Toward a stochastic speech act model of email behavior

John W. Mildinhall*
Department of Experimental Psychology
University of Bristol
Bristol, United Kingdom
J.Mildinhall@bris.ac.uk

Jan M. Noyes
Department of Experimental Psychology
University of Bristol
Bristol, United Kingdom
J.Noyes@bris.ac.uk

Abstract

Human interpersonal face-to-face interaction can be considered in terms of successions of speech acts. These are utterances which contain an intention, and the act of creating an utterance causes the intention to be directed toward the recipient. In this study the language contained in emails is considered in a similar manner. Due to the relatively asynchronous nature of email communication, communicators tend to generate packets of speech acts longer than those used in face-to-face communication in order to communicate efficiently. In this paper, the structure of these packets is examined by comparing probability transition matrices of speech act categories using multidimensional scaling and hierarchical cluster analysis. Results indicate that a ten-cluster system of email classification may represent a possible taxonomy of email intentions. In addition, speech act content in email communication can significantly predict membership of external categories, such as whether the email has a business or a personal nature. The use of an email intention classification system may enable future higher-level analysis of email relationships in terms of language intentions.

1 Introduction

In addition to being a system of information delivery, language is also used by humans as a way of acting upon others (Speech Act theory; Austin, 1975; Searle, 1969; Searle & Vanderveken, 1985). When individuals communicate, the act of passing information on is not just undertaken to edify the recipient of the act of communication, but also to influence their behavior in some manner. These two uses of language can coexist in the same utterance, and the nature of the surface *form* can differ from the underlying *intention*. For example, the utterance "Please can you open the door?" has the form of a question, but the intention of guiding the recipient's behavior.

The use of computer-mediated communication places constraints on the ability to generate utterances. In particular, the asynchronous nature of email inhibits the use of *grounding* language behavior (Clark & Brennan, 1991) such as acknowledgments and interjections. The lack of synchronicity engenders a monological style which consists of a unidirectional stream of several utterances. In order to adapt to this medium, communicators by necessity must develop a method of partitioning their speech acts. Given the maturity of email as a communication medium, we hypothesize that an email culture has emerged with several implicit stereotypical "templates" or conventions for email composition.

The identification of a parsimonious system for classification of emails by language intentions could enable a wider understanding of email behavior in terms of organizations and social groups. For example, interpersonal email relationships could be classified on the basis of multiple email classifications.

1.1 Verbal Response Modes

Verbal Response Modes (VRM; Stiles, 1978) is a system of general language classification widely used in psychology. Originally developed to model therapist-client relationships, it has now been used in many different settings which require the analysis of language. Verbal Response Modes differentiate between intentions using a system based on three dichotomous principles: *Source of experience*, *frame of reference*, and *presumption about experience*. Any utterance can be

*corresponding author

classified "speaker" or "other" for each of these three principles, resulting in a 2x2x2 classification system (See Table 1). Each utterance is coded twice, once for *form*, and once for *intent*.

2 Method

2.1 Data

This study required a large corpus of natural email data. At the time of writing, the only such corpus freely available was the Enron set. This corpus was posted on the website of the US Federal Energy Regulation Commission in 2003. It consists of the contents of the email folders of 150 Enron employees who were involved in the investigation into the large-scale fraud which resulted in the bankruptcy of Enron Corporation in December 2001. In its raw form, the Enron corpus contains approximately 1.5 million emails covering a four-year period. The data used in this study was a random sample of 144 emails from a subset of the Enron Corpus produced by Jabbari et al. (2006) and provided by Dugdale (2007; personal communication). The emails were filtered to avoid duplicates, and to remove junk emails. In addition the emails had also been independently classified into four categories; Close Personal, Core Business, Inter-employee relations (emails which concerned personal relationships between colleagues), and Routine Administration. There were 36 emails in each of the four categories, and overall inter-rater reliability for annotation was 94%.

2.2 Procedure

The data was coded using Stiles' (1978) VRM, using two coders. Cohen's Kappa for inter-rater reliability was satisfactory, $\kappa = 0.76$. While the VRM method codes each utterance twice, for form and intention, only intention data was used in this experiment in order to simplify data analysis. The data was recorded as a set of sequences where one sequence represented one email in the data set.

Zero- and first-order transition probability distributions for VRM codes were calculated for each email using a Matlab algorithm which had been designed for this purpose. The resulting 288 matrices were normalised using a Laplace correction in order to avoid difficulties associated with sparsely populated transition probability matrices. Kullback-Leibler (K-L) divergence (Kullback & Leibler, 1951) was calculated for each possible pair of zero-order, and then first-order transition probability distributions. K-L divergence is a measure of difference between two probability distributions which enables the use of statistical techniques designed for dyadic distance data. For probability distributions P and Q of a discrete random variable, the K-L divergence of Q from P is defined to be:

$$D_{KL}(P \mid\mid Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (1)$$

Two dissimilarity matrices were calculated for each possible combination of the m^{th} and n^{th} probability distribution for the zero- and first-order transitions according to:

$$K_{mn} = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (2)$$

Following this, both matrices were made symmetric, and subjected to non-metric multidimensional scaling (MDS; Kruskal & Wish, 1978). Scree plots and Shepard plots indicated that a two-dimensional solution created a satisfactory fit for both zero- and first-order matrices. This resulted in a four-dimensional solution. These four sets of co-ordinates were subjected to hierarchical cluster analysis, using Ward's (1963) method. This algorithm seeks to minimise information loss, and is suitable for large datasets.

3 Results

Applying this method to the dataset yielded a maximum of 10 clusters. The MDS solutions for K-L divergence matrices for zero- and first-order transition probabilities are shown in Figure 1, along with cluster membership. In addition, K-nearest means territory maps are plotted to show the boundaries of the clusters along the equivalent axes. The characteristics of each cluster are shown in Table 2.

Classification Analysis

In order to test the value of the MDS solution as a predictor of category membership, a direct discriminant analysis was performed. The four MDS-derived dimensions, two dimensions for the zero- and two for the first-order data, were used as predictors. For the 144 cases, evaluation of assumptions of linearity, normality, multicollinearity and singularity were satisfactory. Three discriminant functions were calculated, with a combined χ^2 (12, n=144) = 47.36, $p < .001$. After the first discriminant function was removed, the remaining functions did not significantly predict group membership, χ^2 (6, n=144) = 4.66, $p > .05$. Canonical $R^2 = .51$ for the first discriminant function; therefore, this function accounts for about 51% of the total relationship between predictors and groups. The first discriminant function accounted for 91.4% of the between-group (explained) variance.

Table 1: Taxonomy of Verbal Response Modes (Stiles, 1978)

Source of experience	Presumption about experience	Frame of Reference	Mode
Speaker	Speaker	Speaker	DISCLOSURE (D) Reveals thoughts, feelings, perceptions, or intentions.
		Other	EDIFICATION (E) States objective information
	Other	Speaker	ADVISEMENT (A) Attempts to guide behavior; suggestions, commands, permission, prohibition.
		Other	CONFIRMATION (C) Compares speaker's experience with other's; agreement, disagreement, shared experience or belief.
Other	Speaker	Speaker	QUESTION (Q) Requests information or guidance.
		Other	ACKNOWLEDGMENT (K) Conveys receipt of or receptiveness to other's communication; simple acceptance, salutations.
	Other	Speaker	INTERPRETATION (I) Explains or labels the other; judgments or evaluations of other's experience or behavior.
		Other	REFLECTION (R) Puts other's experience into words; repetitions, restatements, clarifications.

Table 2: Email cluster characteristics

Cluster	n	Category membership	Description
1	11	Mostly Personal or Inter-employee relations	High in Disclosure, some Advisement.
2	36	Mixed	High in Disclosure, mixed structure, some Edification, some Advisement.
3	8	Mostly Personal	High in Disclosure, also Edification, some Question.
4	30	Mixed	Mixed structure, commonly Advisement, Edification. Some Interpretation, Question and Disclosure.
5	16	Mixed	High in Edification, also Disclosure, some Advisement.
6	8	Personal, Core Business or Inter-employee relations	High in Edification, also Disclosure, some Advisement.
7	4	Personal, Core Business or Routine Admin	High in Advisement, also Edification. Typically consists of instructions.
8	21	Core Business, Inter-employee relations, or Routine Admin	Short Edification, some Disclosure, some Advisement. Often without opening Acknowledgement.
9	3	Core Business	Large, Almost exclusively Edification. Long newsletters.
10	7	Mostly Core Business or Routine Admin	High in Edification, some Advisement. Medium-length newsletters and instructions.

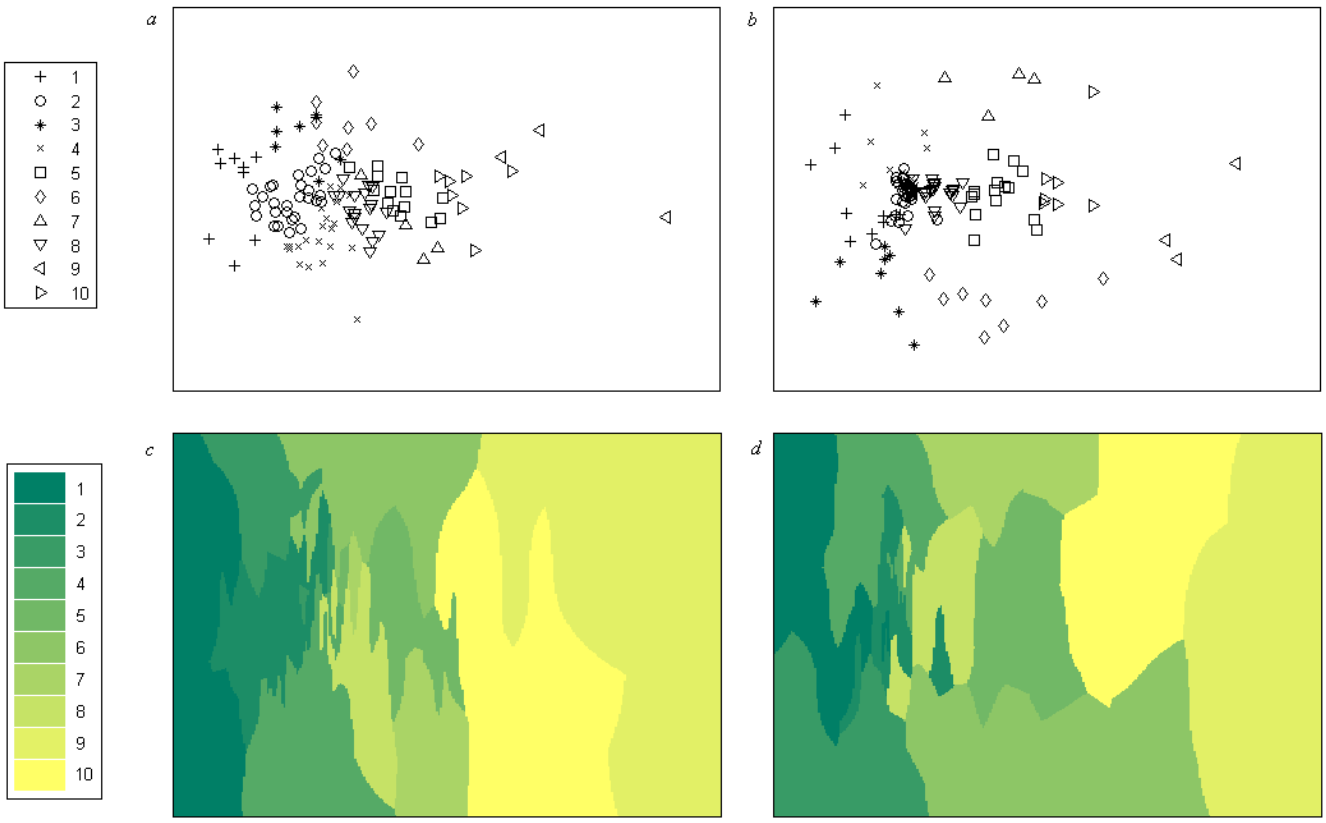

Figure 1: *a* and *b*: Multidimensional Scaling plots calculated for zero- and first- order Kullback-Leibler divergence matrices, showing ten categories yielded from Hierarchical Cluster Analysis. *c* and *d*: k-nearest neighbours territory plots for the equivalent dimensions, where k=3.

As shown in Figure 3, the first discriminant function maximally separates Close Personal and Inter-Employee Relations emails from Core Business and Routine Administration emails. The structure (loading) matrix of correlations between predictors and discriminant functions, as seen in Table 3, suggests the best predictors for distinguishing between the email categories are Dimension 1 and Dimension 2 of the MDS solution, and that the length of the sequence has little predictive value. Of the total sample of 144 emails, 58 (40.3%) were categorized correctly, compared with 25% who would be correctly classified by chance alone. Table 4 shows the confusion matrix for the classification results using the discriminant function with the four categories of emails. Routine Administration emails and Close Personal emails were more likely to be classified correctly than Core Business emails and Inter-Employee Relations emails. The most frequent misclassifications were between Core Business and Routine Administration emails, and between Close Personal and Inter-employee Relations emails. The stability of the classification procedure was checked by a cross-validation run. For the cross-validation cases, the correct classification rate was 35.4%. This suggests a reasonably consistent classification scheme.

Table 3: Structure matrix of correlations between predictors and functions calculated in Discriminant Analysis.

	Function		
	1	2	3
0 order dimension 1	-0.75	-0.18	-0.17
1st order dimension 1	-0.58	-0.58	0.13
0 order dimension 2	0.53	-0.40	-0.25
1st order dimension 2	-0.44	0.59	0.66

4 Discussion

The use of VRMs to describe emails yields quantitative indices which describe the way in which the communicators relate to each other at the point in time of that email. Statistical classification of these indices

Table 4: Confusion matrix in percentages for the four email categories from Discriminant Analysis.

	Close Personal	Core Business	Inter-Employee Relations	Routine Admin
Close Personal	44.4	16.7	33.3	5.6
Core Business	5.6	33.3	22.2	38.9
Inter-Employee Relations	41.7	13.9	30.6	13.9
Routine Admin	5.6	30.6	11.1	52.8

shows a structure that describes clusters of these relationships in a way which minimises information loss. The qualitative nature of the clusters shows some to be easily interpretable, such as Clusters one and nine (respectively, personal emails reflecting feelings and long emails giving news) and some which are less interpretable. This is due to the inclusion in the analysis of first-order transitions as well as (zero-order) frequencies, which access the utterance-to-utterance sequences of VRMs. For example, table 2 and figure 1 show that Clusters 5 and 6 are fairly similar in terms of the overall frequencies of individual VRMs, but are differentiated on the basis of the first-order sequences of VRMs. The difference between the sequences of these clusters is reflected in a differing email category makeup (cluster 5 is mixed, while Cluster 6 contains no Routine Admin emails).

The VRM indices also predict general email categories substantially above chance level. The results in this paper imply that VRM use is fairly strongly predictive of a business-personal dimension. Further work currently being undertaken may strengthen this claim. Jabbari et al. (2006) produced an automatic classifier

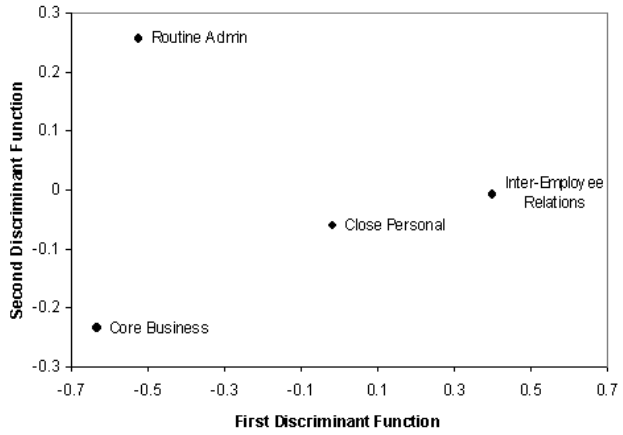

Figure 2: Group centroids for four categories of emails plotted on the first two discriminant functions.

which achieved a high level of accuracy (93%) in discriminating between business and personal emails, but the ability of their classifier is likely to be diminished with other corpuses due to its exclusive reliance upon distinguishing words. The discovery that the use of language intentions in email can access the broad category of that email is an important result, as the human assessment of an email in terms of these categories is likely to involve high-level cognitive processing of schemas (making use of assumptions and stereotypes) in order to succeed. While the training of a word-counting classifier algorithm works well within a corpus, it is unlikely that it reflects the process which humans undertake in order to classify. The use of a more general scheme of intentions combined with context-independend distinguishing words may yield a more versatile method of email classification.

Limitations and future work

This work analyses a relatively small set of emails from a single organization, and the generalizability of this data is not known. While the set of emails used is likely to be reasonably representative in terms of setting the limits of the *intention space*, the certainty of placement of internal boundaries has not been ascertained. Further work currently being undertaken will indicate the reliablility of this system of clustering.

Recent developments in the field of automatic tagging of Verbal Response Modes (Lampert, Dale & Paris, 2006) have given rise to the possibility of the use of VRMs in email analysis on a large scale. This taxonomy of emails has the potential to be used to describe intention relationships in terms of the intentions of multiple emails, and in turn describe social networks in terms of intention relationships. Such a stochastic model of email behavior could prove to be advantageous for security, anti-spam applications and marketing, as well as providing opportunities to improve end-user experience.

Acknowledgments

The authors would like to acknowledge funding and support for this work from Government Communication Headquarters (GCHQ), Cheltenham, United Kingdom.

References

Austin, J. L. (1975). *How to do things with words.* Oxford: Clarendon Press.

Clark, H., & Brennan, S. (1991). Grounding in Communication. In L. Resnick, J. Levine & S. Teasley (Eds.), *Perspectives on Socially Shared Cognition (pp. 127-149).* Washington, DC: American Psychological

Association.

Jabbari, S., Allison, B., Guthrie, D., Guthrie, L. (2006) Towards the Orwellian nightmare. *Proceedings of the COLING/ACL 2006 Main conference poster sessions*, 407-411

Kruskal, J. B., & Wish, M. (1978). Multidimensional Scaling: *Sage University Paper series on Quantitative Application in the Social Sciences, 07-011.* Beverley Hills and London: Sage Publications.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79-86.

Lampert, A., Dale, R., & Paris, C. (2006). Classifying speech acts using Verbal Response Modes. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, 34-31.

Searle, J. R. (1969). *Speech acts.* Cambridge: Cambridge University Press.

Searle, J. R., & Vanderveken, D. (1985). *Foundations of illocutionary logic.* Cambridge: Cambridge University Press.

Stiles, W. B. (1978). *Manual for a taxonomy of verbal response modes.* Chapel Hill, NC: Institute for research in social science, University of North Carolina at Chapel Hill.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.