# Computational Measures for Language Similarity across Time in Online Communities

**David Huffaker Joseph Jorgensen Francisco Iacobelli Paul Tepper Justine Cassell**

Northwestern University

{d-huffaker, josephj, f-iacobelli, ptepper, justine}@northwestern.edu

## Abstract

This paper examines language similarity in messages over time in an online community of adolescents from around the world using three computational measures: Spearman's Correlation Coefficient, Zipping and Latent Semantic Analysis. Results suggest that the participants' language *diverges* over a six-week period, and that divergence is not mediated by demographic variables such as leadership status or gender. This divergence may represent the introduction of more unique words over time, and is influenced by a continual change in subtopics over time, as well as community-wide historical events that introduce new vocabulary at later time periods. Our results highlight both the possibilities and shortcomings of using document similarity measures to assess convergence in language use.

## 1 Introduction

While document similarity has been a concern in computational linguistics for some time, less attention has been paid to change in similarity across time. And yet, while historical linguists have long addressed the issue of divergence or convergence among language groups over long periods of time, there has also been increasing interest in convergence (also referred to as entrainment, speech accommodation, or alignment) in other areas of Linguistics, with the realization that we have little understanding of change in very short periods of time, such as months, in a particular conversational setting, between two people, or in a large group.

The Internet provides an ideal opportunity to examine questions of this sort since all texts persevere for later analysis, and the diversity in kinds of online communities ensures that the influence of social behavior on language can be examined. Yet there has been very little work on language similarity in online communities.

In this paper we compare the use of three separate tools to measure document or message similarity in a large data set from an online community of over 3,000 participants from 140 different countries. Based on a review of related work on corpus similarity measures and document comparison techniques (Section 2.2), we chose Spearman's Correlation Coefficient, a comparison algorithm that utilizes GZIP (which we will refer to as "Zipping") and Latent Semantic Analysis. These three tools have all been shown effective for document comparison or corpus similarity, but never to our knowledge have any of them been used for document similarity over time, nor have they been compared to one another. Even though each of these tools is quite different in what it specifically measures and how it is used, and each has been used by quite different communities of researchers, they are all fairly well-understood (Section 4).

## 2 Related Work

In the next sections, we review literature on language similarity or convergence. We also review literature on the three computational tools, Spearman's Correlation Coefficient (SCC), Zipping, and Latent Semantic Analysis (LSA).

### 2.1 Language Similarity in Computer-mediated Communication

In dyadic settings, speakers often converge to one another's speech styles, not only matching the choice of referring expressions or other words, but also structural dimensions such as syntax, sound characteristics such as accent, prosody, or phonol-



*Proceedings of the Analyzing Conversations in Text and Speech (ACTS) Workshop at HLT-NAACL 2006*, pages 15–22,
New York City, New York, June 2006. 

ogy, or even non-verbal behaviors such as gesture (Brennan & Clark, 1996; Street & Giles, 1982).

Some scholars suggest that this convergence or entrainment is based on a conscious need to accommodate to one's conversational partner, or as a strategy to maximize communication effectiveness (Street & Giles, 1982). Others suggest that the alignment is an automatic response, in which echoic aspects of speech, gesture and facial expressions are unconscious reactions (Garrod & Anderson, 1987; Lakin, Jefferies, Cheng, & Chartrand, 2003). In short, conversational partners tend to accommodate to each other by imitating or matching the semantic, syntactic and phonological characteristics of their partners (Brennan & Clark, 1996; Garrod & Pickering, 2004).

Many studies have concentrated on dyadic interactions, but large-scale communities also demonstrate language similarity or convergence. In fact, speech communities have a strong influence in creating and maintaining language patterns, including word choice or phonological characteristics (Labov, 2001). Language use often plays an important role in constituting a group or community identity (Eckert, 2003). For example, language 'norms' in a speech community often result in the conformity of new members in terms of accent or lexical choice (Milroy, 1980). This effect has been quite clear among non-native speakers, who quickly pick up the vernacular and speech patterns of their new situation (Chambers, 2001), but the opposite is also true, with native speakers picking up speech patterns from non-native speakers (Auer & Hinskens, 2005)

Linguistic innovation is particularly salient on the Internet, where words and linguistic patterns have been manipulated or reconstructed by individuals and quickly adopted by a critical mass of users (Crystal, 2001). Niederhoffer & Pennebaker (2002) found that users of instant messenger tend to match each other's linguistic styles. A study of language socialization in a bilingual chat room suggests that participants developed particular linguistic patterns and both native and non-native speakers were influenced by the other (Lam, 2004). Similar language socialization has been found in ethnographic research of large-scale online communities as well, in which various expressions are created and shared by group members (Baym, 2000; Cherny, 1999).

Other research not only confirms the creation of new linguistic patterns online, and subsequent adoption by users, but suggests that the strength of the social ties between participants influences how patterns are spread and adopted (Paolillo, 2001). However, little research has been devoted to how language changes over longer periods of time in these online communities.

### 2.2 Computational Measures of Language Similarity

The unit of analysis in online communities is the (e-mail or chat) message. Therefore, measuring entrainment in online communities relies on assessing whether or not similarity between the messages of each participant increases over time. Most techniques for measuring document similarity rely on the analysis of word frequencies and their co-occurrence in two or more corpora (Kilgarriff, 2001), so we start with these techniques.

Spearman's Rank Correlation Coefficient (SCC) is particularly useful because it is easy to compute and not dependent on text size. Unlike some other statistical approaches (e.g. chi-square), SCC has been shown effective on determining similarity between corpora of varying sizes, therefore SCC will serve as a baseline for comparison in this paper (Kilgarriff, 2001).

More recently, researchers have experimented with data compression algorithms as a measure of document complexity and similarity. This technique uses compression ratios as an approximation of a document's information entropy (Baronchelli, Caglioti, & Loreto, 2005; Benedetto, Caglioti, & Loreto, 2002). Standard Zipping algorithms have demonstrated effectiveness in a variety of document comparison and classification tasks. Behr et al. (2003) found that a document and its translation into another language compressed to approximately the same size. They suggest that this could be used as an automatic measure for testing machine translation quality. Kaltchenko (2004) argues that using compression algorithms to compute relative entropy is more relevant than using distances based on Kolmogorov complexity. Lastly, Bendetto et al. (2002) present some basic findings using GZIP for authorship attribution, determining the language of a document, and building a tree of language families from a text written in different languages. Although Zipping may be a conten-

tious technique, these results present intriguing reasons to continue exploration of its applications.

Latent Semantic Analysis is another technique used for measuring document similarity. LSA employs a vector-based model to capture the semantics of words by applying Singular Value Decomposition on a term-document matrix (Landauer, Foltz, & Laham, 1998). LSA has been successfully applied to tasks such as measuring semantic similarity among corpora of texts (Coccaro & Jurafsky, 1998), measuring cohesion (Foltz, Kintsch, & Landauer, 1998 ), assessing correctness of answers in tutoring systems (Wiemer-Hastings & Graesser, 2000) and dialogue act classification (Serafin & Di Eugenio, 2004).

To our knowledge, statistical measures like SCC, Zipping compression algorithms, or LSA have never been used to measure similarity of messages over time, nor have they been applied to online communities. However, it is not obvious how we would verify their performance, and given the nature of the task – similarity in over 15,000 e-mail messages – it is impossible to compare the computational methods to hand-coding. As a preliminary approach, we therefore decided to apply all three methods in turn to the messages in an online community to examine change in linguistic similarity over time, and to compare their results. Through the combination of lexical, phrasal and semantic similarity metrics, we hope to gain insight into the questions of whether entrainment occurs in online communities, and of what computational measures can be used to measure it.

### 2.3 The Junior Summit

The Junior Summit launched in 1998 as a closed online community for young people to discuss how to use technology to make the world better. 3000 children ages 10 to 16 participated in 1000 teams (some as individuals and some with friends). Participants came from 139 different countries, and could choose to write in any of 5 languages. After 2 weeks online, the young people divided into 20 topic groups of their own choosing. Each of these topic groups functioned as a smaller community within the community of the Junior Summit; after another 6 weeks, each topic group elected 5 delegates to come to the US for an in-person forum. The dataset from the Junior Summit comprises more than 40,000 e-mail messages; however, in the current paper we look at only a sub-set of these data – messages written *in English* during the 6-week *topic group* period. For complete details, please refer to Cassell & Tversky (2005).

## 3 The Current Study

In this paper, we examine entrainment among 419 of the 1000 user groups (the ones who wrote in English) and among the 15366 messages they wrote over a six-week period (with participants divided into 20 topic groups, with an average of 20.95 English writers per group). We ask whether the young people's language converges over time in an online community. Is similarity between the texts that are produced by the young people greater between adjacent weeks than between the less proximally-related weeks? Furthermore, what computational tools can effectively measure trends in similarity over time?

### 3.1 Hypotheses

In order to address these questions, we chose to examine change in similarity scores along two dimensions: (1) at the level of the individual; and (2) across the group as a whole. More specifically, we examine similarity between all pairs of individuals in a given topic group over time. We also compared similarity across the entire group at different time periods.

As depicted below, we first look at pairwise comparisons between the messages of participants in a particular topic group within a given time period, $T_k$ (one week). For every pair of participants in a group, we calculated the similarity between two documents, each comprising all messages for a participant in the pair. Then we averaged the scores computed for all topic groups within a time period $T_k$ and produced $P_{Tk}$, the average, pairwise similarity score for $T_k$. Our first hypothesis is that the average, pairwise similarity will increase over time, such that:

$$P_{T1} < P_{T2} < P_{T3} < P_{T4} < P_{T5} < P_{T6}$$

For our second set of tests, we compared all messages from a single time period to all messages of a previous time period within a single topic group. Our hypothesis was that temporal proximity would correlate with mean similarity, such that the messages of two adjacent time periods would exhibit more similarity than those of more distant

time periods. In order to examine this, we perform two individual hypothesis tests, where $M_k$ is the document containing all the messages produced in time period $T_k$, and S(X,Y) is the similarity score for the two documents X and Y.

a) $S(M_k, M_{k-1}) > S(M_k, M_{k-2})$
b) $S(M_k, M_{k-1}) > S(M_k, M_1)$

Finally, we posit that SCC, Zipping and LSA will yield similar results for these tests.

## 4 Method

To prepare the data, we wrote a script to remove the parts of messages that could interfere with computing their similarity, in particular quoted messages and binary attachments, which are common in a corpus of email-like messages. We also removed punctuation and special characters.

### 4.1 Spearman's Correlation Coefficient

SCC is calculated as in Kilgarriff (2001). First, we compile a list of the common words between the two documents. The statistic can be calculated on the *n* most common words, or on all common words (i.e. *n* = total number of common words). We applied the latter approach, using all the words in common for each document pair. For each document, the *n* common words are ranked by frequency, with the lowest frequency word ranked 1 and the highest ranked *n*. For each common word, *d* is the difference in rank orders for the word in each document. SCC a normalized sum of the squared differences:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

The sum is taken over the *n* most frequent common words. In the case of ties in rank, where more than one word in a document occurs with the same frequency, the average of the ranks is assigned to the tying words. (For example, if words $w_1$, $w_2$ and $w_3$ are ranked 5th, 6th and 7th then all three words would be assigned the same rank of $\frac{5+6+7}{3} = 6$).

### 4.2 Zipping

When compressing a document, the resulting compression ratio provides an estimate of the document's entropy. Many compression algorithms generate a dictionary of sequences based on frequency that is used to compress the document. Likewise, one can leverage this technique to determine the similarity between two documents by assessing how optimal the dictionary generated when compressing one document is when applied to another document. We used GZIP for compression, which employs a combination of the LZ77 algorithm and Huffman coding. We based our approach on the algorithm used by (Benedetto, Caglioti, & Loreto, 2002), where the cross-entropy per character is defined as:

$$\frac{\text{length}(\text{zip}(A+B)) - \text{length}(\text{zip}(A))}{\text{length}(B)}$$

Here, *A* and *B* are documents; $A+B$ is document *B* appended to document *A*; zip(*A*) is the zipped document; and length(*A*) is the length of the document. It is important to note that the test document (B) needs to be small enough that it doesn't cause the dictionary to adapt to the appended piece. (Benedetto, Caglioti, & Loreto, 2002) refer to this threshold as the crossover length. The more similar the appended portion is, the more it will compress, and vice versa. We extended the basic algorithm to handle the extremely varied document sizes found in our data. Our algorithm does two one-way comparisons and returns the mean score. Each one-way comparison between two documents, *A* and *B*, is computed by splitting *B* into 300 character chunks. Then for each chunk, we calculated the cross entropy per character when appending the chunk onto *A*. Each one-way comparison returns the mean calculation for every chunk.

We fine-tuned the window size with a small, hand-built corpus of news articles. The differences are slightly more pronounced with larger window sizes, but that trend starts to taper off between window sizes of 300 and 500 characters. In the end we chose 300 as our window size, because it provided sufficient contrast and yet still gave a few samples from even the smallest documents in our primary corpus.

### 4.3 Latent Semantic Analysis (LSA)

For a third approach, we used LSA to analyze the semantic similarity between messages across different periods of time. We explored three imple-

mentations of LSA: (a) the traditional algorithm described by Foltz et al (1998 ) with one semantic space per topic group, (b) the same algorithm but with one semantic space for all topic groups and (c) an implementation based on *Word Space* (Schutze, 1993) called *Infomap*. All three were tested with several settings such as variations in the number of dimensions and levels of control for stop words, and all three demonstrated similar results. For this paper, we present the Infomap results due to its wide acceptance among scholars as a successful implementation of LSA.

To account for nuances of the lexicon used in the Junior Summit data, we built a semantic space from a subset of this data comprised of 7000 small messages (under one kb) and 100 dimensions without removing stop words. We then built vectors for each document and compared them using cosine similarity (Landauer, Foltz, & Laham, 1998).

## 5 Results

The tools we employ approach document similarity quite differently; we therefore compare findings as a way of triangulating on the nature of entrainment in the Junior Summit online community.

### 5.1 Pairwise Comparisons over Time

First, we hypothesized that messages between individuals in a given topic group would demonstrate more similarity over time. Our findings did not support this claim; in fact, they show the opposite. All three tests show slight convergence between time period one and two, some variation, and then divergence between time periods four, five and six.

Spearman's Correlation Coefficient demonstrates a steady decline in similarity. As shown in Figure 1, the differences between time periods were all significant, $F_{(5,1375)} = 21.475$, $p<.001$, where $N=1381$ ($N$ represents user pairs across all six time periods).

Zipping also shows a significant difference between each time period, $F_{(5,1190)} = 39.027$, $p<.001$, $N=1196$, demonstrating a similar decline in similarity, although not as unwavering. See Figure 2.

LSA demonstrates the same divergent trend over time, $F_{(5,1410)} = 27.139$, $p<.001$, $N=1416$, with a slight spike at $T_4$ and $T_5$. While the dip at time 3 is more pronounced than SCC and Zipping, it is still consistent with the overall findings of the other measures. See Figure 3.

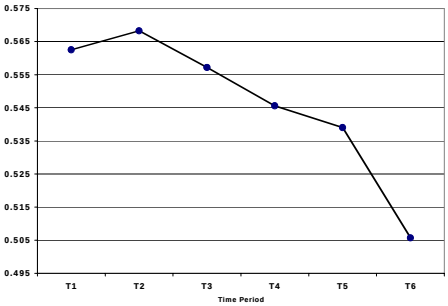


Figure 1. Spearman's Correlation Coefficient Similarity Scores for all Pairwise comparisons, $T_1 - T_6$

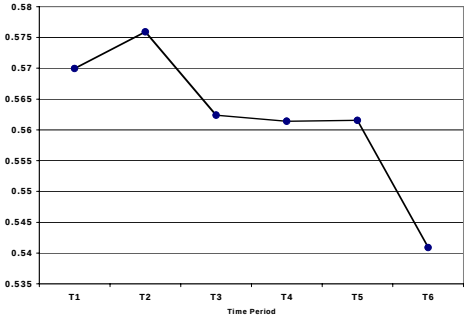


Figure 2. Zipping Similarity Scores for all Pairwise comparisons, $T_1 - T_6$

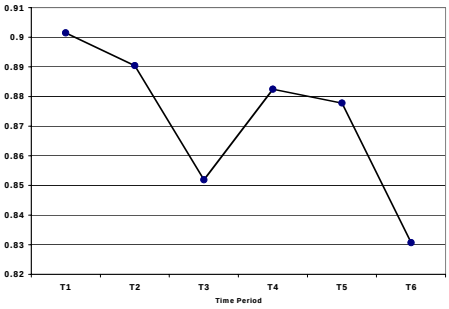


Figure 3. LSA Similarity Scores for all Pairwise comparisons, $T_1 - T_6$.

Because of these surprising findings, we examined the influence of demographic variables, such as leadership (those chosen as delegates from each topic group to the in-person forum), gender, and the particular topic groups the individuals were a part of. We divided delegate pairs into (a) pairs where both individuals are delegates; (b) pairs where both individuals are non-delegates; and (c) mixed pairs of delegates and non-delegates. Similarly, gender pairs were divided into same-sex (e.g., male-male, female-female) and mixed-sex

pairs. For topic groups, we re-ran our analyses on each of the 20 topic groups separately.

Overall, both leaders and gender pairs demonstrate the same divergent trends as the group as a whole. However, not all tests showed significant differences when comparing these pairs.

For instance, Spearman's Correlation Coefficient found a significant difference in similarity between three groups, where $F_{(2,273)} = 6.804$, $p<.001$, $n=276$, such that delegate-delegate pairs demonstrate higher similarity scores than non-delegate pairs and mixed pairs. LSA found the same result, $F(2,280) = 11.122$, $p<.001$ $n=283$. By contrast, Zipping did not find this to be the case, where $F_{(2,226)} = 2.568$, $p=.079$, $n=229$.

In terms of the potential effect of gender on similarity scores, Zipping showed a significant difference between the three groups, $F_{(2,236)} = 3.546$, $p<.05$, $n=239$, such that female-female pairs and mixed-sex pairs demonstrate more similarity than male-male pairs. LSA found the same relationship, $F_{(2,280)} = 4.79$, $p<.005$ $n=283$. By contrast, Spearman's Correlation Coefficient does not show a significant between-groups difference, $F_{(2,273)} = .699$, $p=.498$, $n=276$.

In terms of differences among the topic groups, we did indeed find differences such that some topic groups demonstrated the fairly linear slope with decreasingly similarity shown above, while others demonstrated dips and rises resulting in a level of similarity at T6 quite similar to T1. There is no neat way to statistically measure the differences in these slopes, but it does indicate that future analyses need to take topic group into account.

In sum, we did not find leadership or gender to mediate language similarity in this community. Topic group, on the other hand, did play a role, however no topic groups showed increasing *similarity* across time.

### 5.2 Similarity and Temporal Proximity

Our second hypothesis concerned the gradual change of language over time such that temporal proximity of time periods would correlate with mean similarity. In other words, we expect that messages in close time periods (e.g., adjacent weeks) should be more similar than messages from more distant time periods. In order to examine this, we performed two individual tests, in which our predictions can be described as follows: (a) the similarity between texts in one time period and texts in the neighboring time period is greater than texts in one time period, and texts that came two periods previously, $S(M_k, M_{k-1}) > S(M_k, M_{k-2})$; and (b) the similarity between texts in one time period and texts in the neighboring time period is greater than the similarity between texts in one time period, and texts in the very first time period, $S(M_k, M_{k-1}) > S(M_k, M_1)$.

As shown in Table 1, SCC and Zipping tests confirm these hypotheses, while none of the LSA tests revealed significant differences.

Table 1. Temporal Proximity Similarities SCC, Zipping, and LSA, $n=20$ topic groups

| | $S(M_k,M_{k-1}) > S(M_k,M_{k-2})$ | $S(M_k,M_{k-1}) > S(M_k,M_1)$ | $S(M_k,M_{k-2}) > S(M_k,M_1)$ |
|---|---|---|---|
| SCC | .665 > .653† | .665 > .639° | .653 > .639° |
| ZIP | .628 > .608† | .628 > .605† | .608 > .605§ |
| LSA | 9.74 > .971 | 9.74 > .971 | .97166 < .97168 |

Note: *$p<.05$, °$p<.01$, †$p<.001$, §$p = .0525$, one-tailed

## 6 Discussion

This work presents several novel contributions to the analysis of text-based messages in online communities. Using three separate tools, Spearman's Correlation Coefficient, Zipping and Latent Semantic Analysis measures, we found that across time, members of an online community diverge in the language they use. More specifically, a comparison of the words contributed by any pair of users in a particular topic group shows increasing *dissimilarity* over the six-week period.

This finding seems counter-intuitive given work in linguistics and psychology, which shows that dyads and communities converge, entrain and echo each other's lexical choices and communication styles. Similarly, our own temporal proximity results appear to indicate convergence, since closer time periods are more similar than more distant ones. Finally, previous hand-coding of these data revealed convergence, for example between boys and girls on the use of emotion words, between older and younger children on talk about the future (Cassell & Tversky, 2005). So we ask, why do our tools demonstrate this divergent trend?

We believe that one answer comes from the fact that, while the young people may be discussing a more restricted range of topics, they are contributing a wider variety of vocabulary. In order to examine whether indeed there were more unique

words over time, we first simply manually compared the frequency of words over time and found that, on the contrary, there are consistently fewer unique words by $T_6$, which suggests convergence. However, there are also fewer and fewer *total* words by the end of the forum. This is due to the number of participants who left the forum after they were not elected to go to Boston. If we divide the unique words by the total words, we find that the *ratio* of unique words consistently increases over time (see Figure 4). It is likely that this ratio contributes to our results of divergence.

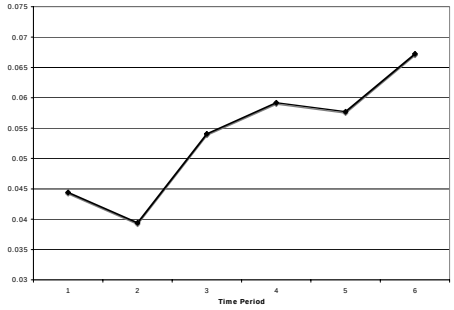


Figure 4. Ratio of Unique to Total Words, $T_1 - T_6$

In order to further examine the role of increasing vocabulary in the Junior Summit as a whole, we also created several control groups comprised of random pairs of users (i.e., users that had never written to each other), and measured their pairwise similarity across time. The results were similar to the experimental groups, demonstrating a slope with roughly the same shape. This argues for convergence and divergence being affected by something at a broader, community-level such as an increase in vocabulary.

This result is interesting for an additional reason. Some users – perhaps particularly non-native speakers or younger adolescents, may be learning new vocabulary from other speakers, which they begin to introduce at later time periods. An increasingly diversified vocabulary could conceivably result in differences in word frequency among speakers. This leads us to some key questions: to what extent does the language of individuals change over time? Is individual language influenced by the language of the community? This is heart of entrainment.

In conclusion, we have shown that SCC, Zipping and LSA can be used to assess message similarity over time, although they may be somewhat blunt instruments for our purposes. In addition, while Zipping is somewhat contentious and not as widely-accepted as SCC or LSA is, we found that the three tools provide very similar results. This is particularly interesting given that, while all three methods take into account word or word-sequence frequencies, LSA is designed to also take into account aspects of semantics beyond the surface level of lexical form.

All in all, these tools not only contribute to ways of measuring similarity across documents, but can be utilized in measuring smaller texts, such as online messages or emails. Most importantly, these tools remind us how complex and dynamic everyday language really is, and how much this complexity must be taken into account when building computational tools for the analysis of text and conversation.

### 6.1 Future Directions

In future work, we intend to find ways to compare the results obtained from different topic groups and also to examine differences among individual users, including re-running our analyses after removing outliers. We also hope to explore the interplay between individuals and the community and changes in language similarity. In other words, can we find those individuals who may be acquiring new vocabulary? Are there "language leaders" responsible for language change online?

We also plan to analyze words in terms of their local contexts, to see if this changes over time and how it impacts our results. Furthermore, we intend to go beyond word frequency to classify topic changes over time to get a better understanding of the dynamics of the groups (Kaufmann, 1999).

Finally, as we have done in the past with our analyses of this dataset, we would like to perform a percentage of hand-coded, human content analysis to check reliability of these statistical methods.

## Acknowledgements

Thanks to members of the Articulab, Stefan Kaufmann, Stefan Wuchty, Will Thompson, Debbie Zutty and Lauren Olson for invaluable input. This research was in part supported by a generous grant from the Kellogg Foundation.

## References

Auer, P., & Hinskens, F. (2005). The role of interpersonal accommodation in a theory of language change. In P. Auer, F. Hinskens & P. Kerswill

(Eds.), *Dialect change: The convergence and divergence of dialects in European languages* (pp. 335-357). Cambridge, MA: Cambridge University Press.

Baronchelli, A., Caglioti, E., & Loreto, V. (2005). Artificial sequences and complexity measures. *Journal of Statistical Mechanics: Theory and Experiment, P04002*, 1-26.

Baym, N. K. (2000). *Tune in, log on: Soaps, fandom, and online community*. New York: Sage Publications.

Benedetto, D., Caglioti, E., & Loreto, V. (2002). Language trees and zipping. *Physical Review Letters, 88*(4), 1-4.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482-1493.

Cassell, J., & Tversky, D. (2005). The language of online intercultural community formation. *Journal of Computer-Mediated Communication, 10*(2), Article 2.

Chambers, J. K. (2001). Dynamics of dialect convergence. *Journal of Sociolinguistics, 6*(1), 117-130.

Cherny, L. (1999). *Conversation and Community: Chat in a Virtual World*. Stanford: Center for the Study of Language and Information.

Coccaro, N., & Jurafsky, D. (1998, November 1998). *Towards better integration of semantic predictors in statistical language modeling.* Paper presented at the International Conference on Spoken Language Processing (ICSLP-98), Sidney, Australia.

Crystal, D. (2001). *Language and the Internet*. New York: Cambridge University Press.

Eckert, P. (2003). Language and adolescent peer groups. *Journal of Language and Social Psychology, 22*(1), 112-118.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998 ). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes, 25*, 285-307.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition, 27*, 181-218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*(1), 8-11.

Kalthchenko, A. (2004, May 2-5, 2004). *Algorithms for estimation of information distance with application to bioinformatics and linguistics.* Paper presented at the Canadian Conference on Electrical and Computer Engineering (CCECE 2004), Niagara Falls, Ontario, Canada.

Kaufmann, S. (1999). *Cohesion and collocation: Using context vectors in text segmentation.* Paper presented at the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6*(1), 97-133.

Labov, W. (2001). *Principles of linguistic change* (Vol. 2: Social Factors). Oxford: Blackwell Publishers.

Lakin, J. L., Jefferies, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior, 27*(3), 145-162.

Lam, W. S. E. (2004). Second language socialization in a bilingual chat room: Global and local considerations. *Language Learning & Technology, 8*(3), 44-65.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Milroy, L. (1980). *Language and social networks*. Oxford: Blackwell Publishers.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*(4), 337-360.

Paolillo, J. (2001). Language variation on internet relay chat: A social network approach. *Journal of Sociolinguistics, 5*(2), 180-213.

Schutze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann Publishers.

Serafin, R., & Di Eugenio, B. (2004, July 21-26, 2004). *FLSA: Extending latent semantic analysis with features for dialogue act classification.* Paper presented at the 42nd Annual Meeting for the Association of Computational Linguistics (ACL04), Barcelona, Spain.

Street, R. L., & Giles, H. (1982). Speech accommodation theory. In M. E. Roloff & C. R. Berger (Eds.), *Social cognition and communication* (pp. 193-226). London: Sage Publications.

Wiemer-Hastings, P., & Graesser, A. C. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments, 8*(2), 149–169.