ON THE IMPACT OF COMMUNITY STRUCTURE ON SELF-ORGANIZING LEXICAL NETWORKS

ALEXANDER MEHLER

Computational Linguistics, Bielefeld University, Universitätsstraße 25, Bielefeld, D-33615, Germany

Alexander.Mehler@uni-bielefeld.de

This paper presents a simulation model of self-organizing lexical networks. Its starting point is the notion of an association game in which the impact of varying community models is studied on the emergence of lexical networks. The paper reports on experiments whose results are in accordance with findings in the framework of the naming game. This is done by means of a multilevel network model in which the correlation of social and of linguistic networks is studied.

1. Introduction

There is an overwhelming evidence for the exceptionality of social and linguistic networks which are known for their S*mall* W*orld* (SW) property (Watts & Strogatz, 1998; Blanchard & Krüger, 2004): other than random graphs, SW-networks do not only have short geodesic distances, but also a high degree of cluster formation. Steyvers and Tenenbaum (2005) relate this property with the time and space complexity of *linguistic* networks where it is seen to guarantee efficient memory storage and retrieval. On the other hand, Newman (2003) reports on assortativity in *social* networks where agents with alike connectivity patterns tend to be linked.

So far, simulation models of language evolution make hardly use of these findings. Rather, they rely on unrealistic community models in which for an increasing number of iterations all agents tend to communicate with each other with equal probability. That is, F*ully* C*onnected* G*raphs* (FCG) are implicitly assumed as community models where the smaller the number of agents, the less rounds are needed to complete their connections. Conversely, if the number of rounds is small but the population is large, agents communicate only with a small number of other agents so that random graphs emerge. Anyhow, FCGs are unrealistic due to their topology, while random graphs lack the clustering of social networks.

Recently, there had been efforts to utilize more realistic community models in language simulation. This has been done in the framework of the naming game (Steels, 1998; Baronchelli, Felici, Loreto, Caglioti, & Steels, 2006) in which agents collectively learn a meaning function $f : V \rightarrow M$ from a set of words to a set of objects. As namings are seen to be independent, M is reduced to a

single object. In this scenario, Baronchelli et al. (2006) start with a community model where sender and listener are always randomly chosen among all agents. Baronchelli, Loreto, Dall'Asta, and Barrat (2006) use instead the SW-model of Barabási and Albert (1999) (i.e. the BA-model) in which agent connectivity obeys a power law. They show that under this regime, language convergence is slowed compared to FCGs. See Dall'Asta, Baronchelli, Barrat, and Loreto (2006b) for an extensive discussion of the impact of the topology of agent networks on the naming game. This includes memory complexity which in the BA-community model turns out to be less. Dall'Asta, Baronchelli, Barrat, and Loreto (2006a) complement this picture by starting from an agent network based on Watts & Strogatz's SW-model and also report an acceleration of the convergence process in conjunction with a reduction of memory load. See also Lin, Ren, Yang, and Wang (2006) who use SWs with homogeneous node degree distributions to separately study the effect of agent clustering. Further, Barr (2004) considers a set of words and of objects whose mapping is learned in a FCG community in comparison to a geometric community model which corresponds to a k-regular graph (Mehler, 2007a).

All these approaches combine a *structured* community model with an *unstructured* meaning space. That is, the set-theoretic naming game does not consider meaning-based associations of lexical items which span *lexical networks*. Thus, we lack a simulation model which studies the impact of *social* agent networks on the emergence of *linguistic* lexeme networks. This paper presents such a model. Our basic hypothesis is that the topology of the agent network does not only have an impact on the process of language change (e.g. by reducing its time and space complexity), but also on the topology of the lexical network being learned. In other words: during language evolution, social network structure imprints on linguistic network structure – at least on the level of topological characteristics. The paper presents a simulation model in support of this hypothesis. In order to do this we invent the notion of an association game which complements the notion of a naming game from the point of view of lexical networks. The paper is organized as follows: Section 2 presents the simulation model and defines association games. Section 3 shows the impact of the community structure on self-organizing lexical networks. Finally, Section 4 concludes and prospects future work.

2. A Three-Level Simulation Model of Self-organizing lexical networks

The basic idea of our approach is to start form a three-level simulation model of lexical networks. In this so called N^3 model, a lexical network is learned by interacting agents subject to their neighborhood relations. More specifically, we distinguish the level of text aggregates (generated by the agents) from the underlying community network and the lexical network as output by the multiagent learning. That is, *agent*, *text* and *lexeme network* are the 3 levels of the N^3 model:

1. The agent network is the independent variable. By analogy with the naming game we start from a model of intra-generational language change (Niyogi,

2006). Thus, we suppose that during the run of a game agents have stable neighborhoods – solely affected by the random choice of interactants.
2. The lexical network is the dependent variable. Its evolvement is observed in terms of small world characteristics where the size of the underlying lexicon is seen to be fixed during the same run.
3. Finally, the intermediary text level bridges the gap between the social and the language network and, thus, conveys information from the social topology to its linguistic counterpart.

A three-level network is exemplified by scientific communication where networking occurs on the level of the scientists involved (i.e. a collaboration network), on the level of the documents being generated (spanning a citation network) and on the level of the shared ontology manifested by these documents. Evidently, networking on any of these levels correlates with structure formation within the other two. In this paper we look on linguistic networking from the point of view of social networking thereby studying ontology formation subject to constraints of the underlying language community (as, e.g., in wiki-based systems).

In order to simulate this dynamics we now present a model of social networking, of lexical networking and of text generation & processing.

2.1. *Agent Networking*

Agent communities P are represented as undirected graphs $G(P) = (P, E)$. In order to vary $G(P)$ as an independent variable we implement three graph classes:

- Random graphs $G_{\mathrm{rand}}(P)$ are based on power law-like degree distributions of agent connectivity.
- k-regular graphs $G_{\mathrm{reg}}(P)$ are graphs in which each vertex has exactly the same number of neighbors, that is, the same degree k.
- Finally, small world graphs $G_{\mathrm{sw}}(P)$ combine a power law-like degree distribution with a high cluster value and short average geodesic distances.

The first two classes were used by Watts and Strogatz (1998) to introduce their SW-model. Both are unrealistic in terms of social networking: random graphs lack the clustering of social networks, while small average geodesic distances are absent from regular graphs. Nevertheless, random graphs share the distance property with SWs, while regular graphs have by definition high cluster values. Random and regular graphs are referred to as baseline community models. That is, we expect that communities of the sort of $G_{\mathrm{rand}}(P)$ and $G_{\mathrm{reg}}(P)$ lead to deficient lexical networks when underlying a language game. We suppose that this is due to their disputable status as models of social networks – *in contrast to SWs*. In this paper, we generate SW-agent networks based on the approach of Mehler (2007a). It outputs connected graphs with high cluster values, short geodesic distances, power law-like node connectivities and assortative mixing of node degree – *in accordance with what is known about social networks* (Newman, 2003).

2.2. *Lexical Networking*

The language learned by the community P is the dependent variable in our model. As explained in Section 1, we focus on lexical networks as target languages which are represented as undirected graphs. We utilize L*atent* S*emantic* A*nalysis* (LSA) (Landauer & Dumais, 1997) as a learning theory of lexical associations in order to induce the edge set of these graphs. As LSA is a *single* agent model of usage-based meaning, we reconstruct it in terms of multiagent learning (Mehler, 2007a). This is done by means of an iteratively computable lexical association measure which is updated per text unit: For a lexicon V and a sequence $S_n = \langle x_1, \ldots, x_n \rangle$ of n texts, the association of two lexical items $v_i, v_j \in V$ is computed as

$$\alpha(v_i, v_j, S) = \sum_{k=1}^{n} \left(f_{ik} \frac{k}{F_{ik}} \right) \left(f_{jk} \frac{k}{F_{jk}} \right) = \sum_{k} k^2 \frac{f_{ik} f_{jk}}{F_{ik} F_{jk}} \tag{1}$$

where F_{ik} is the number of texts in S_k in which v_i occurs and f_{ik} is the frequency of v_i in x_k. In accordance with models of human text processing, $\alpha(v_i, v_j, S_n)$ is sensitive to the order of texts in S_n. Next, we endow each agent $a \in P$ with this learning model so that he can learn lexical associations subject to the communication situations to which he participates. After t iterations of the language game, that is, after processing sequence S_t, this leads to a *distributed semantic space*

$$M_t(P) = \{(V, E_a^t, \omega_a^t) \,|\, a \in P\} \tag{2}$$

in which each agent $a \in P$ has his own meaning space $M_t(a) = (V, E_a^t, \omega_a^t)$ with edge set E_a^t and weighing function $\omega_a^t(v_i, v_j) = \alpha(v_i, v_j, S_a^t)$. Note that lexicon V is common to all agents while the sequence S_a^t of the texts processed by agent a at time t is specific to a. For a text x_t processed at time t by agent a we write

$$M_{t-1}(a) \overset{x_t^i}{\longmapsto} M_t(a) \quad \text{or} \quad M_t(a) = x_t^i(M_{t-1}(a)) \tag{3}$$

where i indicates how often a processes x_t at time t. Thus, at time t the memories $M_t(a)$ of agents may differ dependent on the text sequences S_a^t they have processed till t. This model resembles the one of Hashimoto (1997). The difference is that we concentrate on syntagmatic associations, optimize the model for iterative computability and clarify the topological characteristics of $M_t(P)$.

2.3. *Association Games*

Now we define *Association Games* (AG) which generate distributed semantic spaces $M(P)$ based on community models $G(P)$. That is, AGs are mappings

$$G(P) \mapsto M(P) \tag{4}$$

from social to linguistic networks. They define an association task in which the sender produces a text x to mask the prime word he used to generate x and where

the listener has to identify the prime. A round of an AG looks as follows: starting from a randomly chosen sender $a_S \in P$, all neighbors of a_S in $G(P)$ are picked as listeners a_L each getting a separate text (Zollman, 2005). For such a listener a_L, the sender is masking the word v_+ he used to prime the lexical constituents of his output text x_t so that the listener a_L has to find out which word the sender had in mind when producing x_t. The listener processes x_t and tells the sender his guess v_- so that a_S can decide whether he was understood or not. A single round of the AG is successful if both sender and listener associate the same or related words with the same input text. This scenario resembles the children's game "*I spy with my little eye, something beginning with ...*". The difference is that in the association game not denotations, but lexical primes are guessed using texts as underspecified descriptions thereof and where agents learn the underlying priming relations (i.e. lexical connotations) by playing the game.

More formally: starting from a sender a_S at round t and a randomly chosen prime v_+, a text of length l is generated by collecting a subset of l nearest neighbors of v_+ in $M_{t-1}(a_S)$. Initially, lexical neighbors are picked at random. Note that we suppose fixed text lengths for the whole run of a game. Note further that texts are represented as multi-sets so that types $v \in V$ may recur. Next, the listener uses x_t to activate a subspace in his memory $M_{t-1}(a_L)$ and, based thereon, to context-prime a guess v_-. This is done by an inverse function of text generation which finds the "centroid" among the constituents of x_t and their neighbors in $M_{t-1}(a_L)$. After uttering v_-, the sender evaluates this guess by the geodesic distance $L(v_+, v_-)$ in $M_{t-1}(a_S)$. Here, we start from the hypothesis that any text generation/processing reinforces the associations being manifested in the output/input text so that the sender is "his first recipient", while the listener always tries to "understand" his input. Now, a successful round is rewarded by reinforcing memory update, while otherwise this reinforcement is ommitted:

$$\forall a \in \{a_S, a_L\}\colon M_t(a) = \begin{cases} x_t^2(M_{t-1}(a)) : L(v_+, v_-) \leq r \\ x_t^1(M_{t-1}(a)) : \text{otherwise} \end{cases} \tag{5}$$

r is a further parameter of the model where $r = 0$ means that $v_+ \stackrel{!}{=} v_-$.

So what does it mean now to speak about terminological alignment via association games? Under a local perspective this means that if sender and listener align their lexical associations as they continually communicate they finally play the game more and more successful. Under a global perspective it means that if the AG is successfully played by the community P as a whole, this leads to a lexical network which – as we hypothesize – has the SW-property subject to the SW-property of the agent network $G(P)$. This is evaluated in the next section.

3. Experimentation

We test our hypothesis about the imprint of social on linguistic structure by varying the community model with random, small world and 4-regular graphs using

100 agents. We consider a lexicon of 500 words and set the threshold of the summary language to 0.375. That is, an association between two words is seen to belong to the target language if at least 37.5% of the agents share it. Further, we set the size of texts to 5 tokens and $r = 2$. Finally, we compute $500,000$ iterations of the AG per community model and average over 50 runs. Figure 1 exemplifies a run based on a SW-like agent network. For growing iteration we see gradually evolving a connected graph which (as explained below) results in a SW-like lexical network – starting from a completely disconnected graph.

So what happens to the topology of lexical networks if the community model is varied? This is answered in Figure 2. We start from the fraction of words in the largest connected component (2.a) and observe that a connected graph of all words evolves based on the SW- and on the random community. However, in the former this happens faster whereas the regular graph-community lacks such a component. Next, Figure 2.b shows the cluster coefficient (Watts & Strogatz, 1998). We observe that in the SW-community the lexical network has a much larger degree of clustering – comparable to the values observed in wiki-based networks (Mehler, 2007b). In fact, in the random community-based lexical network clustering is much lower – not to mention the regular graph-community. Figure 2.c completes this picture: the average geodesic distance is smaller and emerges faster in the SW-community based lexical network compared to its random counterpart. However, the regular community-based lexical network seems to have the smallest distance value. Actually, this is due to the fact that in 2.c L is always computed for the largest connected component. Thus, in 2.d we normalize L by assuming that unconnected agents are separated by $|V| - 1$ edges. Now, the random and regular graph-based agent networks are both outperformed by their SW-counterpart.

In summary, we observe an imprint of social on linguistic topology, though not an isomorphic one: the SW-community model results in a SW-lexical network; random and a regular graph-based community models do not. Moreover, in the latter case the lexical networks do not share properties with their social counterparts: the regular agent network has, *per definitionem*, a high cluster value, but not the linguistic network based on it. These observations confirm a strong impact of *social* on *linguistic* networking as motivated by complex network theory. To the best of our knowledge this has not been evaluated by a multiagent simulation model of lexical networks so far.

4. Conclusions

We have introduced association games as a framework to study the self-organization of lexical networks. We have shown that the topology of the underlying agent community has a strong impact on these networks. This has been done in terms of intra-generational language change. The co-evolvement of social and linguistic networks in inter-generational language evolution is object of future

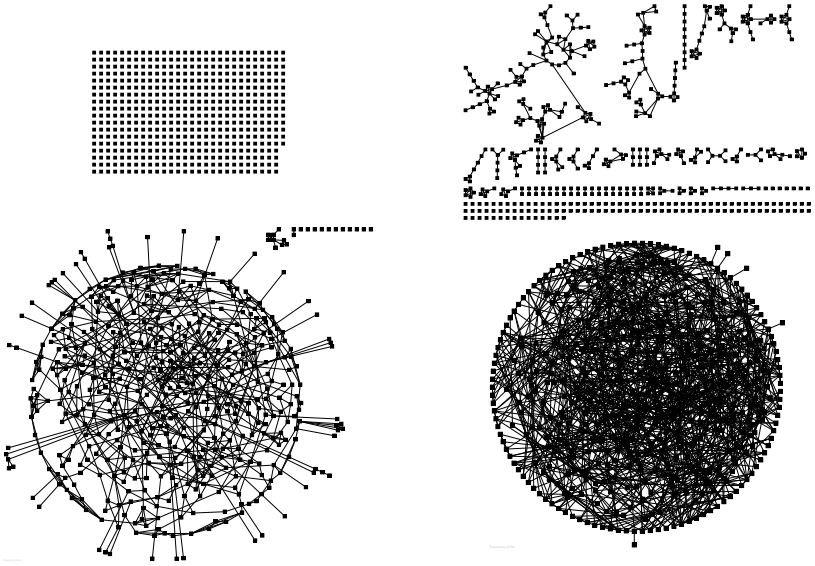

Figure 1. *Having a look at the dynamics of gradually evolving lexical networks:* snapshots after 1,000, 25,000, 50,000 and 300,000 rounds of an association game.

work. This also includes an integration of the naming and the association game.

References

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509-512.

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, *P06014*.

Baronchelli, A., Loreto, V., Dall'Asta, L., & Barrat, A. (2006). Bootstrapping communication in language games: strategy, topology and all that. In *Proc. of Evolang6, 12-15 April 2006, Rome* (p. 11-18).

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*(6), 937-962.

Blanchard, P., & Krüger, T. (2004). The cameo principle and the origin of scale free graphs in social networks. *Journal of statistical physics*, *114*(5-6), 399-416.

Dall'Asta, L., Baronchelli, A., Barrat, A., & Loreto, V. (2006a). Agreement dynamics on small-world networks. *Europhysics Letters*, *73*, 969.

Dall'Asta, L., Baronchelli, A., Barrat, A., & Loreto, V. (2006b). Non-equilibrium dynamics of language games on complex networks. *Physical Review E*, *74*, 036105.

Hashimoto, T. (1997). Usage-based structuralization of relationships between words. In *ECAL97* (p. 483-492).

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem. *Psychological Review*, *104*(2), 211-240.

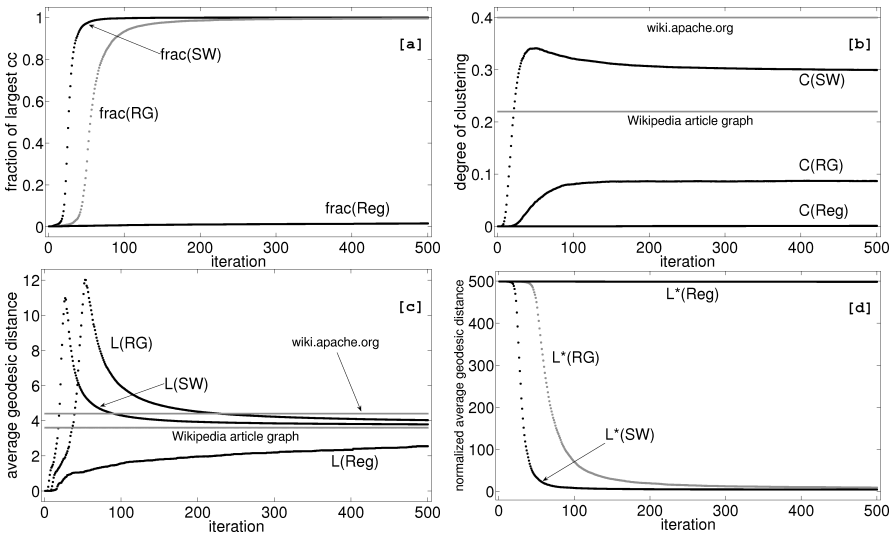

Figure 2. Dynamics of SW-characteristics in the association game: 500,000 iterations are performed per class of agent network: *Small World* graphs (SW), *Random Graphs* (RG) and *Regular graphs* (Reg). Values are averaged over 50 runs of the game and the respective type of agent network. (a) The fraction of words belonging to the largest connected component. (b) The cluster coefficient C of the lexical network. (c) The average geodesic distance of the largest connected component of the lexical network. (d) The normalized average geodesic distance regarding the entire lexicon.

Lin, B.-Y., Ren, J., Yang, H.-J., & Wang, B.-H. (2006). *Naming game on small-world networks: the role of clustering structure.*

Mehler, A. (2007a). Evolving lexical networks. A simulation model of terminological alignment. In A. Benz, C. Ebert, & R. van Rooij (Eds.), *Language, Games, and Evolution. Workshop at ESSLLI 2007* (p. 57-67).

Mehler, A. (2007b). Large text networks as an object of corpus linguistic studies. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook of the science of language and society.* Berlin/New York: De Gruyter.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167-256.

Niyogi, P. (2006). *The computational nature of language learning and evolution.* Cambridge: MIT Press.

Steels, L. (1998). The origins of ontologies and communication conventions in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, *1*(2), 169-194.

Steyvers, M., & Tenenbaum, J. (2005). The large-scale structure of semantic networks. *Cognitive Science*, *29*(1), 41-78.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440-442.

Zollman, K. J. S. (2005). Talking to neighbors: The evolution of regional meaning. *Philosophy of Science*, *72*, 6985.