Niche as a determinant of word fate in online groups

Eduardo G. Altmann\textsuperscript{1,2,3}, Janet B. Pierrehumbert\textsuperscript{1,4} \& Adilson E. Motter\textsuperscript{1,5}

\textsuperscript{1}Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA
\textsuperscript{2}Instituto de Física, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, Brazil
\textsuperscript{3}Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany
\textsuperscript{4}Department of Linguistics, Northwestern University, Evanston, IL 60208, USA
\textsuperscript{5}Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA

Patterns of word use both reflect and influence a myriad of human activities and interactions. Like other entities that reproduce and evolve, words rise or decline depending upon a complex interplay between fitness and environment. Using Internet discussion communities as model systems, we show that the word niche, defined as the extent of the word’s association with specific people and topics, is a strong determinant of changes in word frequency. Previous, a posteriori, studies have indicated that word frequency is a correlate of word success at historical time scales. Our longitudinal analysis reveals that the word niche is far more important than word frequency in the dynamics of the entire vocabulary at shorter time scales, as the language adapts to new concepts and social groupings. We also identify endogenous versus exogenous value as a further component in the fate of words, and demonstrate the force of this distinction in the rise of novel words. Our results indicate that short-term nonstationarity in word statistics is strongly driven by individual proclivities, including inclinations to provide novel information and to project a distinctive social identity.

Much information about the fabric of modern human society has been gleaned from large-scale records of human communications activity, such as time stamps and network structures for email exchanges, mobile phone calls, and Internet activity [1–4]. But the flow of words has the potential to be even more informative. Words characterize both external events and otherwise unobservable mental states. They tap into the variety of experience, knowledge, and goals of different interacting individuals. The word stream is information-dense, because the number of distinct words and expressions is so great. The lexicon of a literate adult is estimated to contain over 100,000 distinct items [5], and it continues to grow as new words are encountered [6].

Records of the linguistic transactions within a community provide an ongoing statistical sampling of the vocabulary of a language. The sample at any time reflects both the social context (who is speaking, and to whom) and the topical context (what they are speaking about). But the language dynamics does not just passively mirror the context. Language adapts to new circumstances and needs through lexical innovation [7]. Large data sets available from the Internet provide an unprecedented opportunity to study the dynamics of words, as well as phrases and tags [8–10]. Here, we explore lexical fluctuations in relation to both individuals and topics by analyzing records of Usenet groups. Created over one decade before the World Wide Web, the Usenet groups were amongst the first systems for worldwide exchange of messages on the Internet. Usenet archives reveal the rise of “Netspeak”, the language nowadays widely used on the Internet and in telephone text messages [11]. The groups we studied, rec.music.hip-hop and comp.os.linux.misc, were selected for their
great lexical creativity. In these datasets, users serve as proxies for individuals, and threads as proxies for topics (see Methods). Our study goes beyond the analysis of user activity in Usenet groups [12], and focuses instead on the content of the messages.

It is known that word frequency is a factor in frequency dynamics on historical time scales [13, 14], a finding that is expected from models of language learning across human generations [15]. We identify and quantify a new factor – the dissemination of words across individuals (users) and topics (threads) – and demonstrate that it is a much more powerful determinant of word fate than word frequency is at shorter time scales. We find that poorly disseminated words are more likely to experience a frequency reduction than widely disseminated words. The results support comparisons between the dynamics of the linguistic system and other social dynamics, such as the spread of opinions or the popularity of news items, videos, and music [16, 17]. They also suggest analogies between word niches and ecological niches of species in population biology [18–20]. These analogies are strengthened by a case study of novel words with rising frequency, in which we compare a set of words for products and public figures, whose use is strongly influenced by exogenous factors, to a set of slang words, whose use more reflects factors endogenous to the sociolinguistic system. The force of this distinction in word dynamics mirrors its force in other behaviours [21, 22]. Finally, we explore the correlations between individuals and topics as dimensions of word dissemination. The two dimensions are shown to be separable, and individual choices prove to be more important than topic in determining patterns of word usage. These results highlight the importance of individuality in use of language, and suggest limits on the role of social influence and conformity.

Dissemination of words across users and threads

If everyone knew the same words, and chose to use them at random with their given frequencies, the dissemination of words across users would be the result of a Poisson process. We are interested in the extent to which the actual number of users of each specific word deviates from this baseline model. We define the measure of dissemination of each word $w$ across users as

$$D^U_w = \frac{U_w}{\tilde{U}(N_w)}$$

where $N_w$ is the number of occurrences of the word in the dataset, $U_w$ is the actual number of users whose posts include word $w$ at least once, and $\tilde{U}$ is the expected number of users predicted by the baseline model. The latter is determined from $\tilde{U} = \sum_{i=1}^{N_U} \tilde{U}_i$, where $N_U$ is the number of users and $\tilde{U}_i$ is the probability that user $i$ used $w$ at least once when all the words in the text are shuffled randomly (see Methods). Dissemination across threads is analogously defined as

$$D^T_w = \frac{T_w}{\tilde{T}(N_w)}$$

where $T_w$ is the number of threads in which the word appears, and $\tilde{T}$ is the corresponding expected value from the baseline model. In the rest of this paper, we will focus on the properties of the measures $D^U_w$ and $D^T_w$, or $D^U$ and $D^T$ for notational simplicity.

The expected value of $D^U_w$ is 1 for a word of any frequency that is distributed randomly across users. $D^U > 1$ indicates over-disseminated words and $D^U < 1$ indicates concentrated
or clumped words. Most generally $\frac{1}{N_w} \leq D_U \leq \min\{N_w, N_U\}/\bar{U}$. Due to the discreteness close to the lower bound, we set a threshold $N_w > 5$ for the computation of $D_{U,T}$. The few dozen most frequent words (mainly common function words) are also omitted from our analysis, because $D_U$ is not informative when $N_w$ is too large compared to the number of users. Figure 1 shows results on the expected statistical fluctuation around $D_U = 1$ for randomly distributed words in a representative window of each Usenet group, as determined by a Monte Carlo simulation. The upper and lower extremes of the fluctuation depend on frequency, but only slightly.

The dissemination with respect to threads $D_T$ is closely related to the residual inverse document frequency ($r$-IDF), a measure used in text processing to characterize the extent to which a word is associated with particular documents. $\text{IDF}$, defined as the reciprocal of the number of documents in which the word occurs, is strongly influenced by word frequency. Residual IDF addresses this artifact by taking the difference $r$-IDF $= \log(\tilde{T}) - \log(T)$, where $\tilde{T}$ is approximated using a Poisson baseline model with equal document lengths. When this condition holds, $-\log(D_T) = r$-IDF. The measure $D_T$ is a generalization of $r$-IDF that remains valid when the lengths of the documents are very unequal, as for the present dataset.

$D_U$ and $D_T$ as predictors of word fate

To explore the changes over time in the statistical attributes of words, we begin by partitioning each dataset into non-overlapping half-year windows. Figure 1 displays the behaviour of $D_U$ within a representative half-year window for both groups. Most words are significantly clumped. At all word frequencies, the median $D_U$ falls below the 10th percentile for random fluctuation of the expected value under the baseline model. For words with $\log_{10} f < -3.5$, $D_U$ varies considerably and is not correlated with $f$. Words with $\log_{10} f > -3.5$ are extremely high-frequency words, and comprise less than 0.5% of all distinct words in this window. But even these words are somewhat clumped. These findings are reproduced in all half-year windows for both Usenet groups, as summarized in SI, Fig. S1. They provide the user counterpart to prior observations of clustering of words in documents and in time [8, 23–25].

We now examine $D_U$ as a predictor of frequency change for words over two-year periods. We first note that $D_U$ is strongly related to the likelihood that a word with $N_w > 5$ at a window $t_1$ falls below this threshold in a window $t_2$ taken two years later. This is illustrated for both Usenet groups in Fig. 2ad. The finding is so statistically robust that it is reproduced for every choice $t_1$ and $t_2 = t_1 + 2$ years, in both groups. The same pattern is also mirrored in the frequency changes of words that are above the $N_w > 5$ threshold at both $t_1$ and $t_2$. Within this group of words in the selected window of comp.os.linux.misc, $D_U$ is a strong predictor of whether the word rose or fell in frequency (Fig. 2b), and likewise in the selected window of rec.music.hip-hop (Fig. 2e). The consistency of this pattern over all windows may be seen by comparing $\Delta \log_{10} f$ for words with $D_U = 0.4$ and with $D_U = 1.0$, values that span the well-populated portion of the range in $D_U$. Words with the former value tend to decline in frequency ($\Delta \log_{10} f$ is negative), while for words with the latter value, $\Delta \log_{10} f$ is near zero or positive. There is no $t_1, t_2$ pair for either dataset in which the effect is reversed (Fig. 2cf).

This far, our analysis has focused on $D_U$. In sociolinguistic parlance, we have considered
the “indexicality” of words, that is the extent to which words are associated with individuals or types of people. Now, let us also consider $D^T$, our measure of “topicality” (dissemination over topics). The results just described for $D^U$ also hold for $D^T$ (see SI, Figs. S1 and S2). What is the relative importance of these factors in predicting $\Delta \log_{10} f$? As Table 1 shows, $D^U$ is more important than $D^T$. Both are more important than $\log_{10} f$, whose importance is comparatively slight (see also SI, Fig. S3).

Words change over time not just in their frequency, but also in their dissemination. A signal aspect of changes in $D^U,T$ is a strong negative correlation with $\Delta \log_{10} f$. For comp.os.linux.misc, the correlations of $\Delta \log_{10} f$ with $\Delta D^U$ and $\Delta D^T$ are $-0.54$ and $-0.40$, respectively; for rec.music.hip-hop, $-0.55$ and $-0.39$, respectively. These negative correlations arise because words that rise in frequency, without a concomitant increase in the number of users and/or topics, thereby become less disseminated. Low $D^U,T$ puts them at risk of declining in frequency thereafter. Just as a population that explodes in a narrow ecological niche may well crash later, it appears that repetitive communications from the same few people are more discounted than emulated by others. Overall, fluctuations in $f$ driven by variability in user behaviour and topic dominate the statistical behaviour in our dataset, with the result that patterns similar to those in Fig. 2 are observed by relating $D^U$ at $t = t_2$ to $-\Delta \log_{10} f$. These large, short-term fluctuations add an important new dimension to the study of the long-term dynamics of language, as any novel expression must survive in the short term to survive in the long term.

**Case study: Rising slang and product words**

A new word, like a new species, must establish itself in a niche to survive in the language. The survival rate of lexical innovations is not known, but any successful innovation must have overcome fluctuations in $f$ that risked driving it to an early extinction. We now present a case study of successful innovations. For each group, we selected two sets of novel words. Each word was not used during the first years of the group, and was consistently used for at least some years thereafter (see SI, Sec. S2). The first set is designated as P-words because they refer to products (such as Gnome, a desktop environment introduced in 1998) and public figures (such as Eminem, a rapper popular from the late 1990’s). Exogenous factors contribute strongly to their use. The S-words exemplify slang words and other novel vernacular language. They were selected with the aid of on-line dictionaries of Internet and Usenet terms (see SI, Sec. S2). We consider the dynamics of these words to be more dominated by factors endogenous to the social and linguistic systems. Paired lists were frequency matched to the extent possible.

Figure 3 compares the dynamics of the P-words and S-words. Typical words in both groups have vacillating frequencies over time (Fig. 3ab). In Fig. 3cd, we look at their behaviour in a frequency-$D^U$ space. As indicated by the horizontal boxplots, the P-words and S-words are located in the frequency region below $\log_{10} f = -3.5$, in which the frequency is not correlated with $D^U$. Trajectories over time for the example words are superimposed, beginning when the word first crosses $N_w > 5$. In contrast to the example S-words, the example P-words begin with very low $D^U$ values, and rise greatly in frequency before becoming widely disseminated. The vertical boxplots show that P-words have overall lower $D^U$ than S-words, over all words and times (though both fall below the median of all words). The contrast in $D^U$ is replicated for the early rising period of each word (see Fig. 3d).
Significant clumping in \( D^U \) is expected for S-words, because choices of vernacular language such as *lol* (*laughing out loud*) and *prolly* (*probably*) reflect the individual’s construction of social identity [27, 28]. Our finding that P-words are even more clumped in \( D^U \) reflects the fact that the on-line behaviour of individuals is shaped both by their interests [29] and by the social groupings that grow from sharing interests [30].

The low \( D^U \) of the S-words and P-words would tend to predict a decline in frequency (see above), but instead the frequencies of these particular words rose. For P-words, the rise is exogenously driven by the intrinsic fitness of the word’s referent. The concept of fitness can also be connected to \( D^T \) and to the behaviour of S-words, because fitness can involve complex interactions of social and non-social factors; recent studies that shed light on these interactions include work on music downloads [17] and on the success of YouTube videos and stories on the news portal Digg [31, 32]. In fact, all of our principal observations about the relationship of \( D^U \) to P- and S-words are also true for \( D^T \). This suggests that each word occupies a niche in a human conceptual space that is defined simultaneously by people and by ideas, because it is defined by people who share ideas. The measures \( D^U \) and \( D^T \) quantify how dispersed the word niche is in the space of individuals and topics. Do these two dimensions reduce to just one underlying dimension through a high correlation of people and ideas? Or are the dimensions separable, even if related through complex interactions? We take up these questions rigorously in the next section.

### Factoring the relative contributions of individuals and topics

We have shown that most words, including both highly indexical words such as slang words and highly topical words such as products, are significantly concentrated in both \( D^U \) and \( D^T \). We have sketched some reasons for these dimensions to be positively correlated. How can we rigorously evaluate their separability and relative importance? To address this issue, we consider new measures that effectively factor indexicality and topicality as contributors to \( D^{U,T} \), and we standardize the datasets to eliminate distributional artifacts.

We first introduce \( \hat{D}^U \) as a modification of \( D^U \) in which \( \hat{U} \) in Eq. (1) is calculated from a baseline model that shuffles the words only within threads in the time window, rather than across all users and all threads. Analogously, we introduce \( \hat{D}^T \) as a modification of \( D^T \) in which \( \hat{T} \) in Eq. (2) is calculated from a baseline model that shuffles the words only within posts of the same user. These new quantities provide a direct measure of the extent to which individuals and topics contribute to the concentration of words observed above. While \( D^U \) reveals whether the word is clumped or over-disseminated by comparing the actual dissemination with that obtained by “erasing” all the structure, \( \hat{D}^U \) maintains the structure of the threads and considers randomization of words across users within them. If \( \hat{D}^U \) is significantly closer to 1 than \( D^U \) is, then topics must strongly influence the individuals’ choice of words. Analogously, the role of individuals can be confirmed by comparing the extent to which \( \hat{D}^T \) is closer to 1 than \( D^T \) is.

To ensure that users and threads serve as comparable proxies of individuals and topics, we randomly trim the dataset to eliminate the differences in their distributions. For each window, the trimming scheme standardizes the user contribution per thread and the size of all posts, matches the number of users and threads, and approximately matches the distribution of posts per user and per thread. The trimmed comp.os.linux.misc (rec.music.hip-hop) dataset remains large enough for our statistical analysis, with an average of 4,593 (1,503)
posts and 2,383 (585) users and threads per half-year window, and an overall average of 77.6 (51.2) words per post.

The exact distributions of values of $D^U$ and $D^T$ change with the trimming. Trimming generally increases $D^U$ and $D^T$ for the words that survive, but the trends and all conclusions from previous sections still stand. For example, the overall median $D^U$ changes from 0.71 to 0.87, and the overall median $D^T$ changes from 0.73 to 0.89, for the comp.os.linux.misc group. The relative differences in both groups remain essentially unchanged, which means that the measures $D^{U,T}$ provide meaningful comparisons even when the distributions are not streamlined. However, the trimmed set offers the advantage of providing exact and non-artifactual information about the correlations between the measures.

Table 2 displays the important correlations amongst the original and modified measures. The correlation between $D^U$ and $D^T$ is positive, confirming the expectation that indexicality and topicality are related. But it is far less than 1, suggesting that $D^U$ and $D^T$ contribute substantially different information. The measures $D^U$ and $D^T$, as well as $D^{T}$ and $D^{U}$ are positively correlated, as expected because these are related measures by definition. Finally, the negative correlation between $D^{U}$ and $D^{T}$ is a confirmation that these quantities partially factor $D^U$ and $D^T$ and hence provide the information they are designed to provide.

We now use the trimmed dataset and modified measures to further test the relative importance of indexicality and topicality. As shown in Fig. 4ac, $\hat{D}^U$ and $\hat{D}^T$ are statistically larger than $D^U$ and $D^T$, but they remain smaller than 1. This confirms that most words are clumped with respect to both users and threads. Overall, $D^U$ is smaller than $D^T$, indicating that words are generally more concentrated with respect to users than to threads. This observation is rigorously confirmed by the fact that $\hat{D}^U$ is smaller than $\hat{D}^T$ to the same extent. Both in the aggregate (Fig. 4ac), and (statistically) for individual words (Fig. 4bd), the effect of threads on users is smaller than the effect of users on threads.

The most striking effect shown in Fig. 4ac is the large number of words with small $D^U$ in comparison to $D^T$. After trimming, over all windows, the comp.os.linux.misc (rec.music.hiphop) dataset has 5,335 (1,808) words with $D^U < 0.4$, versus 1,657 (337) words with $D^T < 0.4$. The list of words with $D^U < 0.4$ but $D^T > 0.4$ includes both very common words and highly topical words. In comp.os.linux.misc, example words include *imagination*, *coffee*, *angst-ridden*, and *saukrates* (a rapper); in rec.music.hiphop, examples include *regards*, *baptized* and *tauri* (a Hungarian Warcraft server). It is interesting that such words are even more distinctive to individuals than to topics. A contributing factor to this clumpiness is the use of formulaic expressions. Such expressions, which are found in signature blocks as well as in other conventionalized communications such as greetings and insults, often have quite idiosyncratic lexical choices.

Altogether, we have strong evidence that the lexical make-up of the threads is strongly determined by the individual users. This speaks against the possibility that the topic dictates the vocabulary, and equally against the possibility that mutual imitation causes strong convergence in lexical choices as people interact in the discussion. This is a striking result. It contrasts with the major thrust of research on modeling the evolution of lexical systems, which is to explain convergence in the community [33, 34]. This suggests that individuals may be more autonomous in their choices of words than in a wide range of other behaviours, from yawning and gait [35] to complex conscious decisions like the decision to purchase a product or to vote [36]. Given that individuals use different words to talk about the same topic, that word concentration over users is more extreme than over threads, and that $D^U$ is the strongest predictor of $\Delta \log_{10} f$, the heterogeneity of people emerges as the single
strongest factor in lexical diversity, both at any particular time and over time.

**Discussion and conclusions**

We have introduced two new quantities, $D^U$ and $D^T$, as measures of the dissemination of words with respect to individuals and topics, and used them to characterize the vocabulary of two online discussion groups over a period of more than a decade. We found that almost all words are concentrated with respect to both individuals and topics, and that at short-term (two-year) time scales, concentration in a $D^{U,T}$ niche is a strong determinant of word fate. $D^U$ and $D^T$ are separable components, and both trump word frequency. However, $D^U$ trumps $D^T$.

Word frequencies over time reflect a replicator dynamic, that is, a dynamic in which the words are reproduced by being copied through imitation [33, 34, 37]. The imitation process includes both learning and use. Word learning is facilitated by variety in the context of use [38], and rates of word use are in turn subject to great vacillations over time, as a reflex of shifting user behaviour and shifting topics. For a lexical innovation to survive in the language, it must avoid an absorbing boundary near $f = 0$, at which it is used so rarely that no one can learn it. Our investigation of the relationship between $\Delta \log_{10} f$ and $\Delta D^{U,T}$ shows that a key to success is increasing $D^{U,T}$ hand-in-hand with increasing $f$. The success of the P-words in our case study can be understood by considering that exogenous fitness assisted their rise. The similar rise of the S-words suggests that their indexical value acts in part as a type of fitness.

Word frequency affects word fate at historical time scales when different forms compete to express the same meaning [13, 14]. Why did frequency not prove to be important in the dynamics of the whole vocabulary, as studied here? The language system has strong functional pressures for words to be distinct from each other, in both form and meaning [33, 34, 37, 39]. Assuming that almost all words are learned with unique meanings, and that replication has low error rates, it follows that most words do not have a direct competitor for the same meaning. This picture presents strong parallels to the exclusion principle in evolutionary biology, which states that occupying distinct niches protects species from competition [18]. Distinct languages are similarly predicted to survive only if they are spoken by distinct populations [40]. Diversity therefore depends on the diversity and viability of the individual niches. For biological species, the size of the geographical range and the species duration are correlated [19, 20]. In studies of the lexicon, the individual words assume the role of species, and we have shown that the size of the word niche is associated with stability in word frequency.

We found that $D^U$ and $D^T$ are positively correlated, but still provide distinct information. A positive correlation is expected because individuals have characteristic interests. Further mechanisms contributing towards this correlation result from the participation of individuals in social and geographical structures. These in turn can cause clumping in topics, as shown, for example, by profiling of the Internet for software products. Structures in the social network can even contribute directly to product adoption, because the usefulness of many products (such as high-tech innovations) can depend on the number of neighbors who already use the product [16, 41]. These same mechanisms pertain to other words, insofar as concepts and opinions resemble products.

We suggest that other mechanisms limit the correlation between $D^U$ and $D^T$, and explain
the striking degree to which individuals were found to use different words in discussing the same topic. The variety in human social identities is thought to provide an impetus for innovation in modes of expression [27, 28, 42]. Clusters within a connected social network can hinder lexical convergence [42, 43]. The fundamental principles of discourse call for one to strike a balance between anchoring contributions in what the listener already knows, and providing novel and relevant information [44]. Online discourse can be viewed as a collective exploration of the conceptual world [45]. The most engaging and fruitful discourse is discourse in which people cooperate in differentiating themselves and what they say.

Methods

Data set. Usenet group archives are available at [http://groups.google.com](http://groups.google.com). The smallest unit of text is the post. Each post is attributed to a user and belongs to a thread (as defined by an initial post and all replies to it). We focus on two Usenet groups from their first post through 2008-03-31: (i) comp.os.linux.misc, which concerns Linux operating systems, includes 128,903 users and 140,517 threads beginning 1993-08-12; (ii) rec.music.hiphop, which is devoted to hip-hop music, has 37,379 users and 94,074 threads beginning 1995-02-08. The activity of users in Usenet groups is bursty [25] and heterogeneous [12]. In the comp.os.linux.misc group, for example, the average user contributes 5.4 posts and remains active for 249.3 days, but the most persistent users have more than 1,000 posts over more than 10 years. The average thread has 4.9 posts and is active for 4.5 days, but the longest threads have more than 1,000 posts over 3 years. See SI, Sec. S1 for information about preprocessing of the text.

Baseline model. The expected number of users $\tilde{U}$ in Eq. (1) is calculated by assuming that all words are randomly shuffled, while holding constant the number of users and the number of words per user. Let $N_w$ be the number of occurrences of the word $w$, $m_i$ be the total number of words contributed by user $i$, and $N_A \equiv \sum_i m_i = \sum_w N_w$. The probability that the $j + 1$ th occurrence of $w$ does not belong to user $i$ is given by $(1 - m_i/(N_A - j))$. The probability $\tilde{U}_i$ that user $i$ used word $w$ at least once is calculated as the complement of the probability of not using it:

$$\tilde{U}_i = 1 - \prod_{j=0}^{N_A - 1} \left(1 - \frac{m_i}{N_A - j}\right) \approx 1 - e^{-f_w m_i},$$

where the approximation is valid for $m_i/N_A \ll 1$ and $f_w \equiv N_w/N_A \ll 1$. This corresponds to a Poissonian baseline model with a fixed probability of using $w$ given by the observed word frequency $f_w$. The error in the approximation is smaller than 0.1% for the datasets we consider. This approximation was used in all calculations involving the untrimmed datasets, while the exact relation was used for the trimmed datasets. An analogous procedure is used for the calculation of the expected number of threads $\tilde{T}$.

Acknowledgments

E.G.A. was supported by the Northwestern Institute on Complex Systems and a Max Planck Society Otto Hahn Fellowship, J.B.P. by JSMF Grant No. 21002061, and A.E.M. by
NSF Grant No. DMS-0709212 and a Sloan Research Fellowship.

[1] Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks. Proc. Natl Acad. Sci. USA 104, 7332–7336 (2007).
[2] González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. Nature 453, 779–782 (2008).
[3] Malmgren, R. D., Stouffer, D. B., Motter, A. E. & Amaral, L. A. N. Poissonian explanation for heavy tails in e-mail communication. Proc. Natl Acad. Sci. USA 105, 18153–18158 (2008).
[4] Seshadri, M. et al. Mobile call graphs: Beyond power-Law and lognormal distributions. Proc. 14th ACM SIGKDD Intl Conf. Knowl. Disc. Data Min., 596–604 (2008).
[5] Kuiper, K. Knowledge of language and phrasal vocabulary acquisition. Behav. Brain Sci. 29, 291–292 (2006).
[6] Davis M. H. & Gaskell, M. G. A complementary systems account of word learning: Neural and behavioural evidence. Phil. Trans. R. Soc. B 364, 3773–3800 (2009).
[7] Muman, J. Lexical Creativity, Texts and Contexts (John Benjamins, Amsterdum, 2007).
[8] Kleinberg, J. Bursty and hierarchical structure in streams. Data Min. Knowl. Dis. 7, 373–397 (2003).
[9] Cattuto, C., Loreto, V., Pietronero, L. Semiotic dynamics and collaborative tagging. Proc. Natl Acad. Sci. USA 104, 1461–1464 (2007).
[10] Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. Information retrieval in folksonomies: Search and ranking, Lec. Notes Comput. Sc. 4011, 411–426 (2006).
[11] Crystal, D. Language and the Internet (Cambridge Univ. Press, Cambridge, UK, 2006).
[12] Fisher, D., Smith, M. A. & Welser, H. T. You are who you talk to: Detecting roles in Usenet Newsgroups. Proc. 39th Annual Hawaii Intl Conf. Syst. Sci. 3, 59 (2006).
[13] Pagel M., Atkinson A. & Meade A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449, 717–720 (2007).
[14] Lieberman, E., Michel, J. B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. Nature 449, 713–716 (2007).
[15] Fontanari, J. F. & Perlovsky, L. I. Solvable mill model for the distribution of word frequencies. Phys. Rev. E 70 (4), 042901 (2004).
[16] Watts, D. & Dodds, P. Influentials, networks, and public opinion formation, J. Consum. Res. 34, 441–458 (2007).
[17] Salganik, M. L., Dodds, P. S. & Watts D. J. Experimental study of inequality and unpredictability in an artificial cultural market. Science 311, 854–856 (2005).
[18] Hardin, G. The competitive exclusion principle. Science 131, 1292–1297 (1960).
[19] Foote, M., Crampton, J. S., Beu, A. G. & Cooper, R. A. On the bidirectional relationship between geographic range and taxonomic duration. Paleobiology 34, 421–433 (2008).
[20] Jablonski, D. Mass extinctions and macroevolution. Paleobiology 31, 192–210 (2005).
[21] Sornette, D. Endogenous versus Exogenous Origins of Crises, in Extreme Events in Nature and Society, Albeverio, A., Jentsch, V. & Kantz, H. (Eds.) (Springer, Berlin, 2005), pp. 95–119.
[22] Crane, R. & Sornette, D. Robust dynamic classes revealed by measuring the response function of a social system. Proc. Natl Acad. Sci. USA 105, 15649–15653 (2008).
[23] Church, K. W. & Gale, W. A. Poisson mixtures. Nat. Lang. Eng. 1, 163–190 (1995).
[24] Serrano, M. A., Flammini, A. & Menczer, F. Modeling statistical properties of written text. PLoS ONE 4 (4), e5372 (2009).
[25] Altmann, E. G., Pierrehumbert, J. B. & Motter, A. E. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS ONE 4 (11), e7678 (2009).
[26] Kruskal, W. Relative importance by averaging over orders. Am. Stat. 41, 6–10 (1987).
[27] Milroy, L. Language and Social Networks (Blackwell, Oxford, UK, 1980).
[28] Eckert, P. Linguistic Variation as Social Practice (Blackwell, Oxford, UK, 2000).
[29] Fortunato, S., Flammini, A., Menczer, F. & Vespignani, A. Topical interests and the mitigation of search engine bias. Proc. Natl Acad. Sci. USA 103, 12684–12689 (2006).
[30] Schifanella, R., Barrat, A., Cattuto, C., Markines, B. & Menczer, F. Folks in folksonomies: Social link prediction from shared metadata. Proc. 3rd ACM Intl Conf. Web Search Data Min. 10271–10280 (2010).
[31] Lerman, K. & Hogg, T. Using a model of social dynamics to predict popularity of news. Proc. 19th Intl Conf. WWW, 621–630 (2010).
[32] Szabo, G. & Huberman, B. A. Predicting the popularity of online content. Commun. ACM (to appear, 2010).
[33] Steels, L. The synthetic modeling of language origins. Evol. Commun. 1, 1–34 (1997).
[34] Komarova, N. L. & Nowak, M. The evolutionary dynamics of the lexical matrix. Bull. Math. Biol. 63, 451–484 (2001).
[35] Dijksterhuis, A. & Bargh, J. A. The perception-behavior expressway: Automatic effects of social perception on social behavior. Adv. Exp. Soc. Psychol. 33, 1–40 (2001).
[36] Nickerson, D. W. Is voting contagious? Evidence from two field experiments. Am. Polit. Sci. Rev. 102, 49–57 (2008).
[37] Kirby, S., Cornish, H. & Smith, K. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proc. Natl Acad. Sci. USA 105, 10681–10686 (2008).
[38] Blythe, R.A., Smith, K. & Smith, A. D. M. Learning times for large lexicons through cross-situational learning. Cogn. Sci. 34, 620–642 (2010).
[39] Baronchelli, A., Gong, T., Puglisi, A. & Loreto, V. Modeling the emergence of universality in color naming patterns. Proc. Natl Acad. Sci. USA 107, 2403–2407 (2010).
[40] Abrams, D. M. & Strogatz, S. H. Modeling the dynamics of language death. Nature 24, 900 (2003).
[41] Rogers, E. M. Diffusion of Innovations (Free Press, New York, NY, 2003).
[42] Labov, W. Principles of Linguistic Change: Social Factors (Blackwell, Oxford, UK, 2001).
[43] Lu, Q., Korniss, G. & Szymanski, B. K. The Naming Game in social networks: community formation and consensus engineering. J. Econ. Interact. Coord. 4, 221–235 (2009).
[44] Grice, H. P. Logic and Conversation, in Studies in the Way of Words (Harvard Univ. Press, Cambridge, MA, 1989), pp. 1–138.
[45] Cattuto, C., Barrat, A., Baldassarri, A. & Schehr, G. Collective dynamics of social annotation. Proc. Natl Acad. Sci. USA 106, 10511–10515 (2009).
Fig. 1: Relationship of frequency $f$ to $D^U$. a,b, The results are shown for half-year windows centered on 1998-01-01 for the comp.os.linux.misc group (a) and the rec.music.hip-hop group (b). Red solid line: running median for all words with $N_w > 5$. Red dashed lines: 10th and 90th percentiles for the same words. Blue dashed lines: 10th and 90th percentiles around the expected value of $D^U$ for randomly distributed words, determined by Monte Carlo simulations with 100 independent shufflings of the text. Black line: analytically calculated ceiling $D_{max}^U = N_w/\bar{U}$ (floor effects and the other ceiling, $D_{max}^U = N_U/\bar{U}$, do not pertain within the scale of the figure). The median empirical $D^U$ is systematically below the 10th percentile of the estimated random variation. The relationship of median $D^U$ to $f$ is nearly flat up to $\log_{10} f = -3.5$.

Table 1: Importance of user, thread, and frequency in word dynamics. Importance is calculated as the fraction of the variance $R^2$ in $\Delta \log_{10} f$ explained by each factor (using the methodology of Ref. [26]), and is combined over all window pairs $t_1, t_2 = t_1 + 2$ considered in Fig. 2. The range of words is restricted in $f$ to avoid artifactual correlations for small and large $f$ (see SI, Fig. S3).

| Group          | $D^U$ | $D^T$ | $\log_{10} f$ | All three simultaneously |
|---------------|-------|-------|---------------|--------------------------|
| comp.os.linux.misc | 9.9%  | 3.8%  | 0.2%          | 15.2%                    |
| rec.music.hip-hop | 21.2% | 5.1%  | 0.4%          | 29.2%                    |
**Fig. 2:** $D^U$ as a predictor of falling below threshold and of frequency decay. The analysis is performed over half-year window pairs $t_1$ and $t_2$ separated by two years for the comp.os.linux.misc and rec.music.hip-hop groups. 

- **a, d.** Fraction of words with $N_w > 5$ in $t_1$ that fall to $N_w \leq 5$ in $t_2$. Histogram in gray: results from selected window pairs centered on $t_1 = 1998-01-01$ and $t_2 = 2000-01-01$. Red line: average over different non-overlapping window pairs with $t_1$ ranging from 2006-01-01 through the (rounded off) beginning of the group at $t_i$, and $t_2 = t_1 + 2$ years. The probability of falling below threshold goes down as $D^U$ increases.

- **b, e.** Scatter plots of all words with $N_w > 5$ in both windows (12,883 words for comp.os.linux.misc, 12,237 words for rec.music.hip-hop). Values on y-axis: log-frequency change $\Delta \log_{10} f \equiv \log_{10} f(t_2) - \log_{10} f(t_1)$. Red lines: running median, 10th percentile, and 90th percentile. Words with rising frequency appear above and words with falling frequency appear below $\Delta \log_{10} f = 0$. Examples of words with large frequency changes are highlighted. The probability of frequency decay is greater for words with low $D^U$.

- **c, f.** Summary of the dominant pattern in (b,e) over all non-overlapping windows with $t_1$ ranging from $t_i$ to 2006 and $t_2 = t_1 + 2$. Median values of $\Delta \log_{10} f$ at $D^U = 0.4$ and $D^U = 1$ are shown for each pair of windows.

Table 2: Correlations between dissemination measures. Correlations are represented by their averages ± standard deviations calculated over non-overlapping half-year windows. All words with $N_w > 5$ were considered within each window of the trimmed datasets.

| Group                | $(\hat{D}^U, D^U)$ | $(\hat{D}^T, D^T)$ | $(D^U, D^T)$ | $(\hat{D}^U, \hat{D}^T)$ |
|----------------------|---------------------|---------------------|-------------|-------------------------|
| comp.os.linux.misc   | 0.82 ± 0.07         | 0.67 ± 0.04         | 0.54 ± 0.12 | −0.30 ± 0.01            |
| rec.music.hip-hop    | 0.94 ± 0.02         | 0.83 ± 0.10         | 0.44 ± 0.09 | −0.23 ± 0.11            |
Fig. 3: Dynamical behaviour of words in the comp.os.linux.misc and rec.music.hip-hop groups. a,b, Number of occurrences of example P- and S-words together with the total number $N_A$ of all words in each half-year window centered at $t$. Example words: P-word *Gnome*, a software product; S-word *lol* ("laughing out loud"); P-word *Eminem*, a rapper; S-word *iirc* (“if I recall correctly”). Curves are normalized by the maximum number of occurrences per window reached over all windows: $5.2 \times 10^6$ for all words; $1,360$ for *Gnome*, and $115$ for *lol* (a); $4.3 \times 10^6$ for all words; $2,510$ for *Eminem*, and $56$ for *iirc* (b). c,d, Relationship of $D_U$ to frequency. Blue and black curves: evolution of example P-words and S-words over time. Red line: median word behaviour as in SI, Fig. S1. Boxplots: distribution of the mean frequency $f$ (solid, horizontal), mean dissemination $D_U$ (solid, vertical), and mean dissemination $D_U$ in the rising period (open, vertical) for all P- and S-words. The mean is calculated over all words with $N_w > 5$ within the corresponding window.
Fig. 4: Summary statistics of the dissemination measures in the comp.os.linux.misc and rec.music.hip-hop groups. a,c, The box and whisker plots indicate the median, the quartiles, and the octiles for the collection of all non-overlapping windows of the trimmed datasets. b,d, Corresponding statistics for $\hat{D}^{U,T} - D^{U,T}$ estimated from individual words. The statistics includes all words with $N_w > 5$ within the corresponding windows, with occurrences in different windows being counted independently.

**Supplementary Information**

S1. Filtering of data

In order to avoid spam present in the database, we eliminate all threads composed by a single post unless the author posted at least 5 times in threads with more than one post. Inside each post we consider only new inputs to the text, i.e., we omit parts of the text quoted from previous posts. Signature blocks, texts systematically placed at the end of posts by some users, were not removed because their content is deliberately chosen by the user. The use of signature blocks in Usenet is analogous to the use of formulaic expressions in other linguistic genres (e.g., greetings, farewells, sales transactions) that legitimately affect token frequency for constituent words.

Words are taken to be strings of characters separated from other strings by white space. In addition to space, tab, and newline, the character underscore (_) as well as all punctuation marks (.,!,?,:;,,,) are treated as white spaces. However, apostrophe (‘) and hyphen (-) are not. This means that web and email addresses are broken up into their component parts, whereas expressions such as weren’t and e-mail are treated as single words. Lines starting with [http:] are eliminated beforehand. Capitalization is removed, so that instances of the same word in sentence-initial and sentence-medial position are tabulated together. Strings consisting entirely or partly of non-alphanumeric characters other than $\$ and @ (e.g., #,%,&,*) are removed. No further lemmatization of purely alphabetic strings was imposed, with the result that all related words (e.g., singular and plural) are treated as distinct.

S2. Selection of target words

We are interested in words that first became popular during the lifetimes of the groups. Target words are selected that have negligible levels of use during the first 2.5 years of
each group, and substantial use during the group’s heyday. As shown in Fig. 3ab (main
text), in more recent times we observe a clear reduction in the activity of the groups and
a deterioration in the informativeness of posting. To avoid the selection of words used
exclusively during this period, we require that at least 40% of target word uses occur prior
to the time when the activity on the group fell to under a quarter of its peak level. This
cutoff falls in 2005 for the rec.music.hip-hop group and 2007 for the comp.os.linux.misc
group.

All target words have over 80 distinct users. In addition, we avoid the inclusion of words
that are used predominately by single individuals, as we are interested in words that rose
in the community more generally. Therefore, the following heuristics are adopted: no more
than 20% of occurrences in a single month, no more than 40% of occurrences by a single
user, and no more than 80% of occurrences by five users.

P-words are identified on a case-by-case basis among the most frequent words satisfying
these criteria. S-words are identified with the help of dictionaries of Internet vocabulary and
are selected from words with more than 100 appearances over the life time of the database.
The following dictionaries of Usenet terms and Internet slang (words) were used to identify
words of interest (the Internet lists were retrieved on 2009-01-29): David Crystal’s list of
abbreviations, pp. 85-86 [S1]; the jargon file 433 [S2]; the Wiktionary appendix on English
Internet Slang [S3]; and the Internet Slang Dictionary [S4]; This leads to comparable counts
for P-word and S-words in both groups.

[S1] Crystal, D. *Language and the Internet* (Cambridge Univ. Press, Cambridge, UK, 2006).

[S2] Chester County InterLink, [http://www.ccil.org/jargon/](http://www.ccil.org/jargon/) Retrieved Jan. 29, 2009.

[S3] Wiktionary, [http://en.wiktionary.org/wiki/Appendix:Internet_slang](http://en.wiktionary.org/wiki/Appendix:Internet_slang) Retrieved Jan. 29, 2009.

[S4] Internet Slang Dictionary & Translator, [http://www.noslang.com/dictionary](http://www.noslang.com/dictionary) Retrieved Jan. 29, 2009.
Fig. S1: **Relationship of frequency** $f$ to $D^U$ and $D^T$. **a-d**, The results are shown for both the comp.os.linux.misc group (**a**, **c**) and the rec.music.hip-hop group (**b**, **d**). The running median shown in Fig. 1 (main text) is calculated in all half-year windows. The blue color code indicates densities in the range of $10^{-4}$ (white) to 1 (dark blue) obtained by combining all running medians, while the red line indicates the median of the resulting, combined distribution.

Fig. S2: **$D^T$ as a predictor of falling below threshold and of frequency decay.** This figure is the $D^T$-counterpart of Fig. 2 (main text).
Fig. S3: Frequency as a predictor of falling below threshold and of frequency decay. This figure is the $f$-counterpart of Fig. 2 (main text). The dashed green lines in (b,e) indicate the minimum possible $\Delta \log_{10} f$ for a given $\log_{10} f(t_1)$, due to the threshold $N_w > 5$ imposed at $t_2$. The analysis in Table 1 (main text) includes only the range $\log_{10} f_{\text{min}} < \log_{10} f < \log_{10} f_{\text{max}}$. The range is truncated at $\log_{10} f_{\text{max}} = -2.52$, because for words above this frequency, $N_w$ is so large compared to the number of users or threads that $D$ is not informative (see the main text). The range is truncated at $\log_{10} f_{\text{min}} = -4.61$ for comp.os.linux.misc ($\log_{10} f_{\text{min}} = -4.52$ for rec.music.hip-hop) because below these cutoffs the exclusion of words falling under the threshold (i.e., $N_w \leq 5$) introduces artifacts in the relationship to $\Delta \log_{10} f$ (c.f. the relationship of the dashed green lines to the 10th percentile line). Specifically, $f_{\text{min}}$ was chosen for each dataset so that the percentage of words falling below the threshold ($N_w \leq 5$) would be less than 5% of the words with $\log_{10} f_{\text{min}} < \log_{10} f < \log_{10} f_{\text{max}}$. 
