Contextuality and Beyond: Investigating an Online Diary Corpus

Laura Teddiman

Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, University of Alberta
Edmonton, Alberta
Canada T6G 2E7

Abstract

Heylighen & Dewaele's (2002) F-score, a measure of formality developed based on categorical frequencies of word types, is used as a starting point for an investigation of an online diary corpus. Comparisons are made between results in the main corpus of diary entries, a smaller corpus of diary comments, and with previously calculated F-scores for similar types of data (Nowson, Oberlander & Gill, 2005). While the overall F-score is similar in these two corpora, results show that internal make-up of the categories upon which the calculation is based can differ. This suggests that while the F-score is a good measure of formality/contextuality and is useful in distinguishing between genres on a large scale, more detailed analyses are required to more completely describe and situate genres with respect to one another.

Introduction

The proliferation of blogs has given rise to a rich source of lexical data. This emergent form of New Media provides researchers with the opportunity to observe the growth of a new set of genres. There has been much discussion as to where and how blogs should be situated as a genre or set of genres, and where they might fall in comparison to more established genres. In this paper, we focus on personal online diaries, as opposed to academically or politically oriented blogs that provide commentaries, for example.

Heylighen & Dewaele (2002) present an elegant means of quantitatively measuring the relative formality of texts, the F-score, which is calculated by subtracting the relative frequencies of linguistic elements (categories) that rely more heavily on context for disambiguation (e.g., pronouns) from the relative frequencies of explicitly informative word categories (e.g., nouns), such that:

> F = 0.5(noun frequency + adjective freq. + preposition freq. + article freq. – pronoun freq. – verb freq. – adverb freq. – interjection freq. + 100).

For Heylighen & Dewaele (2002), flexibility in meaning is a key feature of contextuality. The interpretation of a pronoun, for instance, will vary depending on the surrounding context within which it is situated to a greater extent than a similar noun. For example, the sentence *It was smaller than she expected* does not offer much information to the reader. The pronouns *it* and *she* are undefined, with no disambiguation possible from within this sentence. If, however, the sentence is preceded by *Jean looked at the dog*, then it becomes apparent that *she* refers to *Jean* and *it* refers to *the dog*. The definitions of the proper noun *Jean* and the determiner-noun combination *the dog* rely less on the surrounding context than their pronominal counterparts. More formal texts, therefore, are those with greater specificity in and of themselves. They should not rely on shared information not present in the text (e.g. situational knowledge) to transmit their full meaning. Such formal texts tend to correspond to 'formality' in a more colloquial sense, so that scientific and technical writings are highly formal, magazine articles are less so, and casual conversations are among the least formal (and most informal).

Nowson, Oberlander, and Gill (2005) used the F-score to determine the relative formality of a blog corpus compared to a subset of genres available in the British National Corpus (BNC). In their calculations, the Blog genre had a score of 53.3, marking it as more formal than email and school essays, but less formal than written biographies. The current study expands on this research by adding a complementary corpus of diary entries and one of comments, and further explores the internal structure of the F-score categories. Newman and Teddiman (under review) show that the distribution of first person pronouns (*I, me, my, we, us, our*) can be used to distinguish between Online Diary style of writing and the genres represented in the BNCBaby, a 1,000,000 sampling of the whole BNC. These pronouns behave differently in the Online Diary Corpus, casual conversation, fiction, newspapers, and academic writing. The F-score is not sensitive to these category internal distinctions, so it is on pronouns that we focus first, before moving on to a brief investigation of emphatics.

The Current Study

In the current study, we are interested in the make-up of the tagged categories used in the calculation of the F-score. For example, the category *pronouns*, which is considered be a contextually biasing category, consists of words such as *I, you, it*, and so on. The F-score does not distinguish between category members, such as *I* and *you* that might

prove useful in further categorizing text. It is these elements to which we turn our attention.

The Diary Corpus

The Diary corpus was constructed through the random selection of publicly available online diaries hosted by LiveJournal (http://www.livejournal.com). Data were collected between July 2006 and December 2006. The corpus represents 100 users, 50 from each of the United States and the United Kingdom. Each text sample contains approximately 2000 words, with the entire corpus containing 204,997 words. Hyper-text mark-up and timestamps are not included in word counts. Given that sampling was random, more women are represented as authors within the corpus (65%). Mean age was 24.7. Although both the United States and the United Kingdom are represented in this sample, there were no significant differences based on dialect in this research. Results from the whole corpus are therefore reported together.

Approximately one year after initial data collection, comments to journal entries posted in the original corpus were collected, with 159088 words collected in total. The corpus is imperfectly balanced, as responses were not equal to each journal or to each entry. The average number of words in the comments per diary was 2376, although this number is boosted by a single string of comments in one blog that contained over 60,000 words. Without this series of comments, the average number of words falls to 1620 per journal.

The raw texts were tagged for part of speech using GOTagger07 (Kazuaki, 2007). Texts were then searched for POS tags using Wordsmith Tools 4.0 (Scott, 2004), and the results manually inspected. Word frequencies from the BNC were collected using the VIEW interface (Davies, 2004-) when required.

Results & Discussion

F-score values were calculated for both the main diary corpus and the comments corpus. For both, F = 55.5. The F-score generated by the current data is slightly higher than the blog-related F-score reported by Nowson et al. (2005), at 53.3. The variation in F-score may be a result of differences in the composition of the blog corpus, and it does not affect relative ranking of the Diary corpus along the genre hierarchy they describe. The Diary and Comments corpora fall between School Essays and Biographies when compared to F-scores calculated from data in the BNC. The first result, then, is a replication of Nowson et al.'s findings. This Diary corpus shares many features of more formal writing, for example in the relative number of nouns and prepositions used by authors.

It is important to note, however, that even though the Diary corpus and the Comments corpus have the same F-score, they do not necessarily share the same linguistic features. This is particularly true in an examination of the pronouns used in both corpora. While the overall percentage of pronouns in the corpora is about equal (11.7 in the Diaries, 11.4 in the Comments), the identities of the pronouns vary. In both corpora, *I* is the most frequent pronoun, occurring approximately 37 times per thousand words in each. The relative frequency of *I* should not be surprising, given the self-referential nature of diary writing. With respect to *I*, the diary corpus patterns more closely with S_Conv, the spoken demographic subcorpus of the BNC, in which *I* occurs 42 times/1000 words (compare to Email: 20/1000 words; Biography 12/1000 words; School essays 8.8/1000 words). First personal pronouns are often considered to be markers of an interpersonal focus (Biber, 1998: 225), a factor more closely related to conversation, and by extension, to contextuality rather than formality. This result does not invalidate the F-scores for our corpora, but it does suggest an interesting interplay between categorical frequencies and relative genre similarities.

Second person pronouns, which, according to Biber (1988: 225) "indicate a high degree of involvement with the addressee," tell a different story. *You* occurs 7 times/1000 words in the Diary corpus, but 20 times/1000 words in the Comments corpus. From a discourse perspective, this indicates that respondents to diary entries are directly addressing the author and that they intend for their responses to be personally valuable. While the text of the comments corpus itself is more formal/less contextual than spoken conversation, for this particular linguistic feature, it patterns as closely with speech (*you:* 33.6/1000 words) as with its closest F-score neighbours (email: 5.5/1000 words; biography: 2.7/1000 words; school essays: 3.8/1000 words). The pattern displayed by the possessive pronoun *your* is more pronounced, with the Comments corpus and Spoken demographic subcorpus patterning together (4.1 & 3.6/1000 words, respectively) compared to the Diary corpus (1.6/1000 words) and its F-score neighbours, which occur at 1.1 instances/1000 words or lower.

The patterning of pronouns differs, sometimes greatly, between the two collected corpora, and does not always support greater relative text formality or lesser contextuality, as might be expected by the overall F-score.

Other features of the text are also problematic from the perspective of text formality. The informal emphatics *just* and *really* (Biber, 1988: 241) are as frequent in the Diary and Comments corpora as in spoken conversation, occurring in this these corpora more often than in email, biographies, or school essays (e.g., *just*: 4.31/1000 Diary, 4.87/1000 S_conv, 1.1/1000 Biography). In order to capture such behaviour, it is necessary to go beyond the large scale, category-based contextuality measure.

Conclusions

The Contextuality measure proposed by Heylighen and Dewaele (2002) can accurately separate genres, and returns similar results for the same genre given a different set of data. However, given its scope, the F-score cannot be sensitive to differences that are active within its categorical

determinants; the F-score can successfully differentiate between genres, even if it cannot precisely determine *why* they are different. If the goal is to discriminate between genres, then the F-score is an admirably capable measure. However, if the goal is to understand the nature of emergent internet genres, such as online diaries, then it is necessary to also explore the linguistic activity of specific categories (e.g., pronouns) and items (e.g., *I*) within and across text types. From this perspective, the F-score is a useful starting point for linguistic research, but it cannot be an endpoint.

The results reported here raise the question of how F-scores should be interpreted, given that the behaviours observed are not constant while the F-score essentially remains so. Should personal blogs be considered a separate genre from their comments? The F-Score measure tells us that they are very similar, but there are properties unique to both which might be used to argue for at least some degree of subcategorization. However, the F-score does provide a starting point that may be used to further blog research, for example, by testing whether or not there are significant differences in formality between personal journals and other types of blogs. One might expect, for example, that blogs discussing scientific information might as a group show higher levels of formality than personal journals. The F-score could prove a useful measure to situate differing blog types across a wide range of genres, while also allowing for comparisons with existing genres.

References

Biber, D. 1988. *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Davies, Mark. 2004-. BYU-BNC: The British National Corpus. Available online at http://corpus.byu.edu/bnc.

Heylighen, F., & Dewaele, J.-M. 2002. Variation in the contextuality of language: an empirical measure. *Foundations of Science, 7,* 293–340.

Kazuaki, G. 2007. GoTagger07. Available online: http://uluru.lang.osaka-u.ac.jp/~k-goto/use_gotagger_e.html

Newman, J. & Teddiman, L. Under review. First person pronouns in online diary writing. Submitted to the Handbook of Digital Discourse, edited by Rotimi Taiwo.

Nowson, S., Oberlander, J., and Gill, A.J. 2005 Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pp. 1666-1671.

Scott, M. 2004. *WordSmith Tools.* Version 4. Oxford: Oxford University Press.