The Wheres and Whyfores for Studying Textual Genre Computationally

Jussi Karlgren
SICS - Swedish Institute of Computer Science
Box 1263
S – 164 29 Kista, Sweden
jussi@sics.se

Abstract

This brief paper gives an example of statistical stylistic experimentation and argues for more informed measures of variation and choice and more informed measures of readership analysis to be able to posit dimensions of textual variation usefully.

Variation in text

Texts are much more than what they are about. Authors make choices when they write a text: they decide how to organize the material they have planned to introduce; they make choices between synonyms and syntactic constructions; they choose an intended audience for the text. Authors will make these choices in various ways and for various reasons: based on personal preferences, on their view of the reader, and on what they know and like about other similar texts.

A style is a consistent and distinguishable tendency to make some of these linguistic choices. Style is, on a surface level, very obviously detectable as the choice between items in a vocabulary, between types of syntactical constructions, between the various ways a text can be woven from the material it is made of. It is the information carried in a text when compared to other texts, or in a sense compared to language as a whole. This information — if seen or detected by the reader — will impart to the reader a predisposition to understand the meaning of text in certain ways. Or, more roughly put, style is the difference between two ways of saying the same thing.

So, the variation in a text or differences between texts that is not primarily topical, that has not to do with meaning, is stylistic. Naturally, demarcation of stylistic variation to topical variation is impossible. Certain meanings must or tend always to be expressed in certain styles: legal matters tend to be written in legal jargon rather than hexameter; car ownership statistics in journalistic or factual style. The impossibility of drawing a clean line between meaning and style has lead to much browbeating among stylisticians and linguists, and discussion about if there in fact are any formally distinguishable styles at all (Enkvist 1973).

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Genre is a vague but well-established term. It is used in many ways in many different fields loosely related to each other, some more formal than other. For most purposes, genres can be understood as groupings of documents that are a) stylistically consistent and b) intuitive to accomplished readers of the communication channel in question. Genres are dependent on context: for a business newspaper such as the Wall Street Journal the genre palette will be different from that of texts found on the World Wide Web or distributed by the Book of the Month club.

In most computational stylistics and in fact in the minds of most readers, genre has mostly been equated with or based on text source: texts from some organization are categorized together with texts from similar organizations, with little or no regard for text usage. (Karlgren & Cutting 1994, e.g.) Examples are categories such as Wall Street Journal text archive, personal letters, technical documentation, cookbooks, sometimes even on the history of a specific collection, and further down to the level of individual choice, as in authorship studies of various types (Mendenhall 1887, e.g.). Stylistic variation can be based on the *functional styles* that occur in the collection at hand — as opposed to *individual styles* or sources of text (Vachek 1975).

Variation is not incidental

Most importantly, textual variation is not incidental but an integral part of the intended and understood communication: the content and form of the message cannot be divorced. People - at least skilled readers - are quite adept at distinguishing content of one kind from another, and very early on, readers learn to distinguish genres.

- No, that is for kids!
- Seems too difficult.
- No, I need a manual.
- What?! No pictures? No way!

Genres, while vague and undefined, are well-established and talked about. They are a useful and functional level of categorization.

What then is characteristic of a textual genre? When asked, people will respond by describing situations in which the genre is relevant, experiences of reading a genre, con-

tent, and in some cases specific lexical or other linguistic features of the text.

Concrete example: Newsprint and its subgenres

Newsprint, while a special register of human language use in itself, is composed of several well-established subgenres. While newsprint is – together with scientific and intellectual text – over-represented as an object of empirical philological study as compared to other forms of human linguistic communication, it possesses some quite useful qualities for systematical study of linguistic characteristics, chief among them that the texts are most often intended to be the way they are. The textual material has passed through many hands on its way to the reader in order to best conform to audience expectations; individual variation – worth study in itself – is not preserved to any appreciable extent in most cases; the audience is well trained in the genre. In short, the path is well-trod and as such easy to study. Newsprint may sound to be a homogenous mass of text, but it contains several well-established subgenres. In the present study one year of Glasgow Herald is being studied, and many of the texts are tagged for "ARTICLETYPE" – cf Table 1.

<ARTICLETYPE>		n
tagged		17467
advertising	522	
book	585	
correspondence	3659	
feature	8867	
leader	681	
obituary	420	
profile	854	
review	1879	
untagged		39005
total		56472

Figure 1: Sub-genres of the Glasgow Herald.

Observed differences between genres

It is easy enough to establish that there are observable differences between the genres we find posited in the textual material. Trawling the (morphologically normalized) texts for differences, comparing each identified category in Table 1 with the other categories we find for a one month sample of text some of the most typical words for each category as per Table 2. "Typical" is here operationalized as words with a document frequency deviating from expected occurrence as assessed by χ^2. This sort of analysis can be entertaining, at times revealing, but cannot really give us any great explanatory power. Even the handful of example terms shown in Table 2 are clearly colored by the subject matter of the subgenres and only to some extent reveal any stylistic difference between them.

<ARTICLETYPE>	Typical words
advertising	provide, available, service, specialist, business
book review	novel, prose, author, literary, biography, write
correspondence	(Various locations in Scotland), overdue, SNP
feature	say, get, think, put, there, problem, tell
leader	evident, government, outcome, opinion, even
obituary	church, wife, daughter, survive, former
profile	recall, career, experience, musician
review	concert, guitar, piece, beautifully, memorable

Figure 2: Typical words in sub-genres of one month of the Glasgow Herald.

What are readers aware of?

Starting from first principles, a better approach may be to ask people what they are aware of as differences between newsprint genres. In questionnaires or interviews, readers typically respond by answering in terms of *utility* of the genre rather than text characteristics, in terms of perceived *quality* of the text in some prescriptive terms, or in terms of *complexity* of the text and subject matter. On follow-up questioning readers will bring up subjective qualities such as *trustworthiness* of the text or further discuss *readability* or other specifics related to complexity, both lexical and syntactic (and specifically making claims in both dimensions, typically claiming that leader articles are difficult and long-winded or that sports features are inane and simplistics). These perceptions, again, may be interesting, elucidating, and entertaining to discuss, but are difficult to encode and put to practical use. While we as readers are good at the task of distinguishing genres, we have not been intellectually trained to do so and we have a lack of meta-level understanding of how we proceed in the task. We need to model human processing at least on a behavioural level at least to some extent. We need to be able to address the data on levels of usefulness, and we need to observe people using documents and textual information to understand what is going on. Clustering data without anchoring information in use will risk finding statistically stable categories of data without explanatory power or utility.

Assumed differences between genres

A more linguistically informed approach is to start from established knowledge (or established presumption, as it were) and to work with a priori hypotheses on qualities of textual variation. This proves to be rather unproblematic. We know from past studies and judicious introspection that interview articles contain quotes and that leader articles contain argumentation; both are to be expected to contain pronouns and

overt expressions of opinion. Table 3 shows some such measurements – with indicated significance (better than 95%) assessed for each sub-genre compared to the entire collection by Mann-Whitney U. The explanatory power of this table – again, while interesting in some respects and worth discussion – is rather low, and application of the findings for prediction of genre for unseen cases will be unreliable. A more abstract level of representation is necessary to be able to extract useful predictions and to gain understanding of what textual variation is about.

<ARTICLETYPE>	p	dem	say & think	op & arg arg	cpw
advertising	-	.	+	.	+
book review	+	+	+	+	-
correspondence	-	-	-	-	+
feature	+	+	+	+	-
leader	-	+	.	+	+
obituary	-	-	.	-	-
profile	+	+	+	+	-
review	-	-	-	-	+

+	Significantly higher values
-	Significantly lower values
.	Non-significant value distribution
p	Personal pronouns
dem	Demonstratives: "that" &c.
s & t	Verbs of utterance and "Private" verbs[1]
op arg	Opinion[2] and argument[3]
cpw	Characters per word

Figure 3: Some measurements of linguistic items per sub-genre of one year of the Glasgow Herald.

Dimensions of variation

The above discussion gives purchase to the following claims:

- Readers are aware of genres;
- Writers and editors are aware of genres;
- Genres are to some extent arbitrary and most decidedly not independent of subject matter;
- Readers are *not* aware of most of what formal differences there are between genres;
- It is easy to find and measure formal differences between genres in terms of observable linguistic items;
- These measurements are not very interesting nor immediately useful.

Several authors have posited underlying dimensions of variation within the space of textual variation, where genres can be found as regions in terms of such dimensions – Involved vs Informed, Narration vs Argumentation, Personal vs Impersonal etc. To study such underlying dimensions of variation is the main object of study for computational stylistics today, and a worthwhile goal – a goal which can be accomplished at least to some extent using purely formal methods, without studying text or readership.

But the ingoing measurements have hitherto been hampered by lack of useful tools. Even now, most measurements made in the literature are purely lexical or only locally syntactical. This form of study needs better tools for measuring textual variation on the levels authors and readers are processing information, not on the level of whitespace separated character strings. Only if useful target measures of pertinence, relevance, and utility can be found for evaluation, couched in terms derived from study of readership, and if real measures of variation and choice, couched in informed terms of linguistic analysis rather than processing convenience, can we hope to accomplish anything of lasting value in terms of understanding choice, textual variation, and reading.

References

Enkvist, N. E. 1973. *Linguistic Stylistics*. The Hague, Netherlands: Mouton.

Karlgren, J., and Cutting, D. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International Conference on Computational Linguistics*, volume 2, 1071–1075. Kyoto, Japan: ICCL.

Mendenhall, T. 1887. The characteristic curves of composition. *Science* 9:237–249.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartvik, J. 1985. *A comprehensive grammar of the English language*. London, England: Longman.

Vachek, J. 1975. Some remarks on functional dialects of standard languages. In Ringbom, H., ed., *Style and Text — Studies presented to Nils Erik Enkvist*. Stockholm, Sweden: Skriptor.