Characterizing Language Usage in Formal and Informal Webpage Text Corpus

Virendra Singh Nirban, Birla Institute of Technology and Science, India
Meenakshi Raman, Birla Institute of Technology and Science, India

Abstract

Language is an integral part of any culture. English has changed substantially in the last 1500 years of its use, reflecting patterns of contact with other languages and the changing communication needs of people. The significant changes in structure and use of language can be attributed to the effect of medium of communication. English is also the lingua franca of modern corporate world. With the advent of technologies that provide platforms for creation of virtual communities, the scope for influence of medium on language is substantial. Moreover technologies like World Wide Web and Internet has provided infinite communication avenues to corporate world, individual entrepreneurs and professionals to reach a wider clientele base for business. The ever increasing online user population is creating a vast amount of English text corpus which substantially reflects the patterns of language usage. This paper presents a perspective on the influence of computer-mediated communication on language usage in formal and informal context with special reference to empirical studies conducted using webpage text corpus. The issues considered in the analysis include language usage, special aspects of CMC, user population and the overall influence of communication technology on language and its repercussions in society.

Keywords: Computer-Mediated Communication, Empirical Analysis, Webpage Corpus, Culture, Discourse Analysis.

Introduction

The World Wide Web (WWW) has provided a new way of communication to human beings and can be compared to a magic vehicle which effortlessly transports one from one country to another, one culture to other speaking different varieties of the same language. As we take a whirlwind tour of the world through the websites across the globe, we might expect to find significant differences in terms of language use, and structural elements; the truth is that we can't ignore the similarities in the same. In one of his interviews, David Crystal said:

> The World Wide Web provides a platform to virtually all the styles which have so far developed in the written language: newspapers, scientific reports, bulletins, novels, poems, prayers. Indeed, it's introducing us to new styles of written expressions which none of us have ever seen before – the animated and the abbreviated language specifically. The web is truly a part of a new linguistic medium, more dynamic than traditional writing, and more permanent than traditional speech. It's often been said the Internet is a revolution - yes, indeed, but it's also a linguistic revolution (2001).

Evans, B Mary, et.al, studied the degree of formality and tone deviations in English Language websites. They found that the tone was significantly more formal in websites from countries where English is a second language compared with countries where English is a first language (2005: 849). Some authors assert that a personal, informal tone improves the credibility of written messages. For example, Coney and Steehouder maintain that personal tone such as "In the British Library, you will find…" is better than impersonal, third person tone like "The British Library provides…" because personal tone is more inviting (2000: 332). However, other authors assert that such a tone may not be effective in all situations, and warn that readers from different cultures judge tone differently. For example, Ferraro in his book *The Cultural Dimensions of International Business* states that the personal, informal tone preferred by North Americans can offend people from more formal cultures (2000). Similarly, Hodge in his book *Global Smarts: The Art of Communicating and Deal Making Anywhere in the World* recommends that Americans who visit other countries speak "more formally when they are abroad than when they are in the U.S." (2000: 67). These authors suggest that communicators need to take culturally-based differences in rhetorical expectations into account when crafting their messages. No doubt, the companies trust

Internet to communicate. And English has become the lingua franca of the Internet based communication. Webpage is the most common form of CMC. Even before checking emails, engaging in IRC, blogs or an online conference, the user needs to visit the webpage of the service provider. Webpages have come a long way to become the new face of today’s organization. So it becomes important for a company to present itself through static (sometimes dynamic also) content in form of text, images and colors to communicate effectively.

At the same time, Personal webpages present a new channel for the masses. Hosting a personal web page is convenient, affordable, and allows people to present a multi-mediated self, using audio-visual components and text to communicate to potential mass audience. Authoring personal webpage has become a popular type of Internet use. Internet Service Providers and online hosting portals like Geocities etc., aid this trend. As both informative and expressive tools, personal webpages allow individuals to transcend from consumers of media content to media producers (Dominick 1999: 651). Authors of personal webpages are not merely sharing information with others; they are also engaged in establishing a sense of self on virtual terrain. The personal home page is a media product very heterogeneous in nature. The volume of personal home pages varies between one document and hundreds of files, the number of external links ranges from zero to more than a thousand, the spectrum of functions and topics, of text types and language styles is large and has been only partially studied (Doring 2002). Smith, Siltanen and Hosman (1998: 30) have looked at how powerful and powerless language styles affect evaluations of a speaker's authoritativeness, sociability, and similarity to the receiver, while Adkins and Brashers (1995: 292) examined the impact of such language styles on attractiveness, credibility, and persuasiveness in CMC. Some studies of perceptions of CMC language variables have focused on politeness/impoliteness or grammar use (Jessmer and Anderson 2001: 10). These studies suggest that variations in language styles influence the audience's perception of the writer.

Organizations need to assess whether the effectiveness of their business can be enhanced through CMC. The users of WWW need to address whether the personal communication revolution happening all around them in the form of personal space on the Internet affect language adversely. Similarly, it is important for the researchers to study the impact of such communication revolutions. They need to analyse whether the language used in WWW follows standard constructions and usage rules or there is a mutation happening to the way language is used in today's Internet age.

Methodology

For the study on webpages, twenty five webpages each from Professional and Personal Context were selected based on the premise that it had enough textual content rather than hyper textual content. For professional webpages, different industry domains such as telecommunications, information technology, academics, health, business process outsourcing, manufacturing and entertainment have been selected at random. For personal webpages the selection has been random varying from individual professionals like doctors and engineers to personal blog pages to family webpages and personal diaries. It must be noted that the webpages in this study have not been restricted to one single screen content. In many instances a single web page has run for more than one screen page. After the selection, basic observations have been made on the number of paragraphs, their length in terms of words and sentences, number of sentences and word count. Once this has been done, other observations have been made and the procedure extended to the data compilation stage.

We have used the "coding and counting" paradigm of classical content analysis which is a specific CMDA approach and it scrutinizes the online communication behavior through the lens of language, and its interpretations are grounded in observations about language and language use. For webpages analysis, coding categories and their analysis parameters such as vocabulary usage (use of collocations, multiword chunks, cohesive markers, abstract and concrete words); grammar usage (use of tense forms, active and passive voice, relative clauses, conditional clauses, modal verbs, use of pronouns); communicative functions (informing/announcing, acknowledging, persuading /offering services); conversational tone (greeting and identification, acknowledge, question, request, encourage for more communication, thanksgiving / final greeting); deficiencies (deviations in paragraph structure, sentence length, spellings and other grammatical flaws, mutations and the use of emoticons) have been established. The data for this study has been produced naturally (i.e., participants have created online discourse for their own purposes in real world environment), and the text has been retrieved from live websites or online archives. The context was created by classifying the data into Formal and Informal macro categories. After compilation, Descriptive Analysis has been used to explain the basic features of the data gathered from the study and it helped in providing the statistical summaries of the sample.

Findings

The study revealed that the language used in professional webpages follow a plain style. Collocations are used as an important lexical item for achieving economy of expression; appropriate cohesive markers are used for connectivity of message chunks; present perfect and simple past tense are used to convey a continuous and sustained growth; specific pronouns were used to convey a sense of belonging and involvement with the user; authors of professional webpages used sentence variety and paragraph cohesion to achieve effective style of composition. On the other hand, in case of personal webpages, multiword chunks, personal pronouns, interactional act identifiers, past perfect tense, conditional clauses, high and inappropriate use of conjunctions, etc., were observed to mark the style as conversational and informal. The analysis of the webpages from professional and personal contexts highlighted use of language according to the context to a large extent. However, there are elements of style which were ignored or violated while writing for the web. This section discusses these deviations from the traditions style of writing.

Web masters of the organizational webpages are very careful in terms of the **spellings** of words. Not a single spelling mistake could be observed in the sample webpages. Apart from the content draft going through several hands, this could also be attributed to the software utilities like spell check which is run on the text to check the mistakes. On the other hand, personal webpages were replete with spelling mistakes. This factor can be attributed to carelessness, deficiency in English language, or typographic inabilities of the authors. But the fact that Personal webpages are contributing to bad spelling skills can’t be ignored. Figure 1 show several examples of spelling mistakes.

… natural talent to **aquire** and absorb … (PP_11)
… donation bin on **cemetary** road … (PP_10)

… I also want to teach at a college or **unversity**. (PP_17)

... I like many types of music but my **favurite** singer/writer is Morrissey. (PP_9)

Figure 1. Spelling Mistakes

Starting a sentence with lower case is another frequent occurrence that was observed in the analysis. This indicates a lack of attention toward punctuation in running text. Figure 2 exemplifies the same.

… i was inspired by corrinne and terry's LJ. (PP_10)

… as i grew my oratary skills were promptly recognised in local competitions and school level as well.. (PP_23)

… it was a 1 bedroom. (PP_6)

Figure 2. Sentence start

Grammatical constructions that generally appear in pairs were also incorrectly used as shown in Figure 3.

… become a forum for exchange of ideas **not only** on road transport **but** in general management and overall transport policies as well. (CIRT)

People **are needing** my help on a lot of things. (PP_7)

On & on & on. (PP_9)

The hard part is we learn from our mistakes, not **everyone elses.** (PP_17)

Figure 3. Sentence start

Mutations of all kinds were observed in abundance in the content of webpages. These include compression, abbreviation, typographical extremes and transliteration. Emoticons are symbols of expressions such as smile, laugh, cry, love and shout created using special characters. Almost in all cases they appear in Personal space and almost zero occurrence in Professional context. Compression of words happens when the spelling of the word is mutated to make it small (most of the time). Authors indulge in typographic extremes when they want to express happiness, sadness, frustration etc. The same effect can be achieved through emoticons also. Informal inclusions (from general purpose conversation) also appeared frequently in personal webpages. Table 1 provides a summary of the overall observations in language usage in professional and personal webpages.

Table 1. Summary of observation in language usage

Language Markers	Professional Webpages	Personal Webpages
Collocations	High	-------------------

Multiword Chunks	---------------	High
References	High (Demonstrative)	High (Personal References)
Conjunctions	Appropriate use	High and Inappropriate
Tense	Present perfect and simple past	Past perfect
Voice	Active	Passive
Conditional Clause	Limited	High
Modals	Used strong modals	Used all modals
Pronouns	‘us’, ‘our’, ‘we’ and ‘you’	I’, ‘mine, ‘he’, ‘she’, ‘his’, ‘her’
Sentences and Paragraphs	Variety in sent. And paragraph lengths	Monotonous, unorthodox compositions.
Purpose of comm.	Focused on functional goals	Focused on conversational goals

Discussion

All the webpages that were studied under the Professional context were found to follow a *strict pattern of content presentation*. The content was clearly divided into sections like introduction followed by vision/mission statement, achievements, future plans, societal engagements, etc. The sequence of this content and the use of font and typography were also similar in most cases. Regarding the webpages that were studied in the Personal context, they reflected *innovative and unorthodox presentation style* to convey the content. In some cases, the flow was more acceptable than that of others and in cases of personal diaries, there was no flow at all. For example, writing about the day’s event should ideally start from morning; but it was observed that authors would put event write ups in random order. Instead of following text-heavy presentation, some authors decided to put up their online content in a style laden with graphics.

The webpages analysed under Professional context highlighted a more *rigid language pattern.* The content was found to include all elements such as collocations, a marker for business discourse, abundant use of cohesive linkers to facilitate smooth flow, functional marker for the purpose of informing, announcing, offering, persuading, etc. Moreover the webpages in this category were properly classified into various sections for the purpose of readability. In case of personal webpages, it was the author who determined the content of the page. Though this fact appeared to be logical because that web space belonged to the individual, the style again suffered. The authors of personal webpages engaged in creative and innovative writing in all forms. All styles including intimate, obscure, ambiguous, and laden with font variations, typographic extremes, graphics-heavy contents and transliteration could be found in the personal webpages.

The webpages studied under the Professional category showed ample evidence of carrying *tangible content* which included the organization's short and long term goals, targets, achievements, customer satisfaction, concrete plans, market share and public opinion. Other things such as customer support and feedback were also present in a simple and actionable fashion. In terms of language, use of concrete words, active voice, appropriate functional markers and domain specific vocabulary were observed to have helped in achieving this goal. On the other hand, in case of Personal webpages it was mostly story telling. Whether it was about individual, family, or profession, etc., the author invariably engaged himself in writing about his own experience – good and bad - which turned out to be more abstract.

One of the most disturbing and frequently recurring observations was the *use of fragments* in Personal webpages. It was one of the stark differences between language construct in Professional webpages and Personal webpages. In the former case, the sentences were properly constructed with all structural elements, such as verb, subject, preposition, conjunction, punctuation etc., intact, while in the latter, users took enough liberty to play with the linguistic structure of the sentences and engaged in language abuse. As a consequence, the clarity of content suffered in Personal webpages imprinting a poor impression. The absence of nonverbal cues and asynchronicity of webpages gave rise to mutations and invention of typographical extremes and emoticons. While typographic extremes are created by mutating the spellings and adding extra special characters, the emoticons are constructed entirely of special characters. They help in conveying feelings and emotions such as anger, frustration, love, smile, sadness, happiness etc. However, it is a matter of concern that a growing number of users are using these mutations, especially emoticons in their online communication even in formal context.

Conclusion

Organizational webpages display the use of plain English style (in asynchronous environment) wherein they maintain a neutral tone, varying but a good sentence length, taking care of message logic and clarity by using appropriate language aspects such as connectives and conjunctions. On the other hand, personal webpages use a conversational (Informal) tone with less concentration on English appropriateness, unvarying and a lower average sentence length, minimal appropriate connectives and less message clarity by using of unorthodox language constructions. These observations though simple and direct, establish the widely believed notion that organizational communication is more formal than personal communication. The results indicate that the new media of webpages do have some negative influence particularly in personal context. At the same time the new mode add its own flavor to communication like all other modes do, and provides us a more dynamic way of reaching people across space and time.

Acknowledgements

This study wouldn’t have been possible without the support of numerous people with whom I interacted during the project. I extend my gratitude to Prof. Meenakshi Raman for her guidance and supervision. I am thankful to my colleagues, students and all the web users who have contributed to the text corpus.

References

Adkins, M., and D.E. Brashers. (1995). The power of language in computer- mediated groups.*Management Communication Quarterly 8, 289-322.*

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

Crystal, David. "English or Babel? Monolingual or multilingual." www.abc.net.au. 24 March 2001. Radio National: Lingua Franca. 14 August 2006 <http://www.abc.net.au/rn/arts/ling/stories/s264971.htm

Coney, M. and M. Steehouder. (2000). Role playing on the Web: guidelines for designing and evaluating personas online. *Technical Communication. 47, 327–340.*

Dominick, J. (1999). Who do you think you are? Personal home pages and self-presentation on the World Wide Web. *Journalism and Mass Communication Quarterly. 76, 646-658.*

Doring, Nicola.(2002). Personal Home Pages on the Web: A Review of Research. *Journal of Computer Mediated Communication. 7.3.*

Evans, B Mary, Alive A. AcBride and Matt Queen. (2005). Tone Formality in English-Language University Websites Around the World. *IEEE International Professional Communication*
Conference Proceeding. Limerick, Ireland, July 10-13, 2005. Piscataway, N.J, IEEE, 846-850.

Ferraro, G. P. (2000). *The Cultural Dimensions of International Business*. New Jersey: Prentice Hall.

Hodge, S. (2000). *Global Smarts: The Art of Communicating and Deal Making Anywhere in the World.* New York: Wiley.

Jessmer, S. L., & Anderson, D. (2001). The effect of politeness and grammar on user perceptions of electronic mail. *North American Journal of Psychology. 3.2,* 1-15.