Construction of the Gmane corpus for examining the diffusion of lexical innovations

Kyle Marek-Spartz
University of Minnesota
Minneapolis, Minnesota, USA
mare0132@umn.edu

Paula Chesley
University of Alberta
Edmonton, Alberta, Canada
pchesley@ualberta.ca

Hannah Sande
University of Minnesota
Minneapolis, Minnesota, USA
sande570@umn.edu

ABSTRACT

Large-scale linguistic corpora, complete with information about speakers' social networks as well as demographic and temporal information, allow for empirical validation of complex theories about the social interactions and linguistic properties leading to large-scale language change. We present ongoing work on the diffusion of lexical innovations using a corpus we have compiled from the Gmane electronic mailing list archive, a publicly available dataset of 13,494 mailing lists and 117,606,370 messages to date. Focusing initially on a single list, we derive a social network for actor-speakers, give lexical and network statistics, and empirically categorize tie strength across speakers. Initial explorations of the Gmane corpus suggest suitability for research on language change.

Author Keywords

Social network, SNA, weak ties, diffusion of innovations, language change, linguistic corpora

ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors

General Terms

Languages, human factors, algorithms, experimentation

INTRODUCTION

In *The Stuff of Thought*, Steven Pinker uses trends in baby names as a stepping point to discuss large-scale language composition and change. He concludes that such phenomena are inherently unpredictable: "The naming of babies, and of things in general, is another example in which a large-scale social phenomenon – the composition of a language – emerges unpredictably out of many individual choices that impinge upon one another" [22, p. 322]. Yet research across various disciplines – linguistics [20, 21, 7], sociology [12, 24], and physics [26] – suggests that we can in fact predict aspects of these complex phenomena using the concept of WEAK TIES within speakers' social networks. Crucially though, the extent to which diffusion of linguistic phenomena is similar to diffusion of non-linguistic phenomena is unclear, and the effect of weak tie interactions is a challenge for current sociolinguistic paradigms (see [20, p. 363-373] and [21, p. 1-9]).

A tie between two actor-speakers (henceforth SPEAKERS) in a social network indicates exchange or sharing of resources, social support, or information [13]. Factors affecting tie strength include frequency of contact, duration of the association, level of intimacy, and kinship [13]; tie strength is multivariate. Strength is generally broken down into three types of interpersonal ties: absent, weak, and strong. According to Granovetter [12], absent ties may consist of non-existent, quasi-negligible, or non-social affiliations, such as semi-regular "Hello/good-bye" small talk with the same bus driver. Yet some researchers see such ties as weak: for example, Milroy and Milroy [20, 372-373] hypothesize that new phonetic variants were diffused by employees of a Belfast store in the interface between Protestants and Catholics, by virtue of their weak tie relationships with store clientele. Social or professional acquaintances are generally considered weak ties, as they do not pass the typical tests of strong-tiedom, such as lending money. Intuitively, strong ties constitute close, bi-directional relationships between speakers and typically consist of relationships between family members and close friends.

Since the likelihood that strong ties will have mutual acquaintances is greater than chance, strong ties are associated with INTERLOCKING personal networks, in which a set of individuals all interact with each other [12]. By their redundancy, interlocking networks are not optimal for the diffusion of information [24, 5]. Weak ties serve as bridges between strong-tie personal networks and thus are essential for the wider diffusion of information. Some sociolinguistic research (e.g. [20, 18, 21]) notes that weak ties – although not necessarily stated as such in the case of [18] – may have contributed to the diffusion of new phonetic variants. Yet the Milroy studies [20, 21] argue for the importance of weak ties from a more theoretical perspective, as with their data these authors are not able to show persuasively the effects of weak ties in the diffusion of linguistic innovations.

Most of the sociological research on weak ties has been done on the diffusion of non-linguistic information, such as new

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.

methods of weed control [23] or learning about new job opportunities [12]. While linguistic theory may benefit from such research, the extent to which the diffusion of linguistic innovations echoes that of non-linguistic innovations is unclear. First, according to the levels-of-processing effect [10], we would expect information about a new job opportunity to be processed deeper than a particular phonetic variant, or perhaps a particular lexical choice, used by one's interlocutor. This deeper processing would result in a stronger memory trace. The relatively weak memory trace for linguistic innovations may mean more repetition is necessary for speakers to adopt them. This repetition may come from the redundancy of interlocking personal networks. However, in sociological studies, RADIAL personal networks, where individuals are linked to one focal individual but not to each other, favor diffusion more than the interlocking networks [24, p. 338]. In short, interlocking personal networks may be more beneficial for the diffusion of less consciously perceived phenomena such as linguistic innovations.

Second, as opposed to the adoption of non-linguistic innovations such as a new weed spray, adoption of linguistic innovations is not a binary phenomenon. The change could follow a pattern of LEXICAL DIFFUSION (see e.g. [6]), whereby language change happens at differing rates according to lexical properties such as frequency. Furthermore, during a change, the same lexical items are possibly in free variation between the old and new forms. Thus for the diffusion of linguistic innovations, individual change is likely more protracted than for the diffusion of non-linguistic innovations. These issues imply that models for the adoption of linguistic innovations are more nuanced than models for non-linguistic innovations, and that linguistic data must be empirically examined when assessing the effects of tie strength in social networks.

In light of these issues, this paper represents our first steps toward bridging a gap between sociolinguistic and sociological research as these fields relate to language change. We detail ongoing work on the empirical observation of the diffusion of lexical innovations within the Gmane corpus, a corpus of 117,606,370[1] time-stamped, user-identifiable messages from 2001-2012. We see a tripartite approach as necessary to the study of language change and social networks: information must be obtained about the speakers, about the linguistic innovations themselves, and about the speakers' communities or networks. Focusing first on network properties, the present work discusses empirical categorization of tie strength in a social network derived from the Gmane corpus. We show the distribution of dyadic ties for one list in the Gmane corpus, the Corpora List, and categorize strong and weak ties according to the empirical distribution of tie strength of speaker dyads. We target the weak ties in the network as a next step in the examination of the diffusion of lexical innovations, although it is an open question as to whether tie strength in our network corresponds to tie strength of more traditional social networks. First, we focus on the details of constructing the Gmane corpus.

[1]Figures about the corpus date from May 2012; total Gmane data is generally increasing by over 30,000 messages per day.

METHOD

Corpus construction

How does language change propagate through a social network? Adequately answering this question requires a corpus with the following properties:

1. Repeated interactive linguistic output from speakers (this implies a longitudinal component);
2. The ability to create or exploit existing network structure from speaker interactions;
3. Demographic information about the speakers;
4. A free and public dataset for ease of replication of results;
5. High-quantity, diverse data.

Currently, relatively few linguistic corpora have properties 1-3 above (although see [3] for phonetic data). Concerning (5), linguistic innovations are expected to be inherently rare, and to study their diffusion patterns we need to see several occurrences of them. Suppose that 50-100 longitudinal occurrences suffice to properly study diffusion patterns. This is an estimate based on the work of [8], in which one new lexical borrowing occurs for approximately every 1,000 words in a French newspaper corpus. Assuming similar rates as well as a lower bound on the number of occurrences necessary for studying diffusion patterns, we would need approximately 50,000 words for every linguistic innovation to be studied. That is, if we wanted to study 1,000 lexical innovations, a longitudinal corpus of approximately 50,000,000 words would already be necessary. Furthermore, it is unclear how many speakers suffice to create a sufficient social network for studying large-scale language change, but a larger number of speakers should lead to greater understanding of such phenomena.

With these matters in mind, we considered using the following corpora: the NUS SMS (texting) Corpus [14], Twitter data (see [11]), the Enron corpus ([17], [19]), Usenet ([1], [2]), and Google Groups. Of particular interest was examination of diverse network structures, as exemplified by both one-to-one and one-to-many communication models. It is likely that one-to-one communication, typically private between two speakers such as in instant message exchange, is different in both form and function than one-to-many communication, which is often public and "broadcast-like" in nature [4]. Unfortunately, the NUS SMS corpus lacks demographic information and only explores the one-to-one communication model. Twitter data provide good demographic information, such as geographic location [11], but only consist of one-to-many interactions. Furthermore, the length imperatives in Twitter and text data (140 characters per tweet/text) no doubt lead to shortened lexical innovations, but these shortenings are perhaps forced and might not recur outside of the particular tweet/text.

Next, the Enron email dataset was an attractive option. The unrestricted length of emails could lead to more naturally produced language, which in turn could lead to more diverse lexical innovations. Yet with only 200,399 messages and

158 speakers, it was unclear how much headway we could make in examining the diffusion of lexical innovations with this corpus; furthermore, the opportunities for observing diverse social networks were limited.

As opposed to the previous one-to-one corpora which were private and not one-to-many, Usenet newsgroups provide a hybrid of the one-to-one and the one-to-many communication style: after an initial group post, a response can be a specific reply to the initial poster, in which case the group is effectively cc'd. These newsgroups are structured according to content, offering a community-oriented mode of communication and social network in which the topic of discussion entails a shared interest between the group members. Importantly, users can be members of multiple groups. Usenet servers typically provide an interface via the Network News Transfer Protocol (NNTP). NNTP allows users to check for new messages, and download messages from groups. This makes for a nearly ideal interface for gathering a corpus. Unfortunately, limitations on Usenet content quality would likely affect results obtained from the corpus. Users typically pay for Usenet access, so the Usenet sample could be too biased toward higher levels of socio-economic status. Additionally, today much of Usenet consists of binaries and filesharing. Finally, different Usenet servers each have different subsets of Usenet content, which limits the reproducibility of results. Google Groups archives much of Usenet as well as mailing lists (together, there are nearly 10 million as of May 2012) in a publicly available web interface and does not include many of the binaries groups, so use of this data potentially resolves the content limitations of Usenet. However, it has collection limitations: there is no programmatic way to export Google Groups data comparable to Usenet's NNTP.

Consideration of Usenet and Google Groups led to Gmane.org, which archives electronic mailing lists and provides free and public access. This access is available in multiple forms, including email, web, RSS, and most importantly, NNTP. Unlike Google Groups, it does not archive any Usenet newsgroups, but it is similar to Usenet's newsgroups with its community-oriented structure, organization, and preferred mode of communication. Gmane's 13,494+ lists span a variety of topics, most heavily dominated by technology-related fields; such fields are conducive to lexical creativity [2]. These lists also span a variety of structures: some lists are unidirectional, read-only and/or are used primarily for announcements and thus have few speakers. Other lists have bidirectional communications with hundreds of speakers. As such, the communities created by such lists are fairly diverse.

As a representative example of a Gmane mailing list, our readers may be familiar with the Corpora List[2]. The Gmane archives of this mailing list span from November 2003 to the present, although the list itself dates from 1995. This mailing list is professional in nature; users are spread worldwide and skew toward being highly educated. Researchers working on empirical approaches to natural language, including natural language processing, use the list for multiple purposes: to post queries about linguistic resources; to post information about jobs, conferences, or other events; or to discuss more theoretical or philosophical concerns. In our data, there are 3,508 senders, considered as speakers, in this list, with approximately 5 messages/day, 4.6 participants/day, and about 3 new topics ("threads") per day. Relationships between subscribers can span from insignificant, never having interacted before, to relatively close colleagues or collaborators, as shown in the following addendum to a threaded reply concerning desired characteristics of a corpus: "P.S. Hi R—, it's good to see you on the list and I hope all is well. I hope to see you again in Seoul." Several messages in this list make reference to the community served by this list, e.g. "I was just trying to get something done, and when it took P— some time to reply, I thought of asking the corpora community for a solution."

The messages in the Gmane corpus are formatted as emails, according to the relevant standards (RFC 822, 2822, etc.). They are typically plain-text with readily parseable headers, encoding information such as the topic, sender email, referenced messages, and time and date sent. The message content is structured according to user input and often consists of a salutation, a message, one or more sections of replied-to content, and a signature. As replied-to content is not written by the message's sender, we filter such content by removing text after a right-angle bracket > at the beginning of lines. Ideally, signatures would also be discarded, since these are artificially repeated, and do not typically propagate from user to user. These are harder to parse out efficiently, although we are actively exploring such directions.

The remaining lexical content is tokenized and typified ignoring stop words, or common artificially repeated words. For each remaining string type in a message, the sum of its tokens in the message is noted in a relational database. We then calculate the number of types and tokens in messages, or number of tokens given a type, or most importantly, the distribution of words over time, by referencing the message time.

Construction of the social network

After construction of the corpus, a social network was then created following the criteria below:

1. Each unique email address was considered a unique speaker, or vertex in the network.
2. The DIRECT REPLIES amongst speakers were considered as ties, or undirected edges.

While we cannot say for certain whether each address actually corresponds to a unique speaker – e.g. if multiple users access the same administrative account email – it is reasonable to believe this assumption holds true in the majority of cases. Ties were calculated as follows: for each pair of speakers A and B, the weight of the undirected edge, or tie, between the two is the number of direct replies from either A to B or B to A. For example, if B responds to A, a tie is created. If A responds in turn, the strength of that

[2]Available at http://dir.gmane.org/gmane.science.linguistics.corpora.

tie is increased. Measuring ties through direct replies is a simple method that, crucially for language change, ensures a respondent has read the initial message. We then ran a lowess smoother [9] over the empirical distribution of direct replies, enabling visual inspection of tie strength. Importantly, in the entire Gmane corpus, a speaker is connected to other speakers in any mailing list he or she posts to, which allows for large-scale network analysis across thousands of mailing lists.

RESULTS

As a test case, we chose to work with the Corpora List, described above, for its extended longitudinal data, its diversity of message types, and its familiarity. Basic statistics for this list are given in Table 1; as is seen, messages on average contain 501 tokens.

To date, we have not evaluated the accuracy for our process of filtering tokens, and some over- or under-filtering may be occurring. Another issue concerns thread construction: speakers can request for their message to not be archived using a special message header. Using a naive threading algorithm, the absence of these messages can result in two threads where there should be only one. In our data there were 10 messages that were discarded by Gmane, either because of this header or because they were spam. Future work will account for these missing messages by implementing a more robust threading algorithm.

In this network, a clear distinction emerges between core and outlier component interactions. This trend was so pronounced in the interactions with tie strength $= 1$ that the network structure of the core component was not visually apparent. Opting for a visualization of the core network, Figure 1 shows that this core/outlier trend is still prevalent when considering only ties with strength ≥ 2, as well as showing the internal structure of the core component. The data presented below however concern interactions from all speakers.

	Count	
Speakers	3,508	
Threads	9,306	
Messages	15,635	
Types	135,597	
Tokens	7,841,792	
	Mean	Median
Tokens/message	501.521	258
Types/message	183.962	137
Novel types/message	8.672	1
Tokens/type	57.831	1
Messages/thread	1.680	1

Table 1. Basic statistics for the Corpora List.

The number of direct replies between unique speaker dyads follows a roughly exponential distribution and is given in Figure 2. Few dyads have 10 or more direct replies; the majority of speaker dyads have fewer than 10. Intuitively, these latter interactions represent weak (or absent) ties. In contrast, the long tail of the distribution can be seen to represent strong ties. The lowess-smoothed distribution makes a visual distinction between tie strength of fewer than 10 on one hand and 10 and over on the other, corresponding to this strong-tie/weak-tie bifurcation. However, these results need to be validated across additional mailing lists to make a statistical distinction between strong and weak ties, and to ensure that the category of absent but quasi-negligible ties should not be included as a tie category.

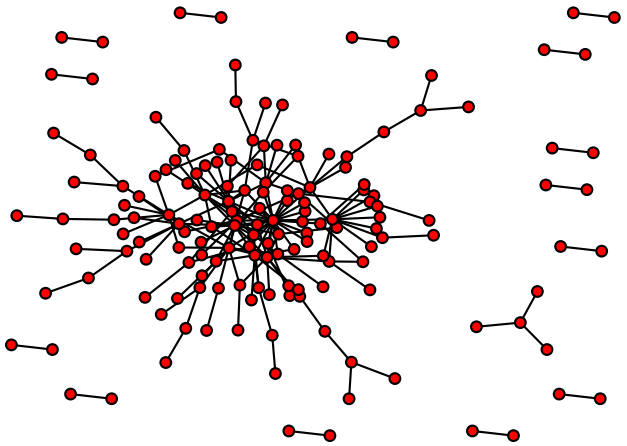

Figure 1. The social network of the Corpora mailing list in the Gmane corpus; each node represents a speaker. Ties between speakers indicate that more than two direct replies have taken place between the dyad.

In qualitative examination of our data we note a distinction between the language speakers use for strong and for weak ties. Examples (1) and (2) below come from the same speaker, but (1) is a response to a weak tie, while (2) responds to a strong tie. The former includes a salutation with a greeting and a closing, and the request is indirectly formulated, with a question mark as punctuation. In contrast, there is no greeting nor closing in the strong-tie reply, and the comments are quite direct.

(1) Hi A—,

That's an interesting comment...

I'd be interested to see your exact results?

Best regards, R—

(2) M—,

You seem to have missed my point in this discussion. Which was...:

...

Your reply is a case in point. You simply ignored [my previous point].

Further analyses are needed to ensure that such differences are statistically robust, and to see if politeness strategies relate to the diffusion of lexical innovations.

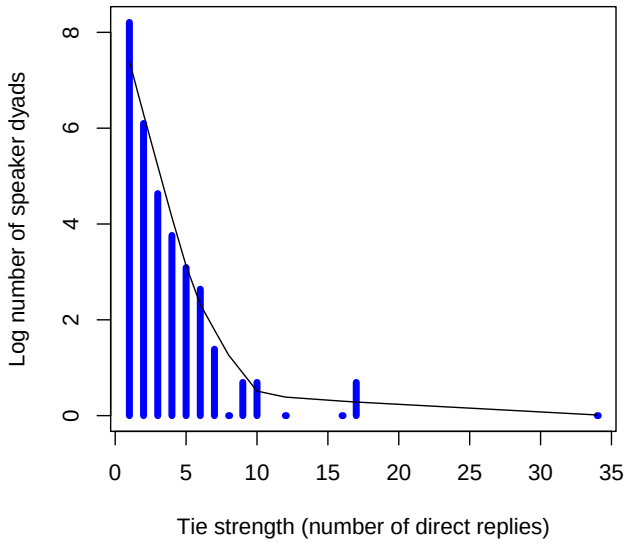

Figure 2. Histogram of tie strengths in the Corpora List. Tie strength is determined by number of direct replies. The black line represents the lowess-smoothed empirical distribution.

DISCUSSION AND FUTURE WORK

This paper documents the use of the Gmane corpus for ongoing work examining the diffusion of linguistic innovations, particularly with respect to network structure. We hypothesize that, similar to the diffusion of non-linguistic innovations, a key component in the diffusion of lexical innovations is the strength of weak ties [12]. To test this hypothesis, and to dig deeper into large-scale language change, we have developed the Gmane corpus, a large-scale corpus of mailing lists dating from 2001-2012. As a first step in examining the role weak ties play in the diffusion of lexical innovations, we used the number of direct responses to previous messages amongst speakers in the Corpora List as a measure of tie strength. Examination of the lowess-smoothed empirical distribution of ties between dyads shows two broad types of ties, strong and weak, in this mailing list. However, we need to examine tie strength across multiple lists to obtain a statistically robust two-way distinction of tie categories.

We are not the first to focus on the importance of weak ties in language change; previous research [20, 21, 18] has mentioned such a concept. However, as Milroy and Milroy note, it is difficult to observe the importance of weak ties within the quantitative sociolinguistic variationist paradigm, or even by neighborhood/ethnographic studies, as weak ties can often consist of brief interactions outside one's neighborhood or work. For this reason, a corpus study, with a corpus annotated for social network information, appears to be an ideal method of inquiry. At first glance, the Gmane corpus seems an appropriate corpus; however, it must be emphasized that, as an internet corpus, correspondence between the diffusion of linguistic innovations in the Gmane corpus and the diffusion of linguistic innovations in the physical world is unclear. We now have a distinction between various types of ties with which to lead our inquiry of the diffusion of lexical innovations; different ties may need to be examined if, for example, what consists of a strong tie in the Gmane corpus turns out to function more like a weak tie in offline networks. Questions of speaker centrality to the network, or rate of endogenous/exogenous posting rates, will also be taken into account in future work.

As an example of related research that could make use of the Gmane corpus, the diffusion of lexical innovations in the corpus could offer the possibility of testing the hypothesis of lexical diffusion (see e.g. [6]), and of further understanding the role of speakers' social networks in lexical diffusion. Similar to phonetic change, morphosyntactic change may affect different lexical items or lexical bundles at different times. Since our corpus is text-based, it is an excellent resource for examining the role of lexical diffusion in morphosyntactic innovation (see [25] on the unclear role of lexical diffusion in morphosyntactic change). Lexical diffusion is also difficult to observe with "brick-and-mortar" sociolinguistic studies, because lexical diffusion is posited as a gradual change [6]. We hypothesize that in cases such as lexical borrowings, lexical diffusion throughout a community of speakers may fall out from the diffusion of lexical innovations throughout a social network, again with pivotal importance placed on weak ties.

The far-reaching question we aim to answer in looking at the effects of social network constraints on the diffusion of lexical innovations is, "How does language change spread throughout a community of speakers?" This question can be broken down into questions about the speakers (e.g., who is doing the innovating and driving the spreading?), about the innovations themselves (e.g., what is the time course of a lexical innovation?), and about the community or network (e.g., what is the structure of networks that are propitious to lexical innovations?). Concerning the speakers, we are soliciting demographic information such as sex, age, geographic location, and community type (urban/suburban/rural). Depending on response rate, we may also infer this information using a program such as JGAAP [16]; we can automatically predict both the sex and education level of a speaker in using authorship attribution techniques [15]. As for the lexical innovations, we can discover if, when a new word is uttered, there are predictable patterns for where and when we will see it again, and how frequent it will be. Does a general pattern across time emerge for all new words? In [8], there is a distinctly bimodal pattern for lexical innovations approximately 10 years later – perhaps a similar pattern emerges with continuous time data.

Finally, the methods and findings focusing on network structure that we present here, particularly our work on weak ties, will be scaled up to the entire Gmane corpus, and other ways to determine network structure will be explored. Additional tie measures such as indirect replies or list co-membership could be used to perhaps improve our methods for modeling tie strength. Intuitively, these constitute weaker bonds than

direct replies, and if in future work these criteria would be incorporated into creating ties, different weights could be assigned for each criterion. Additional network structures beyond weak ties could be considered, such as structural holes [5].

The collection of future data from Gmane is in store. We are taking steps to make the Gmane corpus publicly available, as we believe it has the potential to add to the growing body of linguistic corpora containing social network structure amongst speakers. Ultimately, we feel that results from many linguistic corpora containing social network information, each with diverse types of linguistic output, conceptualizations of network structure, and relationships amongst speakers, will lead to greater understanding – and hence predictability – of the large-scale social phenomena of language change.

ACKNOWLEDGEMENTS

We wish to thank D. Kyle Danielson, Brian Reese, Edward Stronge, and Morten Warncke-Wang for conceptual aide and/or reading a previous version of this paper. Lars Magne Ingebrigtsen has been instrumental in data compilation. KMS was supported with a University of Minnesota Undergraduate Research Opportunity Program (UROP) grant.

REFERENCES

1. Altmann, E., Pierrehumbert, J., and Motter, A. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS 1 4*, 11 (2009).
2. Altmann, E., Pierrehumbert, J., and Motter, A. Niche as a determinant of word fate in online groups. *PLoS 1 6*, 5 (2011).
3. Bane, M., Graff, P., and Sonderegger, M. Longitudinal phonetic variation in a closed system. In *Papers from the 46th regional meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago, 2010/in press.
4. Baron, N. *Always On: Language in an Online and Mobile World*. Oxford University Press, Oxford, 2008.
5. Burt, R. *Structural holes: The social structure of competition*. Harvard University Press, 1995.
6. Bybee, J. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change 14* (2002), 261–290.
7. Chesley, P. You know what it is: Learning words through listening to hiphop. *PLoS ONE 6*, 12 (2011).
8. Chesley, P., and Baayen, R. H. Predicting new words from newer words: Lexical borrowings in French. *Linguistics 48*, 6 (2010), 1343–1374.
9. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74* (1979), 829–836.
10. Craik, F., and Lockhart, R. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior 11*, 6 (1972), 671–684.
11. Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. A latent variable model for geographic lexical variation. In *EMNLP*, ACL (2010), 1277–1287.
12. Granovetter, M. The Strength of Weak Ties. *American Journal of Sociology 78*, 6 (1973), 1360–1380.
13. Haythornwaite, C. Strong, Weak, and Latent Ties and the Impact of New Media. *The Information Society 18* (2002), 385–401.
14. How, Y., and Kan, M. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of Human Computer Interfaces International (HCII)*, vol. 5 (2005).
15. Juola, P. *Authorship Attribution*. Now Publishers Inc., Boston, 2008.
16. Juola, P., Sofko, J., and Brennan, P. A prototype for Authorship Attribution Studies. *Literary and Linguistic Computing 21* (2006), 169–178.
17. Klimt, B., and Yang, Y. The enron corpus: A new dataset for email classification research. *Machine Learning: ECML 2004* (2004), 217–226.
18. Labov, W. *Principles of Linguistic Change: Social Factors*. Wiley-Blackwell, Malden, MA, 2001.
19. McCallum, A., Wang, X., and Corrada-Emmanuel, A. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research 30* (2007), 249–272.
20. Milroy, J., and Milroy, L. Linguistic Change, Social Network and Speaker Innovation. *Journal of Linguistics* (1985), 339–384.
21. Milroy, L., and Milroy, J. Social Network and Social Class: Toward an Integrated Sociolinguistic Model. *Language in Society* (1992), 1–26.
22. Pinker, S. *The Stuff of Thought: Language as a Window into Human Nature*. Viking Adult, New York, 2007.
23. Rogers, E. *Social change in rural society: A textbook in rural sociology*. Appleton-Century-Crofts, New York, 1960.
24. Rogers, E. *Diffusion of Innovations*. Free Press, New York, 2003.
25. Tottie, G. Lexical diffusion in syntactic change: frequency as a determinant of linguistic conservatism in the development of negation in English. In *Historical English Syntax*, D. Kastovsky, Ed. Mouton de Gruyter, Berlin, 1991, 439–468.
26. Watts, D. J. *Six Degrees: The Science of a Connected Age*. Vintage Books, New York, 2004.