Analysis of Linguistic Features Associated with Point of View for Generating Stylistically Appropriate Text

Nancy L. Green
Department of Mathematical Sciences
University of North Carolina at Greensboro
Greensboro, NC 27402 USA
nlgreen@uncg.edu

Abstract

We describe our preliminary work on analyzing a corpus of clinical genetics patient letters. In this genre, a single letter is intended to serve multiple functions and is designed for audiences with different backgrounds, information needs, and affective responses to the information. The goal of the analysis is to identify stylistically-related features for natural language generation. Our hypothesis is that sets of these features are associated with different points of view.

Introduction

We are studying a corpus of clinical genetics patient letters written by genetic counselors to their clients. According to Baker et al. (2002), the typical patient letter, one to two pages in length, summarizes issues discussed during the counselor's meeting with the client, e.g., interpretation of the client's test results or genetic risk, counseling to cope with the emotional impact of information, as well as explanation of clinical genetics concepts relevant to the client's case. While the addressee of the letter is the client, intended secondary audiences include the addressee's family members and (in case the client is the parent/guardian of a pediatric patient) staff members at the patient's school or daycare. In addition, the letter is intended to provide medical documentation for health care providers. The goal of our study is to identify stylistically-related features to guide linguistic realization and content selection in a natural language generation (NLG) system to be used as a letter writing tool for genetic counselors.

Previous NLG research on stylistic variation has viewed style as a constant property within a document and as defining a genre (Hovy 1990, DiMarco and Hirst 1993). After informal review of letters in the corpus, we noted that, perhaps because of the multiple intended functions and audiences, within a single letter (and in some cases within a single sentence) more than one writing style can be observed. Our initial hypothesis is that each style (i.e. coherent set of realization choices) is associated with a different perspective assumed by the author, e.g., a counseling perspective in which realization choices serve to mitigate negative affective responses, or a medical perspective (e.g. that of the referring doctor or genetics clinic) in which they serve a medical documentation function.

For example, in sentence (2) below the writer uses the referring doctor's perspective in reporting the reason for the referral to the author's clinic. (The number in parentheses identifies the sentence; the letter's identifier, *VCF*, is given in parentheses at the end of the excerpt. In the corpus, capitalized words in brackets have been substituted for original text to maintain client confidentiality but convey the gist of the original text. In this domain, *proband* refers to the person who is the focus of a genetic study, i.e., the patient.)

> (2) [DOCTOR] asked us to evaluate [PROBAND] to determine if [HIS/HER] delays in development and [SPECIFIC TYPE OF BIRTH DEFECT] were due to a recognizable genetic condition. (letter VCF)

When speaking from this doctor's perspective, the writer's description of the patient's symptoms is direct, precise and uses words that may have negative connotations to the client (the proband's mother), e.g., *delay*. In contrast, when the writer assumes the genetic counselor's perspective, the wording is designed to mitigate the client's possible negative emotions. Two stylistic choices expressing the voice of the counselor in sentence (14) below are use of the phrase *altered form* instead of *mutation*, and use of a qualitative indicator of probability (*could*).

> (14) [PROBAND] could have inherited an altered form of a gene from both you and [HIS/HER] father that caused [HIS/HER] birth defects and learning problems. (letter VCF)

In summary, we hypothesize that in addition to a representation of what must be said, our NLG system must be provided with a specification of perspective in order to be able to generate stylistically appropriate text in this application. We have identified a set of potential perspectives. Next, we plan to manually annotate the corpus for perspective and candidate linguistic features, and then apply analysis techniques to identify correspondences. This paper justifies our hypothesis by outlining a set of perspectives and some of the features potentially associated with each that we have identified by qualitative analysis so far.

Perspectives in Corpus

Based upon our review of letters in the corpus[1] and on information about genetic counseling, e.g. (Wilson 2000), we have identified the following perspectives:

- author: the letter writer, writing on behalf of the genetics clinic; although this role is performed by the genetic counselor, we shall distinguish this perspective from that of the genetic counselor. (ex. *This letter will summarize…*, *We hope this information is helpful…*)
- client: person(s) who met with the counselor and principal addressee(s) of letter; usually the proband or a member of the proband's family. (ex. *You expressed concern that …*)
- referring doctor: doctor who referred the patient to the clinic; usually not a member of the clinic staff. (ex. [*DOCTOR*] *asked us to evaluate ...*)
- clinic: genetics clinic that tested the patient; the writer/genetic counselor is a member of the clinic staff. (ex. *We obtained a blood sample …*)
- genetic counselor: the genetic counselor who met with the client(s). (ex. *It is important to remember that [PROBAND'S] problems could still be caused by genetic alteration*.)
- educational: basic background knowledge about clinical genetics as portrayed in text books for health care professionals. (ex. *In some instances, children ... inherit two copies of an altered gene...*)
- research: research information on clinical genetics, i.e., newer or more specialized than the educational perspective. (ex. *Individuals with VCF often have ...*)

Although originally developed for the automated analysis of narrative (Wiebe 1994), and later applied to analysis of attitude in newspaper articles (Wilson and Wiebe 2003), the model of psychological point of view (POV) provides a framework for our own study. That model defines a private-state relation whose components include an *experiencer*, an *attitude*, and the *object* of the private state. For example in sentence (2, VCF) repeated below, the experiencer, identified as [*DOCTOR*], is the referring doctor, the attitude could be interpreted as *believes it to be possible that*, and the object corresponds to what is expressed as *the proband's delays in development and* [*SPECIFIC TYPE OF BIRTH DEFECT*] *were due to a recognizable genetic condition.*

> (2) [DOCTOR] asked us to evaluate [PROBAND] to determine if [HIS/HER] delays in development and [SPECIFIC TYPE OF BIRTH DEFECT] were due to a recognizable genetic condition**.**
>
> (3) During your appointment on [DATE], we obtained a blood sample from [PROBAND]
>
> (4a) In addition to the routine chromosome study,
>
> (4b) in which a microscopic study of the 46 chromosomes is done,
>
> (4c) a special analysis of the long arm of chromosome 22 (22q11)
>
> (4d) by a technique called fluorescence in situ hybridization (FISH)
>
> (4e) was done to test for Velocardiofacial syndrome (VCF).
>
> (5) Individuals with VCF often have [SPECIFIC TYPE OF BIRTH DEFECT] and learning problems. (letter VCF)

This excerpt illustrates several other points. First, as noted in (Wiebe 1994), experiencer and attitude need not be stated explicitly. In (3), the experiencer, signalled by *we*, is the clinic and the attitude could be interpreted as *knowledge shared by experiencer and addressee*. In (4), the experiencer could be interpreted as the clinic again, although it was not explicitly signalled; (4) *continues* (Wiebe 1994) the experiencer of the current POV. The second point illustrated is how subtle differences in attitude can be signalled by both content selection and wording. In contrast to how we interpreted the attitude in (3), we claim that the explanatory information provided in (4b) and (4d) suggests that attitude in those phrases should be represented something more like, *knowledge that the experiencer believes the addressee does not have.* This change in attitude is associated with a shift in tense; the explanatory information in (4b) and (4d) is presented in the present tense while the rest of (2) through (4), a narration of the patient's history and clinic visit, is presented in the past tense. Finally, according to our interpretation the experiencer in (5) is the research perspective, rather than the preceding experiencers in this excerpt, namely, the referring doctor and clinic. This

[1] Currently, the corpus contains 16 patient letters given to us by genetic counselors and a letter published in (Baker et al., 2002); we are continuing to collect more letters.

change in experiencer is marked also by a shift to the present tense.

Identifying Associated Features

Table 1 (at the end of the paper) shows for each perspective defined above, some candidate associated features that we have identified by manual inspection of the corpus. The second column provides a brief description of semantic criteria for identifying perspective. The third column lists the typical forms used for referring to each type of experiencer when explicitly given. Note that first person plural pronoun forms (abbreviated as *1p-plural* in the table) such as 'we' may be used to refer to several categories of experiencer. The fourth column lists characteristic forms for conveying probability, and is discussed in the following section. The last column lists several other types of possible features.

Probability

In a previous study of this corpus (Green 2003), we manually tagged both qualitative and quantitative indicators of probability and frequency. Examples of qualitative indicators are modal verbs (e.g. *can, could*) and frequency adverbs (e.g. *often*) and quantifiers (e.g. *many*). Quantitative indicators are phrases including numbers, e.g. *about 1 in every 3000*, and *approximately 80%.*. That study determined that the ratio of probability statements to the number of sentences was high, which is not surprising due to the inherent uncertainty in diagnosis and prediction in clinical genetics.

Table 1 (column four) shows the probability indicators that seem to be associated with each perspective. As can be seen, the use of probability indicators characterizes several perspectives. However, indicators that we refer to as *Mendelian numbers* in the table, i.e. idealized ratios that characterize a Mendelian inheritance model such as 1/4, 1/2, and 3/4 (which can be expressed as percentages also), seem to be associated with the educational perspective. Other numeric indicators not characteristic of the Mendelian model seem to be associated with the research perspective. For example, in sentences (4) and (7) of letter NF in the corpus, which present a research perspective, the probability indicators (underlined) use non-Mendelian numbers:

(4) Neurofibromatosis 1 (NF) is a progressive disorder of the skin and nervous system that occurs <u>in about 1 in every 3000</u> individuals.

(7) <u>Approximately 80%</u> of individuals affected with NF have mild to moderate symptoms. (letter NF)

However, as illustrated by (5) in letter VCF above, the writer may use a qualitative indicator (*often*) even when speaking from a research perspective.

The tagged probability indicators were used in (Green 2003) to encode conditional probability statements conveyed in the text. For example, (7, letter NF) could be paraphrased as

P("mild to moderate symptoms" | NF)
= "approximately 80%".

In this case, the POV notions of object and attitude overlap with what we were attempting to characterize then; the attitude corresponds to the belief that the probability is approximately 80%, and the object corresponds to the belief encoded on the left hand side of this equation. However, the POV notion of attitude is broader than what we were characterizing in that work. For example, as illustrated in our discussion of sentence (4, letter VCF) above, attitude may encompass the distinction between shared and non-shared knowledge.

Mitigating Negative Affective Responses

According to writing guidelines for genetic counselors (Baker et al. 2002), writers should consider possible negative emotional effects (grief, fear, shame, etc.) that information in a letter may have on the client; the guidelines therefore suggest using "non-stigmatizing", "value-free" language, e.g. *alteration* instead of *mutation*. The corpus shows that the writers often do employ this strategy, especially when writing from the counseling or educational perspective. However, it seems that the writers do not always follow this strategy when writing from the referring doctor's or research perspective (what we refer to as use of *technical genetics/medical words* in Table 1). It is plausible that use of these technical words is required in some cases to achieve the medical documentation function.

Also, Baker et al. (2002) describe a number of writing strategies in this genre for managing "distance" between patient and information, between the staff and information, or between the staff and the patient. To increase distance between the patient and information that might be viewed negatively by the patient, one recommended strategy is to state the information in general terms instead of in terms of the patient, e.g., *Children with this condition tend to lose their hearing*, instead of *Nisha is likely to lose her hearing*. We have observed this strategy in the corpus, but we represent it as a shift from the genetic counselor perspective to the educational perspective. This is reflected in Table 1 in the features in column five called *reference to population* and *universal*, as opposed to the feature called *direct reference to proband/family.*

Conclusion

We identified a set of perspectives and some of the features potentially associated with each in a corpus of genetic counseling patient letters. Our goal is to analyze

how perspective is reflected in linguistic realization and content selection options for natural language generation.

Acknowledgments

This work is supported by the National Science Foundation under CAREER Award No. 0132821.

References

Baker, D.L., Eash, T., Schuette, J.L., and Uhlmann, W.R. 2002. Guidelines for Writing Letters to Patients. *Journal of Genetic Counseling* 11(5):399-418.

DiMarco, C. and Hirst, G. 1993. A Computational Theory of Goal-Directed Style in Syntax. *Computational Linguistics.* 19(3):451-500.

Green, N. 2003. Towards an Empirical Model of Argumentation in Medical Genetics. In *Proceedings of IJCAI 2003 Workshop on Computational Models of Natural Argument*, 39-44.

Hovy, E. 1990. Pragmatics and Natural Language Generation. *Artificial Intelligence* 43:153-197.

Wiebe, J. M. 1994. Tracking Point of View in Narrative. *Computational Linguistics* 20(2):233-288.

Wilson, T. and Wiebe, J. 2003. Annotating Opinions in the World Press. In *Proc. of 4th SIGDial Workshop*.

Wilson, G.N. 2000. *Clinical Genetics: A Short Course*. New York: Wiley-Liss.

Experiencer	**Topics**	**Reference to experiencer**	**Probability**	**Other cues**
author	about letter, greetings, closing	pronoun (1p-plural)		"this letter", formulaic language (e.g. "it was a pleasure ")
client	issues brought up by client at meeting	proband/family by name or pronoun (2p 3p)		"you asked whether"
referring doctor	reason for referral to clinic, symptoms-findings, candidate diagnosis	doctor by name	indirect (e.g. "due to")	"[DOCTOR] asked us" , "referred by [DOCTOR] to", past tense, technical genetics/medical words
genetics clinic	clinic visit, procedures/tests performed, results	pronoun (1p-plural)		past tense
genetic counselor	interpretation of results or risk, recommendations	pronoun (1p-plural)	qualitative, Mendelian numbers	emphasis ("still", "it is important"), value-free genetics/medical words, direct reference to proband/family
educational	background information about clinical genetics		qualitative, Mendelian numbers	value-free genetics/medical words, reference to population (e.g. "the parents") or universal (e.g. "we", "everyone")
research	research information on clinical genetics		qualitative, quantitative	technical genetics/medical words, reference to population (e.g. "Individuals"), present tense

Table 1. Preliminary table of features. The righthand three columns list some candidate linguistic features associated with perspective in first column. The second column provides semantic criteria for identifying perspective.