the appocolypse run fir your lives Apr. 1
6	o such ha ha ha So as I am on CoH usually tis the two chaps on there I have most intera
7	sics. And gossip in the corridor at lunchtime. 'tis a hard life... Btw, Rachy, I finally have y
8	ties euro rock! 31 Mar 2004 mood: cheerful 'Tis my last day working here at the Universit
9	... Which I listened to on the way over there. 'Tis odd, but 'tis Stephin Merritt. Also, she g
10	stened to on the way over there. 'Tis odd, but 'tis Stephin Merritt. Also, she got me The St
11	pose, random postings about a random life. 'Tis the last few days of singleness for this p
12	nough bitching, moaning, and complaining :P 'Tis time to do... errr... good question. What
13	pose, random postings about a random life. 'Tis the last few days of singleness for this p
14	hing happens. nobody listens. nothing .... tis' nothing but silence.... i deal wieht everyt
15	ed talking to her, but nothing else came of it. Twas a good night though, all said and done.
16	e tramp - so nice :) Long bus wait home but twas a freaking legend night :DDD! Current
17	ertime. Oct. 13th, 2007 Good meet Yah twas a very good meet for me. Hopped on a
18	having a great time and being so uninhibited. Twas awesome. Today was just lovely. We
19	rough a very interesting microbiology lecture. Twas fun, knowing i didnt have to pay attenti
20	Chris and Sarah's wedding on the weekend, twas good. Stayed with Dave and Suz, in wh
21	p again about 11:50am to my phone ringing. Twas my gran. She was saying Hi and hopin
22	body by appearing at Christmas lunch today. Twas nice to see him. We had to finish Chr
23	ou notice how Rob smiled when we hugged? 'Twas rather.. Like your mom... Anyways..
24	ixandrea for referring me to MAS club nights. 'Twas quite nice to chill out there after a god-
25	nd that's when we met all the Marine dudes. 'Twas good stuff, and the movie, Final Destin

Figure 8.15 Concordance lines for 'TIS and 'TWAS

8.2 Variation in pragmatic features

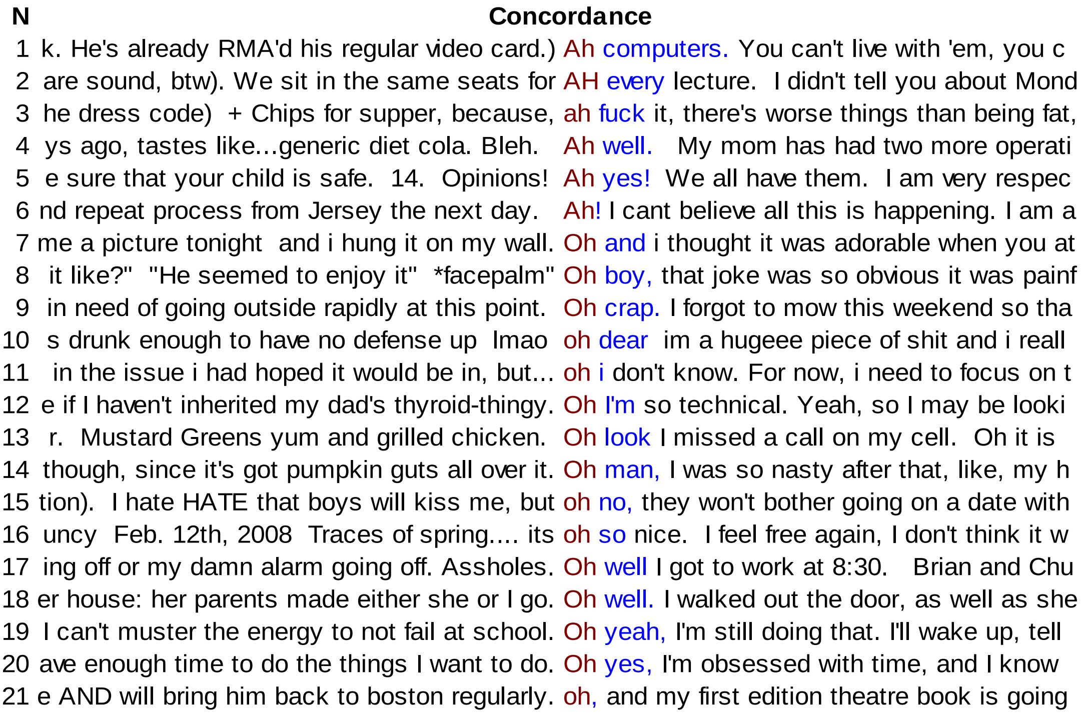
Apart from the lexical (including the orthographic and semantic aspects) and grammatical aspects which have already been described so far, there is another aspect which is also very important for our understanding of the relationship between linguistic variations and identity representation in personal blogs: the pragmatic aspect. Due to the constraints of space, I will only focus on pragmatic features which are related to the use of pragmatic markers. As has already been discussed in Chapter 2, there is no general consensus on what should be categorized as pragmatic markers in existing literature. According to Carter and McCarthy (2006, p. 208), "pragmatic markers are a class of items which operate outside the structural limits of the clause and which encode speakers' intentions and interpersonal meanings." These include: discourse markers (regarding the speaker's intentions concerning discoursal organization, structuring and monitoring), stance markers (concerning the speaker's stance or attitude towards the message), hedges (being less assertive in formulating the message), and interjections (reflecting affective responses and reactions to the discourse). Mainly following this definition, I will report on bloggers' use of three pragmatic features: discourse markers, interjections, and vague language.

8.2.1 Discourse markers

Carter and McCarthy (2006) define discourse markers as "words and phrases which function to link segments of the discourse to one another in ways which reflect choices of monitoring, organization and management exercised by the speaker" (p. 208) According to them, the most frequent discourse markers in everyday informal spoken English are the single-word items *anyway*, *cos*, *fine*, *good*, *great*, *like*, *now*, *oh*, *okay*, *right*, *so*, and *well*, and phrasal and clausal items such as *you know*, *I mean.* (p. 214). Space does not allow me to report on each and every word on this list. What is going to be presented here is bloggers' use of eight markers: *OK, oh* (*ah*), and *yeah* (*yes*). As these markers are typically used in spoken discourse, their presence in personal blogs (a written genre) is itself worth commenting. What is more important, by using these markers, the bloggers can achieve many pragmatic functions which may not be easily achievable otherwise. Considering that the functions of almost all these items have been well-described in existing literature, I am not going to discuss their specific uses unless it is really necessary. More focus will be put on their distributions in the entries contributed by bloggers from different groups and the potential link between the use of such markers and identity representation.

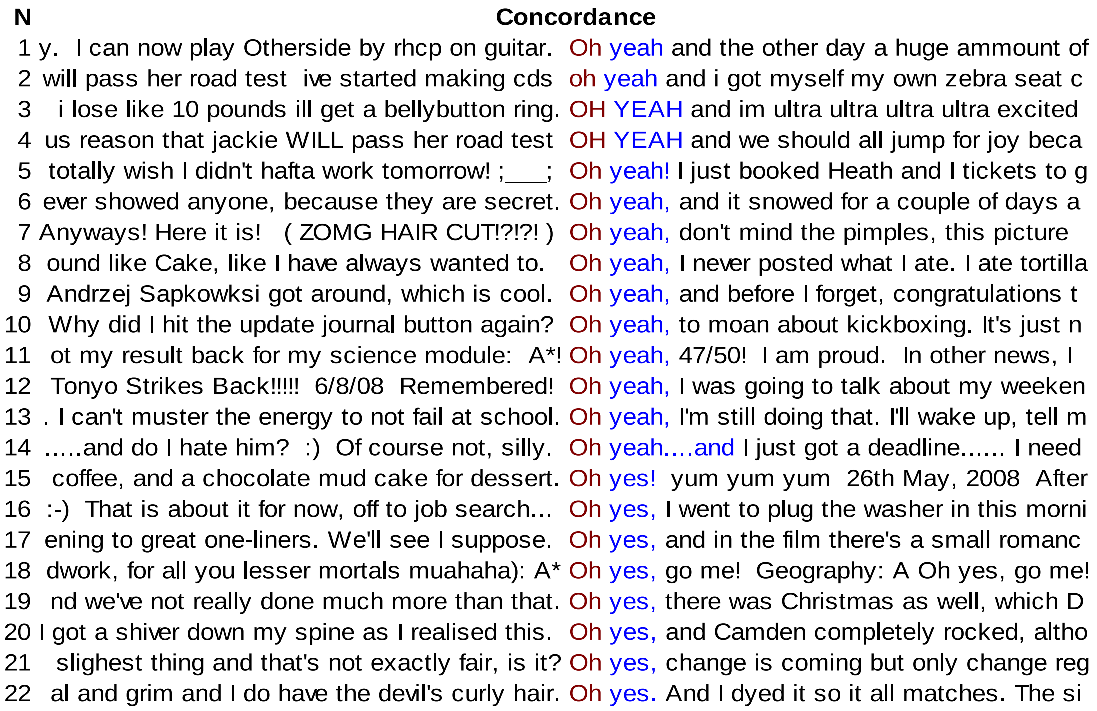
8.2.1.1 Oh (ah)

The core meaning of the marker *oh* and its variant *ah* is to express "surprise." Two very commonly observed uses are: creating an unexpected diversion in the conversation and expressing emotions (for instance being happy, angry or disappointed about something). The former is more of discoursal organization nature and the other is more of pragmatic nature. There are 601 instances of *oh (ah)* in the blog corpus for this research, half of which are cases of *oh (ah)* being used as a stand-alone marker. The other half are cases of *oh (ah)* being used with collocates. The most frequent collocates include: *well* (117 occurrences), *and* (86), *yeah (yes)* (45), *god* (and variants such as *gosh, goodness, lord*) (25), *boy* (or *man*) (13), *dear* (12), and *no* (11). Among the 117 cases of *oh (ah) well*, 57 cases are more of discoursal nature in that they are related to topic diversion or topic expansion. When *oh (ah) well* is used this way, it is often followed by a comma. The rest 61 cases are more of pragmatic nature and are used to express the meaning of "I don't care," a new sense which has recently started to gain currency. When this collocation is used in this sense, it often appears in the form of a stand-alone clause, that is, it will be followed by a full-stop. The second most common collocate of *oh (ah)* is the conjunction *and*. This collocation is almost exclusively used as a discourse marker for topic expansion, though the adding of *oh* or *ah* implies a sense of afterthought. The rest collocates listed above are almost all related to emotion expression. Figure 8.16 shows a flavor of how *oh (ah)* has actually been used by bloggers.



N	Concordance
1	k. He's already RMA'd his regular video card.) Ah computers. You can't live with 'em, you c
2	are sound, btw). We sit in the same seats for AH every lecture. I didn't tell you about Mond
3	he dress code) + Chips for supper, because, ah fuck it, there's worse things than being fat,
4	ys ago, tastes like...generic diet cola. Bleh. Ah well. My mom has had two more operati
5	e sure that your child is safe. 14. Opinions! Ah yes! We all have them. I am very respec
6	nd repeat process from Jersey the next day. Ah! I cant believe all this is happening. I am a
7	me a picture tonight and i hung it on my wall. Oh and i thought it was adorable when you at
8	it like?" "He seemed to enjoy it" *facepalm" Oh boy, that joke was so obvious it was painf
9	in need of going outside rapidly at this point. Oh crap. I forgot to mow this weekend so tha
10	s drunk enough to have no defense up lmao oh dear im a hugeee piece of shit and i reall
11	in the issue i had hoped it would be in, but... oh i don't know. For now, i need to focus on t
12	e if I haven't inherited my dad's thyroid-thingy. Oh I'm so technical. Yeah, so I may be looki
13	r. Mustard Greens yum and grilled chicken. Oh look I missed a call on my cell. Oh it is
14	though, since it's got pumpkin guts all over it. Oh man, I was so nasty after that, like, my h
15	tion). I hate HATE that boys will kiss me, but oh no, they won't bother going on a date with
16	uncy Feb. 12th, 2008 Traces of spring.... its oh so nice. I feel free again, I don't think it w
17	ing off or my damn alarm going off. Assholes. Oh well I got to work at 8:30. Brian and Chu
18	er house: her parents made either she or I go. Oh well. I walked out the door, as well as she
19	I can't muster the energy to not fail at school. Oh yeah, I'm still doing that. I'll wake up, tell
20	ave enough time to do the things I want to do. Oh yes, I'm obsessed with time, and I know
21	e AND will bring him back to boston regularly. oh, and my first edition theatre book is going

Figure 8.16 Concordance lines for OH/AH



```
 N                                    Concordance
 1 y.  I can now play Otherside by rhcp on guitar.  Oh yeah and the other day a huge ammount of
 2  will pass her road test  ive started making cds  oh yeah and i got myself my own zebra seat c
 3   i lose like 10 pounds ill get a bellybutton ring. OH YEAH and im ultra ultra ultra ultra excited
 4  us reason that jackie WILL pass her road test  OH YEAH and we should all jump for joy beca
 5  totally wish I didn't hafta work tomorrow! ;___;  Oh yeah! I just booked Heath and I tickets to g
 6  ever showed anyone, because they are secret. Oh yeah, and it snowed for a couple of days a
 7 Anyways! Here it is!   ( ZOMG HAIR CUT!?!?! )  Oh yeah, don't mind the pimples, this picture
 8  ound like Cake, like I have always wanted to.  Oh yeah, I never posted what I ate. I ate tortilla
 9  Andrzej Sapkowksi got around, which is cool.  Oh yeah, and before I forget, congratulations t
10  Why did I hit the update journal button again? Oh yeah, to moan about kickboxing. It's just n
11  ot my result back for my science module:  A*! Oh yeah, 47/50!  I am proud.  In other news, I
12  Tonyo Strikes Back!!!!!  6/8/08  Remembered!  Oh yeah, I was going to talk about my weeken
13 . I can't muster the energy to not fail at school. Oh yeah, I'm still doing that. I'll wake up, tell m
14 .....and do I hate him?  :)  Of course not, silly.  Oh yeah....and I just got a deadline...... I need
15  coffee, and a chocolate mud cake for dessert. Oh yes!  yum yum yum  26th May, 2008  After
16 :-)  That is about it for now, off to job search...  Oh yes, I went to plug the washer in this morni
17  ening to great one-liners. We'll see I suppose.  Oh yes, and in the film there's a small romanc
18  dwork, for all you lesser mortals muahaha): A* Oh yes, go me!  Geography: A Oh yes, go me!
19  nd we've not really done much more than that. Oh yes, there was Christmas as well, which D
20 I got a shiver down my spine as I realised this.  Oh yes, and Camden completely rocked, altho
21  slighest thing and that's not exactly fair, is it? Oh yes, change is coming but only change reg
22  al and grim and I do have the devil's curly hair. Oh yes. And I dyed it so it all matches. The si
```

Figure 8.17 Concordance lines for OH YEAH (YES)

There are 44 instances of *oh* collocating with *yeah* or *yes*, a sample of which is presented in Figure 8.17. The collocation *oh yeah* (*yes*) is mainly used to achieve two functions: signaling that what follows escaped from the blogger's mind for some reason but gets remembered (see lines 1-2, 7, 10, 16, and 22) and expressing excitement or happiness (see lines 3-4, 11, 15, and 18).

If we take a closer look at the distribution of the use of *oh (ah)* among blogger groups, we will find that younger bloggers (those aged below 25) use this particle more often than older bloggers (those aged from 25 to 40), with the former accounting for 63% of the total occurrences whereas the latter taking up around 37%. Female bloggers seem to use this particle more often than male bloggers. On the whole, female bloggers have contributed around 61% of the total occurrences whereas the male ones only account for 39%. Table 8.1 lists the details.

Table 8.1 Distribution of OH/AH across groups

Age Group	British bloggers			American bloggers		
	Male	Female	Subtotal	Male	Female	Subtotal
15-17	n/a	49	49	51	40	91
18-19	19	59	78	12	25	37
20-24	27	33	60	32	33	65
25-29	21	18	39	16	20	36
30-34	13	19	32	20	20	40
35-40	16	17	33	10	31	41
Total	96	195	291	141	169	310

8.2.1.2 Ok (okay)

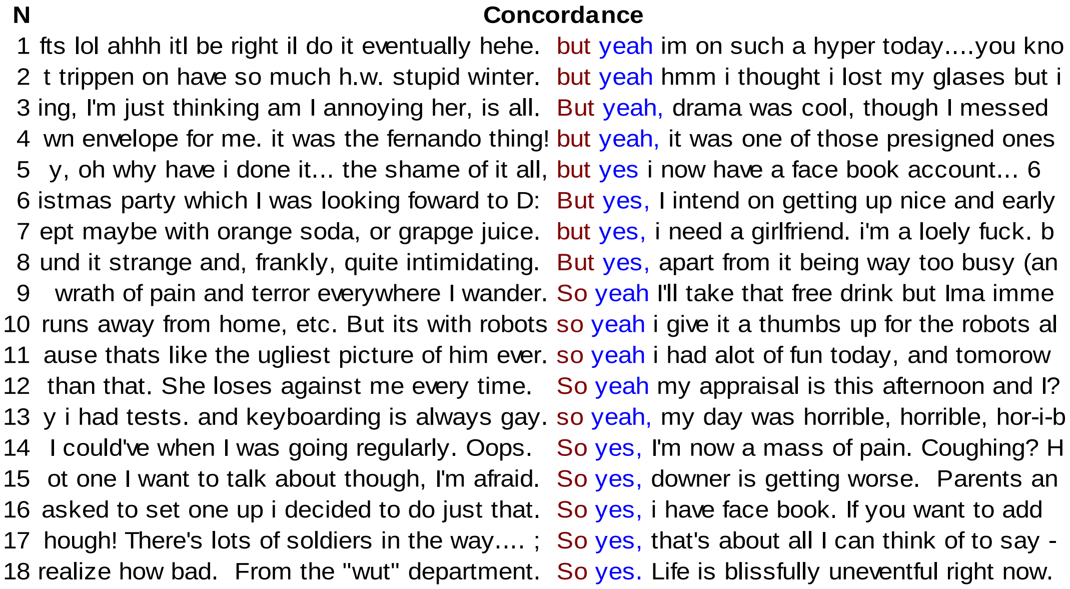
The particle *ok* and its variant *okay* are extremely common in spoken discourse. There are 246 cases of *ok* used as a marker to achieve discourse and pragmatic functions. Table 8.2 lists the distribution of *ok* (*okay*) across different blogger groups. From this table we can see that this marker is more frequently used by female bloggers. Among the total occurrences 160 are from female bloggers, accounting for around 69%; male bloggers have only contributed 31% of the occurrences. Age-wise, bloggers from the younger groups outnumber those from the older groups in their use of *ok or okay* if we do not take the region variable into account, with the former contributing 55% and the latter 45% of the total occurrences. If we look at the British and American bloggers separately, we will see a different picture. The pattern that female bloggers outnumber their male counters can be observed in both regional groups. For the British group, female bloggers outnumbered male ones by 60% to 40% in terms of percentage of total occurrences. For the American group, on the other hand, females outnumber males by 69% to 31%. Age-wise, however, the pattern is quite different. British bloggers from the younger age groups (those below 25) have contributed twice as many instances as bloggers from older age groups (those aged from 25 to 40). The pattern for the American bloggers, however, is reversed: bloggers from the older age groups (53%) outnumber those from the younger groups (47%). Considering the American origin of the particle *ok (or okay)* and its status as an icon of Americanism, it makes sense to observe younger British bloggers tend to use it more often than bloggers from the older age groups. As mentioned elsewhere in this chapter, young British people seem to be more willing to identify with American English in certain aspects.

Table 8.2 Distribution of OK/OKAY across groups

Age Group	British Bloggers			American Bloggers		
	Male	Female	Subtotal	Male	Female	Subtotal
15-17	n/a	17	17	11	9	20
18-19	11	18	29	3	18	21
20-24	14	9	23	15	11	26
25-29	6	4	10	10	16	26
30-34	4	2	6	2	18	20
35-40	6	11	17	4	27	31
Total	41	61	102	45	99	144

8.2.1.3 Yeah (yes)

According to Biber et al. (1999, p. 1089), *yeah* and *yes* are two typical response forms which are used as brief and routinized responses to a previous remark by a different speaker. *Yeah* is treated as canonical in conversation English, where it is considerably more frequent than *yes*. What is interesting here in this research is that blog entries are not conversation, so why are bloggers still using such responses forms? As mentioned earlier, *yeah and yes* can collocate with *oh* to express excitement and mention a topic which temporarily escaped the blogger’s mind and there are around 44 occurrences of such use. In fact, *yeah* and *yes* are often used to collocate with two other words *so* and *but*, to express new meanings which are not achievable when they are used alone. *So yeah* (*yes*) can be used both to conclude remarks and to initiate remarks. As a way of concluding a statement, it is used when relating a past event and teller is unsure or too lazy to think of a good way to conclude. It can also be used when it is assumed that nothing more can be said to adequately explain what is happening, or when the user just feels lazy or embarrassed about what is being said. As an initiator of remarks, it often appears at the initial position of a paragraph and the remarks following it are actually a sort of conclusion of what has already been talked about in the preceding utterances. The collocation *but yeah* (*yes*) performs similar functions yet the meaning is slightly different. In informal speech, *but yeah* is often used as a silent gap filler when the speaker does not know what to say next. Here in the blog corpus, *but yeah* is often used for a change of topic. There are 97 instances of *so yeah* (*yes*) and *but yeah* (*yes*) in the corpus, a sample of which is presented below in Figure 8.18.



N	Concordance
1	fts lol ahhh itl be right il do it eventually hehe. but yeah im on such a hyper today....you kno
2	t trippen on have so much h.w. stupid winter. but yeah hmm i thought i lost my glases but i
3	ing, I'm just thinking am I annoying her, is all. But yeah, drama was cool, though I messed
4	wn envelope for me. it was the fernando thing! but yeah, it was one of those presigned ones
5	y, oh why have i done it... the shame of it all, but yes i now have a face book account... 6
6	istmas party which I was looking foward to D: But yes, I intend on getting up nice and early
7	ept maybe with orange soda, or grapge juice. but yes, i need a girlfriend. i'm a loely fuck. b
8	und it strange and, frankly, quite intimidating. But yes, apart from it being way too busy (an
9	wrath of pain and terror everywhere I wander. So yeah I'll take that free drink but Ima imme
10	runs away from home, etc. But its with robots so yeah i give it a thumbs up for the robots al
11	ause thats like the ugliest picture of him ever. so yeah i had alot of fun today, and tomorow
12	than that. She loses against me every time. So yeah my appraisal is this afternoon and I?
13	y i had tests. and keyboarding is always gay. so yeah, my day was horrible, horrible, hor-i-b
14	I could've when I was going regularly. Oops. So yes, I'm now a mass of pain. Coughing? H
15	ot one I want to talk about though, I'm afraid. So yes, downer is getting worse. Parents an
16	asked to set one up i decided to do just that. So yes, i have face book. If you want to add
17	hough! There's lots of soldiers in the way.... ; So yes, that's about all I can think of to say -
18	realize how bad. From the "wut" department. So yes. Life is blissfully uneventful right now.

Figure 8.18 Concordance lines for BUT/SO YEAH (YES)

Apart from collocating with other words to express new discourse or pragmatic meanings, *yeah* and *yes* are more frequently used alone to express other discourse or pragmatic meanings. There are 443 instances of *yeah* and *yes* used alone. Whatever specific meaning these two words are used to convey, their presence in a written discourse inevitably increases the interactive nature of the discourse and makes the resultant discourse more like talking. Table 8.3 summarizes the distribution of *yeah* and *yes* across different blogger groups for this research. From this table we can observe that younger bloggers (those aged from 15 to 24) tend to use *yeah* and *yes* more often than older bloggers (those aged above 25), with the former contributed 61% of the total occurrences and the latter 39%. Gender-wise, female bloggers outnumber male bloggers: the former have contributed 58% of the occurrences and the latter 42%.

Table 8.3 Distribution of YEAH/YES across groups

Age Group	British Bloggers			American Bloggers		
	Male	Female	Subtotal	Male	Female	Subtotal
15-17	n/a	45	45	37	38	75
18-19	31	42	73	21	32	53
20-24	29	21	50	32	26	58
25-29	16	27	43	21	30	51
30-34	20	7	27	23	24	47
35-40	9	18	27	5	30	35
Total	105	160	265	139	180	319

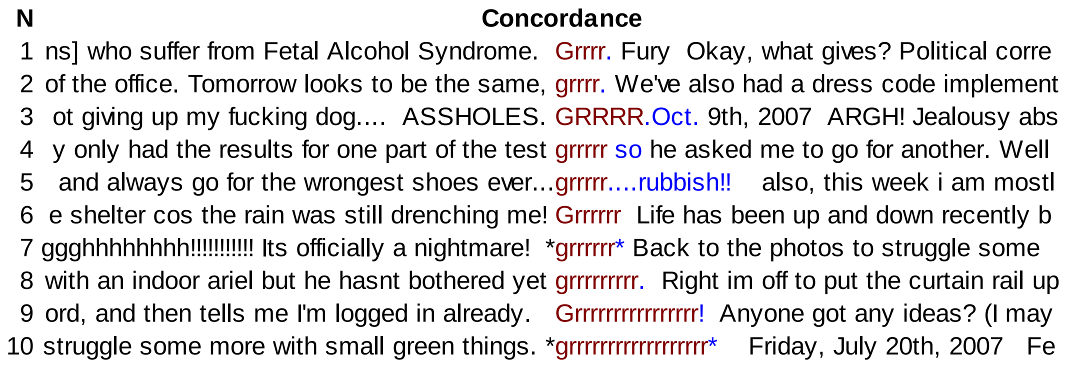
8.2.2 Interjections

According to Carter and McCarthy (2006), interjections are exclamative utterances used to “express positive or negative emotional reactions to what is being or has been said or to something in the situation.” They are “especially common in spoken language and *rare in writing except in written representations of speech*” (p. 224 ; my italics). Blogging is a written genre yet a consultation of the word list for the EBC reveals that interjections are not “rare.” Table 8.8 lists the interjections appeared in the blog corpus and their raw frequencies. This list is by no means exhaustive but it has captured most of the interjections in the corpus. One frequently used interjection, *oh*, is not included in the list because it is more commonly used as a discourse marker and it has already been discussed in the previous section.

Table 8.4 List of interjections in the blog corpus

Item	FRQ	Item	FRQ	Item	FRQ	Item	FRQ
yay	139	ow	13	ooo	5	jeh	1
*ugh (argh, urgh)	102	goodness	12	yipee	3	lawks	1
god	89	ew	12	pft	3	meep	1
wow	82	gosh	11	rawr	2	nomnomnom	1
*blah(bleh)	56	oops	11	jeebers	2	ooer	1
grr	52	phew	11	byah	1	psssht	1
*gah	22	*geez (jeez)	11	holey moley	1	wahey	1
woohoo	19	doh(duh)	10	hrm	1	yikes	1
hooray	16	ouch	9	hurr	1	yowza	1
woo	14	squee	9	hurumph	1	Total	729
Items marked with * have variant spelling forms other than those listed in the brackets.							

From Table 8.4 we can see that interjections are not rare in terms of both tokens and types in the EBC. If we take a look at the ten most frequently used interjections, we will soon find that all these words are closely related to emotion expression. *Yay*, for instance, is often used as an exclamation of pleasure, approval, elation, or victory. *Ugh* (also in the forms of *argh*, *aargh*, and *urgh*) is used to show displeasures or disgust. *Gah* and its variants *gargh* and *guh* are used to denote frustration and/or excitement. *Wow* is used to express wonder, amazement, or great pleasure. *Blah* and its various spelling forms (*bleh*, *blech*, *blargh*, *bleargh*, *blegh*, *blergh*, *bleugh*) are used to expression frustration or depression. *Grr* is usually used to indicate anger or frustration. *Hooray* is used to express delight and excitement). *Woohoo* is another term for showing excitement. Almost all of these interjections can be spelled with one or more letter repeated to strengthen the intensity of emotional expression. For instance, the term *grr* can be spelled as having as many r's as the blogger feels like; the more repeated letters, the stronger the term is intended to be (See Figure 8.19 below for a flavor of that).



```
 N                                    Concordance
 1  ns] who suffer from Fetal Alcohol Syndrome.  Grrrr. Fury  Okay, what gives? Political corre
 2  of the office. Tomorrow looks to be the same, grrrr. We've also had a dress code implement
 3   ot giving up my fucking dog....  ASSHOLES. GRRRR.Oct. 9th, 2007  ARGH! Jealousy abs
 4   y only had the results for one part of the test grrrrr so he asked me to go for another. Well
 5    and always go for the wrongest shoes ever...grrrrr....rubbish!!    also, this week i am mostl
 6   e shelter cos the rain was still drenching me! Grrrrrr  Life has been up and down recently b
 7 ggghhhhhhhh!!!!!!!!!!! Its officially a nightmare!  *grrrrrr* Back to the photos to struggle some
 8  with an indoor ariel but he hasnt bothered yet grrrrrrrrr.  Right im off to put the curtain rail up
 9  ord, and then tells me I'm logged in already.   Grrrrrrrrrrrrrrrr!  Anyone got any ideas? (I may
10  struggle some more with small green things. *grrrrrrrrrrrrrrrrrr*    Friday, July 20th, 2007   Fe
```

Figure 8.19 Concordance lines for GRR

Another observation we can make from Table 8.4 is that bloggers are quite innovative in their invention of new interjections. Quite a number of interjections used by bloggers cannot be found in conventional dictionaries but they can be found in Urbandictionary.com (UD for short), an online slang dictionary compiled by netizens. For a better view of these items and their possible meanings, I list them in the following table (Table 8.5).

Table 8.5 New interjections

Item	Meaning	FRQ	Info Source
woohoo	showing excitement	13	UD
ew (eeew)	disgusting or not-good; or cool/awesome	12	UD
squee	showing excitement	9	UD
geez	Jesus	7	UD
pft (pfftt, pffft)	showing shock, surprise, disgust, or anger	3	UD
jeebers (jeheebers)	Jesus	2	UD
rawr	expressing personal feelings	2	UD
byah	showing excitement	1	UD
holey moley	expression of surprise	1	UD
hrm	similar to hmm	1	UD
hurr	oh, so obvious	1	UD
hurumph	an expression of frustration and despair	1	UD
jeh	yeah or yeh	1	UD
lawks	an expression of surprise	1	UD
meep	ouch or uh oh	1	UD
nomnomnom	sound made when eating something	1	UD
ooer	wow	1	UD
psssht	showing disgust, aggravation or disbelief	1	UD
wahey	an expression of surprise, and/or of joy.	1	UD
yikes	expressing shock	1	LDCE
yowza	an exclamation of surprise	1	UD

Aside from the great difference in terms of interjection tokens used, bloggers from different age groups have also displayed slightly different preference for interjection types, as can be seen from Table 8.6 below.

Table 8.6 Favorite interjections for younger and older bloggers

Bloggers (aged from 15 to 24)		Bloggers (aged from 25 to 40)	
Item	Frequency	Item	Frequency
yay	96	yay	43
*ugh(argh, urgh)	70	grr	33
god	61	*argh (ugh, urgh)	32
wow	51	wow	31
*bleh (blah)	34	god	28
grr	19	*blah (bleh)	22
*gah(guh)	14	woohoo	14
*ew	11	oops	9
woo	11	*gah(guh)	8
gosh	9	hooray	8
hooray	8	ouch	8

The interjection inventory for male and female bloggers from the same age groups does not vary too much as can be seen from Tables 8.7 and 8.8. The major difference lies in the number of tokens. Orthographic differences can also be observed for some interjections. For instance, female bloggers tend to use the spelling *ugh* more whereas the male bloggers from the same age groups prefer the form *argh* though other variants are also used.

Table 8.7 Favorite interjections for bloggers aged below 25

Females (aged from 15 to 24)		Males (aged from 15 to 24)	
Item	Frequency	Item	Frequency
yay	68	yay	28
*ugh(argh, urgh)	52	wow	23
god	41	god	20
wow	28	*argh (ugh, urgh)	18
*blah(bleh)	24	Bleh	10
grr	12	grr	7
*gah(guh)	10	ew	6
woo	10	*gah(guh)	4
gosh	6	phew	4
hooray	6	gosh	3

Table 8.8 Favorite interjections for bloggers aged above 25

Females (aged from 25 to 40)		Males (aged from 25 to 40)	
Item	Frequency	Item	Frequency
yay	31	wow	16
grr	23	yay	12
*ugh(argh, urgh)	21	*argh(ugh, urgh)	11
god	17	god	11
*blah(bleh)	15	grr	10
wow	15	*Blah(bleh)	7
woohoo	9	ouch	5
*gah (guh)	6	woohoo	5
oops	6	hooray	4
goodness	5	phew	4

From what has been presented above, we can see that male and female bloggers have demonstrated different practices in their use of interjections. Female bloggers appear to be more willing to transplant typical oral features into blogging to help fulfill the function of emotion expression. Bloggers from the younger age group have also displayed a similar tendency.

8.2.3 Vague words

The importance of vague language in human communication has been noted by many scholars (e.g. Carter & McCarthy, 2006; Channell, 1994; Crystal & Davy, 1975; Overstreet, 1999; Stenström et al., 2002). According to Stenström et al. (2002), vagueness in language, which is said to be essential for communication to be adequate, is characterized by its close relation to the degree of formality (or more accurately informality) of the situation. The more informal the situation the more vagueness there will be. Carter and McCarthy (2006, p. 202) have also pointed out that "being vague is an important feature of interpersonal meaning and is especially common in everyday conversation." According to them, vague language is often used to perform two important functions. First, it helps soften expressions so that they do not appear "too direct or unduly authoritative and assertive." Second, it is "a strong indication of an assumed shared knowledge and can mark in-group membership." As Stenström and colleagues point out, it is very difficult to arrive at a precise definition of vagueness as there are numerous ways of being vague in language. Vagueness can be expressed by vague words and expressions such as approximators (e.g., *around*, *about*, (-)*ish*, *or so*), vague quantifiers (such as *loads of*), frequency adverbs (such as *seldom*), general extenders (or set markers) (such as *and something, and stuff, and everything*), and placeholders (such as *thing* and *thingy*). It can also be expressed by implicatures. Channell (1994) creates a list of 51 vague expressions. I am not going to follow Channell's full list of vague expressions but rather select a number of such expressions from the list, focusing more on the less conventional (that is, newer and more unconventional) ones and add in words and expressions which are found to be interesting by other researchers such as Overstreet (1999) and Stenström et al. (2002). Table 4.90 lists the major vague words and expressions identified from the blog corpus based on this criterion. One thing which distinguishes this list from Channell's list is the former's inclusion of word forms such as *kinda* and *sorta* which are variants of *kind of* and *sort of* respectively. It is somewhat odd for such forms to be absent from Channell's list and even that of Stenström et al. as both studies are about daily English conversations. One possible reason would be all cases of *kinda/sorta* were transcribed into *kind of* and sort *of*. The items listed in Table 8.9 can be roughly categorized into approximators (*like*, *around, or so,* and *ish*), placeholders (*thing*), set markers or general extenders (*and stuff, or anything, or everything, or whatever, and stuff, and all that*, etc.), and hedges (*kinda/kind of, sorta/sort of*).

Table 8.9 List of vague expressions in the blog corpus

Vague expression	Frequency	Vague expression	Frequency
kinda	183	and everything	32
kind of	155	thingy (thingies)	30
thing	145	and things	20
like	116	sorta	20
or something	106	and shit	19
around	91	and all that	17
or so	84	ish	13
and stuff	68	and the like	11
sort of	67	and that	8
or anything	51	and crap	5
loads of	45	and all that (jazz)	4
or whatever	34	and all that (shit)	2
	Total		1,326

Altogether, 1,326 occurrences of the above-listed vague words and expressions have been identified from the whole blog corpus. Judging from the overall raw frequencies of tokens of vague words and expressions, no significant differences could be found between British and American bloggers, as can be seen from the almost identical overall relative frequency of vague words and expressions (see Table 8.10 below for details).

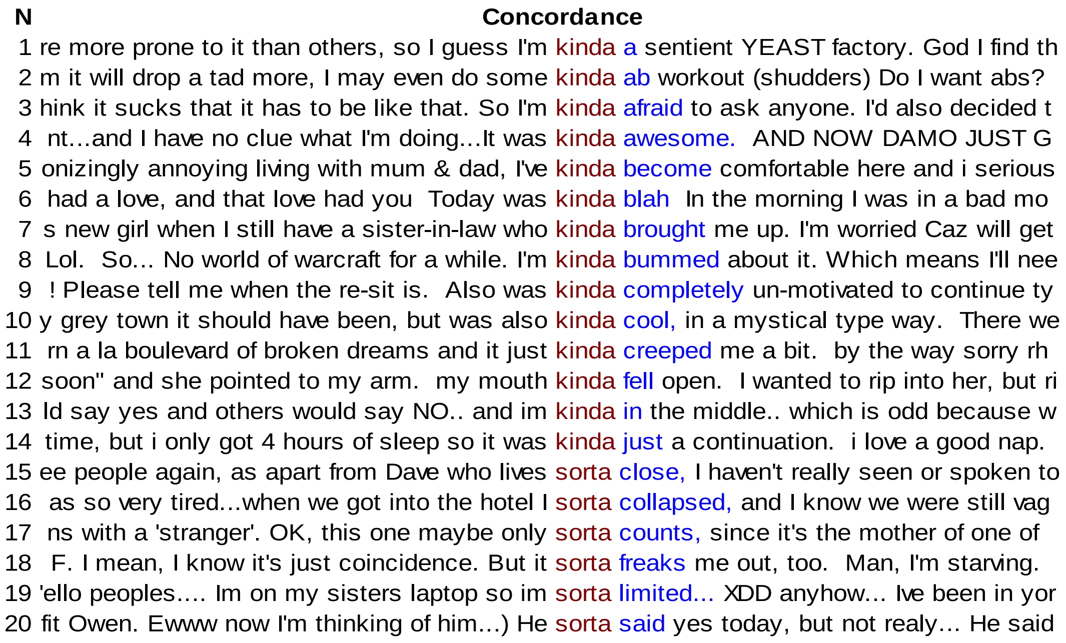
Table 8.10 Overall distribution of vague words across groups

Age Group	Total	British bloggers			American bloggers		
		Male	Female	Subtotal	Male	Female	Subtotal
15-17	268/*36.5*	n/a	106/*39.2*	106/*39.2*	84/*38.6*	78/*31.6*	162/*34.9*
18-19	278/*24.5*	92/*29.8*	70/*20.8*	162/*24.3*	59/*24.6*	57/*25.3*	116/*24.9*
20-24	269/*19.1*	73/*20.4*	62/*18.8*	135/*19.6*	76/*22.2*	58/*15.4*	134/*18.6*
25-29	209/*17.5*	54/*21.1*	43/*14.7*	97/*17.6*	56/*19.3*	56/*15.8*	112/*17.3*
30-34	168/*14*	54/*18.4*	22/*7.9*	76/*13.3*	51/*16.5*	41/*13.1*	92/*14.8*
35-40	134/*11.1*	26/*10.2*	39/*12.5*	65/*11.5*	32/*11.7*	37/*10.0*	69/*10.7*
Grand Total	1,326/*19.3*	299/*20*	342/18.8	641/*19.3*	358/*21.4*	327/*17.3*	685/*19.2*

*Numbers in italics are relative frequency per 10k words

The overall gender difference is not great, either, with overall relative frequency of the male bloggers (20.8 per ten thousand words) slightly higher than that of the female bloggers (18.03 per ten thousand words). Age-wise, the overall pattern is that the relative frequency decreases with the increase of age, meaning that the younger the bloggers the more vague words and expressions they tend to use. The teens groups (the 15-17 and the 18-19 groups) have used at least twice as many vague words and expressions than the older adult groups (the 30-40 and the 35-40 groups). These two patterns hold for both British bloggers and American bloggers. The significant differences between teens and older adults in terms of vague expression employment seem to be echoing a statement made by Stenström et al. that “in the teenage world it is cool to be vague, and it is cool to demonstrate that one cannot be bothered to be precise” (2002, p. 88). As blogging is a genre where communication is conducted through sharing the blogger’s daily life experiences, the precision or accuracy of the information offered is not the main issue. The information is just a means for an end.

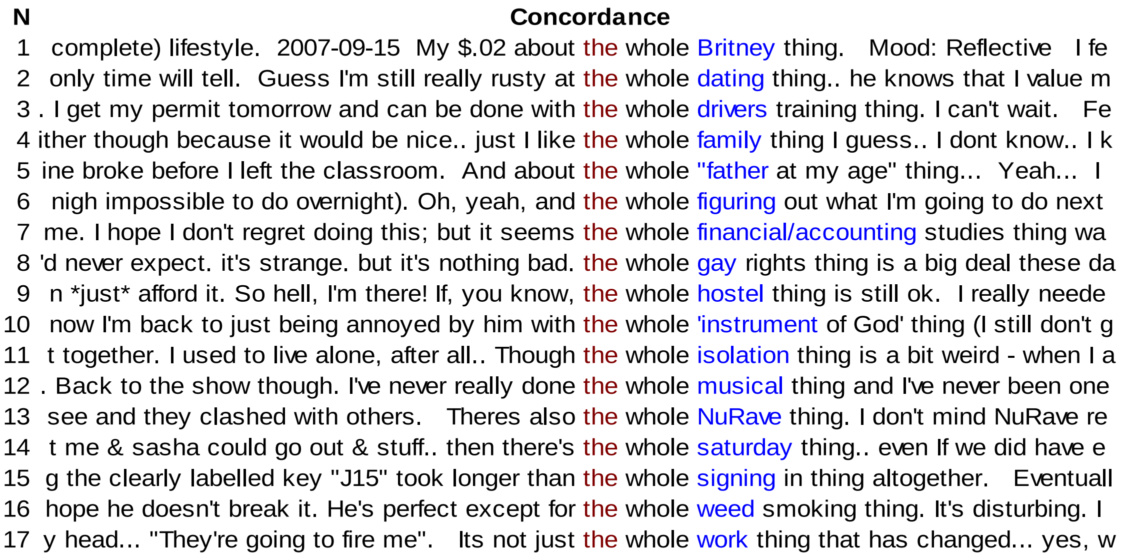
The overall distribution of vague language use reveals certain patterns but it cannot tell us which vague words or expressions are preferred by bloggers from which groups. To get information in this regard, we may need to take a look at how different vague words are actually being used by bloggers. First, let us take a look at the hedges (*kind of/kinda* and *sort of/sorta*). *Kind of* and *sort of* are often used preceding verbs and adjectives to downtone the assertiveness of a segment of discourse. *Kinda* and *sorta* are the phonetic spellings for kind of and sort of. Figure 8.20 shows how these two groups of words were actually being used in the blog corpus.



```
N                                   Concordance
 1 re more prone to it than others, so I guess I'm kinda a sentient YEAST factory. God I find th
 2 m it will drop a tad more, I may even do some kinda ab workout (shudders) Do I want abs?
 3 hink it sucks that it has to be like that. So I'm kinda afraid to ask anyone. I'd also decided t
 4  nt...and I have no clue what I'm doing...It was kinda awesome.  AND NOW DAMO JUST G
 5 onizingly annoying living with mum & dad, I've kinda become comfortable here and i serious
 6  had a love, and that love had you  Today was kinda blah  In the morning I was in a bad mo
 7  s new girl when I still have a sister-in-law who kinda brought me up. I'm worried Caz will get
 8  Lol.  So... No world of warcraft for a while. I'm kinda bummed about it. Which means I'll nee
 9  ! Please tell me when the re-sit is.  Also was kinda completely un-motivated to continue ty
10 y grey town it should have been, but was also kinda cool, in a mystical type way.  There we
11  rn a la boulevard of broken dreams and it just kinda creeped me a bit.  by the way sorry rh
12 soon" and she pointed to my arm.  my mouth kinda fell open.  I wanted to rip into her, but ri
13 Id say yes and others would say NO.. and im kinda in the middle.. which is odd because w
14  time, but i only got 4 hours of sleep so it was kinda just a continuation.  i love a good nap.
15 ee people again, as apart from Dave who lives sorta close, I haven't really seen or spoken to
16  as so very tired...when we got into the hotel I sorta collapsed, and I know we were still vag
17  ns with a 'stranger'. OK, this one maybe only sorta counts, since it's the mother of one of
18  F. I mean, I know it's just coincidence. But it sorta freaks me out, too.  Man, I'm starving.
19 'ello peoples.... Im on my sisters laptop so im sorta limited... XDD anyhow... Ive been in yor
20 fit Owen. Ewww now I'm thinking of him...) He sorta said yes today, but not realy... He said
```

Figure 8.20 Concordance lines for KINDA/SORTA

Sometimes, it is difficult to observe the association between the use of vague expressions and bloggers' identity representation just from the overall distribution across blogger groups. Nevertheless, a particular pattern of the vague expression use may reveal something which may be masked by taking all the occurrences as a whole. One example from this corpus is the use of *thing* as placeholders. Placeholders are expressions used when people cannot remember the name of a person or thing. Such words have little or no semantic meaning and should rather be interpreted pragmatically (Stenström et al., 2002, p. 94). There are 145 occurrences of *thing* used as a placeholder in the corpus, 59% of which come from bloggers aged below 25. Female bloggers have used this sense of the word slightly more often than male bloggers, with the former's average relative frequency (per ten thousand words) slightly greater than that of the latter (2.2: 1.9). No differences could be observed between American bloggers and British bloggers. Nevertheless, a running of concordance lines of this usage reveals a pattern which takes the form of "the whole … thing." Figure 8.21 shows how this pattern works. From these concordance lines we can see that this structure is quite flexible in that the slot between *the whole* and *thing* can be filled with a word (see lines 1 and 2), a phrase (lines 3 and 5), and even a whole clause (line 6).



N	Concordance
1	complete) lifestyle. 2007-09-15 My $.02 about the whole Britney thing. Mood: Reflective I fe
2	only time will tell. Guess I'm still really rusty at the whole dating thing.. he knows that I value m
3	. I get my permit tomorrow and can be done with the whole drivers training thing. I can't wait. Fe
4	ither though because it would be nice.. just I like the whole family thing I guess.. I dont know.. I k
5	ine broke before I left the classroom. And about the whole "father at my age" thing... Yeah... I
6	nigh impossible to do overnight). Oh, yeah, and the whole figuring out what I'm going to do next
7	me. I hope I don't regret doing this; but it seems the whole financial/accounting studies thing wa
8	'd never expect. it's strange. but it's nothing bad. the whole gay rights thing is a big deal these da
9	n *just* afford it. So hell, I'm there! If, you know, the whole hostel thing is still ok. I really neede
10	now I'm back to just being annoyed by him with the whole 'instrument of God' thing (I still don't g
11	t together. I used to live alone, after all.. Though the whole isolation thing is a bit weird - when I a
12	. Back to the show though. I've never really done the whole musical thing and I've never been one
13	see and they clashed with others. Theres also the whole NuRave thing. I don't mind NuRave re
14	t me & sasha could go out & stuff.. then there's the whole saturday thing.. even If we did have e
15	g the clearly labelled key "J15" took longer than the whole signing in thing altogether. Eventuall
16	hope he doesn't break it. He's perfect except for the whole weed smoking thing. It's disturbing. I
17	y head... "They're going to fire me". Its not just the whole work thing that has changed... yes, w

Figure 8.21 Concordance lines for THE WHOLE...THING pattern

There are 17 cases of this pattern in the corpus, of which 65% come from bloggers aged below 25. Taking the 25-29 group into account, the percentage will add up to 78%. That is to say, the placeholder of *thing* in the pattern “the whole … thing” is a better age marker than the word in isolation. No gender difference could be observed. Another item *thingy*, which is actually a variant of the word *thing*, is also a good age marker. There are 30 instances of this word in the corpus and 23 (around 80%) of them are from bloggers aged below 25.

Due to the confinement of space, I am not going to give further details about the actual uses of different vague words and expressions. From what has been presented above, we can see that bloggers’ variation in the use of vague language is often associated with age. The younger the bloggers are the more vague language they are going to use.

8.3 Chapter summary

The grammatical and pragmatic features discussed in this chapter have something in common: almost all of them are directly linked to daily informal speech. By transplanting oral features directly into a writing genre, bloggers, especially younger bloggers, are actually deviating from the conventional writing norms and establishing a new one which they find suitable for publishing their own thoughts and feelings while at the same time communicating with their intended audience. The presence of these features also reveals something about the blogging genre as both a platform of information sharing and a tool for social communication. Meanwhile, the very existence of these features in personal blogs shows that bloggers are actually applying the strategy of bricolage (that is, taking whatever linguistic materials available, be it oral features or written features, formal or informal) in getting their meaning across and their emotions expressed. Moreover, these features are also carriers of bloggers' identities, as they can disclose information about bloggers' age, gender, or country of origin, as we will see in the next chapter.

Chapter 9 Variations and Identity Representation

This chapter discusses the relationship between linguistic variations and bloggers' identity representation, focusing on age, gender, and regional identities.

9.1 Introduction

In the preceding four chapters (Chapters 5 to 8), I have presented a rather detailed description about the variations in different aspects of bloggers' language use, ranging from non-conventional orthographic representations of existing words, creative exploitation of word-formation strategies, use of neologisms pertaining to IT and newly emergent Internet culture, use of slanguage words, preferences for semantic domains, to morpho-syntactic and pragmatic features. What we can conclude from such description is that linguistic variations do exist in bloggers' language use. However, we are yet to explain why these variations exist. People may attribute the presence of these variations to the personal nature of blogging as a genre and the potential that the blog entry could be used as a component of a bigger interactive discourse consisting of the blog entry and comments related to it (which are conversational in nature). To a certain extent, this kind of explanation makes senses. It can explain in part why the language of English blogging is different from conventional English speech and writing, but it could not explain the differences present in the blog entries produced by bloggers from different groups. We need to go beyond the genre perspective and look for other factors which may have contributed considerably to shaping the linguistic variations in personal blogs. The blogosphere as a virtual space gives an illusion that bloggers are living in a world which only exists on the Internet. The seemingly intangible blogging community is actually deeply rooted in the material world where the bloggers come from. The linguistic practices in the blogging community have a great deal to do with who the bloggers are, how old they are, where they are from, what language(s) they speak, and what social roles they assume in the so-called meat space. The only difference is that bloggers wear masks and they enjoy greater freedom of self-expression. In other words, the blog entries reflect bloggers' identities in the material world. Possibilities that people fake (or rather play with) their identities do exist, but being able to fake their identities consistently demands a basic familiarity with how people of a particular identity act and express themselves. That is mostly a marked situation. According to Huffaker and Calvert (2005), teenage bloggers tend to take blogs as an extension of their real life identities rather than a place to pretend. Van Doorn et al. (2007) also find that blog authors tend to present themselves in almost exclusively 'real life' categories thus "leaving no room for the construction of gender identities that bear no relationship to their offline lives" (p. 156). If bloggers do take blogs as an extension of their real life identities, we should be able to identify the link between linguistic variations and social factors such as age, gender, social roles, and regional factors.

9.2 Age-related identity representation

9.2.1 Age and non-conventional orthographic representation of words

As demonstrated in Chapter 5, among all the strategies that bloggers in the EBC have employed to realize linguistic variations, non-conventional orthographic representation of existing words is a major one. As mentioned earlier, orthographic variation (regardless of the strategies involved) is a result of deviating from the established norm represented by conventional writing regulations. The employment and tolerance of deviated forms in blogging has actually become a means for bloggers to represent their own identities. This function of non-conventional orthographic representation of common words is well illustrated from the following remarks of Mark Sebba (2003):

> The symbolic value of deviations thus becomes much greater than it would be if the practice of spelling were not so normative. With the relaxation of norms about swearing, it is no longer possible to shock an audience by using the word *bloody* on the English stage, but it is still possible to offend readers by spelling words 'incorrectly' in print. *Orthography -highly visible, and a part of the physical image of language -is an ideal site for ideological struggle and rebellion of various kinds* (Sebba, 2003, pp. 151-152. My italics.).

Generally speaking, the more orthographically engineered forms (OEFs) a blogger (or blogger group) uses, the more distantly deviated they are from the established writing norm which is expected to be identified with by all members of the speech community. If we recall our discussion about the nature of identity and its development in Chapter 2, we know that people behave differently at different developmental stages of their life, and adolescents and young adults tend to be more rebellious against established social norms. If this statement is true, we can expect a difference in behavior between bloggers from different age groups. That is, the younger the bloggers are, the less compliant with the established writing norm they will be. One way of obtaining such information from the EBC is to calculate the distribution of OEFs (that is, the total number of non-conventional contracted forms, abbreviations, letter repetition words, e-paralinguistic words, misspellings, and phonetic spellings) bloggers have used. By observing the distributions of the total number of OEFs across bloggers from different age groups, we can obtain some insights about whether and to what extent age plays a role in forming the orthographic variations. Table 9.1 shows the distribution of OEFs across different blogger groups.

Table 9.1Distribution of OEFs across age groups

Blogger Group	Raw FRQ	Text Size	Relative FRQ
15-17	1,325	73,479	180.3
18-19	1,689	113,278	149.1
20-24	1,396	140,675	99.2
25-29	950	119,535	79.5
30-34	754	119,672	63.0
35-40	667	120,945	55.1

What we can observe from Table 9.1 is that the younger the group is the more unconventional orthographic representations of existing words they use. The density of unconventional orthographic forms employed by the teens groups (both the mid-teens and the late-teens) is three times that of mature adult groups (the 30-34 group and 35-40 group). This tendency holds for both the British and the American bloggers, as Table 9.2 shows. It is not surprising that the mid-teens group ranks the top and the older adults group locates at the bottom of the list. This seems to be echoing Eckert's (1997) finding that adults tend to be more conservative in their language use than younger age groups, though it is not sure whether this conservatism should be attributed to the pressure for use of standard language at work place as Eckert claims. To put it in a different way, orthographic variation is more closely linked to teenagers.

Table 9.2 Distribution of OEFs across age & regional groups

Age Group	UK	US
15-17	194.1	172.3
18-19	142.0	159.3
20-24	117.8	81.5
25-29	96.4	65.1
30-34	84.1	43.6
35-40	72.8	39.6

What Tables 9.1 and 9.2 have presented is an overall (or rather an aggregated) picture of bloggers' employment of non-conventional orthographic representations. This kind of presentation reveals certain patterns about bloggers' realization of orthographic variations but it runs the risk of masking certain features of individual groups, especially those in terms of the preferred linguistic strategies for realizing orthographic variations. As described in Sections 5.2 to 5.3, there are six major strategies for bloggers to realize orthographic variations. Although all of them can be used to achieve stylistic function of being informal (that is, they can all be used as marker of informality), each strategy actually involves different degrees of manipulation and effort on the part of the bloggers and thus may be preferred by bloggers from different age groups. Among these six categories, four are closely related to bloggers' age. They are: unconventional contracted forms, letter repetition, e-paralinguistic words, and phonetic spellings.

The use of unconventional contracted forms is a strategy which involves arguably the least effort on the part of the bloggers. What they need to do is simply omit the apostrophe. Table 9.3 lists the top five blogger groups which have employed this strategy. From this table, we can see that all the five groups are the teens. The other two teens groups (the female late-teens bloggers) are also among the top nine, with the British female late-teens group ranking the sixth with a normalized frequency of 41.9 and their American counterparts ranking the ninth with a normalized frequency of 39.5. The six groups which have used the fewest non-conventional contracted forms are all aged from 25 to 40 (see Table 9.4).

Table 9.3 Groups using most non-conventional contracted forms

Blogger Group	Frequency	Relative frequency
uk_m_18-19	212	64.1
us_m_18-19	151	63.0
us_f_15-17	144	58.4
uk_f_15-17	142	52.5
us_m_15-17	98	45.0

Table 9.4 Groups using fewest non-conventional contracted forms

Group	Frequency	Relative frequency
us_f_25-29	26	7.3
us_m_25-29	21	7.2
us_m_30-34	10	3.2
us_f_30-34	3	1.0
us_f_35-40	3	0.8
us_m_35-40	1	0.4

Bloggers' use of words with non-conventional letter repetition also seems to be related to their age. As pointed out in Section 5.2.3, word-forms created out of non-conventional repetition do not follow any fixed pattern, suggesting an impromptu and playful nature aside from the intended accentuation of a particular word. Table 9.5 and Table 9.6 list the blogger groups which have used most letter repetition words and those that have used fewest such words respectively.

Table 9.5 Groups using most letter repetition words

Blogger Group	Frequency	Relative frequency
us_f_15-17	106	43.0
uk_f_15-17	87	32.2
us_f_18-19	59	26.2
us_m_15-17	52	23.9
uk_f_18-19	45	13.4

From Table 9.5 we can see that the teens groups, again, take all the top five positions and the female bloggers are the dominant ones. Table 9.6 lists the ten blogger groups with the lowest letter repetition frequencies. These ten groups cover all the four groups aged from 35 to 40 (the oldest groups), the two male groups aged between 30 and 34, the two male groups aged between 20 and 24, one male group and one female group from the age group of 25-29. The frequency for all these groups is below four occurrences per ten thousand words.

Table 9.6 Groups using fewest letter repetition words

Blogger Group	Frequency	Relative frequency
uk_f_35-40	12	3.8
uk_m_25-29	9	3.5
uk_m_30-34	10	3.4
uk_m_35-40	8	3.1
us_f_35-40	11	3.0
us_m_20-24	9	2.6
us_m_35-40	5	1.8
us_f_25-29	6	1.7
uk_m_20-24	6	1.7
us_m_30-34	5	1.6

The third feature which is found to be related to blogger age is the use of e-paralinguistic words. The so-called e-paralinguistic words are actually word-forms that bloggers employed to mimic laughter in textual means. Just like letter repetition words (some of them are actually used to mimic lasting laughter), e-paralinguistic words are deliberate efforts in infusing oral discourse features into a written genre. To a certain extent, employing this strategy implies a more distant deviation from the established norm of conventional writing. Table 9.7 lists the top five groups which have used the greatest number of e-paralinguistic words. This list is quite similar to that of Table 9.5, suggesting a close link between the use of e-paralinguistic words and teenagers, especially female ones.

Table 9.7 Groups using most e-paralinguistic words

Blogger Group	Frequency	Relative frequency
us_m_15-17	38	13.9
uk_f_15-17	46	13.7
us_f_18-19	38	12.9
us_f_15-17	32	12.6

Table 9.8, on the other hand, just shows a reverse relationship between the frequency of e-paralinguistic word-forms and the increase of age. All the five groups on the table fall within the age range from 30 to 40.

Table 9.8 Groups using fewest e-paralinguistic words

Blogger Group	Frequency	Relative frequency
uk_f_30-34	7	2.6
uk_m_35-40	7	2.5
uk_m_30-34	6	2.3
us_m_35-40	4	1.1
us_m_30-34	3	1.0

The use of phonetic spellings is also closely related to blogger age. Table 9.9 lists the top six groups with highest frequency of phonetic spellings. From this list we can see that once again the mid-teens are among the top. Although the top position is taken by the British female early adult group (the British female 20-24 group), but that is because one of the bloggers in this group used an overwhelmingly amount of phonetic spellings in her entries. In other words, this is an extreme case. If we exclude this extreme case and look at the rest five groups on the list, we will notice the dominance of teenage bloggers.

Table 9.9 Groups using most phonetic spellings

Blogger Group	Frequency	Relative frequency
uk_f_20-24	211	88.0
uk_m_30-34	66	25.8
uk_f_15-17	83	24.6
us_m_15-17	67	24.6
us_f_18-19	72	24.5
us_f_15-17	62	24.4

Table 9.10 lists the four groups with the lowest frequency of phonetic spellings. The normalized frequencies of these groups are all below eight. No conclusive remarks can be made from this table, although we can still see the shadow of age behind bloggers' employment of phonetic spellings.

Table 9.10 Groups using fewest phonetic spellings

Blogger Group	Frequency	Relative frequency
us_m_30-34	23	7.3
us_m_25-29	22	7.1
uk_f_35-40	11	5.1
us_m_35-40	16	4.3

9.2.2 Age and emergent Internet culture

According to Bucholtz (2000), youth culture is often taken as a resource for teenagers and young adults to draw on in the construction and display of their identities, due to its diverse and rapidly changing stylistic practices. Identifying oneself with celebrities or other sources of fandom is an important part of the emergent Internet culture. As described in Section 6.6, quite a number of neologisms originated from emergent culture on the Internet. From the distribution of these words among texts produced by different bloggers, we can obtain some clues about their age-related identity. Among the 50 tokens of neologisms related to "fan culture" (see Table 6.10), 86% were contributed by bloggers aged below 25, with the mid-teens accounting for 36% of the tokens, the late-teens 16%, and the early adults (the 20-24 group) 34%. Young people, especially the teens and young adults, are normally the creators and the supporters of subculture. By subculture, it means something which has not yet been recognized by the main stream culture and something which is unconventional and deviant from the established norms. Embracing subculture is again an important marker of youth identity.

9.2.3 Age and the use of slanguage in blogging

As mentioned in Chapter 2, many studies have indicated that slanguage use is particularly associated with young people. For instance, Finegan (2004) finds that slang is especially popular among teenagers and college students and Bucholtz (2000) contends that slang is the most noticeable linguistic component of youth-based identities. By and large, the distribution of slanguage words in the EBC seems to be associated with age groups: the younger the group age, the more slanguage words (in terms of tokens per ten thousand words) are used. The only exception is the 25-29 group, which has a slightly higher relative frequency of slanguage use than the younger group aged from 20 to 24. Table 9.11 shows the details.

Table 9.11 Slanguage and blogger age (total)

Age Group	Slang Tokens	Per_10K Words	Sample Size
15-17	789	107.4	73,479
18-19	950	83.7	113,278
20-24	1,003	71.3	140,675
25-29	918	76.8	119,535
30-34	751	62.8	119,672
35-40	598	49.4	120,945
Total	5,009	72.6	687,584

If we take a closer look at the relative frequencies of the slanguage words used by different age groups, we will soon find striking differences between the teens groups (the 15-17 group and the 18-19 group) and the older adult groups (the 30-34 group and the 35-40 group). The difference between the two age groups at the middle ground is very small. The same patterns also hold for British bloggers and American bloggers, as can be seen from Tables 9.12 and 9.13.

Table 9.12 Slanguage and blogger age (UK)

Age Group	Slang Tokens	Per_10k Words	Sample Size
15-17	259	95.7	27,053
18-19	507	75.9	66,764
20-24	476	69.2	68,748
25-29	401	73.0	54,966
30-34	354	61.8	57,309
35-40	233	41.2	56,606
Total	2,230		331,446

Table 9.13 Slanguage and blogger age (US)

Age Group	Slang Tokens	Per_10k Words	Sample Size
15-17	530	114.2	46,426
18-19	443	95.2	46,514
20-24	527	73.3	71,927
25-29	517	80.1	64,569
30-34	397	63.7	62,363
35-40	365	56.7	64,339
Total	2,779		356,138

All these tables show that teenage bloggers tend to use a greater number of slanguage words than the older generation, i.e., bloggers aged from 30 to 40. This observation echoes the findings in existing literature (e.g., Barbieri, 2008; Holmes, 1992, 2001) that the use of swear words and slang is very common among teenagers and young adults, but it will be less frequently observed from the discourse of old people. This is not out of expectation since adolescence is best known as the years when young people start to become aware of, experiment with, and seek their own identities. For adolescents, an important way of trying to be "themselves" is to violate social taboos and use "their own language" as a means of provocation and as a means of keeping the older generation outside, while at the same time strengthening the bonds within their own peer group (Stenström et al., 2002, pp. 67-68). This rise of nonconformity can be seen in the "ad o les cent peak"- the rise in nonstandard language use by teenagers (see Labov, 2001, pp. 101-120), a peak which flattens out as teenagers become older (Kiesling, 2004, p. 299). Meanwhile, adolescence is also a transitional period which is full of excitement, puzzlement, frustration, and rebellion. As Chambers (2003) has pointed out, the transition from childhood to adulthood is often, almost characteristically, accompanied by extremism. The reason is simple: adolescence requires a purposeful divergence from adult norms in favor of alternative norms. Teenage bloggers' preference for slanguage may well be a reflection of their turbulent and hyper-active nature and their needs in expressing strong emotions, be it excitement or anger.

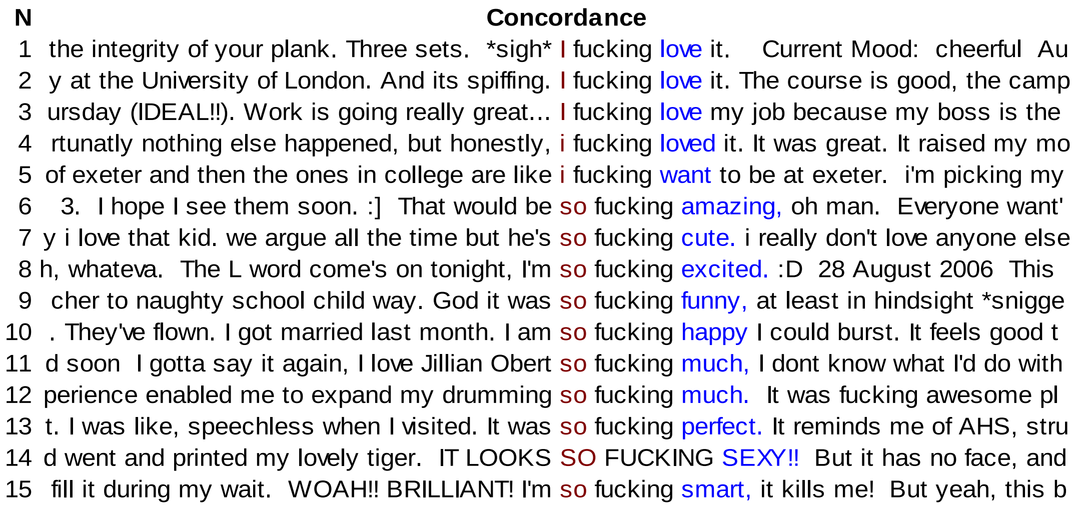
In order to understand why so many slanguage words are used in personal blogs and to what extent they can be regarded as identity markers, we need to examine the use of these words from a more general perspective by looking at their functions. According to Eble (1996), slang can be used to achieve three major functions. First, it moves the discourse towards the informality end along the formal-informal continuum. Second, it identifies members of a group. Third, it opposes established authority (p. 119).

One defining nature of slanguage is its deviation from the established norm, linguistic or otherwise. Looking from the identity representation point of view, the established norm represented by standard, formal speech and writing is actually a collective identity imposed on its members by the society. Slanguage, on the other hand, is a rebellion against the norm for achieving certain effects, for instance, making the language opaque to outsiders or creating shortcuts for the ease of communication. In other words, the use of slanguage is language users' efforts in breaking the formal and semantic constraints imposed by the norm. Moreover, the use of slanguage is basically a spoken phenomenon. Using slanguage in writing will undoubtedly move the written discourse towards the informal end of the continuum. One more factor, which is also very important, is the playfulness sometimes embedded in the slanguage words and phrases. Many a time, language users deliberately choose to deviate from the norm by playing with it. This playfulness is also regarded as an ingredient of informality.

The second function can be viewed as the default function of slang. Slang identifies activities, events and objects that have become routine for those involved, and it has an important function in creating rapport in the work or recreational environment (Allan & Burridge, 2006). Using the same slang as other members of a group could simply be an effort of trying to identify with people who share similar personality, interest or hobbies, or similar experiences. It is quite normal for individuals to want to identify with some people but be different from some others. In other words, slang may well be used as an identity marker, consciously or unconsciously. For example, when teenagers nowadays use “inverted language” (use certain words simply to mean the opposite, e.g., using ‘*wicked*’ to mean ‘*very good, excellent, or cool*’ and ‘*insane*’ for ‘*very good*’), they are actually showing they are different from the older generation. This way of using language is definitely deviant from the established norm of language use but means no harm, thus not antisocial. In fact, slanguage has always been associated with groups, though the popularity of Internet-based communication over the past decade may have changed the kinds of groups with which people may choose to identify and the purposes of doing so. The potential interconnectedness among different groups facilitated by the Internet-based communication may have caused the diminishing of the traditional group-identifying functions of slang for the population at large. Instead, language users may be using slanguage to identify with a style or an attitude rather than a specific group. In other words, for some people the use of slanguage words and expressions may well be a stylistic marker rather than an in-group identity marker in its narrow sense. As blogging is a special genre in that it is both a platform for sharing one’s own experiences with friends and one which is also open for public access (unless the blogger locks it only to friends), the usual practice of sharing and maintaining a constantly changing in-group vocabulary in real life groups (for excluding out-groupers) becomes less important in online situations due to the anonymous nature of the latter. One of the main reasons for the rapid and constant change of slanguage in face-to-face situations is that slanguage users want to fence off the outsiders. In Internet-based communication such as blogging, this purpose can be easily achieved by adopting nicknames and controlling the accessibility of their blog entries. Shielded by this anonymity, the necessity of changing the shibboleth is greatly reduced. Instead, being able to set/start (or at least follow) a linguistic trend becomes more important. Like stylish clothing and modes of popular entertainment, fashionable slanguage is able to gain quick group acceptance. Being fashionable is an important means for young people to seek identity identification and a sense of belonging. Just like the knowledge of electronic games, IT gadgets, hit TV programs, box office movies, and most popular songs which is a sign of social awareness and a part of the identities of particular groups, the mastery of current slang has the same function, especially for young people. This helps explaining why certain new slanguage words can be frequently observed from entries of almost all blogger groups. That said, I do not mean that people no longer use slang for building in-group solidarity. As Eble (1996) rightly points out, "small groups that desire social solidarity – fraternities, dormitories, sports teams – continue to invent and maintain linguistic forms that serve as shibboleths. Slang provides with users with automatic linguistic responses that assign others to either an in crowd or an out crowd" (p. 122). Even in the virtual reality which only exists on the Internet, people flock as groups, after all.

The third function of slanguage is that it can be used to deliberately express irreverence, i.e. to deliberately flout social conventions. Slanguage used to be associated with antiestablishment or antisociety and that is also why some scholars (e.g., Halliday, 1975) categorize it as antilanguage, meaning the "language of antisociety." It is very true that certain social groups do use slangs as in-group recognition devices and purportedly disguise meanings from out-groupers and they do this either because they are conducting antisocial or illegal activities or because their behaviors are suppressed by the mainstream social or cultural norms (e.g., homosexuality in Asian cultural environments). Nevertheless, it does not follow that the concept of ingroupness should be understood as carrying negative connotation at all times.

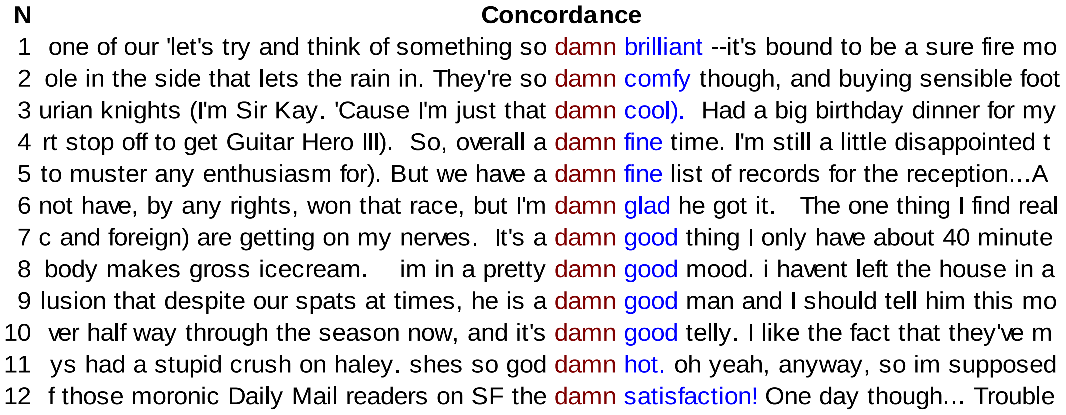
Aside from the three basic functions mentioned above, slanguage can also be used to express feelings and emotions. To a considerable extent, this function is of greater importance than the above three in personal blogs. The reason is simple: personal blogs are a genre of recording bloggers' daily experiences, reflections, feelings, and emotions. In this kind of writing, the author's attitudes towards the subject matter or audience is very important. According to Allan and Burridge (2006), slanguage reveals a lot about its users in this regard. Slanguage can be used to show familiarity with what is being referred to, or at least familiarity with the group that uses this term. To describe something as *wicked* or *insane* is more than saying that thing is good; it expresses connotations that the conventional language does not convey. If we say that using proper slang (such as *awesome*, *cool*, *wicked* or *insane*) is an indirect way of expressing speaker or author attitudes, using dirty words should be taken as a more direct way of expressing feelings and emotions. This is especially the case when they are used to express strong emotions such as excitement, disappointment, anger and hatred. Dirty words are often associated with negative connotation, but this is not always the case. Some dirty words are often used as intensifiers to amplify the semantic meaning which they are modifying. It is the modified item which normally determines whether the modifiers are meant to express negative connotations. One typical example is the word *fucking*, which is very commonly used by personal bloggers from almost all age and gender groups. This word is often associated with negative connotation, but it can also be used to modify adjectives or verbs with positive connotations, as Figure 9.1 shows.



```
  N                                Concordance
  1  the integrity of your plank. Three sets.  *sigh* I fucking love it.    Current Mood:  cheerful  Au
  2  y at the University of London. And its spiffing. I fucking love it. The course is good, the camp
  3  ursday (IDEAL!!). Work is going really great... I fucking love my job because my boss is the
  4  rtunatly nothing else happened, but honestly, i fucking loved it. It was great. It raised my mo
  5  of exeter and then the ones in college are like i fucking want to be at exeter.  i'm picking my
  6    3.  I hope I see them soon. :]  That would be so fucking amazing, oh man.  Everyone want'
  7  y i love that kid. we argue all the time but he's so fucking cute. i really don't love anyone else
  8  h, whateva.  The L word come's on tonight, I'm so fucking excited. :D  28 August 2006  This
  9  cher to naughty school child way. God it was so fucking funny, at least in hindsight *snigge
 10  . They've flown. I got married last month. I am so fucking happy I could burst. It feels good t
 11  d soon  I gotta say it again, I love Jillian Obert so fucking much, I dont know what I'd do with
 12  perience enabled me to expand my drumming so fucking much.  It was fucking awesome pl
 13  t. I was like, speechless when I visited. It was so fucking perfect. It reminds me of AHS, stru
 14  d went and printed my lovely tiger.  IT LOOKS SO FUCKING SEXY!!  But it has no face, and
 15  fill it during my wait.  WOAH!! BRILLIANT! I'm so fucking smart, it kills me!  But yeah, this b
```

Figure 9.1 Concordance lines for FUCKING

Another example is the word *damn*, which is also more commonly associated with words of negative connotations. The following concordance lines (Figure 9.2) show how they are used to modify words with positive connotations as well.



```
  N                                Concordance
  1  one of our 'let's try and think of something so damn brilliant --it's bound to be a sure fire mo
  2  ole in the side that lets the rain in. They're so damn comfy though, and buying sensible foot
  3  urian knights (I'm Sir Kay. 'Cause I'm just that damn cool).  Had a big birthday dinner for my
  4  rt stop off to get Guitar Hero III).  So, overall a damn fine time. I'm still a little disappointed t
  5  to muster any enthusiasm for). But we have a damn fine list of records for the reception...A
  6  not have, by any rights, won that race, but I'm damn glad he got it.   The one thing I find real
  7  c and foreign) are getting on my nerves.  It's a damn good thing I only have about 40 minute
  8  body makes gross icecream.    im in a pretty damn good mood. i havent left the house in a
  9  lusion that despite our spats at times, he is a damn good man and I should tell him this mo
 10  ver half way through the season now, and it's damn good telly. I like the fact that they've m
 11  ys had a stupid crush on haley. shes so god damn hot. oh yeah, anyway, so im supposed
 12  f those moronic Daily Mail readers on SF the damn satisfaction! One day though... Trouble
```

Figure 9.2 Concordance lines for DAMN

The word *bloody* can also be used to achieve similar effects but less often (see Figure 6.10 for concordance lines). As Allan and Burridge (2006) point out, despite that swearing is normally an emotive reaction to anger, frustration, or something unexpected, it is possible to be used as intensifiers to modify actions or qualities which the speaker finds desirable. Even insulting terms can be used in informal writing for the author to indicate a bond of friendship with the audience. This is a phenomenon transplanted from spoken language where "the use of normally abusive address forms or epithets are uttered without animosity and reciprocated without animus for indicating a bond of friendship" (p. 87).

The use of slanguage may also tell us something about the intended audience. According to Jay (1992, p. 139), in spoken situations a speaker is more likely to use "off-color language" in the company of members of the same gender. Maybe that can explain why female bloggers from the older adult groups use greater number of slanguage words than their male counterparts: they may be writing for same-gender audience. Unfortunately, this research is not designed in a way that such claims could be verified. This can be an issue for future studies.

To summarize, slanguage can be used to achieve a variety of functions. As a fundamentally spoken feature, its appearance in personal blogs increases the informality of the discourse, which in turn marks off personal bloggers from non-bloggers and possibly the blogging self from the non-blogging self in terms of writing styles. Just like the users of slanguage in spoken language, bloggers are also trying to use slanguage as a means of representing themselves and their intended audience. Younger bloggers (male and female alike), especially the mid- and late-teens, tend to use greater density of slanguage words than mature adults. Bloggers have displayed both collective and individual identities through their slanguage use.

9.2.4 Age and grammatical features

As mentioned in Chapter 8, morpho-syntactic and syntactic rules are more established and thus less likely to be manipulated without creating a sense of oddness. Nevertheless, it does not follow that these rules will not be challenged by bloggers. Among the five new or less conventional grammatical features examined in the current study, two are closely related to blogger age. One is the new usage of the plural form marker *–s/z* and the other is the use of accusative case of the first person singular pronoun (*me*) in the subject position of a clause or sentence. The previous chapter has presented a detailed description about the expansion of the usage of the plural form marker from its conventional function of attaching to nouns only into attaching to words of a variety of parts of speech (verbs, adverbs, inserts, and so on.) Among the 127 cases of this new usage identified from the EBC (see Figure 8.2 for concordance lines), 100 are from bloggers below 25, accounting for 79%. Among the 100 cases, 73 are from the teens groups, taking up 73%. Bloggers aged from 25 to 40 only contributed 27 cases, occupying around 21%. Among those words that carry this new usage, *anyway* is a typical example. There are two plural forms of this word in the EBC: *anyways* and *anywayz*. There are 104 occurrences of the plural forms of *anyway,* among which 83 occurrences are from bloggers aged below 25, accounting for 80%. Occurrences from bloggers aged from 30 to 40 only take up less than 11%. In other words, the new usage of plural form marker can be considered as a marker of young age. Figure 9.3 gives a flavor of how the pluralized forms have been used.

```
 N                                  Concordance
  1   some of my friends, i havent met that many!  anyways before we left we asked tony again i
  2  astard!  I'm gonna continue copying my blogs anyways before Doctor Who comes on.  Pea
  3  st. Sad, I know, but they looked really funny.  Anyways college tomorrow. Sian and Manda
  4   ughing as he took it. I must have looked bad. Anyways didnt get chips and went back in to
  5 - which has been really, REALLY hard on me. Anyways - we are leaving on Fri afternoon an
  6 recipe is off and a little too weak for my liking. anyways have a good weekend.   Jun. 4th, 20
  7   e still got scared. so yea im really tired.     so anyways here is the link to the scary maze g
  8 and play guitar hero. always a good time. well anyways im gonna go cause i have hair dye d
  9 ever. but like i said no one believes in me. but anyways i don't know what else to say and i'
 10  your birthday, but then they throw one for you anyways (or when Jaws II comes out even th
 11   k to normality and finding a job to kill time lol anyways time to go cos i dont wanna get into
 12  after time, but then I is only human (hopes so anyways)  But all that, along with seeing so
 13  tried to sleep but it proved almost impossible. Anyways, I woke up at 8 when my alarm wen
 14 erable in my job.  Can we say "circular"?  But anyways, I have to stay there til xmas; our bo
 15   e patriots cause they had a perfect season.   anyways, it was awesome. at the end, when
 16 ou have trouble even giving away SDTVs now. Anyways, my parents will like that, they gave
 17  s is the front of the booklet;  back of booklet;  anyways. i was super happy. :)   & omg, my f
 18  n't even been bothered, but I've been too busy anyways.  Tuesday, I'll probably play some m
 19   my job. Therefore I feel I am partly to blame.  Anyways...I think I'm done for now.      Octob
 20 in the way of foreplay.  What the hell was that anyways?  I had no freaking idea.  I only have
 21   not what I had in mind for this summer...  but anywayz why is college so damn expensive, i
 22    But it has no face, and looks really scay D:  Anywayz, I couldn't bring it home yet, 'cause
```

Figure 9.3 Concordance lines for ANYWAYS/Z

Another grammatical feature which is closely related to blogger age is using *me* in the subject position of a clause or sentence. There are around 78 occurrences of *me* being used in subject position or as a determiner. Among the cases where *me* is used as part of the subject, 59 cases are in the structure of *me and*. 76.3% of them are from bloggers aged from 15 to 24. Again, we can see the shadow of age in the use of unconventional grammatical features. It seems that younger bloggers tend to transplant oral linguistic features directly into their informal writing, probably to achieve informality and reduce the social distance between them as bloggers and the intended readers. This phenomenon is not unique to blogging; it may well be an extension of similar tradition in online chat discourse which has its roots in daily speech, of course.

One more grammatical feature which is closely related to blogger age is the use of *like* as a quotative complementizer. If we take a look at the age groups of the American bloggers who have used the "*be like*" expression, we will find that 85% of the occurrences are from the younger generation (those aged from 15-24). In other words, the quotative use of the "*be like*" expression can be taken as a marker of young people. The British bloggers as a whole use this feature much less frequently than their American counterparts. Nevertheless, if we take a look at the age groups of the British bloggers who have used the "*be like*" structure, we will soon find that they are exclusively from the younger generation (aged from 15-24). I would not attribute this difference between American and British bloggers to the slower pace of the grammaticalization of *like* in British English. Rather, I will take this difference as the result of British English speakers attempting to identify with the American fashion in terms of language use. It is more of an identity issue than an issue of pace of grammaticalization. The fact that only younger British bloggers use the "be like" expression indicates part of the nature of youth in wanting to identify with fashion, linguistic or otherwise.

9.2.5 Age and pragmatic features

Age and pragmatic features are often found to be linked with each other in existing literature. The reason is that pragmatic markers are vulnerable to change and young people are found to be more active innovators. According to Tagliamonte (2005), the emergence of a number of new discourse/pragmatic markers which have gained considerable high-profile attention in recent years can be attributed to the linguistic innovation of the younger generation. This is also an indication that examining the bloggers' use of pragmatic markers may reveal their age-related identity. From the description presented in Chapter 8, we can see that bloggers tend to transplant spoken discourse features into their blog entries. Apart from performing intended communicative functions and making the entries sound more informal, the employment of some features can reveal age-related identity of the bloggers. For instance, the use of *so yeah* (*yes*) and *but yeah* (*yes*) can give us some clues about the age of bloggers. Among the 97 occurrences of these two patterns, 71 are from bloggers aged from 15 to 24, accounting for 73%. Bloggers from the older age groups (those above 25) have only contributed 26 occurrences, taking up 27%. Among the 26 occurrences, 19 are from the 25-29 group (taking up around 20% of the total), five are from the 30-34 group (5%), and two are from the 35-40 group (2%). In other words, 93% of the occurrences are from bloggers aged below 30. To a considerable extent, we can take the use of *but yeah* (yes) and *so yeah* (*yes*) as a marker of youth identity.

Some of the vague expressions appear to be related to blogger age as well, for instance, *kind of*, *kinda*, *sort of*, *and sorta.* As mentioned in Section 8.2.3, there are 183 occurrences of *kinda*, 155 occurrences of *kind of*, 67 occurrences of *sort of*, and 20 cases of *sorta* in the whole blog corpus. 65% of the occurrences of *kind of* and 67% of the occurrences of *kinda* come from bloggers aged below 25. If we take the 25-29 age group into calculation, the percentages will add up to 78% for *kind of* and 85% for *kinda*. As far as *sort of* is concerned, only 43% of the occurrences are from bloggers aged below 25. *Sorta*, on the other hand, seems to be preferred by younger bloggers as 80% of its occurrences come from bloggers aged below 25. What we can conclude from the numbers presented here is that *kind of*, *kinda*, and *sorta* can be taken as marker of younger age. Other vague expressions such as *or whatever, and shit/crap, like*, and *and everything* are also closely related to blogger age. 65% of the occurrences of *or whatever*, 79% of the occurrences of *and shit/crap*, 88% of the occurrences of *like* (used as approximators), and 69% of the occurrences of *and everything* come from bloggers aged below 25.

9.2.6 Age and preference for semantic domains

Chapter 7 has offered a rather detailed description about the preference for semantic domains for each blogger group. If we examine the list of preferred semantic domains for each of the six age groups, we can obtain some insights about the social and psychological reality of each group. Bloggers from different age groups are at different developmental stages of their life and they may assume different social roles. Despite the considerable overlap between the list of preferred semantic domains for the mid-teens group and that for the late-teens group, the mid-teens have displayed their uniqueness by more frequent mention of body parts, feelings and emotions (sadness and happiness alike), music, sports, games, and their involvement in oral communication. This seems to have something to do with typical life styles of adolescents at puberty. The overlap between the mid-teens and late-teens tells a lot about some other important aspects of the teenagers' life. Education is undoubtedly an important part, whether they enjoy it or not. People and friends are also of vital importance at this stage of their life when seeking semi-independence (or independence) from the family environment is on the go. Puberty is also a stage when teenagers start to develop a strong interest in the opposite sex and thus they may be eager to explore or talk about intimate relationships with other teenagers. This is also the age to seek their identity by deviating from the established norms, linguistic norms included. According to existing studies, teenagers are the most rebellious in their use of language in spoken settings because they just want to show that they are different from both the younger generations represented by the tweens and the older generation represented by their parents. They seem to have carried this practice into online discourses, as can be seen from the overuse of unconventional linguistic elements, be it unconventional spellings or use of oral discourse markers.

The young adult group (20-24), on the other hand, starts to display certain features or styles of their own due to the change of studying environment and the gaining of more life experiences. For instance, they are more often involved in reading, understanding, and other mental work. Meanwhile, they still share certain features with the late-teens. For instance, many bloggers of this group mention education in general very often because college life is still an important part of their life. Music and bands are also important, although this time they may not be contented to listening to music; they may well be interested in forming their own bands. For many of them, college years is also time for enjoying life a little bit.

For the 25-29 group, the central stage of life has been shifted from school and college onto work place and home. As a consequence, greater mention about work and employment and housing looks quite natural. Meanwhile, health issues have also become something very important. For the 30-34 group, life might have become diversified. The shortest overall list of preferred semantic domains may well be an indicator that they share certain things with one group or another. Nevertheless, we can still feel the busyness of their life from their frequent mention of moving from place to place and the obtaining of new things (new cars, new jobs, and many other new things). For the 35-40 group, life seems to have become more stable: they are financially better off and they are more actively involved in social activities and travelling. From the insights we have obtained from comparisons in terms of preferred semantic domains, we can see a link between what bloggers tend to talk about and the potential age group they fall into. Or to put it in an arguably simplistic way, bloggers are what they write about.

9.2.7 Summary

What has been presented in the previous sections shows that there is a close link between linguistic variation and blogger age. If we take a closer look at those features which are typically associated with younger age, we will see that this link is not coincidental. The frequent use of non-conventional contracted forms, words with unconventional letter repetition, and words expressing electronic paralinguistic features are all examples of bloggers' efforts in playing with the spelling of words. According to Sebba (2003), normally only one spelling of a word which is acceptable in print, but 'authorized' spellings are not the only ones in use. In practice there is room for deliberate deviation from standard spelling. "Adolescents are among the potential users of these opportunities, which allow for the possibility of quiet, but visible rebellion against the authority of spelling" (Sebba, 2003, p. 151). The word "rebellion" may sound a bit too strong. What bloggers of the younger generation are doing may well be demonstrating that they are the owners of the language and they want to look different. In other words, they are actually representing their identity in orthographic means. Young people's desire to construct an identity which belongs to their generation can also be felt in their efforts in identifying themselves with emergent new cultures on the Internet, as newly emergent culture or subculture is often associated with the younger generation. By getting involved in the creation and spread of new Internet culture, bloggers are also representing a part of their identity as trend setters. The use of slanguage, a practice which is mostly observed in spoken discourse and often associated with teenagers and young adults in existing literature, is another feature that is found to be closely linked to younger age in the EBC. Considering the potential of slanguage in expressing ingroupness and strong emotions and the factor that personal blogs are a platform for expressing emotions and realizing social communications, especially for young people, it is quite normal to find bloggers of the younger generations are the ones that have used the greatest density of slanguage words and expressions. The link between younger age and the use of new or less conventional grammatical features is another piece of evidence to show young people's desire to more experimental. From the dominance of younger bloggers in experimenting with attaching new grammatical meaning to the plural marker *–s*, their preference for using the accusative case *me* in the subject position, and their preference for using the qualitative complementizer *like*, we can see that young people are on the edge of language change and they are more willing to identify with new developments. This trend is also noticeable in their use of pragmatic features. Again, younger bloggers are the ones who use new pragmatic markers such as *so yeah (yes)* and *but yeah (yes)*. Younger bloggers are also the ones who use certain vague expressions more frequently than bloggers of the older generation (aged from 30 to 40), echoing the findings of Stenström and colleagues that "in the teenage world it is cool to be vague, and it is cool to demonstrate that one cannot be bothered to be precise" (2002, p. 88). From the preference for semantic domains of bloggers from each age group, we see the influence of age on what bloggers write about. From what we have presented in Chapter 7, we can see that bloggers at different developmental stages of their human life tend to focus on different topics and themes which reflect their social and psychological realities they are facing. In a word, certain aspects of the linguistic variation identified in this research are closely related to bloggers' representation of age-related identity.

9.3 Gender-related identity representation

9.3.1 Gender and non-conventional orthographic representation of words

From the discussion in Section 9.2.1 we can see that non-conventional orthographic representations are very closely related to blogger age. The general tendency is the number of non-conventional forms used decreases as the blogger age increases. Young bloggers, especially the teens, have showed a preference for certain strategies in representing existing words in orthographically non-conventional ways. What I am going to present next concerns whether the use of non-conventional orthographic representations is associated with bloggers' gender. In order to identify that link, we need to take a look at the distribution of non-conventional orthographic forms across gender groups. Table 9.14 lists the distribution of OEFs across gender groups.

Table 9.14 Distribution of OEFs across age & gender groups

Age Group	Male	Female
15-17	155.8	190.6
18-19	143.9	154.4
20-24	97.3	101.1
25-29	80.6	78.5
30-34	53.1	73.1
35-40	52.2	57.4

If we look at the use of OEFs by male and female bloggers from the same age groups, we will find that female bloggers tend to use more OEFs than their male counterparts. There is only one exception: the male bloggers from the 25-29 group tend to use more OEFs than their female counterparts, though the difference between them is very small. What we can observe from Table 9.14 is that male and female bloggers may behave differently in their use of non-conventional orthographic word forms. If we want to find out where they differ, we need to examine the specific linguistic strategies they employed in producing non-conventional word forms.

Among the six linguistic strategies that are employed by bloggers to realize orthographic variations, three are closely related to blogger gender: the use of acronyms and initials, the use of e-paralinguistic words, and the use of letter repetition words. Table 9.15 lists the top five blogger groups which have employed the strategy of acronym and initialism. Unlike the use of non-conventional contracted forms, the use of acronyms and initials does not seem to be associated with a particular age group (or groups). The top five blogger groups come from four different age groups: 35-40, 25-29, 20-24, and 15-17. Nevertheless, they share something in common: they are all female blogger groups. If we look at the groups which have used the fewest acronyms and initials (see Table 9.16 below), we will find a quite different pattern. This time, all the five groups are male groups. The age groups involved include: 20-24, 25-29, 30-34, and 35-40.

Table 9.15 Groups using most acronyms & initials

Blogger Group	Frequency	Relative frequency
uk_f_35-40	68	31.2
uk_f_25-29	76	30.8
us_f_18-19	83	28.3
us_f_15-17	64	25.2
uk_f_15-17	75	22.3

Table 9.16 Groups using fewest acronyms & initials

Blogger Group	Frequency	Relative Frequency
us_m_25-29	30	9.7
uk_m_20-24	31	9.1
uk_m_25-29	15	6.7
us_m_30-34	20	6.4
us_m_35-40	23	6.2

As discussed in Chapter 5 (Section 5.2.2.2), more than half of acronyms and initials fall into two major categories: markers of online discourse and abbreviated noun phrases, with the former accounting for 38% and the latter 15% of all the occurrences. By markers of online discourse, they refer to the initials and acronyms which are often used in online chat (be it public chatting or instant messaging) to express paralinguistic features (e.g., *lol* for laughing), emotions (*wtf* (*what the fuck*) for showing anger and *omg* (*Oh my God!*) for showing surprise), and other shorthands (such as *imo* for *in my opinion*).

Among these online discourse markers, the acronym *lol* is the most commonly used among bloggers. This acronym (and its variants) is one way of representing the paralinguistic behavior of laughing and laughter. Table 9.17 presents the five blogger groups with the highest frequency of acronyms and initials representing laughing and laughter. From this table we can see that the blogger groups with the highest frequencies of *lol* and its variants come from four different age groups: 15-17, 18-19, 25-29, and 35-40. The subtotal of the acronym *lol* and its variants accounts for 54% of the total number of such forms in the whole EBC. What these groups have in common is that they are all female bloggers. This seems to suggest that the acronym *lol* is more closely associated with female bloggers.

Table 9.17 Groups with the highest frequency of LOL

Blogger Group	Frequency	Relative frequency
us_f_18-19	48	21.3
uk_f_25-29	49	16.7
uk_f_35-40	45	14.4
uk_f_15-17	39	14.4
us_f_15-17	27	10.9

Table 9.18 shows the eleven groups whose frequencies of the acronym *lol* and its variants per ten thousand words are below 3. Among these eleven groups, eight are male groups; two are mature and older female groups; and only one female group from the age group of 20-24.

Table 9.18 Groups with the lowest frequency of LOL

Blogger Group	Frequency	Relative frequency
uk_m_35-40	7	2.8
uk_f_18-19	9	2.7
us_f_35-40	8	2.2
us_m_18-19	6	1.8
us_m_35-40	4	1.5
uk_m_18-19	4	1.1
uk_f_30-34	3	1.1
uk_m_30-34	3	1.0
us_m_25-29	2	0.7
uk_m_25-29	1	0.4
us_m_30-34	0	0.0

Apart from the acronym *lol* and its variants, there is another way of representing laughter in personal blogs, that is, onomatopoeia or mimicking laughter. As described in Chapter 5, there are two basic orthographic forms for mimicking laughter in the EBC: *haha* and *hehe*. There are 224 occurrences of the former and its variants and 79 occurrences of the latter and its variants (see Table 5.11 for details). Among the 303 occurrences of haha and hehe and their variants, 190 are contributed by female bloggers, accounting for 63%. If we take a look at the distribution of *haha* and *hehe* separately, we can still find the dominance of females. Among the 224 occurrences of *haha* and its variants, 158 are from female bloggers, accounting for 71%. Only 66 occurrences are from male bloggers, taking up 29%. The occurrences of *hehe* and its variants reveal a different pattern: among the 79 total occurrences, 47 are from male bloggers, accounting for 60%.

As discussed in Section 9.2.1, the use of words with unconventional letter repetition is also closely related to blogger age and the teens in particular. If we examine the distribution of these words across different gender groups, we can also see a link between the use of letter repetition and blogger gender. If we recall Tables 9.5 and 9.6, we can see that non-conventional letter repetition is more frequently used by female teenage bloggers to realize orthographic variation.

9.3.2 Gender and the use of asterisks

As discussed in Section 5.3, the use of asterisks can be used to achieve special communication effects while at the same time add some flavor of performance to the blogging entries. Among the four different uses, the one as action markers is worthy of particular mention here. Among the 266 total occurrences of this use in the corpus, 192 were from female bloggers, accounting for around 72%. Only 74 instances were from the male bloggers, taking up 28%. The same pattern holds for both British bloggers and American bloggers. Within the British blogger group, females contributed 72% and males 28% of the total occurrences respectively. The percentages are the same for the females and the males in the American blogger group. What we can conclude from the distribution reported here is that using the asterisks as action markers is more often associated with female bloggers.

9.3.3 Gender and neologisms related to IT and Internet culture

Another aspect where gender differences may be observed would be bloggers' use neologisms related to IT and Internet culture. As mentioned in Chapter 5, there are 426 new lexical items which are IT-related. Among these items, 241 tokens are contributed by male bloggers and 197 tokens by female bloggers, with the former accounting for 57% and the latter 43% of the total occurrences. Recall the fact that the number of female blogger groups is greater than that of male groups due to the paucity of British-based mid-teens male bloggers. The distribution of IT-related new lexical items across gender seems to be echoing the finding of Argamon, Koppel, Pennebaker, & Schler (2007) that male bloggers tend to use more Internet-related content words than female bloggers. Nevertheless, if we take a look at the distribution of such words across bloggers from different age and gender groups, we can find certain patterns which might be masked by simply doing an overall comparison between males and females. Among the six age groups, three groups have displayed the pattern of males outnumbering their female counterparts in the use of IT-related lexical items (in terms of both absolute frequency counts and relative frequencies per ten thousand words). The 35 to 40 age group and the 30 to 34 group have both displayed much greater differences between male and female bloggers than the younger age groups, with the former group being 60 to 33 and the latter group being 60 to 31. Within the 25 to 29 age group, males outnumber females by 43 to 33 (7.86 to 5.09 in relative frequency per ten thousand words). As for the late teens (the 18-19 group), the absolute frequency ratio between males and females is 29 to 28 (or 5.08 to 4.98 in terms of relative frequency per ten thousand words), which suggests no significant gender differences. The mid-teen groups can be excluded from this comparison as there is no comparison group for the British mid-teens female bloggers. The 20 to 24 age group, however, have displayed a reverse pattern, with female bloggers outnumbering the male bloggers in terms of both absolute frequency counts (42: 38) and relative frequencies (5.94: 5:43), despite that the difference is very small. If we put the 18-19 group and the 20-24 group together, the gap between male bloggers and female bloggers in the use of computer-related new items will be leveled out. This narrow gap may well be a reflection of the importance of computers and the Internet in their daily life of these two groups of people who are mostly college students, regardless of their gender. The next-to-zero difference in the use of IT-related terms can be taken as a marker of generation.

Apart from IT-related neologisms, some game-related terms can also offer some clues about the gendered identity of bloggers. Among the 96 tokens of game-related new lexical items, approximately 69% come from the entries of male bloggers from almost all age groups, which suggests that males are still the dominant consumers of electronic or online games.

9.3.4 Gender and the use of slanguage in blogging

In Section 9.2.3, I have discussed the relationship between the use of slanguage and blogger age and found that younger bloggers, especially the teens, tend to use greater density of slanguage in their blog entries. I have pointed out that it might have something to do with the developmental stage of adolescence which is a transitional period full of excitement, puzzle, frustration, and rebellion. Moreover, teenage bloggers' preference for slanguage may well be the demand of trying to express their strong emotions, be it excitement or anger. Existing studies (e.g., Allen, 1998) concerning slanguage use show that males tend to use more slang than females. In order to find out whether this is true for personal bloggers, I have reorganized the data according to gender in Table 9.19 below. This table shows the raw and relative frequencies of slanguage use of males and females in all six age groups.

Table 9.19 Slanguage and gender (whole)

Age Group	Female			Male		
	Slang Tokens	Per_10K	Sample Size	Slang Tokens	Per_10K	Sample Size
15-17	539	104.2	51,718	246	113.0	21,761
18-19	421	74.9	56,233	527	92.4	57,045
20-24	439	62.1	70,690	563	80.4	69,985
25-29	514	79.3	64,845	400	73.1	54,690
30-34	341	57.5	59,265	408	67.5	60,407
35-40	418	61.2	68,259	178	33.8	52,686

From this table we can see that male teenage bloggers do use more slanguage words than their female counterparts. This is also the case for early adult bloggers aged from 20 to 24. For older blogger groups, on the other hand, there exist two opposing patterns. Male bloggers outnumber female bloggers in terms of slanguage use for the 30-34 group. For the rest two groups, the pattern is reverse: female bloggers use greater number of slanguage words. This is especially the case for the group aged from 35 to 40. If we split the data according to bloggers' country of origin and then compare the slanguage use between male and female bloggers, we will see a slightly different picture. Table 9.20 lists the details about British bloggers' use of slanguage across different age and gender groups. Similar to the overall distribution demonstrated in Table 9.19, both the late-teens group and the early adult group display a tendency of male bloggers using more slanguage than female ones. As for the mid-teens group, there is no way to determine whether male bloggers outnumber their female counterparts due to lack of data for the male group. For the three more mature adult groups, female bloggers outnumber their male counterparts in two groups (the 25-29 group and the 35-40 group). In the 30-34 group male bloggers' use of slanguage is far more than that of female bloggers.

Table 9.20 Slanguage and gender (UK)

Age Group	Female			Male		
	Slang Tokens	Per_10K	Sample Size	Slang Tokens	Per_10K	Sample Size
15-17	259	95.7	27053	n/a	n/a	n/a
18-19	222	65.9	33686	285	86.2	33078
20-24	205	62.2	32971	271	75.7	35777
25-29	217	73.9	29350	184	71.8	25616
30-34	136	48.7	27900	218	74.1	29409
35-40	146	46.8	31209	87	34.3	25397

As far as the American bloggers' use of slanguage is concerned, the distribution patterns across different age and gender groups are slightly different from the overall patterns. Table 9.21 shows the details.

Table 9.21 Slanguage and gender (US)

Age Group	Female			Male		
	Slang Tokens	Per_10K	Sample Size	Slang Tokens	Per_10K	Sample Size
15-17	283	114.7	24,665	247	113.5	21,761
18-19	201	89.1	22,547	242	101.0	23,967
20-24	234	62.0	37,719	293	85.7	34,208
25-29	299	84.2	35,495	218	75.0	29,074
30-34	206	65.7	31,365	191	61.6	30,998
35-40	274	74.0	37,050	91	33.3	27,289

Different from the overall pattern for the mid-teens group, there is only a very small difference between the American male bloggers and the female ones, with the former being outnumbered by the latter. The patterns for the late-teens and the young adult groups are more or less the same as the overall patterns, with males outperforming females in the use of slanguage. As for the three older adult groups, they all display a similar pattern, with varying extent of difference, of course. The oldest group displays the biggest gender difference. One more observation is that the distribution of slanguage use for American male bloggers is neatly associated with their age groups: the younger the group is the more slanguage words are used. The distribution pattern for female bloggers is not straightforward.

9.3.5 Gender and pragmatic features

In Chapter 8 I have presented a description about bloggers’ use of pragmatic features, focusing on such features as discourse markers, interjections, and vague words. All the features discussed in Chapter 8 have their origin in spoken discourse. Some of these features are closely related to blogger age, as have been discussed in Section 9.2.4. As a matter of fact, some of these features can offer us clues about the gender of bloggers as well. One of these features is the use of interjections. There are 729 occurrences of interjections in the EBC, 486 of which are from female bloggers and the rest 243 are from the male bloggers. In other words, female bloggers as a whole have contributed twice as many occurrences of interjection use as their male counterparts. Table 9.22 lists the distribution of interjection use across all age and gender groups. From this table we can see that female bloggers outnumber male bloggers in all age groups and this pattern holds for both British bloggers and American bloggers. This may be indicating that female bloggers are more willing to express their emotions in blogging.

Table 9.22 Distribution of interjection use across groups

Age Group	British Bloggers			American Bloggers		
	Male	Female	Subtotal	Male	Female	Subtotal
15-17	n/a	59	59	29	66	95
18-19	37	75	112	18	35	53
20-24	24	36	60	31	38	69
25-29	19	27	46	28	31	59
30-34	10	36	46	16	25	41
35-40	11	23	34	20	35	55
Total	101	256	357	142	230	372

9.3.6 Gender and preference for semantic domains

As discussed in Section 9.2.6, bloggers' preferences for semantic domains are closely related to their age groups. The reason is that bloggers from different age groups are experiencing different developmental stages of their life which are characterized by different social roles and modes of behavior. In fact, age is not the only factor which shapes bloggers' preferences for semantic domains. Gender is another important factor. According to what I have presented in Chapter 7, male and female bloggers do display different preferences for semantic domains and gender differences can be observed within each of the six age groups. If we compare male and female bloggers on the overall basis, we can find that they tend to focus on talking about different topics. Female bloggers tend to mention self, people, thoughts, feelings, and emotions, physical attractiveness or wellness, and social communication with other people and they use negation more often. This finding is similar to that of Mulac et al. (2001) and Mulac and Lundell (1994) that typical female language features comprise intensive adverbs, references to emotions, uncertainty verbs, negations, and hedges. It is also similar to Thomson and Murachver’s (2001) finding that females make more references to emotion and provide more personal information in email writing. Male bloggers, on the other hand, tend to talk about work and employment, entertainment and sports (the latter of which is often regarded a more masculine hobby), electronic or IT gadgets, and participation in social events.

Within each age group, male and female bloggers tend to display the gendered identities which are particularly associated with that age period. Within the mid-teens group, female bloggers tend to talk more about their thoughts, feelings, dislikes, and emotions whereas the male bloggers focus more on music, sports, crimes, moving from one place to another, and Internet-based written communication. Within the late-teens group, female bloggers tend to talk about arts and crafts, photos, feelings and emotions, school work, and oral communication with other people whereas the male bloggers tend to talk about games, sports, music, drinks and alcohol, and personal relationships. Within the early adult group (the 20-24 group), female bloggers tend to write about people and relationships, health, disease, and food whereas the male bloggers are more interested in writing about TV programs, movies, computers, and entertainment. Within the mid-adult group (the 25-29 group), female bloggers write more often about people and relationships, clothes, personal belongings and so on whereas the male bloggers write a lot more about work and employment, power relations at work place, money matter, music, and sports. Within the mature adult group (the 30-34), the female bloggers tend to write about people, self, body, weight, physical wellness, and thoughts feelings whereas the male bloggers write more about moving from one place to another, participation in social events, and entertainment. For the 35-40 group, the female bloggers show greater interest in such topics as people, family members, education, and communication with people whereas their male counterparts are more interested in talking about entertainment, electronic or IT gadgets, and work and employment. If we take a look at the preferred semantic domains for female bloggers from different age groups we can see some consistency in the kind of themes they tend to talk about. This is the same for all the male bloggers. In other words, preferred semantic domains reflect clues about blogger gender.

9.3.7 Summary

From what we have presented so far, we can see that certain aspects of the linguistic variation identified in this research are closely related to blogger's gender-related identity. The non-conventional orthographic representation of common words is not only related to blogger age but also to their gender. Overall, female bloggers have showed a stronger preference for non-conventional spelling of words. If we take a closer look at the specific features that female bloggers tend to use more often, we can see the string which is pulling behind the scene. As discussed in Section 9.3.1, female bloggers, especially the teens, tend to use initials and acronyms much more frequently than male bloggers. Among the most frequently used initials and acronyms, almost half are online discourse markers and abbreviated forms representing the act of laughing and laughter. If we put together the prominence of female bloggers in using the asterisk as action markers, their use of words with unconventional letter repetition, and their use of initials and acronyms representing electronic paralinguistic features, we can see that female bloggers have demonstrated a tendency of adding the flavor of performance to their blogging texts.

Existing literature (e.g., Argamon et al., 2003) indicates that female language exhibits greater usage of “involved” features. By incorporating paralinguistic features and actions into the blogging texts, the female bloggers (especially the teens) have actually increased the “involvedness” and the vividness of the texts.

This tendency of using “involved” features is less frequently observable in blogs produced by male bloggers, especially bloggers aged from 30 to 40. If we take a look at the use of neologisms related to IT and video or Internet games, we will see a different picture. This time, the male bloggers, especially the more mature adult bloggers (those aged from 25 to 40), become the dominant ones. The gender difference within bloggers of the younger generation (those aged between 15 and 24), however, seems to be leveled out in this regard, possibly reflecting the importance of computers and the Internet in the daily life of this group of people who are mostly students, regardless of their gender.

From the analysis we have presented in this chapter and Chapter 6, we can observe a link between the use of slanguage and the bloggers’ expression of age- and gender-related identities. A more striking difference can be observed between the younger generation represented by the teens and young adult blogger groups (i.e., those aged from 15 to 24) and the older generation represented by bloggers aged 30 and above. The gender difference is also more prominent within the blogger groups which fall into the category of the younger generation (those aged below 25). Younger bloggers tend to use more slanguage words than older bloggers and male bloggers tend to use more slanguage words than females, echoing findings in existing literature that males tend to use more slanguage. For bloggers of the older generation (those aged between 25 and 40), the pattern seems to be reverse: female bloggers tend to use greater density of slanguage words. This is quite different from the findings of existing research. It is quite difficult to explain why this is the case. Nevertheless, two factors might have contributed to female bloggers' more frequent use of slanguage words. One is the anonymous nature of personal blogs as a genre. With the protection of anonymity, female bloggers might be less willing to suppress their desire to use slanguage, especially the strong or dirty words. The second factor may have something to do with their intended readers: they may well be writing to readers of the same gender. According to Jay (1992, p. 139), in spoken situations a speaker is more likely to use "off-color language" in the company of members of the same gender. Considering female bloggers effort in trying to make blogging more like talking, this explanation seems to make sense to a certain extent.

As far as the pragmatic features are concerned, the only feature which seems to display strong association with gender is the use of interjections. As interjections are exclamative utterances used to "express positive or negative emotional reactions to what is being or has been said or to something in the situation (Carter & McCarthy, 2006)." This is actually another strategy to increase the "involvedness" of the blogging texts. This may also be an indirect indicator of females being more interested in expressing emotions.

The gender difference in terms of preferences for semantic domains seems to echo the findings in existing studies concerning computer-mediated communication that females make more references to emotion and provide more personal information whereas male bloggers make more references to material things.

9.4 Linguistic representation of regional identity in blogging

Theoretically and technologically speaking, through blogging we can expand the boundary of the readership from the people around us to anyone whom the Internet can reach. Nevertheless, it is very difficult for us to erase the social and cultural imprints that we carry with the language we use. As Warschauer (2001) points out, while the Internet masks the role of other identity markers such as race, gender, or class, it highlights the role of language. It may not be easy to find out whether you are a male or female, gay or straight but people can immediately notice what language and dialect you are using. A part of this imprint is the regional identity which our language reflects. Compared with conventional orthographic variation which is where regional identity is observable, the use of slanguage words and that of certain grammatical and pragmatic features are more deeply rooted in the history of a particular speech community and are thus better markers of regional identity.

9.4.1 Grammatical features and regional identity

As mentioned in Chapter 8, grammatical variations are more deeply rooted in the historical and cultural development of a speech community. Grammatical rules are normally a part of the collective identity imposed on a particular speech community. They do not lend themselves to easy manipulation on the part of the language users. Rather, they are simply a reflection of the users' collective identity. Among the five grammatical features described in Chapter 8, three have something to do with bloggers' regional identity. These features include: *go/come* plus bare infinitives, *like* as a quotative complementizers, and the use of archaic morpho-syntactic features.

In Section 8.1.3 I have already presented a rather detailed description about the structure *go/come* plus bare infinitives and pointed out that the verbs *go* and *come* may be undergoing the process of grammaticalization. If we take a look at the respective distribution of the pattern "*go/come* plus bare infinitives" and that of "*go/come and* plus bare infinitives," we can see the link between the preference for a particular pattern and part of the bloggers' identity. Among the 158 occurrences of "*go/come* plus bare infinitives," 113 are from American bloggers, accounting for around 72%. Among the 62 occurrences of the pattern "*go/come and* plus bare infinitives," 52 are from British bloggers, taking up around 84%. What we can observe from the distribution of these two patterns is that the pattern "*go/come and* plus bare infinitives" is a marker of Britishness whereas the pattern "*go/come* plus bare infinitives" is a marker of Americanism. From the fact that around 28% of the occurrences (45 occurrences) of "*go/come* plus bare infinitives" are from British bloggers, we can see the American influence. If we take a further look at the distribution of this pattern among British bloggers from different age groups, we find that 71% of the cases are from bloggers aged from 15-24. Those from bloggers aged 30 and above only take up 15%. Here, we can see the influence of age again. By identifying with a more powerful regional variety of English, younger bloggers from the United Kingdom are actually displaying a part of their identity, whether or not they are aware of it. From features such as the "*go/come* plus bare infinitives," we see less intentional effort but more reflection of the collective identity of a particular social or speech community. When the British bloggers chose a more British way of saying things and the American bloggers chose a more American way of saying things, it is just a natural reflection of their cultural backgrounds. The intentionality will only show when members of a particular community starts to identify with members from another community, as the case of British bloggers using American English features reveals.

Another feature which shows strong regional identity of bloggers is the use of *like* as a quotative complementizer. An examination of the distribution of the "*be like*" expression among bloggers shows two tendencies. First, this expression is much more frequently used by American bloggers. Second, it is more frequently used by younger bloggers, that is, those aged from 15 to 24. Out of the 65 occurrences, 55 are from American bloggers, accounting for around 85% whereas the rest ten occurrences are contributed by British bloggers, only taking up around 15%. This seems to be echoing the findings of Stenström and colleagues (2002), though their findings were based on spoken data recorded in 1993.

The use of archaic inflectional forms also displays certain aspect of the region-related identity. If we pool all these archaic forms together and take a look at their distribution among different groups of bloggers, we will find something quite interesting. Among the 61 occurrences of the archaic forms 42 are from British bloggers, accounting for around 69%. Considering the fact that Britain is the source country of the English language and that the Early Modern English period was also the time when the works of William Shakespeare and his contemporaries were written, it is no wonder at all to find British bloggers outnumber their American counterparts in using archaic forms which can be dated back to that period. In other words, using Early Modern English features can be taken a marker of the Britishness.

9.4.2 Slanguage use and regional identity

One of the most salient features of slanguage should be its cultural specificity. Recall the various definitions of slang I have cited in Section 6.7.1. Almost all of them have mentioned the concept of "group," which presupposes the concepts of both “commonality” and “localness.” The so-called commonality is actually what is shared by all members of a particular group (socially definable group) and this commonality is also a marker of localness which distinguishes one group from another. Of course, this is just one side of the coin. On the other side, all those groups live under the same roof of a bigger social community where efforts of constructing a collective identity for all its members are constantly being made. As a consequence, each individual member of a society bears the collective identity of that particular society while at the same time he or she can choose with whom they want to identify. For the collective identity (the one which belongs to the whole community), normally people do not have much choice. For individual and group identities, people can make certain choices but not without constraints. This is especially the case for the linguistic and cultural aspects of people’s life. Just like grains of sand in a desert, they may display certain features due to particular surroundings they are in but they cannot shake off the collective identity of being a grain of sand in that particular desert. If we take a look at the most commonly used slanguage words identified from the British blog components and those from the American blog components, we will observe certain differences between British bloggers and their American counterparts. In other words, by looking at the use of slanguage words, we can identify the Britishness and the Americanisms displayed in the blog entries. Table 9.23 lists the top 20 slanguage words used by British and American bloggers respectively. From this table, two observations can be made. One is that thirteen words appear on both lists but their relative frequencies of some of them are quite different. The other is that each list has seven words which do not appear on the other list. Let us take a look at the different words first. The seven words which only appeared on the British list are: *bloody*, *uni*, *bastard*, *gig*, *arse*, *emo*, and *bloke*. These words are typically British English words. The seven words which only appeared on the American list are: *man*, *dude*, *screw*, *asshole*, *chill*, *bullshit*, and *rock*. Again, these words are more of American nature. Among the words which are on both lists, some are almost equally frequently used by British and American bloggers, for instance, *fuck*, *guy*, *shit*, *hell*, *damn*, and *cool*. For some others like *awesome*, *ass*, *suck*, *bitch*, and *freak*, their relative frequencies in American blog entries are much higher than those in British blogs, again revealing the Americanism of these words.

Table 9.23 Top 20 slanguage words used by British and American bloggers

British Bloggers			American Bloggers		
Item	Tokens	Per 10K	Item	Tokens	Per 10K
fuck*	391	11.8	fuck*	474	13.3
guy	170	5.1	guy	284	8.0
shit*	137	4.1	shit*	270	7.6
hell	116	3.5	suck*	164	4.6
crap*	105	3.2	cool	140	3.9
damn	103	3.1	damn	138	3.9
cool	100	3.0	hell	131	3.7
piss*	93	2.8	awesome	129	3.6
bloody	89	2.7	ass	120	3.4
uni	70	2.1	bitch*	91	2.6
awesome	63	1.9	freak	86	2.4
suck*	61	1.8	piss*	83	2.3
bitch*	47	1.4	crap*	80	2.2
bastard	39	1.2	man	42	1.2
gig	28	0.8	dude	40	1.1
arse	27	0.8	screw	37	1.0
ass	25	0.8	asshole	29	0.8
emo	24	0.7	chill	24	0.7
freak	23	0.7	bullshit	20	0.6
bloke	22	0.7	rock	19	0.5
Subtotal	1733	52.2	Subtotal	2401	67.4

9.4.3 Pragmatic features and regional identity

The use of vague expressions is also able to reveal from which speech community the blogger is or with which speech community the blogger wants to identity. For instance, *or* *whatever, and shit/crap,* and *like* are all good indicators of Americanism. Among the 34 occurrences of *or whatever,* 23 (68%) are from American bloggers. 83 (72%) out of the 116 occurrences of *like* are from American bloggers. Out of the 24 occurrences of *and shit* or *and crap,* 17 (71%) come from American bloggers. Some vague words and expressions are markers of Britishness, for instance, *and all that* and *loads of*. Out of the 23 occurrences of *and all that,* 16 (around 70%) were from British bloggers. 42 out of the 45 occurrences of *loads of* were contributed by British bloggers, accounting for 93%.

9.5 Linguistic representation of individual identity

What has been presented in this chapter so far is all about the collective aspects of bloggers' identities, be they age-related, gender-related, or region-related. As mentioned in Chapter 2, identity is a multi-faceted concept which covers both the collective aspects and the individual aspects. One way of observing individual identity would be to examine the hapax legomena produced by different bloggers. In Chapter 6, I have presented a detailed discussion about the lexicological strategies that bloggers employed to form new lexical items so as to achieve their intended communication purposes or effects. Most of these new lexical items are nonce formation and almost each of them is used to achieve a specific communication effect that a particular blogger intended to. It is beyond the scope of the current research to analyze the function of each and every hapax legomena in helping bloggers to linguistically represent their identities. I will only take bloggers' use of phrasal compounds as an example to show how linguistic representation of individual identity can be realized.

In Section 6.3, I have described bloggers' effort in creating phrasal compounds. In many of these compounds, the bloggers have employed a semantically direct yet lexically round-about way of saying things. Many of these sayings presuppose certain cultural and social identities. For instance, when a blogger uses 'a cop-arrests-hot-woman romance novel' (Example 11, p. 166), he or she presupposes that the readers have a good knowledge of American pop culture represented by Hollywood movies. One popular scene (or theme) of such movies is that a policeman arrests a hot woman and falls in love with her at the first sight. Of course, there are different versions of romance of this kind. It is so frequently presented in Hollywood movies that it has become something like a cultural icon. It is highly likely that the blogger is an American and the intended readers are also Americans or at least people who are familiar with Hollywood movies. Quite a number of examples cited above have similar function of revealing certain aspects of the blogger's identities. For instance, by using expressions like 'a boob-popping-out-incident' and 'walking-around-with-flab-on-show problem' (Example 13, p. 166) the blogger reveals her identity of being a female, probably a young female, as can be told from her concerns over clothes and body shape. By using expressions like 'the 10MB-shared-between-one-thousand-plus-students on campus', the blogger discloses at least one aspect of his or her identities of being a college student. It is very likely that this blogger is a male, as males are generally believed to be more enthusiastic about technical details, though this may sound a bit hind sight. By using 'the biannual let's-clean-out-the-kids'-books-so-we-have-room-for-other-crap' (Example 23, p.168), an expression which seems to be a bit unusual, the blogger is also disclosing her (most probably) identity of a mother and showing the readers a glimpse of her family life (kids, cleaning, storage room, etc.). By using 'the hastily-put-together-and-we're-totally-not-following-it syllabus', the blogger is representing his or her identity as a college student while at the same time discloses what is actually happening in college education. It may also disclose the blogger’s slight resentment of such practice. By using ‘the you're-not-at-home-in-your-sitting-room-so-DON'T-TALK-ALL-THE-WAY-THROUGH-THE-GODDAMN-FILM syndrome’ (Example 17, p.167), the longest phrasal compound ever identified in the EBC data, the blogger also discloses part of her identity and her anger with those people who kept talking while watching a movie. She even employs one of the online discourse strategies of showing anger, that is, all-capitalization, meaning that the capitalized words are actually shouted out. Of course, another way of showing her anger is to hyphenate all these words into one entry, giving the reader a vivid image of somebody who is speaking rapidly, loudly, and angrily. However odd these examples may appear they are indeed used by personal bloggers as ways of representing their identities.

9.6 Chapter summary

From what has been presented in this chapter, we can see that the various aspects of linguistic variation that have been identified from the EBC are actually related to bloggers’ efforts in representing their identities in their blog entries. The prominent presence of non-conventional orthographic representation of common words, the use of slanguage words and neologisms related to emergent Internet culture, the use of new or less conventional grammatical features, and the more frequent use of new pragmatic markers and vague expressions are all examples of bloggers of the younger generation to construct their age-related identity. The different preferences for semantic domains displayed by bloggers from different age groups are also a reflection of the social and psychological realities bloggers are facing. The frequent use of initials and acronyms representing laughing and laughter, the use of words with unconventional letter repetitions, the use of e-paralinguistic words, the use of the asterisk as action markers, and the frequent use of interjections are all important indicators of female bloggers' efforts in increasing the involvedness and vividness of their blog entries. By incorporating these features which are commonly associated with spoken discourse, female bloggers have actually added a flavor of performance to a written genre. From the use of neologisms related to IT and video and Internet games, we can see the shadow of male gender. The use of slanguage has displayed two opposing patterns in gender representation. For bloggers of the younger generation, males outperform the females. For bloggers of the more mature generation, females outperform the males. From the preferred semantic domains, we can identify a consistent difference between male and female bloggers, with the former writing more about material things and the latter more about emotional topics. The variations in certain grammatical and pragmatic features and the use of slanguage words can also tell us something about the regional aspect of bloggers' identities. Apart from reflecting the collective identity of bloggers, linguistic variation is also able to demonstrate bloggers' individual identities, which is more frequently observable in their use of new lexical items of nonce formation. In a word, linguistic variation is a reflection of bloggers intention to represent themselves differently in linguistic ways.

Chapter 10 Conclusion and Implications

This chapter first summarizes the major findings of the research and then presents some of its implications. Following that, it points out the limitations of the research and recommends some directions for future research. It concludes with some final remarks.

10.1 Summary of major findings

The main objective of this research is to investigate, using a Wmatrix-based multi-variable approach, the strategies employed by personal bloggers in realizing linguistic variations and the extent that the employment of these strategies is related to their identity representation. The major findings are summarized as follows:

1. The language of personal blogs as revealed by the EBC (the corpus constructed for this research) has displayed certain features which are different from both spoken and written texts. This has been evidenced by a comparison of the top 20 word forms generated from the EBC with those generated respectively from the spoken and the written texts in the Cambridge International Corpus. One striking difference lies in the ranking of the first person singular pronoun (*I*) and that of the definite article (*the*) on the top 20 wordlist, with the former ranking the first and the latter the second. Further examination of the distribution of these two words across the texts produced by bloggers from different age and gender groups shows that bloggers' use of these two words are related to their expression of age- and gender-related identities. A further comparison between the wordlist generated from the EBC and those from the BNC Sampler Corpus Spoken and the BNC Sampler Corpus Written shows that the language of personal blogs is a

hybrid of speech and writing, as existing studies have already revealed. It is characterized by the substantial presence of self-mentioning, deliberate deviation from the established spelling norms, use of non-conventional grammatical features, and the extensive use of typical markers of online discourses.

2. Bloggers in this research have employed seven major strategies to realize orthographic variations: 1) unconventional contracted forms, 2) abbreviations, 3) letter repetition, 4) e-paralinguistic words, 5) misspellings, 6) phonetic spellings, and 7) innovative use of special symbols like the asterisk. Most of these strategies have been used for two main purposes: as markers of informality and additives to increase the talking and performance flavor of the blog entries. In this way, the bloggers have actually turned blogging into talking and the static silent letters of blog entries into dynamic audible sounds accompanied with paralinguistic features. By deviating from the established norm of conventional writing, bloggers have created a new writing style which is undoubtedly more suitable for the purpose of communicating with people via information sharing. Apart from orthographic variation, bloggers have also displayed variations in terms of lexicological strategies, slanguage use, preference for semantic domains, and the use of grammatical and pragmatic features.
3. Among the various features identified from the EBC, the following are found to be closely related to bloggers' expression of age-related identity: non-conventional contracted forms, words with unconventional letter repetition, words expressing paralinguistic features, the use of slanguage words and neologisms related to emergent Internet culture, the use of new or less conventional grammatical features (such as the new usage of the plural marker, the use of *like* as a quotative complementizer, and the use of accusative case of pronouns in

subject positions), and the use of new pragmatic markers and vague expressions. The different preferences for semantic domains displayed by bloggers from different age groups also reveal a close relation between blogger age and the blogging content which is a reflection of the social and psychological realities bloggers are facing.

4. Certain features are more closely related to bloggers' gender-related identity. The frequent use of initials and acronyms representing laughing and laughter, the use of words with unconventional letter repetitions, the use of e-paralinguistic words, the use of the asterisk as action markers, and the frequent use of interjections are all important indicators of female bloggers' efforts in increasing the involvedness and vividness of their blog entries. By incorporating these features which are commonly associated with spoken discourse, female bloggers have actually added a flavor of performance to a written genre. The use of neologisms related to IT and video or Internet games, on the other hand, are more closely related to the male gender. The use of slanguage has displayed two opposing patterns in gender representation. For bloggers of the younger generation (those below 25), males outperform the females. For bloggers of the more mature generation (those above 30), females outperform the males. From the preferred semantic domains, we can identify certain consistent differences between male and female bloggers, with the former writing more about material things and the latter more about emotional topics.

5. Bloggers' preference for certain slanguage words and grammatical and pragmatic features reflects their regional identities which are more deeply rooted in the history of a particular speech community, although conventional orthographic variation is also able to reveal that kind of information.

6. Apart from reflecting the collective identities of bloggers, linguistic variation is also able to demonstrate bloggers' individual identities, which are more easily observable in their use of new lexical items of nonce formation.

From the summary presented above, we can see that linguistic variations in personal blogs are closely related to bloggers' identity representation. Moreover, identity representation as revealed by linguistic variations in personal blogs calls for a more eclectic approach or theoretical framework, as none of the existing theories or frameworks which have mainly been abstracted from the investigation of speech data is sufficient in explaining bloggers' linguistic practices. From the various strategies bloggers have employed to realize orthographic representations of existing words, we can see that language users' greater attention to their language use may not necessarily lead to a more formal style as what traditional variationists claim. It may well result in a more informal style as has already been demonstrated by this thesis. While blogging, the bloggers may occasionally choose certain linguistic strategies for winning approval from the intended audience and they may well be responsive to their readers as Allen Bell's Audience and Referee Design model (1984; 2001) suggests. Nevertheless, this model cannot explain the fact that personal blogs are also a self-expression means which may have very little to do with how the audience are going to react. As the thesis has already demonstrated, many a time bloggers are actually playing an agentive role in their identity construction and many of their linguistic acts are indeed acts of identity as defined by Le Page and Tabouret-Keller (1985). However, deliberate identity construction efforts do not happen all the time. Bloggers' linguistic practice may well be a reflection of their collective identity, as has also been demonstrated in this thesis. The Community of Practice model may help to explain why certain patterns of behavior (linguistic behavior included) emerge or prevail in the certain blogging communities, but it is very difficult to delimit these communities of practice in the first place. Besides, this model seems to be not really suitable for a study like the current one which intends to describe overall patterns and involve cross-group comparisons. What we need for this kind of research is a more eclectic approach or framework which draws on the advantages of the existing ones and takes into the consideration the uniqueness of the medium and the special features of the blogging genre.

10.2 Implications

10.2.1 Personal blogs, corpus, and identity research

This thesis has demonstrated the feasibility and the power of a Wmatrix-based multi-variable approach in investigating identity representation as revealed by linguistic variation personal blogs. Linguistic variation does not find its expression only in the phonological aspects of a particular language as traditional variationists seem to have been advocating. In an era where Internet-based communication is playing an increasingly important role in people's daily communication, the function of writing as a social communication tool has been greatly elevated. The emergence of new written genres such as emails, online chat, and personal blogs has provided language researchers with new fields for observing linguistic variations. Just like in spoken language where people use different phonological features to achieve different communicative purposes, people also use different written features to fulfill different purposes as this thesis has hopefully demonstrated. Just like phonological variations are often associated with the speakers' identities, linguistic variations in a written genre such as personal blogs should also have a great deal to do with the identity representation of the authors. In fact, personal blogs are a very good place (arguably the ideal place) for observing people's linguistic representation of identities. The reason seems to be quite obvious. Personal blogs are recordings of bloggers' daily life experiences and reflections: they are stories about themselves. Unlike daily face-to-face interactions which are normally constrained by the social distances between the conversing participants, personal blogs entitle the authors to greater power in deciding what kind of relations they want to maintain with the intended or potential readers. The anonymous nature of personal blogs has rendered the constraining force of the established social norms less powerful than in face-to-face confrontations and thus makes it possible for the bloggers to choose whatever means they find suitable to express themselves. As a consequence, bloggers are able to present a self which may be quite different from the self in the meat space if they feel like. In that sense, personal blogs could be a better place for us to observe the truer selves and the real identities which the bloggers want to assume. This is simply incomparable by other research scenarios. A more practical reason is that personal blogs provide the real possibility of studying naturally occurring data which can avoid the observer's paradox issue and other issues which might arise out of poor data quality and inaccurate transcription which are often troubling researchers using naturally occurring spoken data. Personal blogs also lend themselves better to corpus-based analysis as the construction of a blog corpus is relatively easier. Moreover, by constructing a corpus of personal blogs in a principled way, we can conduct a whole range of comparisons which will help reveal the similarities and differences between bloggers from different groups defined according to the researchers' criteria, as demonstrated by this thesis. Theoretically speaking, this kind of comparison can be conducted even on individual blogger level if that's what the researcher desires. With the help of natural language processing tools such as Wmatrix and WordSmith Tools, researchers can conduct comparisons based on bloggers' use of a variety of linguistic features. This kind of analysis is able to reveal what pure qualitative analysis based on an extremely small sample cannot easily achieve, as has already been demonstrated in this thesis. Meanwhile, using a corpus for identity research does not undermine the important role of qualitative analysis. To a great extent, they are complimentary to each other.

10.2.2 The unconventionality of personal blogs as linguistic data

The ready availability of personal blogs as corpus data for identity representation analysis does have great advantages compared with naturally occurring (or elicited) spoken data, but it does not follow such data are always easy to process. One of the defining features of a corpus-linguistic approach is its reliance on language processing tools for the retrieval of linguistic features. From the description presented in Chapters 5 to 8 we can see that the language of personal blogs is different from conventional written language in some important ways. One of the most prominent differences is the substantial presence of unconventional orthographic representations for existing English words in the blog texts (see Chapter 5). The presence of such word forms in a corpus forces us to review or reexamine some of the usual practices in corpus-linguistic analysis. As existing language processing tools are trained on and designed for processing standard language data, when they come to unconventional data such as personal blogs, they are very likely to produce distorted reports about some of the very fundamental aspects of descriptive information.

The first aspect to be affected would be the practice of calculating the type token ratio (TTR) of a text or a piece of discourse. TTR is often used to measure the lexical diversity (or density) of a dataset (though many scholars have found that there are flaws in this kind of measurement). The abundance of unconventional spelling variants is very likely to increase the TTR of the blogging texts, especially when the variant forms are resulted from random creativity. The reason is that each new innovative spelling variant will be taken as a new type and a new token in the calculation of TTR. For instance, if there are 5 occurrences of the word *so* in a text and all of them take the conventional orthographic form, this word will be counted only once for the type but 5 times for the token. If two of the occurrences take the original form and the rest three take the forms of *soo*, *sooo*, and *soooo* respectively, then the type counts will increase by three whereas the token counts will remain the same. A distorted TTR report would give a false impression of greater lexical density. The TTR solely based on the counting of orthographic words still makes certain sense, especially in terms of the stylistic aspect (if the researcher is aware of the unconventionality of orthographic words), but it can no longer reflect the lexical diversity it is meant to reflect. Then the issue of how to treat these unconventional orthographic variants arises: should they be taken as new words or repetition of existing words? If it is the former, then there is no need to modify the language processing tools. Nevertheless, many people will find it hard to agree that we should treat orthographic words this way. If it is the latter, we will have to find out a solution to solve this problem. Considering the various ways available for orthographic manipulation, that solution is not going to be simple. Moreover, this touches upon another issue which is more often related to the status of online discourses as data for linguistic investigation. If online discourses are still considered peripheral or sloppy versions of the so-called conventional or standard language use, then nobody will be really bothered to solve problems arising out of the new language use scenario. As illustrated in the thesis (and many other existing studies), the unconventionality of personal blogs as linguistic data is not the result of bloggers' sloppiness, but rather a strategy they employ to achieve various communicative purposes and represent their identities. Thus, it deserves greater attention from language researchers as well as natural language processing specialists.

The second aspect to be affected is what information we can obtain from the average word length. Word length used to be regarded as an indicator of formality. In the English language, there seems to be a positive association between the length of a word and its degree formality. In online discourses such as personal blogs, this conception is also under challenge. As discussed in Chapter 5, bloggers often employ the strategy of letter repetition to emphasize certain words and sometimes the resultant word forms can be as long as containing 25 letters. For instance, spelling the word *so* as an *S* plus 20 *O*s gives a new word form, but this 21-letter word is not indicating formality but rather the opposite: it is actually an indication of informality. A rough count of the corpus used for this research shows that there are more than 1,200 words or word forms which are above 12 letters. This number does not include those word forms containing repeated letters but with the total number of letters below 13. The unconventional long words increase the average word length of a text but they are not necessarily markers of formality. Long hyphenated words as described in Section 6.3 will cause problems to language processing tools.

The third aspect is related to the treatment of nonce words in a corpus. As personal bloggers enjoy great freedom in choosing unconventional orthographic representations and word-formation strategies, the chances for producing nonce words are higher and the importance of studying nonce formations increases, especially when language change or identity representation is the focus. The formation of nonce words is not as random as we originally thought. Rather, there are also word-formation principles working at the background. Chapter 6 of the thesis has offered an account of many of the nonce formations in the corpus. Many of these words are invented for very specific meaning conveyance and identity representation purposes. They may not be very important for identifying general linguistic patterns but they are very important for identifying individual identity. Thus, they should not be disposed of as mere hapax legomena but should be taken into serious consideration at least for identity-related linguistic enquiries.

Apart from challenging the conventional statistical accounts of language processing tools, the unconventionality of blog data has also increased the difficulty in the part-of-speech and semantic annotation of the data. The difficulty may arise from any of the following aspects: 1) unconventional spelling of words, 2) starting proper nouns with lower case letters, 3) omitting apostrophes, 4) creative use of word-formation processes, 5) slanguage words and expressions, 6) abbreviations, 7) lengthening of words by letter repetition, and 8) infusing special symbols into words for special purposes. Although language processing system developers such as those of Wmatrix have been experimenting with letting their clients use their own lexicon (or the customized lexicon) for semantic tagging, this can only solve some of the problems. Moreover, the eventual tagging results will rely heavily on the comprehensiveness and quality of the client’s lexicon. In addition, more research should be carried out in terms of improving the semantic annotation of linguistic corpora. One important issue is how to solve the problem of annotating polysemous words and words of slanguage nature. This thesis has already demonstrated the potential power of what semantic annotation can offer us. With improved semantic annotation, a corpus-linguistic approach is definitely going to be more illuminating.

10.2.3 Personal blogs and other linguistic studies

Another implication of the current research is that bloggers' employment of various strategies in realizing linguistic variations and representing identities has given rise to some interesting linguistic phenomena which may well have been neglected by linguists due to limited samples. For instance, the phrasal compounds that I have described in Chapter 6 deserve more attention from morphologists or lexicologists. A broader (or maybe a brand-new) definition of compounding and a better model for explaining the presence and the internal structures unconventional compounds are needed if we want to have a better idea about the process of compounding. One thing which is also related to the lexis of personal blogs is the substantial existence of new lexical items. Many of these words are newly emerged vocabulary related to IT, new social networking platforms, online gaming, and Internet subcultures. There are also plenty of new slanguage words, new derivatives, and new coinages. These words are good candidates for research related to language change. For lexicographers, they may need to consider whether to take these new lexical items as potential candidates for dictionary entry selection.

10.2.4 Speech-writing relations revisited

One frequently recurrent theme in this thesis is the presence of oral discourse features in the written genre of personal blogs. As pointed out by many existing studies concerning blogs, the language of blogging tends to display a combination of both spoken and written features. This hybrid nature has often been attributed to the uniqueness of the medium as being a publishing tool and a social communication platform at the same time. It is definitely true that the uniqueness of the medium is playing an important role in shaping its linguistic features. Nevertheless, some other factors, I believe, are also playing an important part in adding unconventionality to the language of blogging. These factors are all related to the speech-writing relations to a great extent. They are: the level of author autonomy that blogging authors are enjoying, the different objects for linguistic manipulation in speech and writing, and the issue of spontaneity.

10.2.4.1 Author autonomy matters

From what has been presented in chapters 5 to 8 we can see that bloggers have displayed a tendency of deviating from the established writing norms whenever it is possible. This is something totally unimaginable in informative writing such as academic discourse. One important reason is that authors enjoy different levels of autonomy in different writing contexts. Blogging is a genre which offers arguably the highest level of autonomy to the contributors (i.e. the bloggers). They can decide on the topic, the language style, the discourse structure, the level of accessibility, the frequency of updating. Basically, whether to observe the established writing norm is a matter of choice rather than a matter of imposition. Bloggers have the final say in almost all the core decisions. This next-to-absolute author autonomy has encouraged bloggers to adopt a more pragmatic approach in expressing themselves. Bricolage is thus a natural option for them: making use of whatever linguistic and non-linguistic means available. Bloggers' linguistic repertoire of daily speech will become the most readily available resource and the easiest object to model on. This explains why there are so many oral features in the blogging discourse. In other words, it is the level of author autonomy which has considerably shaped the linguistic features of the blogging language. Also due to the high level of author autonomy, the diversity of strategies that bloggers have chosen to represent themselves becomes a very natural outcome.

10.2.4.2 Objects for linguistic manipulation

Closely related to the level of author autonomy is the issue of the object for linguistic manipulation. As mentioned in Chapter 2, speech and writing are two fundamentally different media with the former relying mainly on the sound waves and the latter on the orthographic symbols. Thus, in speech, what the speakers are manipulating are the sounds and paralinguistic features whereas in writing (be it key-board mediated or not) the object of manipulation becomes the orthographic representations (be it at the lexical level or discoursal level). Determined by the primacy of speech over writing and the readiness of spoken repertoire as an object for modeling, the orthographic representations are almost exclusively mirroring what normally takes place in speech. Just like the spoken sounds reflect lots of information about the speaker, the written representations in personal blogs also reflect many aspects of the bloggers' identities, for instance, their age, gender, ethnicity, cultural backgrounds, country of origin, among other things.

10.2.4.3 The issue of spontaneity

Speech and writing are often distinguished from each other by the feature of spontaneity. Speech takes place in spontaneity by default whereas writing is generally taken to be non-spontaneous by nature. This is normally the case. Nevertheless, when it comes to blogging, the issue of spontaneity is no longer absolutely irrelevant. It is quite likely that bloggers would sacrifice accuracy, correctness, or completeness so that they are able to capture the flow of thoughts or emotions before they elapse. This might be one of the contributing factors for the substantial presence of misspellings (phonetic spellings included), abbreviated forms, apostrophe-less contractions, neglected upper and lower cases, elliptical sentence structures, and dominance of coordinated structures, and run-on sentences. The spontaneity displayed in blogging, however, may not be the same as that in spoken contexts. It is still imprinted with written language features. For instance, there are no false starts or hesitations in personal blogs, showing the authors' awareness that they are writing things rather than really speaking.

10.2.4.4 Speech and writing as changing concepts

Like many other existing studies concerning Internet-mediated discourses, and blogs in particular, this thesis also reveals a blurred boundary between speech and writing. In fact, this trend of colloquialization of written discourse did not originate from the advent of Internet-based communication and it is not going to stop here either. The convergence of speech and writing in many ways has a great deal to do with the social development of human society which is characterized by a greater tolerance of diversity and a greater respect for individual differences. The information and communication technology has actually helped to accelerate the process of this development. With the general public gaining greater and greater autonomy in expressing themselves in writing (with the help of Internet-based communication platforms), they have not only acquired a good stage for identity representation but also gained the power of having a say in deciding what a piece of writing should be like. By experimenting with new writing styles characterized by substantial presence of spoken discourse features, they are not only challenging the established writing norms but also shaping new norms, new attitudes and expectations about what writing should be like. This will in turn promote the leveling of discourses. Having said that, I am not suggesting that colloquialization of written discourse is the only thing that is happening. In fact, the concepts of speech and writing are changing constantly according to the social and technological developments of human society and they are always influencing each other. The advent of Internet-based communication has opened up another channel for this mutual influence to take place. Maybe it is better for us to take speech and writing as two linguistic resources to draw on for more effective communication.

10.3 Limitations of current research

Despite that the current research has obtained some interesting insights about how personal bloggers are representing their identities via linguistic variation, it has several limitations:

1. The sample size is not big enough. Although a total number of 460 bloggers have been included in this research, considering the number of age and gender groups, it is quite small: only 20 bloggers from each gender group of the same age. Due to the limited access of data, the British component of the corpus does not have the mid-teens male group for comparison with the American component. Within a less constrained time frame and more manpower, it is advisable to have a greater sample size so as to increase the generalizability of the findings and reduce the risk of the findings being skewed by individual informants.
2. The grouping of the bloggers from this research also has room for improvement. It might be more reasonable to leave some gaps between age groups so that overlapping between neighboring groups could be reduced to the minimum.
3. More syntactic and discoursal features such as syntactic complexity, ellipsis should be included so that a fuller picture of linguistic variation and identity representation could be depicted. Again, that is only feasible with more time and team work, as analyzing these features may involve labor-intensive manual annotation.
4. Although language processing software tools such as Wmatrix have played an extremely important role in the current research, due to the unconventionality of

the data, cases of erroneous tagging (or annotation) and miscalculation are inevitable, which might have led to occasional misinterpretations.

10.4 Future research

For future research, the following aspects could be considered:

1. Focusing more on syntactic features and discoursal features. This is a topic which has seldom been tapped in existing studies concerning personal blogs. As existing language process software tools are unable to process syntactic and discoursal features. Computer-assisted human annotation will be necessary if a corpus-linguistic approach is to be adopted. How to realize that is also an important topic for future research.
2. Taking other semiotic features into consideration. As mentioned in this research, personal blogs do not only consist of textual messages. In fact, the semiotic presentations could also be an interesting vantage point for observing identity representation in personal blogs. This can be approached from a multi-modal perspective rather than a more text-oriented approach as I have adopted for this research. A multi-modal corpus-linguistic approach would be a future direction for studying identity representation in personal blogs.
3. A cross-linguistic investigation about linguistic variation and identity representation in personal blogs is worthy of exploration if the researcher is interested in tapping the influence of language difference and cultural differences on identity representations.

10.5 Final remarks

Adopting a Wmatrix-based multi-variable approach supplemented with qualitative analysis, I have conducted a quite comprehensive investigation about how identities are represented or reflected through linguistic variations in personal blogs. By examining bloggers' practice in orthographic representations, lexicological strategies, slanguage use, preference for semantic domains, use of non-conventional grammatical features, and employment of pragmatic features, I have demonstrated the necessity of adopting an eclectic framework in understanding the multi-faceted concept of identity and an eclectic analysis approach in capturing the various linguistic strategies for identity representation in a written genre. My findings have also revealed that deviating from the established writing norms and transplanting oral discourse features into blogging are two major means for bloggers to represent various aspects of their identities.

Bibliography

Ackema, P., & Neeleman, A. (2004). *Beyond morphology: Interface conditions on word formation*. New York: Oxford University Press.

Adams, V. (2001). *Complex words in English*. Harlow: Pearson/Longman.

Al-Sa'Di, R. A., & Hamdan, J. M. (2005). "Synchronous online chat" English: Computer-mediated communication. *World Englishes, 24*(4), 409-424.

Allan, K., & Burridge, K. (2006). *Forbidden words: Taboo and the censoring of language*. Cambridge: Cambridge University Press.

Allen, I. L. (1998). Slang: Sociology. In J. L. Mey (Ed.), *Concise encyclopedia of pragmatics* (pp. 878-883). Amsterdam; New York: Elsevier.

The American Heritage College Dictionary (4th ed.). (2002). Boston: Houghton Mifflin.

Andersen, G. (2000). *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*. Amsterdam/Philadelphia: Benjamins.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321-346.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday, 12*(9). Retrieved September 18, 2008, from http://outreach.lib.uic.edu/ www/issues/issue12_9/argamon/

Barber, C. (1997). *Early Modern English*. Edinburgh: Edinburgh University Press.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics, 12*(1), 58-88.

Baron, N. S. (1998). Letters by phone or speech by other means: The linguistics of email. *Language & Communication, 18*, 133-170.

Baron, N. S. (2002). Language of the Internet. In A. Farghali (Ed.), *The Stanford handbook for language engineers* (pp. 59-127). Stanford: CSLI Publications.

Baron, N. S. (2004). See you online: Gender issues in college student use of Instant Messaging. *Journal of Language and Social Psychology, 23*(4), 397-423.

Bauer, L. (2006). Compounds and minor word-formation types. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 483-506). Malden, MA: Blackwell Publishing.

Bauer, L., & Renouf, A. (2001). A corpus-based study of compounding in English. *Journal of English Linguistics, 29*(2), 101-123.

Bearn, G. C. F. (2000). Differentiating Derrida and Deleuze. *Continental Philosophy Review, 33*(4), 441-465.

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145-204.

Bell, A. (2001). Back in style: Reworking audience design. In P. Eckert & J. R. Rickford (Eds.), *Style and Sociolinguistics* (pp. 139-169). Cambridge: Cambridge University Press.

Biber, D., & Burges, J. (2000). Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics 28*(1), 21.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. New York: Longman.

Blood, R. (2002). *The weblog handbook: Practical advice on creating and maintaining your blog*. Cambridge, MA: Perseus Publishing.

Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM, 47*(12), 53-55.

Bloomfield, L. ([1933] 1984). *Language*. Chicago: University of Chicago Press.

Bolinger, D. L. (1946). Visual morphemes. *Language, 22*(4), 333-340.

Brake, M. (1985). *Comparative youth culture: The sociology of youth cultures and youth subcultures in America, Britain, and Canada*. London: Routledge & K. Paul.

Brinton, L. J. (1996). *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: Mouton de Gruyter.

Bucholtz, M. (2000). Language and youth culture. *American Speech, 75*(3), 280-283.

Bucholtz, M. (2003). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics 7*(3), 398-416.

Cameron, D. (1998). Gender, language, and discourse: A review essay. *Signs, 23*(4), 945-973.

Carter, J. B. (2006). English spelling reform. *Prometheus, 24*(1), 81-100.

Carter, R. (1999). Common language: Corpus, creativity and cognition. *Language and Literature, 8*(3), 195-216.

Carter, R. (2001/2002). A response to Neal R. Norrick. *Connotations, 11*(2-3), 291-197.

Carter, R. (2004). *Language and creativity: The art of common talk*. London: Routledge.

Carter, R. (2007). Response to Special Issue of Applied Linguistics devoted to language creativity in everyday contexts. *Applied Linguistics, 28*(4), 597–608.

Carter, R., & McCarthy, M. (2004). Talking, creating: Interactional language, creativity, and context. *Applied Linguistics, 25*(1), 62-88.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide: Spoken and written English grammar and usage*. Cambridge: Cambridge University Press.

Cerulo, K. A. (1997). Identity construction: New issues, new directions. *Annual Review of Sociology, 23*, 385-409.

Chambers, J. K. (2003). *Sociolinguistic theory: Linguistic variation and its social significance*. Oxford: Blackwell.

Channell, J. (1994). *Vague language*. Oxford: Oxford University Press.

Coates, J. (1993, 2004). *Women, men, and language: A sociolinguistic account of gender differences in language*. Harlow, England: Pearson Longman.

Colley, A., & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology, 21*(4), 380-392.

Collins COBUILD English Dictionary for Advanced Learners (3rd ed.). (2001). Glasgow: HarperCollins.

Collot, M., & Belmore, N. (1996). Electronic language: A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 13-28). Amsterdam: John Benjamins.

Cook, G. (1997). Language play, language learning. *ELT Journal, 51*(3), 224-231.

Cook, G. (2000). *Language play, language learning*. Oxford: Oxford University Press.

Cook, V. (2008). Writing Systems. Retrieved January 10, 2009, from http://homepage.ntlworld.com/vivian.c/index.htm

Coupland, N. (2001). Language, situation, and the relational self: Theorizing dialect-style in sociolinguistics. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 185-210). Cambridge: Cambridge University Press.

Coupland, N. (2007). *Style: Language variation and identity*. Cambridge: Cambridge University Press.

Crain, C. (2008). Pixies, Sheilas, Dirtbags and Cougar Bait: Modern Slang. *The Nation*. Retrieved January 10, 2009, from http://www.thenation.com/doc/20081229/crain

Crystal, D. (1995). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.

Crystal, D. (2001a). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2001b). *Language play*. Chicago: University of Chicago Press.

Crystal, D. (2006). *Language and the Internet* (2nd ed.). Cambridge, UK; New York: Cambridge University Press.

Crystal, D. (2008). *Think on my words: Exploring Shakespeare's language*. New York: Cambridge University Press.

Crystal, D., & Davy, D. (1975). *Advanced conversational English*. London: Longman.

Damaso, J., & Cotter, C. (2007). UrbanDictionary.com. *English Today, 23*(2), 19-26.

Davies, B. (2005). Communities of practice: Legitimacy not choice. *Journal of Sociolinguistics, 9*(4), 557-581.

De Klerk, V. (1990). Slang: A male domain. *Sex Roles, 22*, 589-606.

Eble, C. (1996). *Slang & sociability: In-group language among college students*. Chapel Hill: University of North Carolina Press.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist, 44*(2), 78-89.

Eckert, P. (1997). Age as sociolinguistic variable. In F. Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 151-167). Oxford: Blackwell.

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Oxford: Blackwell.

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology, 21*, 461-490.

Eckert, P., & McConnell-Ginet, S. (1999). New generalizations and explanations in language and gender research. *Language in Society, 28*(2), 185-201.

Eckert, P., & McConnell-Ginet, S. (2003). *Language and gender*. Cambridge: Cambridge University Press.

Edwards, J. (1985). *Language, society and identity*. Oxford: Basil Blackwell.

Ehrlich, S. (1999). Communities of practice, gender, and the representation of sexual assault. *Language in Society, 28*(02), 239-256.

Erickson, T. (1999). Persistent conversation: An Introduction. *Journal of Computer-mediated Communication, 4*(4). Retrieved September 18, 2008, from http://jcmc.indiana.edu/vol4/issue4/ericksonintro.html

Erikson, E. (1956, 2008). The problem of ego identity. In D. L. Browning (Ed.), *Adolescent identities: A collection of readings* (pp. 223-240). New York: The Analytic Press.

Erikson, E. (1959). *Identity and the life cycle: Selected papers by Erik H. Erikson*. New York: International Universities Press.

Erikson, E. (1963). *Childhood and society*. New York: Norton.

Erman, B. (2001). Pragmatic markers revisited with a focus on you know in adult and adolescent talk. *Journal of Pragmatics, 33*(9), 1337-1359.

Fernback, J. (2003). Legends on the net: An examination of computer-mediated communication as a locus of oral culture. *New Media & Society, 5*(1), 29-45.

Finegan, E. (2004). *Language: Its structure and use*. Boston, MA: Thomson Wadsworth.

Fitzpatrick, L. (2008, Tuesday, August 12). Making an arguement for misspelling. *Time*. Retrieved September 18, 2008, from http://www.time.com/time/world/article/0,8599,1832104,00.html

Freed, A. F. (1999). Communities of practice and pregnant women: Is there a connection? *Language in Society, 28*(02), 257-271.

Garcia, A. C., & Jacobs, J. B. (1999). The eyes of the beholder: Understanding the turn taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction, 32*(4), 227-367.

Gerrig, R. J., & Gibbs, R. W. (1988). Beyond the lexicon: Creativity in language production. *Metaphor and Symbol, 3*(1), 1-19.

Giles, H. (2008). Communication accommodation theory. In L. A. Baxter & D. O. Braithwaite (Eds.), *Engaging theories in interpersonal communication: Multiple perspectives* (pp. 161-173). London: Sage Publications.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation*. London: Academic Press.

Gong, W., & Ooi, V. B. Y. (2008). Innovations and motivations in online chat. In S. Kelsey & K. St.Amant (Eds.), *Research handbook on computer mediated communication* (Vol. 1, pp. 917-933). Hershey, PA: Information Science Reference.

Goody, J. (1992). Oral culture. In R. Bauman (Ed.), *Folklore, cultural performances, and popular entertainments* (pp. 12-20). New York: Oxford University Press.

Görlach, M. (1991). *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Grossman, A. L., & Tucker, J. S. (1997). Gender differences and sexism in the knowledge and use of slang. *Sex Roles, 37*(1/2), 101-110.

Gumbrecht, M. (2004). *Blogs as "protected space"*. Paper presented at the Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics: WWW 2004.

Halliday, M. A. K. (1975). Anti-Languages. *American Anthropologist, 78*(3), 570-584.

Herring, S. C. (1994). Gender differences in computer-mediated communication: bringing familiar baggage to the new frontier. Keynote talk presented at the annual convention of the American Library Association, Miami, FL. Retrieved September 18, 2008, from http://cpsr.org/ issues/womenintech/herring2/

Herring, S. C. (2000). Gender differences in CMC: Findings and implications. *The CPSR Newsletter, 18*(1).

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612-634). Oxford: Blackwell Publishers.

Herring, S. C. (2004a). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338-376). New York: Cambridge University Press.

Herring, S. C. (2004b). Content analysis for new media: Rethinking the paradigm. In *New research for new media: Innovative research methodologies symposium working papers and readings* (pp. 47-66). Minneapolis, MN: University of Minnesota School of Journalism and Mass Communication.

Herring, S. C. (2008). Web content analysis: Expanding the paradigm. In J. Hunsinger, M. Allen & L. Klastrup (Eds.), *The international handbook of internet research*: Springer Verlag.

Herring, S. C. (Ed.). (1996). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Amsterdam: Benjamins.

Herring, S. C., Kouper, I., Scheidt, L. A., & Wright, E. L. (2004). Women and children last: The discursive construction of weblogs. In L. Gurak, S. Antonijevic, L. A. Johnson, C. Ratliff & J. Reyman (Eds.), *Into the blogosphere: Rhetoric, community, and culture of weblogs*.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. *Information Technology & People, 18*, 142-171.

Hogan, R. (1991). Engendered autobiographies: The diary as a feminine form. *Prose Studies: History, Theory, Criticism, 14*(2), 95-107.

Holmes, J. (1992, 2001). *An introduction to sociolinguistics*. Harlow, England: Longman.

Holmes, J. (1995). *Women, men, and politeness*. New York: Longman.

Holmes, J. (1998). Women's role in language change: A place for quantification. In Natasha Warner et al. (Eds.), *Gender and belief systems: Proceedings of the Fourth Berkeley Women and Language Conference, 1996* (pp. 313-330). Berkeley: Berkeley Women and Language Group.

Holmes, J. (2006). *Gendered talk at work: Constructing gender identity through workplace discourse*. Oxford: Blackwell.

Holmes, J., & Meyerhoff, M. (1999). The Community of Practice: Theories and methodologies in language and gender research. *Language in Society, 28*(2), 173-183.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-mediated Communication,10*(2). Retrieved September 18, 2008, from http://jcmc.indiana.edu/vol10/issue2/ huffaker.html

Jespersen, O. (1922). *Language, its nature, development and origin*. London: Allen & Unwin.

Johnstone, B. (2000). The individual voice in language. *Annual Review of Anthropology, 29*(1), 405-424.

Karlsson, L. (2006). Acts of reading diary weblogs. *Human IT*, *8*(2), 1-59.

Kegan, R. (1982). *The evolving self: Problem and process in human development*. Cambridge, MA: Harvard University Press.

Kendall, L. (2007). "Shout into the wind, and it shouts back": Identity and interactional tensions on LiveJournal. *First Monday, 12*(9). Retrieved September 18, 2008, from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2004/1879

Kiesling, S. F. (2004). Dude. *American Speech, 79*(3), 281-305.

Kress, G. R. (2003). *Literacy in the new media age*. London: Routledge.

Kroger, J. (2007). *Identity development: Adolescence through adulthood* (2nd ed.). Thousand Oaks, California: Sage Publications.

Kroger, J., & Adair, V. (2008). Symbolic meanings of valued personal objects in identity transitions of late adulthood. *Identity: An International Journal of Theory and Research, 8*(1), 5-24.

Kroskrity, P. V. (1999). Identity. *Journal of Linguistic Anthropology, 9*(1-2), 111-114.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.

Labov, W. (2001). *Principles of linguistic change* (Vol. 2). Oxford: Blackwell.

Lave, J., & Wenger, É. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lawler, S. (2008). *Identity: Sociological perspectives*. Cambridge: Polity.

Le Page, R. B., & Tabouret-Keller, A. (1985). *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge: Cambridge University Press.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 133-149). Amsterdam/New York: Rodopi.

Levine, L. W. (1992). The folklore of industrial society: Popular culture and its audiences. *The American Historical Review, 97*(5), 1369-1399.

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. New York: Routledge.

Loevinger, J. (1976). *Ego development: Conceptions and theories*. San Francisco: Jossey-Bass.

Longman Dictionary of Contemporary English. (2006). Harlow, Essex: Pearson Education.

Maybin, J., & Swann, J. (2007). Everyday creativity in language: Textuality, contextuality, and critique. *Applied Linguistics, 28*(4), 497-517.

McEnery, A., & Xiao, Z. (2003). *Fuck revisited*. Paper presented at the Corpus Linguistics 2003.

McGann, R. (2004). The blogosphere by the numbers. *The ClickZ Network*. Retrieved September 18, 2008, from http://www.clickz.com/showPage.html?page=3438891

Mead, G. H. (1934). *Mind, self and society from the standpoint of a social behaviorist*. Chicago: The University of Chicago press.

Meibauer, J. (2007). How marginal are phrasal compounds? Generalized insertion, expressivity, and I/Q-interaction. *Morphology, 17*, 233-259.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*(4), 857-870.

Mendoza-Denton, N. (2002). Language and identity. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 475-499). Malden, MA: Blackwell Publishers.

Merchant, G. (2005). Electric Involvement: Identity performance in children's informal digital writing. *Discourse: Studies in the cultural politics of education, 26*(3), 301 - 314.

Merriam-Webster's Collegiate Dictionary (11th ed.). (2005). Springfield, MA: Merriam-Webster.

Meyerhoff, M. (2002). Communities of Practice. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 526-548). Malden, MA: Blackwell.

Meyrowitz, J. (1997). Shifting worlds of strangers: Medium theory and changes in "them" and "us". *Sociological Inquiry, 67*(1), 59-71.

Miller, J. E. (2001). Speech and writing. In R. Mesthrie (Ed.), *Concise encyclopedia of sociolinguistics* (pp. 270-276). Amsterdam: Elsevier.

Mondorf, B. (2002). Gender differences in English syntax. *Journal of English Linguistics 30*(2), 158-180.

Mulac, A., Bradac, J. J., & Gibbons, P. (2001). Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research, 27*(1), 121-152.

Mulac, A., & Lundell, T. L. (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects. *Language & Communication, 14*(3), 299-309.

Nardi, B. A., Schiano, D. J., & Gumbrecht, M. (2004). *Blogging as social activity, or, would you let 900 million people read your diary?* Paper presented at the 2004 ACM Conference on Computer Supported Cooperative Work.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

Nevalainen, T. (2006). *An introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

North, S. (2007). 'The voices, the voices': Creativity in online conversation. *Applied Linguistics, 28*(4), 538-555.

Nowson, S., Oberlander, J., & Gill, A. J. (2005). *Weblogs, genres, and Individual differences*. Paper presented at the 27th Annual Conference of the Cognitive Science Society. Retrieved September 18, 2008, from http://www.ics.mq.edu.au/~snowson/papers/nowson-cogsci.pdf

Nystrand, M. (1983). The role of context in written communication. *The Nottingham Linguistic Circular, 12*, 55-65.

Ochs, E. (1993). Constructing social identity: A language socialization perspective. *Research on Language and Social Interaction, 26*(3), 287-306

Ooi, V. B. Y. (2002). Aspects of computer-mediated communication for research in Corpus Linguistics. In P. Peters, P. Collins & A. Smith (Eds.), *New frontiers of corpus research: Papers from the Twenty-First International Conference on English Language Research on Computerized Corpora, Sydney 2000* (pp. 91-104). Amsterdam-New York: Rodopi.

Ooi, V. B. Y., Tan, P. K. W., & Chiang, A. K. L. (2007). Analyzing personal weblogs in Singapore English: the Wmatrix approach. *eVariEng (Journal of the Research*

Unit for Variation, Contacts, and Change in English), 2. Retrieved September 18, 2008, from http://www.helsinki.fi/varieng/journal/volumes/02/ooi_et_al/

Orlowski, A. (2003). Most bloggers “are teenage girls”- survey. *The Register*. Retrieved September 18, 2008, from http://www.theregister.co.uk/2003/05/30/ most_bloggers_are_teenage_girls/

Overstreet, M. (1999). *Whales, candlelight, and stuff like that: General extenders in English discourse*. Oxford: Oxford University Press.

Oxford Advanced Learner's Dictionary (6th ed). (2000). Oxford: Oxford University Press.

Papacharissi, Z. (2002). The virtual sphere: The Internet as a public sphere. *New Media & Society, 4*(1), 9-27.

Peccei, J. S. (1999). Language and age. In L. Thomas & S. Wareing (Eds.), *Language, society and power: An introduction* (pp. 99-115). London: Routledge.

Pedersen, S., & Macafee, C. (2007). Gender differences in British blogging. *Journal of Computer-Mediated Communication, 12*(4).

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, ourselves. *Annual Review of Psychology, 54*, 547-577.

Pennycook, A. (2007). 'The rotation gets thick. The constraints get thin': Creativity, recontextualization, and difference. *Applied Linguistics, 28*(4), 579-596.

Piao, S. S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, T., et al. (2005). *A large semantic lexicon for corpus annotation*. Paper presented at the Corpus Linguistics 2005, July 14-17, Birmingham, UK.

Plag, I. (2003). *Word-formation in English*. Cambridge: Cambridge University Press.

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Unpublished PhD thesis. Lancaster University.

Rayson, P. (2008a). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-550.

Rayson, P. (2008b). Wmatrix: a web-based corpus processing environment: Computing Department, Lancaster University.

Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2*(1), 133-152.

Riley, P. (2007). *Language, culture and identity: An ethnolinguistic perspective*. London: Continuum.

Romaine, S. (2003). Variation in language and gender. In J. Holmes & M. Meyerhoff (Eds.), *The handbook of language and gender* (pp. 98-118). Malden, MA: Blackwell.

Schaap, F. (2004). Links, lives, logs: Presentation in the Dutch blogosphere. *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. Retrieved September 18, 2008, from http://blog.lib.umn.edu/blogosphere/ links_lives_logs.html

Schiano, D. J., Nardi, B. A., Gumbrecht, M., & Swartz, L. (2004). *Blogging by the rest of us*. Paper presented at the Conference on Human Factors in Computing Systems (CHI 2004).

Schilling-Estes, N. (2002). Investigating stylistic variation. In J. K. Chambers, P. Trudgill & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 375-401). Malden, MA: Blackwell Publishers.

Schönfeldt, J., & Golato, A. (2003). Repair in chats: A conversation analytic approach. *Research on Language and Social Interaction, 36*(3), 241-284.

Schwartz, G., & Merten, D. (1967). The language of adolescence: An anthropological approach to the youth culture. *The American Journal of Sociology, 72*(5), 453-468.

Scott, M. (1999). WordSmith Tools (Version 3.00.00).

Sebba, M. (2003). Spelling rebellion. In J. K. Androutsopoulos & A. Georgakopoulou (Eds.), *Discourse constructions of youth identities* (pp. 151-172). Amsterdam: Benjamins.

Shank, G., & Cunningham, D. (1996). Mediated phosphor dots: Toward a post-Cartesian model of computer-mediated communication via the semiotic superhighway. In C. Ess (Ed.), *Philosophical perspectives on computer-mediated communication* (pp. 27-41). Albany, NY: State University of New York Press.

Sinclair, J. M. (2001). Preface. In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. vii-xv). Amsterdam/Philadelphia: Benjamins.

Sinclair, J. M. (2004). *Trust the text: Language, corpus and discourse*. New York, N.Y.: Taylor & Francis.

Sinclair, J. M. (1991). *Corpus, collocation, concordance*. Oxford: Oxford University Press.

Stenström, A.-B., Anderson, G., & Hasund, I. K. (2002). *Trends in teenage talk: Corpus compilation, analysis and findings*. Amsterdam/Philadelphia: Benjamins.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 3-15). Cambridge: Cambridge University Press.

Taboada, M. (2004). The genre structure of bulletin board messages. *Text Technology, 13*(2), 55-82.

Tabouret-Keller, A. (1997, 2000). Language and identity. In F. Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 315-326). Oxford: Blackwell Publishers.

Tagliamonte, S. (2005). So who? Like how? Just what?: Discourse markers in the conversations of young Canadians. *Journal of Pragmatics, 37*(11), 1896-1915.

Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series Friends. *American Speech, 80*(3), 280-300.

Tannen, D. (1990). *You just don't understand: Women and men in conversation.* New York: William Morrow.

Tannen, D. (1995). *Gender and discourse*. Oxford: Oxford University Press.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics 10*(1), 1-13.

Teubert, W., & Čermáková, A. (2007). *Corpus Linguistics: A short introduction*. London: Continuum.

Thompson, N. (2003). *Communication and language: A handbook of theory and practice*. Basingstoke, Hampshire: Palgrave MacMillan.

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology, 40*(2), 193-208.

Tree, J. E. F., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language, 40*(2), 280-295.

Urbandictionary. (2009). http://www.urbandictionary.com/

van Dijck, J. (2004). Composing the self: Of diaries and lifelogs. *Fibreculture, 3*. Retrieved September 18, 2008, from www.journal.fibreculture.org/ issue3/issue3_vandijck.html

van Doorn, N., van Zoonen, L., & Wyatt, S. (2007). Writing from experience: Presentations of gender identity on weblogs. *European Journal of Women's Studies, 14*(2), 143-158.

Vaughan, G. M., & Hogg, M. A. (2005). *Introduction to social psychology* (4th ed.). Frenchs Forest, N.S.W: Prentice Hall.

Warschauer, M. (2001). Language, identity, and the Internet. *Mots Pluriels* No 19. October 2001. Retrieved October 12, 2009, from http://www.arts.uwa.edu.au/ MotsPluriels/MP1901mw.html

Weber, S., & Mitchell, C. (2008). Imaging, keyboarding, and posting identities: Young people and new media technologies. In D. Buckingham (Ed.), *Youth, identity, and digital media* (pp. 25-47). Cambridge, Massachusetts: The MIT Press.

Werry, C. C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 47-63). Amsterdam: Benjamins.

Wiese, R. (1996). Phrasal compounds and the theory of word syntax. *Linguistic Inquiry, 27*(1), 183-193.

Wikipedia: The free encyclopedia. http://www.wikipedia.org

Woolbert, C. H. (1922). Speaking and writing -- a study of differences. *Quarterly Journal of Speech Education, 8*(3), 271-285.