Psychology of Language and Communication 2012, Vol. 16, No. 2

VERSITA OPEN

DOI: 10.2478/v10057-012-0010-z

MICHAŁ B. PARADOWSKI [1], ŁUKASZ JONAK [1,2]
[1] University of Warsaw
[2] National Library of Poland, Warsaw

DIFFUSION OF LINGUISTIC INNOVATION AS SOCIAL COORDINATION

Linguistic creativity is a manifestation of communities' and cultures' innovativeness. The initial results will be presented of an empirical project analysing the character and speed of the social spread of winged words and neologisms in a microblogging site, using the tools of social network analysis applied to big-scale data. Investigating the diffusion of linguistic innovation requires an approach pooling competences from human, social, and computational sciences. Such a complex systems perspective can lead to a deeper understanding of how mutual relations and communication between Internet users impact the cultural evolution of language in time and space, and the shape and dynamics of the interactions themselves, delivering quantitative estimates on the expansion of linguistic expressions and allowing the prediction of future trends and their scale.

Key words: linguistic creativity, neologisms, tags, innovation diffusion, social networks, influence, social coordination

The greatest challenge today, not just in cell biology and ecology but in all of science, is the accurate and complete description of complex systems. Scientists have broken down many kinds of systems. They think they know most of the elements and forces.

The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles.

Edward Osborne Wilson (1998, p. 85)

Address for correspondence: Michał B. Paradowski, University of Warsaw, Institute of Applied Linguistics, Dobra 55, 00-312 Warsaw, Poland. E-mail: michal.paradowski@uw.edu.pl

Setting the scene

Why is an *octomom* ahead of a *n00b?* When and where did *bangster* originate? How/Why did the *seatmate of size* become notorious? How come *carrotmobs* against *vendrification* and your *cookprint* are less numerous than *tweetups* at *notspots*? What makes some of yesterday's isolated neologisms globally accepted, essential parts of tomorrow's vocabulary? Why does another, even though initially widespread, fade into oblivion?

Language is a system of symbols registering categorisation of the world and human creation (how else could we ever learn the difference between e.g. a cup and a mug?). It is also an instrument used to alter reality (cf. e.g. Castells, 2007, 2009), the newspeak of propaganda employed by totalitarian systems on the one hand, and the now much contended recommendations to use the "language of benefits" and "positive thinking" in theories of marketing and psychology on the other).

Defining a complex system

Complexity science (CS) is a highly transdisciplinary, loosely categorised field, utilising the concepts and tools of nonlinear dynamics, the laws of statistical mechanics and probability theory, and numerical simulations, with the goal of capturing emergence, self-organisation, and related phenomena observable in changeable systems. A CS perspective allows us to conceptualise, re-evaluate and explain erstwhile traditionally perceived common phenomena in a new dynamic framework. In CS the same questions – but also methods – arrive on the scene from different fields.

In the first place, a complex system must be distinguished from a complicated one. While no unanimously agreed definition exists of what constitutes a complex system, most researchers agree that in order to be classified thus, a system has to share a number of properties (Paradowski 2009). Normally, we are dealing with a large number of elements interacting via simple local rules. The system is dynamic, constantly evolving and unfolding over time, with emergent properties, and this emergent global behaviour of the system is not a simple product of the sum of the behaviours of its components ("more is different"). Other frequently encountered features are topological diversity and heterogeneity (and directedness[1]), alongside recurrence (feedback loops). Complex systems tend to be highly structured with strong self-organisation (without an orchestrator), resulting in resistance to damage, resilience to failure and high flexibility due to learning and natural adaptability to changing conditions (homeostasis). This at once means their sensitivity to initial conditions, and the fact that ostensibly negligible perturbations and rare events on the local level lead to significant changes in the behaviour of the whole system. Additional observable phenomena may include bifurcation and phase transitions, stability and multistability, hysteresis (with the consequence that the system is

[1] I.e., non-reciprocity of relations.

nondeterministic), scale-freeness – self-similarity, heavy-tailed distribution scaling like a power law, and entropy (many possible pathways of evolution). This means that such systems are typically impossible to solve/verify/predict analytically, hence the fundamental role played by numerical simulation.

Examples of complex systems are ubiquitous. In biology, ontogeny and phylogeny, the spot character of a cheetah, structure of the leaf, ant trails, termite mounds, flashing fireflies, chirping crickets, and predator-prey ecosystems are all underlain by complex systems. So are communities and social behaviours, from cooperation to conflict and riots. Complex are epidemics, nervous and immunological systems, neural networks, the climate and earthquakes, financial markets and price fluctuations, logistics and traffic jams. In the sphere of technology, the Internet, telecommunications infrastructure and power grids, and the World Wide Web instantiate complex networks. Last, but not least, the field where it all started – physics – has long been grappling with analyses and descriptions of non-equilibrium thermodynamics, crystals, boiling liquid, or the nuclear fission reaction.

In linguistics, within an individual, perceptual dynamics and categorisation in speech, the emergence of phonological templates, word and sentence processing, and language acquisition; across society, variations and typology, the rise of new grammatical constructions, semantic bleaching, language evolution in general, and the spread and competition of both individual expressions, and entire languages, are all inherently complex and/or dynamic systems (cf. e.g. Tabor & Tanenhaus, 2001; Van Geert, 2009; The "Five Graces Group", 2009; Winters et al., 2010). More than a hundred papers have already been published dealing with language simulations. However, many of the (especially incipient) *in silico* experiments carried out were grossly inadequate to the scenery of the 21st c. (Paradowski & Jonak, 2012, p. 27f.). The models:

- only allow for Euclidean relationships (whereas nowadays more and more of our linguistic input covers immense distances),
- are 'static' (while mobility is not exclusively a 20th or 21st-c. phenomenon, as evidenced by warriors, refugees, missionaries, or tradespeople),
- assume a limited, identical number of 'neighbours' for every agent,
- presuppose identical perception of a given individual's prestige by each of its neighbours, as well as
- invariant intensity of interactions between different agents,
- most fail to take into account multilingual agents,
- have no memory effect, and
- zero noise (while noise may be a mechanism for pattern change).

To address these limitations, rather than take a modelling outlook, we can tap into one repository of language data nearly perfectly suited to large-scale dynamic linguistic analyses – the Internet. After all, this medium stores data which is virtually unregulated, essentially uncensored, spontaneous, being immediately registered, interconnected, and amenable to relatively easy search and analyses with the use of statistical and concordancing tools.

Language on the Internet

Erstwhile research on language evolution and change focused on large time-scales, typically spanning at least several decades. Nowadays, observable changes are taking place much faster. According to the Global Language Monitor (2009) a new English word[2] is born roughly every 98 minutes (admittedly a rather problematic estimate). As a handful of popular recent expressions recall *alcopop, ambush marketing, Anthropocene, Ardi, bangster, birther, brown state, choice mom, clickjacking, content farm, culturomics, death panel, defriend, deleb, Dogbo, e-vampire, Facebook narcissism, flatform, freemium, fundoo, funemployment, glamping, green washing, grey vote, hacktivist, hurt locker, hyperlocal, intexticated, jai ho, jeggins, kinetic typography, mobama, n00b, octomom, pineberry, planking, quendy-trendy, recessionista, reverse graffiti, robocall, seatmate of size, sexting, slumdog, spot fixing, superinjunction, tab napping, teabagger, tombstoning, tramp stamp, tweetup, unfollow, vook, vuvuzela, wonderstar, yarn bombing,* or *zombie bank*… Particularly useful for multi-angle analyses of language phenomena are Web 2.0 services, with content (co)generated by the users, especially the ones which allow enriching analyses with information concerning the structure of the connections and interactions between the participating users.

The uptake of novel linguistic creations in the Internet has been commonly believed to reflect the focus of attention in contemporary public discourse (suffice it to recollect the dynamics and main themes of status updates on Twitter following the presidential elections in Iran, Michael Jackson's death, Vancouver Olympic Games, and the recent Oscars gala, last July's L.A. earthquake, the Jasmine Revolution – by some also called the "Internet Revolution" – in Tunisia, the developments in Libya, the 2011 Tōhoku earthquake and tsunami, bin Laden's death[3], or the coverage of the notorious 'Dancing Man' YouTube video in mainstream media; see e.g. Gladwell, 2010). However, even where the topics coincide, the proportions in the respective channels of information are divergently different (correlation between the ranking of the same news in mainstream media vs. on the Internet, including blogs and social networks, at a level of a mere .3; e.g. Paradowski, 2010 – just as television ratings cannot be used to predict online mentions; O'Dell, 2010), and not infrequently the top stories in the mainstream press are markedly different than those leading on social media platforms (e.g. PEJ, 2010). The emotive content of comments on different social platforms is also distinctly different (compare e.g. YouTube vs. Flickr; cf. Davies, 2007; Benson, 2010).

Our research project has set out to investigate how mutual relations and communication between Internet users impact the social diffusion of neological tags (semantic shortcuts) in the Polish microblogging site Blip.pl. Diffusion of linguistic innovation depends on the social structure. There is a tradition that can be traced

[2] Mostly open-class words.

[3] When following the announcement of the news Twitter was registering 4*103 related status updates per second.

Table 1. The microblogging site in numbers (at time of data dump)

Users	20k, over half logging on daily
Users in the giant component*	5.5k (density 0.003)
Relations	110k
Tags	38k
Tagged statuses	720k

* By the giant component we here mean the largest connected cluster of vertices in the graph, spanning the majority of the nodes.

back to Gabriel Tarde's ideas about the mechanisms of the spread of inventions in social systems, popularised by conceptualisations of Everett Rogers (1962/2003), that the existence of reputable, well-connected individuals is crucial to the success of diffusion. More recently this assumption has been challenged by an observation that a deciding factor for the final magnitude of the diffusion process is the system-wide distribution of network diffusion thresholds (Watts, 2007). In our study we adopted the latter point of view, assuming that the innovativeness of a social system and its readiness to adopt inventions is a function of the pattern of diffusion thresholds characterising its members. We set out to explore this theme by analysing the diffusion of tags in a microblogging system. We used two datasets. One represented the structure (and structural dynamics dynamics) of the network of communication relations between the users. Each record of this dataset described a directed relation of user B being followed by user A, with a timestamp indicating the creation time of the relation. The other dataset specified the usages of tags, with information on who and when used a given tag. Both datasets span 2 years, starting from the moment of the inception of the service in April 2007. An overview of the dataset's statistics is provided in Table 1.

Tags and social coordination

Blip.pl, not unlike the better-known Twitter, has been designed to allow users to communicate with their subscribers by means of short messages resembling mobile text messages. Blip messages can be delivered to the followers by a number of channels, the most important of which being the user's personalized web page ("dashboard" or "cockpit"). Each follower is "exposed" to the content of the messages that each person she has subscribed to is sending, and him/herself can send messages to his/her followers. In this way on a social level blip.pl constitutes a network of communication relations which allows for the spread of information.

Blip.pl provides a units of information ready to be analysed: "tags". Users can easily make any word of their message a tag, just by preceding it with the "hash"

sign ("#"). This tells the system that the message should be added to the category defined by the given tag. The tag is, then, meta-information about the content of the message. When used for the first time by a user who had not observed it around, it is his/her invention, can spread through the system along the follower – followed ties, and this diffusion process can be analysed.

The intended purpose of tagging systems introduced to various Web 2.0 services was to provide ways of building ad hoc, bottom-up, user-generated classifications (later labelled "folksonomies"; Vander Wal, 2007) of content produced or published within those systems. This was also the reason for introducing tagging into micro-blogging systems such as Twitter or Blip – it permitted adding metadata to users' unorganised status updates, and consequently their thematic classification.

However, the tagging system of Blip became much more than that, as users redefined the meaning and modes of using tags. In the site, tagging is not merely a mechanism for retrospective content classification, but also provides institutional scaffold for on-going communication within the system. From the point of view of *individuals*, using a tag within a status update still provides information about what the update is about, but also implies that it is joining the conversation defined by the tag, and, consequently, subscribing to the rules and conventions governing conversation. In this sense, tags can be thought of as institutions: establishments that regulate and coordinate social conduct – here, mostly communication. From the systemic point of view, tags-institutions define what Blip.pl is about, the meaning of its dynamics, and its culture.

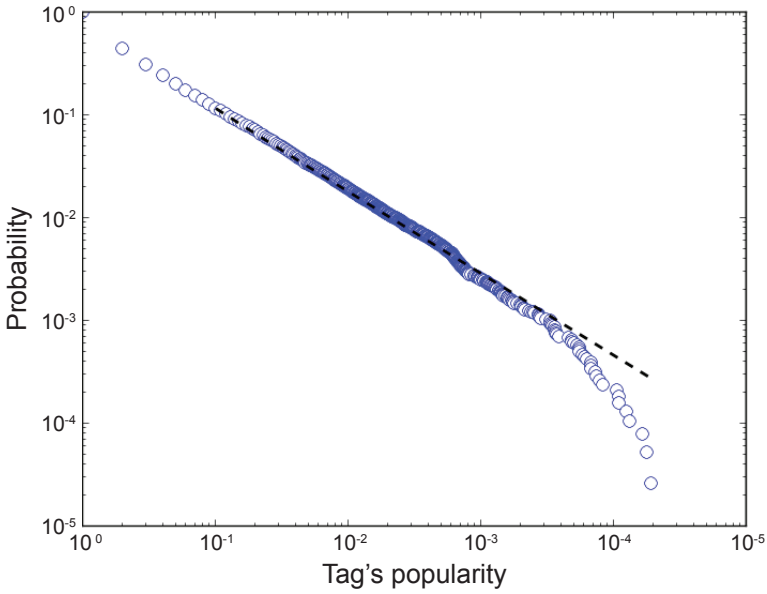
Just like other entities which get reproduced and evolve, tags skyrocket or deteriorate depending on a complex interplay of their intrinsic qualities and their environment. Our intention is to zoom in on the latter, systemic factors.

The long tail of the Blip culture

One of the preliminary results obtained from the data analysis carried out concerns tag popularity, whose distribution scales like a power law, a feature Blip shares with a wide range of natural, technological and socio-cultural phenomena (cf. e.g. Newman, 2005; Clauset et al., 2009). A power law distribution describes a situation where there are very few items in examined system that are very popular and a large number of hardly popular ones, and the frequency of a variable is related to the inverse of the power of its rank. The latter constitute the long tail of the skewed popularity distribution.[4] Our assumption is that at least a considerable proportion

[4] Different growth processes result in different size distributions. A power law (e.g. Zipf's law; Zipf, 1949; whereby logarithmic rank is linear with respect to logarithmic size) should not be confused with Gibrat's (1931) rule of proportionate growth, which claims that a quantity (e.g. size) and its change (e.g. growth rate) are independent, and yields a limiting, log-normal (rather than rank-size) distribution. It has been argued that while the former seems to be a good fit as long as a restriction is imposed on the entities not being too small or exceptionally large (cf. Sornette & Cont, 1997; Gabaix, 1999; Blank & Solomon, 2000) and thus focuses on the long tail, Gibrat's law spans over the entire class of sizes.

Figure 1. Tag popularity distribution in Blip



of popular Blip tags (popularity understood as the total number of a tag's occurrences in all status updates) constitute the "meaning" and structure of the system, its cultural and institutional establishment, while the long tail consists of tags more or less accidental (from the system's perspective). The most popular tags, whose usage exceeds 10,000, can be generic categories such as "slucham" ("listeningto"), or "dobranoc" ("goodnight"), but also tags specific to the service's culture, such as "drogiblipie" ("dearblip"), usage of which is a means of asking the community for help. There are a huge number of tags used only a couple of times. Some of them are unsuccessful variants of more popular categories, some are idiosyncratic to the users who created them, finally there are tags consisting of strings of random characters. The actual distribution of tags' popularity is presented in Figure 1, with the *X*-axis representing popularity and the *Y*-axis the probability that a tag with a given popularity exists in the system. The dashed line represents a fitted power law (exponent alpha = 1.79). The power law model could be adequately fitted only after discarding cases of popularity lower than 10 (as indicated by truncated dashed line), leaving almost 90% of all tags. This shows how extreme the distribution is in terms of the number of items which have not been diffused and hence play only a negligible role in the functioning of the system. The other side of the same issue is the degree to which the probability of a tag's popularity drops away from the model at the "high popularity" side of the figure: very popular tags are even less

likely than predicted by the power law model. Our interests lie in answering questions about the mechanisms which were responsible for the system becoming the way it is in terms of cultural tag composition.

Social influence and diffusion

The most important mechanism we are looking for has to do with diffusion of innovation. Diffusion and creation of novelty has been traditionally assumed to be among the most important social processes (Tarde, 1890). In our case, each of Blip's tags, a potential communication coordinator, had been first created by a user, then spread throughout the system with greater or smaller success. Some of the most successful, most frequently imitated tags have become Blip's culture and structure.

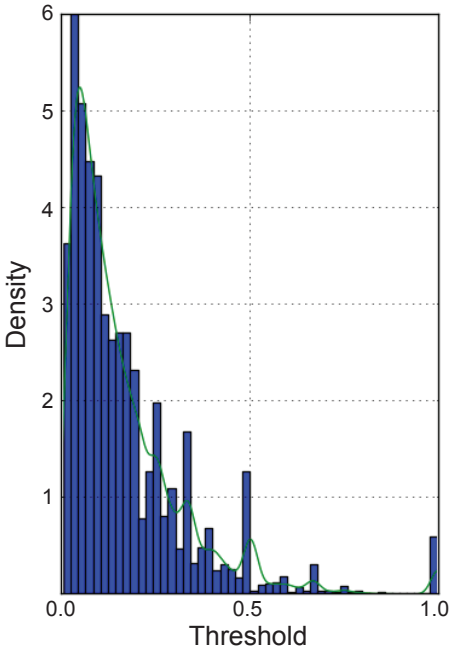
There are a number of theories explaining the mechanisms of diffusion of novelty, and one of our goals is to find out which best fits our data. Memetic theory assumes that ideas (here coded as words-tags) are like viruses which "use" the mechanisms of human minds in order to reproduce. The most successful reproducers would be those optimally adapted to the environment of the human mind – its natural dispositions and the ecosystem of already established ideas (Dawkins, 1976; Dennett, 1990)

The theory of social influence models a situation in which individual behaviour (including adoption of innovation) is contingent on peer pressure. The threshold model of collective behaviour postulates that a person will adopt a given behaviour only after a certain proportion of the people s/he observes have already adopted it. This proportion – the "adoption threshold" – constitutes the individual characteristic of each member of the group (Granovetter, 1978; Valente, 1995).

A third point of view is offered by the social learning theory (Bandura, 1977), which assumes that innovation or behaviour adoption is a result of a psychocognitive process which involves evaluation of other people's behaviour and its consequences. In this case the adoption process is perceived as more reflexive and less automatic than the previous two (Sperber, 2000; Henrich & Boyd, 2002).

The preliminary analysis we conducted involved calculating thresholds for all tag adoptions (i.e., their *first* usages). Following the above-mentioned Valente's network interpretation as a threshold model of group behaviour, we define adoption threshold as network exposure of a given tag at the moment of adoption. In this particular context, exposure means the proportion of people followed who had adopted the tag before a given user did so. So if our focal blip.pl user followed 7 others at the moment of adoption, and if out of those followed users 3 had adopted a tag before s/he did, then his/her exposure (and hence threshold) at the moment of adoption would be 3/7, or approx. 43%. We calculated threshold values for all adoption events. The resultant distribution of the thresholds (values of which range from 0 to 1, 0s being discarded as indicating invention instead of adoption) is considerably skewed, with most of the cases concentrating at the lower end of

Figure 2. Distribution of tag adoption thresholds in Blip



the threshold range, with a median of distribution is 0.11, a mean of 0.16, and a tail of higher values present, as indicated by probability density function of threshold values (Fig. 2)[5]. These prevailing low values mean that in general the users of blip.pl do not look up to their network neighbourhood when deciding whether to adopt a tag or not. For cases where a user follows 10 others or fewer, on average the first contact with a given tag decides whether it will be adopted. This, in turn, suggests that the population of Blip users is generally innovative and/or corroborates the viral model of diffusion, where adoption is decided each time tag and human cognitive apparatus interact, just as being infected is decided each time a virus interacts with a healthy human's immune system. However, we expect other factors (such as tag and user characteristics) to play an important role as well. Our aim is to consider models that include these factors in explaining diffusion mechanisms. Since the importance of the distinction between exogenous (not confined to the medium) and endogenous (idiosyncratic) innovation has already been proven (cf. Altmann et al., 2011), we particularly focus on those neological tags that are not used outside the site, which enables us to treat it as a hermetic, closed-circuit system.

[5] The "humped" feature of the distribution tail stems from the skewed distribution of the variables used to calculate the threshold values. The border values going beyond the 0÷1 range result from the smoothing algorithm of the density function and naturally have no significance.

Implications

Apart from the empirical merit, the results of the current and on-going research can deliver concrete practical solutions and indications for effective social policy and advertising, efficient dissemination of knowledge, coolhunting (tracking of existing cultural trends and prediction of changes), sentiment exploration, historical studies, forensic linguistics (e.g. detection of paedophile and terrorist activity online, facilitating timely prevention), and policies to promote linguistic diversity (sustaining endangered and minority languages). The newly developed research algorithms will be able to inform analyses in other fields of social sciences and humanities, as well as enhance Web 2.0+ algorithms, recommendation systems, and knowledge-oriented information retrieval such as automatic text analysis, classification, fast algorithmic meme aggregation, and event-tracking. Last, but not least, they will have significant practical import for the discipline of artificial intelligence, providing the parameters necessary in the devising of autonomous artificial cognitive systems' learning and (human-machine and machine-machine) interaction mechanisms.

Acknowledgments

The authors thank the reviewers for useful comments and suggestions. All the usual disclaimers apply. The research has been supported by a grant from the Polish Society for Social Psychology, Agora SA and PBI. The authors thank the reviewers for useful comments and suggestions. All the usual disclaimers apply.

References

Altmann, E.G., Pierrehumbert, J.B., & Motter, A.E. (2011). Niche as a determinant of word fate in online groups. *PLoS ONE,* 6 (5): e19009.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Benson, P. (2010). SLA after YouTube: New literacies and new language learning. Inv. talk, Univ. Warsaw.

Blank, A. & Solomon, S. (2000). Power laws in cities population, financial markets and internet sites: Scaling and systems with a variable number of components. *Physica A*, 287, 279-288.

Castells, M. (2007). Communication, power and counter-power in the network society. *International Journal of Communication,* 1 (1), 238-266.

Castells, M. (2009). *Communication power*. New York: Oxford University Press.

Clauset, A., Shalizi, C.R., & Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51 (4), 661-703.

Dawkins, R. (1976). *The selfish gene*. Oxford: Oxford University Press.

Davies, J. (2007). Display, identity and the everyday: Self-presentation through

digital image sharing. *Discourse, Studies in the Cultural Politics of Education*, 28 (4), 549-564.

Dennett, D.C. (1990). Memes and the Exploitation of Imagination. *Journal of Aesthetics and Art Criticism*, 48 (2), 127-135.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, 114, 739-67.

Gibrat, R. (1931). *Les Inégalités économiques*. Paris: Librairie du Recueil Sirey.

Gladwell, M. (2010). Small change: Why the revolution will not be tweeted. *The New Yorker*, October 4, 2010.

Global Language Monitor (2009). Death of Michael Jackson. Retrieved from: http://www.languagemonitor.com/news/death-of-michael-jackson/

Granovetter, M. (1978). Threshold models of collective behavior. *The American Journal of Sociology*, 83 (6), 1420-1443.

Henrich, J., & Boyd, R. (2002). On modeling cognition and culture: Why cultural evolution does not require replication of representations. Journal of Cognition and Culture 2(2), 87-112.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46 (5), 323-351.

O'Dell, J. (2010). Does online buzz mean better TV ratings? Retrieved from: http://mashable.com/2010/06/24/neilsen-vs-social-media/

Paradowski, M.B. (2009). Applying a complexity science approach to analysing and modelling language phenomena. Invited lecture, Higher English Language Seminar, Dept English, Stockholm Univ.

Paradowski, M.B. & Jonak, Ł. (2012). Understanding the social cascading of geekspeak and the upshots for social cognitive systems. In A. Galton & Z. Wood (Eds.), *Understanding and modelling collective phenomena* (pp. 27-32). AISB/IACAP World Congress, 2-6 July 2012, Birmingham, UK.

Paradowski, M.B., Jonak, Ł., & Kuscsik, Z. (2010). Tracking the diffusion of lexical innovation in online social networks. Workshop on Data-Driven Dynamical Networks, l'École de Physique des Houches.

Project for Excellence in Journalism (2010). New media, old media. How Blogs and Social Media Agendas Relate and Differ from Traditional Press. Retrieved from: http://www.journalism.org/analysis_report/new_media_old_media

Rogers, E.M. (2003). *Diffusion of innovations*. New York: Free Press.

Sornette, D. & Cont, R. (1997). Convergent multiplicative processes repelled from zero: Power laws and truncated power laws. *Journal of Physics I*, 7 (3), 431-444.

Sperber, D. (2000). An objection to the memetic approach to culture. In R. Aunger (Ed.), *Darwinizing culture: The status of memetics as a science* (pp. 163-173). Oxford: Oxford University Press.

Tabor, W. & Tanenhaus, M.K. (2001). Dynamical systems for sentence processing. In M.H. Christiansen & N. Chater (Eds.), *Connectionist psycholinguistics* (pp. 177-211). Westport, CT: Ablex.

de Tarde, G. (1890). *Les lois de l'imitation: étude sociologique*. Paris: Félix Alcan.

The "Five Graces Group", Beckner, C., Blythe, R., Bybee, J., Christiansen, M.H., Croft, W., Ellis, N.C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning,* 59 (Suppl. 1), 1-26.

Valente, T.W. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ, Hampton Press.

Van Geert, P. (2009). A comprehensive dynamic systems theory of language development. In K. De Bot & R.W. Schrauf (Eds.), *Language development over the life span* (pp. 60-104). New York/London: Routledge.

Vander Wal, Th. (2007, Feb 2). Folksonomy coinage and definition. Retrieved from: http://vanderwal.net/folksonomy.html

Watts, D.J. (2007). The accidental influentials. *Harvard Business Review*, 85 (2), 22-23.

Winters, M.E., Tissari, H., & Allan, K. (2010). *Historical cognitive linguistics*. Berlin: Mouton de Gruyter.

Zipf, G.K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.