Modeling Lexicon Emergence as Concept Emergence in Networks

Juan Galán-Páez, Joaquín Borrego-Díaz and Gonzalo A. Aranda-Corral

Abstract A model for lexicon emergence in social networks is presented. The model is based on a modified version of classic *Naming Games*, where agents' knowledge is represented by means of formal contexts. That way it is possible to represent the effect interactions have on individual knowledge as well as the dynamics of global knowledge in the network.

1 Introduction

In Complex Systems (CS) research, the majority of recent studies on dynamics in Social Networks (SN) considers two non-exclusive perspectives. On the one hand, agent evolution is studied by considering it as part of a network, where its topology plays a relevant role. On the other hand the global perspective considers the network itself as a dynamic system on which its topology depends on local rules [13]. In the case of emergent properties based on agent interaction, modeling techniques adopt the first one, whilst the second perspective aids to understand the overall behavior of the system by means of global features and parameters.

In the field of Social Networks Analysis, human interactions lead to the formation of different kinds of (emergent) opinions. For example opinion consensus/formation (a product of collective intelligence) is very similar, in some cases, to lexicon emergence—a topic widely studied in CS field—and its modeling can be

J. Galán-Páez · J. Borrego-Díaz
Department of Computer Science and Artificial Intelligence,
University of Seville, Seville, Spain

G.A. Aranda-Corral (✉)
Department of Information Technology, Universidad de Huelva, Huelva, Spain
e-mail: garanda@us.es

exploited for developing social strategies [10]. Sentiment/opinion analysis attempts to analyze the effect and scope of these tendencies. Lexicon emergence is a research line within Language Dynamics (LD) paradigm, where agent modeling is a key tool. LD is a rapidly growing field in CS community that focuses on all processes related with emergence, evolution, change and extinction of languages [12]. The models of LD provides interesting ideas to model and predict similar social processes. The consensus (as an extension of agreement) in LD allows to study which are the key factors that drive lexicon evolution. Thus it could be interesting to adapt those to semantically enriched agent networks. Classic LD does not consider strong semantic features on agents' interactions, and Formal Concept Analysis (FCA) [8] provides a general framework in which semantic features can be added to LD (at object/attribute level) [2].

Collective consensus on concepts (opinions, new ideas, etc.) is a phenomena that is frequent in social media, and in the users own use of (weak) semantic strategies for streamlining this process (for example the use of hash-tags in Twitter or the spreading, mutation and adaptation of *memes* which has complex dynamics [1]). Moreover, sentiment/opinion analysis provides social researchers with new ideas and tools. For example, in [7] authors show how Formal Concept Analysis of twitter stream on a *Trending Topic* provides a global representation of sentiment concepts as well as a kind of new sentiment concepts based on sentiment vocabularies. This kind of study can predict viral meme evolution, for example.

The aim of this paper is to develop FCA-based Naming Games for modeling concept emergence in social networks by applying ideas from [2].

Related Work: In [4] authors design a basic FCA-based agent interaction model, intended (in that paper) for modeling consensus between bookmarking agents in order to estimate how semantic heterogeneity behaves in social bookmarking services. The particular case of NG in Social Networks is an interesting research line in CS research where the net effect on lexicon emergence is studied (see [14] for a nice introduction to the topic). The most used approaches to NG modeling are focused on lexicon emergence and semantics is not usually considered (because the own vocabulary it is not predefined). This paper is focused on the dynamics of the inherent semantic (modeled by FCA) associated to preexistent vocabularies. This basic model was enhanced to be used in NG [2], which is adapted here to social networks in turn. Several variants of NG can be studied by considering different levels of reasoning ability of agents [2]. In [2] authors show how FCA can be used to enhance models of language emergence by enriching the semantics of agent models. In this way semantics self-emergence can be modeled. For instance, an approximation to the study of self-organization and evolution of the language and its semantics could be to consider the community of users as a CS that collectively builds the semantic features of their own lexicon.

A related study is [9], where authors studied the effects of randomness on the competition between strategies in an agent-based model of tag-mediated cooperation evolving on large-scale complex networks.

2 Background

A popular approach in LD consists in modeling agents' interaction by means of the called *Naming Games* (NG) [16], which was created to explore self-organization in LD (emergence of vocabularies, in other words, the mapping between words and meanings). Naming games consist in the interaction between two agents, a speaker and a listener. The information in a language game is local to its participants: other agents are not aware of innovations or adaptations that might have come up during that interaction. Only after this innovation has been used again in other interactions, it can be spread through the population. From the basic model, a number of variants for several and specific models can be considered. The aim of the agent community is to achieve a common vocabulary. In the minimal NG each agent has its own context (object/word) and interacts according to the following steps [12]:

1. The speaker selects an object from the current context.
2. The speaker retrieves a word from its inventory associated with the chosen object, or, if its inventory is empty, invents a new word.
3. The speaker transmits the selected word to the listener.
4. If the listener has the word named by the speaker in its inventory, and that word is associated with the object chosen by the speaker, the interaction is a success. In this case, both players maintain in their inventories only the winning word, deleting all the other words that fitted the same object.
5. If the listener does not have the word named by the speaker in its inventory or the word is associated to a different object, the interaction fails: the listener updates its inventory by adding an association between the new word and the object.

NG have been considered in non-situated and situated models. In the second one, agents are placed in an artificial world, where environmental features as distance between agents or agent's neighborhood can be considered. Situated models are very interesting since the communication among agents does not obey purely random selections: communication takes place between agents that are able to do it, according to the restrictions of the model itself (for example, only neighboring agents can communicate). NG with spatially distributed agents allow modeling the emergence of different language communities by stabilization of the system [17]. This is due to the fact that the "success" of a linguistic innovation dependends on whether the group, as a whole, has adopted it or not. Likewise agent networks can be considered as in this paper.

Our interest in NG is based on their *adaptive nature*; that is to say, NG can produce changes in the lexicon of both, the speaker and listener, as side effect. Thus, agent's lexicon changes during its life within the system. In the case of this work, it is interesting to consider agents' neighbors within the network to see how their interaction model agent's lexicon. Especially in complex networks as for instance the scale-free networks, a kind of network topology very frequent in social networks.

3 Formal Concept Analysis and Implications

According R. Wille, FCA mathematizes the philosophical understanding of a concept as a unit of thoughts composed of two parts: the extent and the intent [8]. The extent covers all objects belonging to this concept, while the intent comprises all common attributes valid for all the objects under consideration. It also allows the computation of concept hierarchies from data tables. In this section, we succinctly present basic FCA elements, although it is assumed that the reader is familiar with this theory (the fundamental reference is [8]).

A formal context is represented as $M = (O, A, I)$, which consists of two sets, O (objects) and A (attributes) and a relation $I \subseteq O \times A$. Two derivation operators, both denoted by $'$, formalize the sharing of attributes for objects, and, in a dual way, the sharing of objects for attributes: Given $X \subseteq O$ and $Y \subseteq A$,

$$X' := \{a \in A \mid oIa \text{ for all } o \in X\} \text{ and } Y' := \{o \in O \mid oIa \text{ for all } a \in Y\}$$

A (formal) concept is a pair (X, Y) such that $X' = Y$ and $Y' = X$. Logical expressions used in FCA are the *implications between attributes*, pair of sets of attributes, written as $Y_1 \rightarrow Y_2$, which is true with respect to $M = (O, A, I)$ according to the following definition. A subset $T \subseteq A$ *respects* $Y_1 \rightarrow Y_2$ if $Y_1 \nsubseteq T$ or $Y_2 \subseteq T$. It is said that $Y_1 \rightarrow Y_2$ holds in M ($M \models Y_1 \rightarrow Y_2$) if for all $o \in O$, the set $\{o\}'$ respects $Y_1 \rightarrow Y_2$. See [8] for more information.

Definition 1 Let $\mathcal{L}$ be a set of implications and L an implication of M.

1. L follows from $\mathcal{L}$ ($\mathcal{L} \models L$) if each subset of A respecting $\mathcal{L}$ also respects L.
2. $\mathcal{L}$ is complete if every implication of the context follows from $\mathcal{L}$.
3. $\mathcal{L}$ is non-redundant if for each $L \in \mathcal{L}$, $\mathcal{L} \setminus \{L\} \not\therefore \models L$.
4. $\mathcal{L}$ is a basis for M if it is complete and non-redundant.

A basis called *Stem Basis* (SB) can be obtained from the *pseudo-intents* (see [8]). Although SB is used for showing the modeling, it is important to remark that SB is only an example of basis. Throughout the paper none specific property of the SB can be used, so it can be replaced by any other (implication) basis.

Roughly speaking, the goal of using FCA in NG is to analyze interactions between agents equipped with (partial) knowledge, in order to study how a collective knowledge emerges. Each interaction is a communicative act where two agents interchange new knowledge. In the experiments carried out, the creation steps have been dropped, as the aim of these models is to induce the emergence of the basis from a preexisting lexicon. For modeling communication between agents, the English lexical database *WordNet*[1] has been chosen as the global knowledge. In *WordNet* nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms called *Synsets*, where each of these expresses a distinct concept. *Synsets* are interlinked by means of conceptual-semantic and lexical relations.

[1] http://wordnet.princeton.edu.

A formal context associated with WordNet It is interesting to consider a real and structured language in order to exploit FCA semantic features, since the future aim of this paper is to apply the model to real world social networks (see [4]). Therefore in this paper experiments are carried out using (subsets of) WordNet. A Formal Context associated with WordNet can be obtained by considering words as objects of the context and synsets as the attributes. Synsets are sets of synonymous words, thus they can be considered as a potential definition (meaning) of each word.

Concept lattice and the basis associated to a lexical database of this size cannot be efficiently computed (in fact, conceptual descriptions of complex systems have complex structure [3]), therefore small subsets of WordNet have been considered. That way it will be possible to evaluate the soundness of the proposed models with respect to the FCA elements associated to these subsets. Likewise, only connected subsets of *WordNet* and only formed by full *synsets* (containing every word referring to that synset) have been considered. A connected subset is one such that if a network interlinking its words and synsets were built, the resulting network would be a connected graph.

Once a WordNet subset has been chosen, the same one will be used in every experiment. Two different formal contexts have been considered in the experiments presented in this work. The *global knowledge* is a formal context containing the full WordNet subset and is the knowledge that agents' community aims to achieve. Each agent has its own formal context (*local knowledge*), which is randomly initialized with some word/synset pairs drawn from the global knowledge.

3.1 Formal Contexts Within a Networked Multiagent System

Definition 2 Let $M = (O, A, I)$ be a formal context and $G = (N, E)$ be a network.

- A **linguistic distribution** d is a function $d : I \times N \to \{0, 1\}$, which induces the relation $I_d = \{(o, a) \in I \ : \ \exists n \in N \mid d((o, a), n) = 1\}$
  The **global context associated with** d is the context $M^d = (O, A, I_d)$
- Given $n \in N$, the vocabulary and the dictionary of n w.r.t d are resp.
  $V^d(n) := \{o \in O \ : \ \exists a \in A\,[\,(o, a) \in I \ \wedge \ d((o, a), n)) = 1]\}$
  $D^d(n) := \{a \in A \ : \ \exists o \in O\,[\,(o, a) \in I \ \wedge \ d((o, a), n) = 1]\}$
  and the formal context associated with n w.r.t. d is $M^d(n) := \left(V^d(n), D^d(n), I^d_n\right)$, where $I^d_n = \{(o, a) \ : \ d((o, a), n) = 1\}$
- The **context network associated with** d and G is a network where each node n is labeled by $M^d(n)$
- Given $t \in [0, 1]$, the t-collective context associated with d is $M^{d,t} := (O, A, I^{d,t})$

$$\text{where } I^{t,d} = \{(o, a) \in I \ : \ \frac{\#\{n \ : \ d((o, a), n) = 1\}}{\#N} \geq t\}$$

Given $\delta \in [0, 1]$, a random linguistic distribution according to δ is a uniform distribution d built by assigning $d((o, a), n)) = 1$ with probability δ. Note that in this case $Prob\left((o, a) \in I_d\right) = 1 - (1 - \delta)^{\#N}$.

Stability of NG can be studied with ideas from knowledge convergence in multi-agent systems (see e.g. [6]). Mainly four parameters are considered:

- $\#N$, the population size (each node is an agent). The values chosen in the experiments are conditioned by the feasibility of the computation of each model.
- δ (the probability for an agent to have within its initial knowledge a pair (object-attribute)) is selected in a value range providing each pair *lemma-synset* to appear in at least one agent from the overall population, with probability $p = 0.95$.
- Given a threshold t (usually $t \geq 0.9$), the convergence (stabilization) test checks whether every existent pair *lemma-synset* is present within $M^{d,t}$.
- The size of the selected subset of WordNet is determined by both, the fact that the subsystem must contain complete synsets and by computational feasibility.

4 Modeling Communication as FCA-based Naming Games

Initial knowledge of two agents that will perform examples of communicative acts is shown in Fig. 1. The scale-free network was randomly generated following the principle of preferential attachment [5]. Each node is an agent that in each step communicates with one of its neighbors (those directly connected in the graph). The communicative process is as follows: Starting with d a linguistic random distribution according to $\delta \in [0, 1]$:

1. The world (network) is randomly built with $\#N$ agents (nodes). Each agent starts with an initial knowledge represented by $M^d(n)$. To obtain successful communicative games, it is necessary that the union of the initial local knowledge of each agent contains approximately all concepts within the global knowledge, hopefully $\bigcup_{n \in N} I_n^d = I$. The probability p for each pair within the selected WordNet subset to appear in the initial local knowledge, of at least one agent, is given by $p = 1 - (1 - \delta)^{\#(N)}$.
   It is recommended carrying out communication games with at least $p > 0.95$, thus the value δ to be considered depends on the number of agents N in the world (i.e. for $\#N = 200$ it is suggested that $\delta \geq 0.015$).
2. In each step, each agent (speaker) chooses randomly a listener agent between its neighbors (adjacent nodes), in order to start a communicative process (request).
3. After the simulation, the collective knowledge $M^{d,t}$ is measured.

Recall that a pair (o, a) can be considered part of the *collective knowledge* $(o, a) \in M^{d,t}$, only for a certain collective knowledge threshold $t = CK_{th}$. Thus CK_{th} denotes the minimum proportion (usually between [0.9, 1]) of agents knowing a certain pair, necessary to consider that pair as part of the collective knowledge.

There are different ways of carrying out the communicative act as well as different ways of measuring the collective knowledge (see [2]).

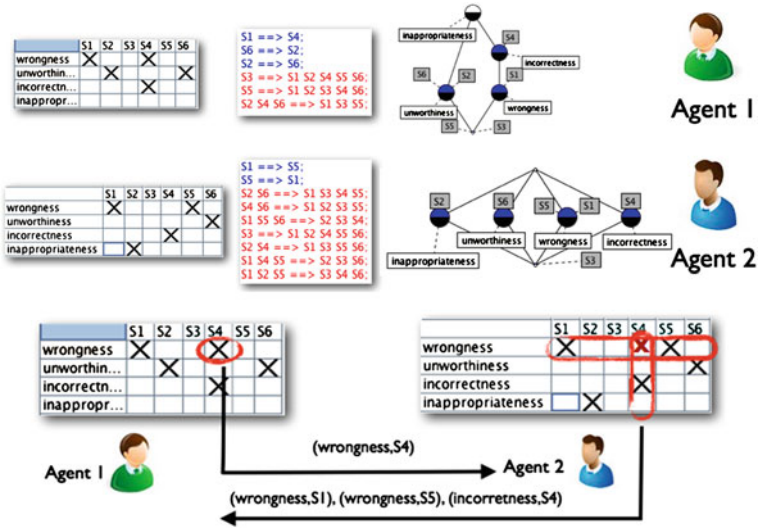

Fig. 1 Agents for the example (*top*) and Intent-extent communicative act (*bottom*)

4.1 Modeling WordNet Emergence by Intent-Extent Games

The basic model is based on the relative similarity between a *formal concept* and a *synset*. In this game, the communicative act consists in direct interchange of lemmas (objects) and synsets (attributes) between the speaker and the listener.

Communicative act: Each step, the speaker s randomly chooses a pair (o_i, a_j) from its local knowledge (formal context) and sends it (request) to the listener. The answer of the listener will consist of two sets $intent(o_i)$ and $extent(a_j)$ (relative to its own formal context). In case the listener does not have any information on o_i or a_i, it will return an empty set and will add the pair (o_i, a_j) to its local knowledge (see Fig. 1). Figure 2 shows four different states of the communicative process for this model.

Collective knowledge emergence: In order to detect and study the emerging knowledge due to agents' interactions in this communication game, the *collective knowledge* can be measured. In each time step the *error rate* between the global and the collective knowledge is measured as the difference between these.

Convergence criteria: the communication game ends when the *collective knowledge* emerged from agents' interactions is equal to the *global knowledge*. It is worthy to note that the convergence rate of the game highly depends on the collective knowledge threshold CK_{th} considered.

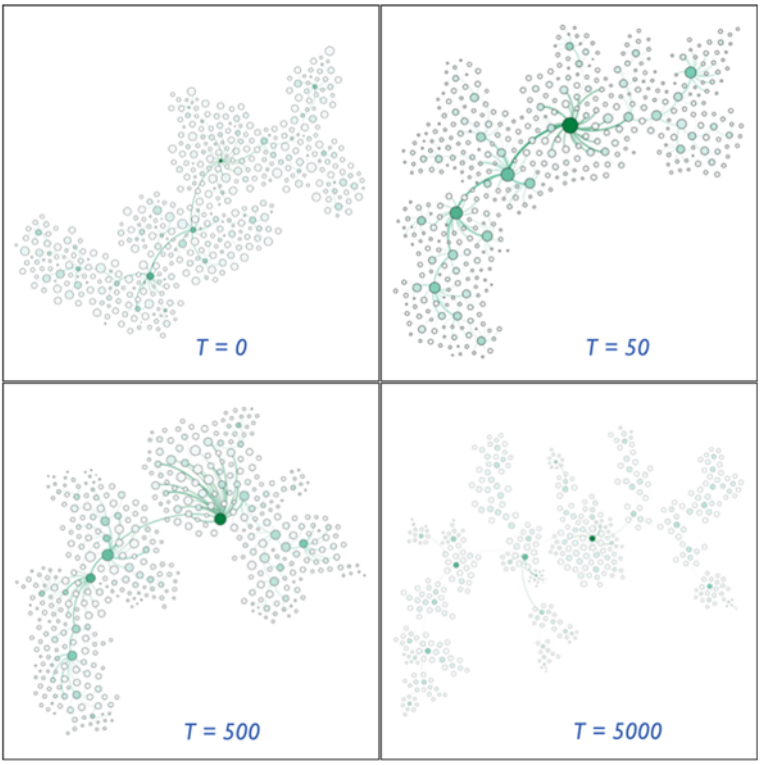

Fig. 2 Evolution of agents knowledge within the network. Node *color* denotes connectivity (darker the colour, higher the degree) and size denotes the amount of knowledge achieved. Edge *thickness* denotes knowledge difference between two agents (thicker the edge, bigger the difference)

4.2 Modeling Emergence by Implication Bases Games

This model aims to exploit the implication bases in knowledge detection tasks. In a first approach the communication process goal is the emergence of the collective knowledge by detecting and eliminating inconsistencies within agents' local knowledge. The communicative process in this case concerns to consistency issues. Each agent will contrast its knowledge with others, in order to detect and fix inconsistencies (see Fig. 3).

Communicative act: The speaker computes the SB of its local formal context $M^d(s)$, randomly chooses a rule L from it and sends it (request) to the listener l. If $M^d(l) \models L$ then it returns a positive answer and finish the communicative act. Otherwise l returns a negative answer from this context and sends to the speaker a counterexample, that is, a pair set $\{o\} \times \{o\}' \subseteq I_l^d$ such that $\{o\}' \therefore \not\models L$ in $M^d(l)$ and the speaker adds it to its local knowledge. This communicative act is similar to one step of the *attribute exploration* (cf. [8]).

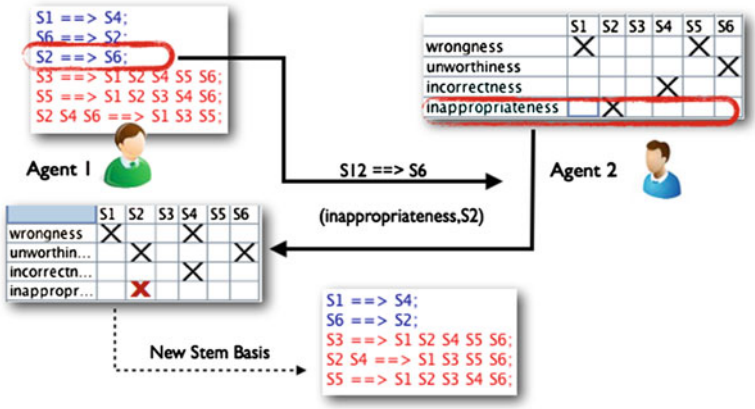

Fig. 3 Communicative act in implication basis game

Collective knowledge emergence: Since the model works with implication basis, collective knowledge has to be considered as the *collective consistent knowledge*. That is, the implication base associated with the collective knowledge has to be consistent with the base associated with the global knowledge. Since in this case, game rules have changed, the collective knowledge is measured by its consistency. In order to estimate the soundness of the emerged knowledge, it should be verified whether the true implications within the collective knowledge can be entailed from the basis of the original vocabulary. Collective implication base has to be computed from the collective knowledge. Then, its consistency is verified according the global implication base.

Convergence criteria: The game ends when it reaches the equilibrium, that is to say, when there are no more inconsistencies between agents' local knowledge. It is worthy to note that from this model, the global knowledge does not emerge (it is not the aim), but knowledge consistent with the global one.

4.3 Modeling WordNet Emergence by Hybrid Games

To model (in terms of language convergence and consistency) a sound language emergence, a hybrid model has been considered. Particularly, to complete the implication bases game, which stabilizes before agents' local knowledge converges to the global one. The consistency-based approach is interesting but does not provide full emergence of collective knowledge. Thus in this hybrid approach the two communicative act formerly presented will be used, one providing consistency (stem basis interactions) and the second providing direct information exchange (intent-extent interactions).

Communicative act: to combine both games, by selecting the communication type with a certain probability q. For low values of q, the agent will communicate mainly by means of the intent-extent based method (*diffusing agent*) and for high values, the agent will use mainly the implication base based method (*conciliatory agent*).

Collective knowledge emergence: In this case the both notions of collective knowledge above mentioned should be considered, in order to evaluate both, consistency and knowledge emergence.

Convergence criteria: game ends when both objectives, the emergence of consistency and the global knowledge, are achieved within agents' collective knowledge.

5 Experiments and Discussion

To study the convergence of agents' collective knowledge for each of the aforementioned NG many experiments have been carried out. The *WordNet* subset (a connected subset) selected for the experiments has a relatively small size (around 400 lemma-synset pairs) due to the high computation time of implication bases for huge formal contexts. Figure 4 shows the results of two of these experiments. Plots show the emergence of collective knowledge tending to the global knowledge (shown as % of the global knowledge). In each experiment we compare the convergence rate for the grid-based world (deeply studied in [2]) and for the network-based world. Figure 4 (top) shows knowledge convergence for the intent-extent game and Fig. 4 (bottom) for the hybrid probability-based game with $q = 0.5$. The mean collective knowledge $f(t)$ (being t the number of simulation steps) follows from $f(x) \sim a + be^{-\alpha x}$.

It should be noted that no experimental results are provided for the game based exclusively on implication bases, due to the fact that the system stabilizes before any concept exceed the collective knowledge threshold (CK_{th}) (which should be high),

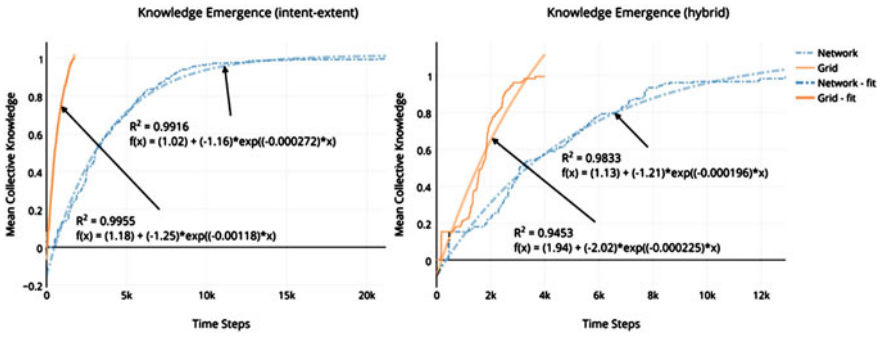

Fig. 4 Convergence of collective knowledge in communication games based on intent-extent (*left*), and hybrid games (*right*)

in order to be considered as collective knowledge. This phenomenon is similar to others in LD simulation (language competing).

Roughly speaking, intend-extend based communication produces collective knowledge fast, but will remain inconsistent (with respect to the global knowledge) until the global knowledge has been achieved. While those based on implication bases do not converge to the global knowledge, but produce partial knowledge consistent with the global one. It is worthy to note that despite the convergence rate being slower using network topology, both scenarios present the same behavior (exponential).

6 Conclusions and Future Work

In this paper a FCA-based model for lexicon emergence (and similar phenomena as sentiment emergence) in social networks is presented. Since FCA provides a solid formal semantic characterization of implicit conceptual structures of agents, the application of this model in social networks could provide some insights on the nature of information spreading and consensus in social networks. For example, the methodology presented here could be useful to model misinformation spreading in social media and networks, by enhancing other methods designed to measure the credibility of information [11]. In [2] authors show that language convergence in situated (mobile) models seems to be governed by the shared vocabulary (the mapping between words and meaning) instead of the shared language. Experiments presented in this paper show how dynamics change in models based in scale-free networks. A great amount of experiments have to be executed to deeply validate these insights.

Moreover, it could be interesting to use association rules (e.g. by Luxenburger basis [15]) instead of implications. This choice is very related with the idea of lexicon mediated by a certain confidence in the relationship.

Acknowledgments Work partially supported by TIC-6064 Excellence project (*Junta Andalucía*) and TIN2013- 41086-P (Spanish Ministry of Economy and Competitiveness), co-financed with FEDER funds.

References

1. Adamic, L.A., Lento, T.M., Adar, E., Ng, P.C.: Information evolution in social networks. CoRR(2014). arXiv:abs/1402.6792
2. Aranda-Corral, G.A., Borrego-Díaz, J., Galán-Páez, J.: Simulating language dynamics by means of concept reasoning. Bio-Inspired Models of Network, Information, and Computing Systems. Lecture Notes Institute for Computer Sciences, Social Information and Telecommunications Engineering, vol. 134, pp. 296–311. Springer (2014)
3. Aranda-Corral, G.A., Borrego-Díaz, J., Galán-Páez, J.: On the phenomenological reconstruction of complex systems-the scale-free conceptualization hypothesis. Syst. Res. Behav. Sci. **30**(6), 716–734 (2013)

4. Aranda-Corral, G.A., Borrego-Díaz, J., Giráldez-Cru, J.: Agent-mediated shared conceptualizations in tagging services. Multimed. Tools Appl. **65**(1), 5–28 (2013)
5. Barabási, A.L., Réka, A.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
6. Chli, M., de Wilde, P.: Convergence and Knowledge Processing in Multi-Agent Systems, 1st edn. Springer, London (2009)
7. Galán Páez, J., Borrego-Díaz, J.: Discovering new sentiments from the social web. CoRR (2014). arXiv:abs/1407.0374
8. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
9. Hadzibeganovic, T., Stauffer, D., Han, X.-P.: Randomness in the evolution of cooperation. Behav. Process. **113**, 86–93 (2015)
10. Kaur, R., Kumar, R., Bhondekar, A.P., Kapur, P.: Human opinion dynamics: an inspiration to solve complex optimization problems. Sci. Rep. **3** (2013)
11. Kumar, K.P., Geethakumari, G.: Detecting misinformation in online social networks using cognitive psychology. Hum. Centric Comput. Inf. Sci. **4**(1), 14 (2014)
12. Loreto, V., Baronchelli, A., Mukherjee, A., Puglisi, A., Tria, F.: Statistical physics of language dynamics. J. Stat. Mech. Theory Exp. **2011**(04), P04006 (2011)
13. Lozano, S.: Dynamics of social complex networks: some insights into recent research. In: Ganguly, N., Deutsch, A., Mukherjee, A. (eds.) Dynamics On and Of Complex Networks, pp. 133–143. Birkhäuser, Boston (2009)
14. Lu, Q., Korniss, G., Szymanski, B.: The naming game in social networks: community formation and consensus engineering. J. Econ. Interact. Coord. **4**(2), 221–235 (2009)
15. Luxenburger, M.: Implications partielles dans un contexte. Mathématiques, Informatique et Sciences Humaines **29** (1991)
16. Steels, L.: A self-organizing spatial vocabulary. Artif. Life **2**(3), 319–332 (1995)
17. Steels, L., McIntyre, A.: Spatially distributed naming games. Adv. Complex Syst. **1**(4), 301–323 (1999)