Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pages 141–147.
Trento, Italy, 30 May – 1 June 2007. Edited by Ron Artstein and Laure Vieu.

Group Dialects in an Online Community

Patrick G. T. Healey
Interaction, Media and Communication Group
Department of Computer Science
Queen Mary University of London
ph@dcs.qmul.ac.uk

Carl Vogel
Computational Linguistics Group
School of Computer Science and Statistics
Trinity College
Dublin
vogel@tcd.ie

Arash Eshghi
Interaction, Media and Communication Group
Department of Computer Science
Queen Mary University of London
eshghi@dcs.qmul.ac.uk

Abstract

Variations in group sub-languages evolve quickly and are a key marker of social boundaries such as those between professions, workgroups, tribes and families. In this paper we present a quantitative analysis of the effects of group structure on language use in naturalistic interaction. The data come from text chat interactions in an online social community. Using statistical techniques developed for the analysis of authorship attribution we use this corpus to test three accounts of the emergence of group sub-languages: a) local coordination mechanisms b) network topology and c) influential individuals. The results indicate that it is influential individuals who have the strongest effects on sub-language convergence.

1 Introduction

Language use is sensitive to a variety of social and cultural factors. Place of residence, education, religion, occupation, hobby, age group, expertise and ethnic origin can all influence people's use of e.g., words, syntax prosody, and style. Communicative alignment –similarity in the forms of language used by participants in an interaction– is consequently a key indicator, for members and analysts alike, of community co-membership (Clark, 1996; Clark, 1998; Gumperz, 1996).

Field studies have shown that communicative alignment indexes social organisation at quite fine-grained resolutions. For example, linguistic homogeneity is a criterion for distinguishing tribal groupings in ethnographic studies of hunter-gather societies (Dunbar, 1993). Communication in institutional environments is often characterised by local, institution-specific, forms of talk (Bergmann and Luckmann, 1994). Distinct sub-languages have been documented within different subgroups in a single workplace (Robinson and Bannon, 1991) and families also frequently develop their own jargon words and idioms.

Communal sub-languages can emerge rapidly. Experimental studies have shown that seman-

tically distinct sub-languages develop in small groups in less than an hour of group interaction (Healey, 1997; Healey et al., 2007). This divergence can interere with the intelligibility of communication across community boundaries (Gumperz, 1996; Shaw and Gaines, 1988). People can also sometimes use their ability to switch between different codes and repetoires as a means of establishing alignment with, or exclusion of, others (Gumperz, 1982).

We can distinguish three logicially independent, but not mutually exclusive, hypotheses that have been suggested to account for group sub-language co-ordination:

1. Local Dialogue Coordination: patterns of co-ordination are explained by local, pair-specific, dialogue mechanisms that are common across interactions (Garrod and Doherty, 1994; Clark, 1996; Healey et al., 2007).

2. Network Topology: patterns of co-ordination are explained by differences in the patterns of interaction amongst the members of a population (Garrod and Doherty, 1994; Healey et al., 2007).

3. Influential Individual: patterns of co-ordination are explained by reference to key individuals who have a disproportionate effect on the language of others in the group (Garrod and Anderson, 1987).

In this paper we investigate how well these hypotheses account for the patterns of language use observed in a text-based online community. The data consists of all the interactions over a three day period in a group of 150 individuals. This provides a unique opportunity to carry out a quantitative analysis of the relationship between of patterns of interaction in this community and similarity of language use.

Although previous work has looked at patterns of interaction and the emergence of group norms in email (Postmes et al., 2000) we believe this is the first quantitative study of conversational interactions across a whole community. The natural, conversational character, of the exchanges (see below) and the scale of the analysis help to address some of the key limitations of case-based and experimental studies of group sub-languages.

First we provide background information about the character of the online environment: 'Walford' and the data used in the analysis. Then we present the statistical technique -unigram statistics developed for forensic linguistics and the results of our analysis.

2 Interaction in Walford

Walford is a text-based online social community or 'talker' that has been established for more than a decade. It has approximately 1500 regular users who are predominantly based in North America and Europe. It emerged as one of the many variants on James Aspene's 'TinyMUD'[1] which was first created in 1989. The environment is structured around a spatial metaphor with rooms, objects, players and exits. Once users have reached a sufficient level of expertise they can create their own rooms, objects and commands (macros).

The residents of Walford have taken advantage of this structure to build up a complex environment. There are shared public spaces such as a high street, a pub, a townhall, a bank, a bus station. There is also a rubbish dump and a network of roads. Although based on a MUD, people's main preoccupation in Walford is with their interactions and social relationships with each other. This is illustrated by a sample of conversation topics from the logs: chocolate, outsourcing, mobile phones, births and deaths in resident's (real) families, relationships (both inside and outside Walford), economics assignments, redundancy and boredom.

A sample conversational sequence from the Walford pub is provided in Excerpt 1. The extract helps to illustrate the conversational character of theses exchanges. Multi-installment turns, clarification questions and ellipsis are common features.

The data analysed in this paper come from a corpus of chat logs collected over approximately one year in 2004-2005. For each person-to-person the ID of the 'speaker', their virtual location, the recipients ID and their virtual location is recorded. In order to protect the anonymity of participants the names of people, characters, places,

[1] MUD stands for multi-user dungeon, from its text-based computer gaming tradition.

Table 1: A Sample Dialogue from the Queen Vic pub in Walford

A:	Yeah dave is a cool guy... Good mechanic... Good guy.
A:	though I wouldn't be surprised if he was a wife/child beater.
B:	he seemed very gentle
B:	but he did drink a lot
B:	he an my dad share war stories now that they've both had their prostates removed
B:	ugh
B:	the last thing you want to hear two old guys chatting about
A:	war = prostate ? or vietnam?
B:	hehe yeah prostate

some commands and the name of the environment have been changed. Users agreed as a condition of use to the system to the logs being used, in anonymised form, for the purposes of research and publication.

3 Methods

A sample of three consecutive days of logs of interaction in Walford were randomly selected for anaylsis. The logs were preprocessed to remove all automatic formating and command names. This yielded a total of 20,043 turns by 150 unique identities from 148 unique locations.

To analyze the data, we applied statistical text classification methods (Van Gijsel and Vogel, 2003; Vogel and Brisset, 2006; Vogel, 2007). This approach draws on research on authorship attribution in forensic linguistics in which there is a preference for methods that do not use content analysis. This helps to ensure more robust inter-judge reliability and for this reason letter n-grams are favoured (Chaski, 1999). Surprisingly, letter unigrams have provided remarkable results to date. Although, it seems counter-intuitive that letter unigrams might be effective in identifying categories such as authorship or genre the relative efficacy of predictive text on mobile phones suggests what is possible. Consider also the way that Scrabble boards distribute their hundred letter tiles across the alphabet differently in German and in English; or notice the fact that a latinate vocabulary will have a noticeable distribution of the letter "Q" (e.g. "horse riding" vs. "equestrian"). These observations indicate how word choice impacts on orthography (poetry and lipograms aside, written text does not involve word choice on the basis of the spelling of words). This approach also has advantages over measures based on shared words. A long tail of words in any corpus corresponds to singleton occurrences and many will appear in one text and not another. In addition, the closed class words will all be shared and differently inflected forms of the same root may appear. Thus, subword analysis is necessary. Letter unigrams are thus a limiting case and, unlike words, constitute a closed class.

Here we use the chi-square divided by degrees of freedom (cbdf) statistic adapted from other work in comparing corpora (Kilgarriff, 2001; Kilgarriff and Salkie, 1996). The idea is to compare the n-gram distributions between two files in any category. The overall similarity between two files is computed as the sum of the chi-square values of each of the n-grams between the two files, relativized to the number of distinct n-gram types compared. A smaller cumulative chi-square value thus indicates a smaller difference between two files (note that this is the opposite of the normal, contrastive, use of the chi-square test in order to locate significant differences). The similarity metric is computed for all pairwise comparisons of relevant files. These similarity metrics can then be used to rank order the files by the categories they comprise. That is, one has a category of all of the texts by a single author, versus all of the other texts. Mann-Whitney tests can then be used to examine whether each file in a category fits best with its natural category or with some other category.

The Walford log is organised as a temporally ordered sequence of turns with speaker ID, location, recipient ID(s) and their location(s) (local or remote). In the analysis reported here, we used speaker ID's as categories. For each speaker the logs were separated into single files corresponding to each continuous sequence of interaction with each recipient group. For example, if A speaks to B for 5 turns then C for 5 turns then B again for 5 turns this creates three files for speaker A. If by contrast they alternate between A and B for 10 turns this creates 10 files for speaker A. Each file thus consists of a contiguous sequence of turns by one speaker to a particular set of recipients.

With this background understood, it is possible to understand that the single-line file containing:

```
************************* ok **************************
```

scores as most dis-similar to a file of 183 lines, with this representative start:

```
i live
```

This approach allows us to explore the relationship between absolute similarity among files and appropriateness in their category (speaker) and also to explore the similarity relationships between categories of speakers partitioned according to who interacted directly or not.

Suppose a speaker has 20 files. It is an open question whether each of the files in that 20 will be most like the other 19 produced by that speaker or more like the other files derived from the log. Further, one wants to know how well the file fits with the sets of files produced by other speakers (categories). In fact, it might fit with a number of speakers' files and to a relatively high degree of significance.

A speaker whose files are most similar to the rest of the same speaker's files is a self-homogeneous speaker. A speaker can also be homogeneous with respect to other speakers' files. The homogeneity of a speaker with respect to another speaker can be measured by the (relativized) number of files produced by the speaker which score as most similar for the category. For a given speaker, A, we calculate:

1. The similarity set: the set of other speakers whose files are reliably ($p > 0.05$) similar to individual files produced by A.
2. The contact set: all of the speakers that spoke to A and all the speakers that A spoke to.

In order to examine further the interrelationship between patterns of interaction and levels of similarity we also adopted the notion of a pivot. A pivot is a speaker who is has a common relationship to at least two other speakers and therefore can 'represent' them. A pivot set is a smallest set which represents all of the speakers. For example, the audience pivot set is the smallest set of recipients that everyone has sent at least one turn to at least one of. Here we distinguish the contact pivot set and the similarity pivot set.

The pivot set can be understood as a contrast with Gärdenfors' analyses of meaning-determining groups (Gardenfors, 1993). A filter, defined on sets of sets (here, the basic entities within the sets are individuals), is a construct smaller than the power set of the set of basic individuals. The entire set of individuals is a member of a filter, but the empty set is not. For any two sets of individuals that are elements of a filter, their intersection is also a member. Further, a filter is monotonically increasing – if there is a set of individuals in the filter, then every containing superset is an element as well. This is a useful construct for explicating various social structures for meaning. Distinctions are available through distinct subsets of individuals. If there is exactly one individual that is common to all sets of individuals in the filter, then that individual can be seen as a 'dictator' of meaning, (Gardenfors, 1993). In thinking of a pivot set, one is considering a set that characterizes a set of sets that is not necessarily a filter – thus, no unique determiner of meaning, and potentially no shared meaning. Thinking of a signature set of sets based on a set of individuals, with a monotonically increasing closure, a pivot set is the smallest set of individuals required to ensure that every set is represented by one individual. A dictator would correspond to a singleton pivot set, the entire set of individuals would constitute a pivot set just if none of the sets of subsets had any elements in common (Babel).

In our analysis, the pivot sets can be treated in terms of contact or by similarity – sets of individuals who communicated directly with each other or sets of individuals comprising similarity equiv-

alence classes.

The *contact pivot set* is just the audience pivot set. This is intended to capture the degree of interconnectedness within the community. If the contact pivot set is large there is a relatively 'dispersed' network of interconnections between residents, if it is small there are a number of 'gatekeeper' or 'funnel' individuals who provide contact points between different, relatively isolated, groups.

The *similarity pivot* set is the smallest set of speakers who are reliably similar to all the other speakers. In effect they represent the degree of differentiation of sub-group 'dialects' in the sample. If the similarity set is large there is relatively little convergence in dialects amongst the residents if it is small there are correspondingly fewer distinguishable 'sub-languages'.

The *non-pivot* set is the speakers who are neither contact pivots nor similarity pivots.

It is worth noting that the similarity based pivot set and the contact based pivot set are logically independent. A population with a relatively dispersed network of interactions could, nonetheless, have a relatively homogenous dialect. Conversely a highly centralised population might nonetheless sustain multiple dialects.

4 Results

The results reported here are based on the first 25 percent of the data set, and consists of the turns of 39 different residents of Walford.[2] This resulted in 547 files, with an average of 14.03 files per speaker.

The first question concerns the degree of overlap between the similarity set and the contact set. Of the 39, 25 had spoken to someone they were similar to, 14 of the 39 did not speak to anyone they were similar to. In the receiving direction, 23 speakers had at least one of their similarity set who had spoken to them and 16 had none of their similarity set among the people who spoke with them.

[2]The three day sample involves comparison of approximately 4,500 files with each other, which yields a space to reason about similarity with about 10 million elements. The combinatorial problem is large but not insoluble. The second author is investigating this complexity problem.

The second question concerns the pivots. In the sample of 39 speakers there are 4 similarity pivots and 27 contact pivots. However, in part because the data analyzed was truncated as the first 25% of the overall three-day log, some of the contact pivots are not present as actual speakers. Thus, the set of contact pivots who were also speakers (and thus provided text that can be measured for similarity, see below) contains seven individuals.

Table 2: Average Self-similarity in Pivot Groups

	N	Self-Homogeneity	Files
Nonpivot	28	0.03	7.36
S-pivot	4	0.15	45.43
C-pivot	7	0.07	5.75

Table 2 shows the average levels of self-homogeneity amongst speakers in the different pivot groups. The Similarity pivots have the highest level of self similarity, the Contact pivots moderate and the Non pivots lowest.

5 Discussion

Prima facie, the results provide evidence that local dialogue mechanisms such as interactive alignment (Pickering and Garrod, 2004), grounding (Clark and Wilkes-Gibbs, 1986) or local repair and clarification (Healey and Mills, 2006; Healey et al., 2007) do not account for the patterns of similarity in language use, as measured by letter unigrams, observed in Walford. If the mechanisms of dialect co-ordination were primarily local then the main locus of influence should be the contact set. However, the results show that residents interact with relatively high proportion of 'disimilar' people. This is indicative that convergence is not primarily mediated by direct contact.

Moreover, it appears that the pattern of interconnections amongst residents or 'network topology' is also a poor predictor of the pattern of sub-language convergence. Although there are a relatively high number of contact pivots (27) – indicating that the network is relatively diffuse or fragmented – there are a relatively low number of similarity pivots (4) indicating a small set of (statistically) distinguishable 'dialects'.

More importantly, in the current data set no individuals were both Similarity pivots and Contact pivots. This is consistent with an 'influential individual' explanation of the emergence of sub-group languages (Garrod and Anderson, 1987). Particular individuals who contribute a high number of turns (but who are not particularly well interconnected with other residents) have a disproportionate influence on the patterns of language use in the group but this influence appears to be mediated indirectly.

It is also interesting that in terms of self-similarity the least homogeneous speakers were the non-pivots. By definition these speakers interacted with fewer people and were least similar to the others. It appears that being, in effect, peripheral nodes in the network correlates with less consistent language use. The speakers with the highest level of self-similarity were the Similarity pivots.

Considered together this analysis suggests that the factors which promote sub-language convergence operate through indirect patterns of influence over successive exchanges rather than through local patterns of influence within interactions.

Overall, this is generally consistent with a version of Putnam's linguistic division of labour (Putnam, 1975) explanation of co-ordination of meaning in which control of language use is effectively deferred to key individuals in a community. In Walford it is unclear whether this is due to sheer persistence and volume of communication or whether, as in Putnam's conjecture, it is a consequence of differences in expertise or perhaps esteem.

A key challenge in this analysis has been to develop techniques that can analyse large networks of communal interaction. Two problems arise, first we want to look at a much smaller grain size than is typical for corpus analysis; turns and groups of turns rather than extended texts. In addition, a clear implication of this work is that we must pay close attention to the pattern of possible direct and *indirect* inter-relationships in a community. This creates a formidable computational problem.

The uni-gram technique has the advantage that it avoids problematic judgements about the form or content or intended force of each contribution. However, it simultaneously raises questions about what is really being measured. It's main virtue for our purposes is as a crude but robust index of similarity. Future work will need to explore how it correlates with other linguistic and pragmatic structures.

Acknowledgements

We are grateful to Graham White for his work on creating the logs analysed in this paper and the residents of Walford for agreeing to participate in this research. Science Foundation Ireland (RFP 05/RF/CMS002) has supported this work.

References

J. R. Bergmann and T. Luckmann. 1994. Reconstructive genres of everyday communication. In U. Quasthoff, editor, *Aspects of Oral Communications*, pages 289–304. Berlin: Mouton de Gruyter.

Carole Chaski. 1999. Linguistic authentication and reliability. In *Proceedings of National Conference on Science and the Law*, pages 97–148.

H. H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Herbert H. Clark. 1996. Communities, commonalities, and communication. In John J. Gumperz and Stephen C. Levinson, editors, *Rethinking linguistic relativity*, pages 324–355. Cambridge: Cambridge University Press.

Herbert H. Clark. 1998. Communal lexicons. In K. Malmkjoer and J. Williams, editors, *Context in language learning and language understanding*, pages 63–87. Cambridge: CUP, 3rd edition.

R. Dunbar. 1993. Coevolution of neocortex size, group size and language in humans. *Behavioural and Brain Sciences*, 16:681–735.

P. Gardenfors. 1993. The emergence of meaning. *Linguistics and Philosophy*, 16:285–309.

Simon C. Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27:181–218.

Simon C. Garrod and Gwyneth Doherty. 1994. Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53:181–215.

John J. Gumperz. 1982. Conversational code switching. In John J. Gumperz, editor, *Discourse Strategies*, pages 59–99. Cambridge: Cambridge University Press.

John J. Gumperz. 1996. The linguistic and cultural relativity of conversational inference. In John J. Gumperz and Stephen C. Levinson, editors, *Rethinking linguistic relativity*, pages 374–406. Cambridge: Cambridge University Press.

P.G.T. Healey and G. Mills. 2006. Participation, precedence and co-ordination in dialogue. In R. Sun and N. Miyake, editors, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1470–1475.

P.G.T. Healey, N. Swoboda, I. Umata, and J. King. 2007. Graphical language games: Interactional constraints on representational form. *Cognitive Science*, 31:285–309.

P.G.T. Healey. 1997. Expertise or expert-*ese*?: The emergence of task-oriented sub-languages. In M.G. Shafto and P. Langley, editors, *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 301–306.

A. Kilgarriff and R. Salkie. 1996. Corpus similarity and homogeneity via word frequency. In *Proceedings of Euralex 96*.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

M. Pickering and S. Garrod. 2004. The interactive alignment model. *Behavioral and Brain Sciences*, 27(2):169–189.

T. Postmes, R. Spears, and M. Lea. 2000. The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3):341–371.

Hilary Putnam. 1975. The meaning of meaning. In K. Gunderson, editor, *Language, Mind, and Knowledge, Minnesota Studies in the Philosophy of Science, 7*. Minneapolis: University of Minnesota Press.

M. Robinson and Liam Bannon. 1991. Questioning representations. In L. Bannon, M. Robinson, and K. Schmidt, editors, *Proceedings of the Second European Conference on CSCW*, pages 219–233. Dordrecht: Kluwer.

Mildred L. G. Shaw and Brian R. Gaines. 1988. A methodology for recognising consensus, correspondence, conflict and contrast in a knowledge acquisition system. In *Third Workshop on Knowledge Acquisition for Knowledge-Based Systems*. Banff.

Sofie Van Gijsel and Carl Vogel. 2003. Inducing a cline from corpora of political manifestos. In Markus Aleksy et al., editor, *Proceedings of the International Symposium on Information and Communication Technologies*, pages 304–310.

Carl Vogel and Sandrine Brisset. 2006. Hearing voices in the poetry of brendan kennelly. In *Varieties of Voice*. 3rd international BAAHE conference. Leuven, 7-9 December 2006.

Carl Vogel. 2007. N-gram distributions in texts as proxy for textual fingerprints. In Anna Esposito, Eric Keller, M. Marinaro, and Maja Brat anic, editors, *The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue*. IOS Press.