Re-conceptualizing Latent Semantic Analysis in Terms of Complex Network Theory. A Corpus-Linguistic Approach

Alexander Mehler & Lorenz Sichelschmidt
Computational Linguistics & Cognitive Linguistics
Bielefeld University, D-33615 Bielefeld, Germany
{Alexander.Mehler,Max.Sichelschmidt}@uni-bielefeld.de

Abstract

Recently, the small world phenomenon has been investigated by example of networks of lexical and textual units. These analyses show a remarkable conformity of the topology of social and biological networks on the one hand and linguistic networks on the other hand. More specifically, this relates to their macroscopic organization as characterized by high cluster values as well as by short average geodesic distances between any randomly chosen pair of nodes in these networks. In this paper, we motivate the development of a model of semantic spaces in accordance with these findings. We also propose to explore necessary conditions to be fulfilled by corpora in order to be judged as reliable data bases for computing lexical network models showing the SW property on their own. In this sense, we aim at contributing to a cognitive linguistic grounding of explorative corpus data analyses.

1 Semantic Spaces in the Light of the Small World-Phenomenon

A central question of cognitive science concerns the kind of memory structure which supports efficient memory organization in the sense of time and space complexity. Recently, Steyvers and Tenenbaum (2005) interpreted the small-world property of association networks as an indicator of efficient information storage and retrieval. They apply *complex network theory* (Newman, 2003) which, for the time being, investigates network topologies in terms of their *Small World* (SW) characteristics (Watts and Strogatz, 1998): Firstly, compared to random graphs, SW graphs have a considerably higher amount of cluster formation. Secondly, compared to regular graphs, any randomly chosen pair of nodes of a SW graph has, on average, a considerably shorter geodesic distance.[1] Steyvers and Tenenbaum (2005) show that lexical association networks as well as reference systems like WordNet share these properties: Whereas the existence of many local clusters is seen to enable effective associations, the existence of short average geodesic distances is seen to guarantee fast information access and search since, under this condition, any information units are separated only by a couple of associations – *irrespective how different they are*. Cf. Motter et al. (2002) and Sigman and Cecchi 2002 who, likewise, interpret the small-world property of association networks as an indicator of efficient information storage and retrieval.

Beyond their clustering and short average geodesic distances, SWs are characterized by the out degrees of their nodes. Barabási and Albert (1999) show that the connectivity of nodes in social-semiotic networks is distributed according to a scale-free power law. They recur to the observation – confirmed by many social-semiotic networks, but not by random graphs (Bollobás, 1985) – that the probability $P(k)$ that a randomly chosen vertex interacts with k other vertices of the same network is approximately $k^{-\gamma}$. According to Steyvers and Tenenbaum (2005), lexical association networks and reference systems have this SW property too.

A central implication of these findings is that they question the cognitive plausibility and adequacy of *Latent Semantic Analysis* (LSA) (Lan-

[1] The geodesic distance of two vertices in a graph is the length of the shortest path in-between.

In: Cognitive-linguistic approaches: What can we gain by computational treatment of data? A theme session at DGKL-06/GCLA-06 (Meeting of the German Cognitive Linguistics Association). 7 October 2006, Munich, Germany.

dauer and Dumais, 1997) and related models which rely on semantic spaces in terms of completely connected weighted graphs as the underlying memory representation format. LSA has been proposed as an approach to learning contiguity and similarity relations of linguistic units from corpora of natural language texts. It has been applied in many areas of computational and cognitive linguistics in order to disambiguate lexical items (Schütze, 1997) or to quantify textual coherence (Foltz et al., 1998). Further, LSA has been integrated into the construction-integration theory (Kintsch, 1998) and is used to automatically learn the meanings of predications and metaphorical speech (Kintsch, 2001).

LSA starts from the hypothesis that the similarity relations of cognitive units result from a *two-stage process of inductive learning* operating on the units' contiguity relations. In the case of lexical items, contiguity relations are equated with co-occurrences. The learning of lexical similarity relations is described as a process of dimensionality reduction which may detect similarities even if the items do not or only rarely co-occur. This and related models (Rieger, 1989; Mehler, 2006b) follow the so called *weak contextual hypothesis* (Miller and Charles, 1991) which says that the similarity of the contextual representations of words contributes to their semantic similarity.

In formal, mathematical terms, a semantic space, whether based on LSA or some related model of semantic spaces, can be described as follows – for the details of these definitions cf. Mehler (2006a) (cf. Leopold (2005) for a related definition):

Definition 1 *Let $C = \{x_1, \ldots, x_n\}$ be a text corpus, $\mathbb{S}$ a segmentation mapping each text $x \in C$ onto an ordered rooted tree $\mathbb{S}(x) = (S(x), E, x, O_1, O_2)$ as a model of its kernel hierarchical structure in the sense of an ordered hierarchy of content objects Renear et al. (1996) and $\mathbb{L}\colon T(C) \to L(C)$ a lemmatization mapping each token $\boldsymbol{a} \in T(C)$ onto its type $a \in L(C)$; $T(C) \subset S(C)$ is the set of tokens and $L(C)$ the set of types of corpus C. O_1 is an order relation mapping the syntagmatic order of all immediate constituents of any segment of x. That is, $O_1(y_i, y_j)$ iff $y_i, y_j \in S(x)$ are immediate constituents of the same $z \in S(x)$ according to $\mathbb{S}$ so that y_i precedes y_j in z. O_2 is the linear order relation induced by the postorder traversal of $\mathbb{S}(x)$.*

We define $S(x)$, $x \in S(x)$, as the set of all segments of x according to $\mathbb{S}$ and $S(C) = \cup_{x \in C} S(x)$. Further, $T(x) \subset S(x)$ is the set of all tokens of x according to $\mathbb{S}$ and $T(C) = \cup_{x \in C} T(x)$. Next, $L(x) = \{a \,|\, \exists \boldsymbol{a} \in T(x)\colon \boldsymbol{a} \models_T a\}$ is the set of all types classifying at least one token in $T(x)$. Thus, $L(C) = \cup_{x \in C} L(x)$. We write S, T and L instead of $S(C)$, $T(C)$ and $L(C)$ if the corpus C is known from the context.

Now we can give an abstract definition of semantic spaces which grasps the varying space models of fuzzy linguistics (Rieger, 1989), LSA (Landauer and Dumais, 1997) and derivations thereof (e.g. Burgess et al. 1999) leaving out the details of their computation:

Definition 2 *Let a corpus C, a segmentation $\mathbb{S}$ and a lemmatization $\mathbb{L}$ be given according to definition (1). Further, let $\mathbb{X}$ be an uncountable set, e.g. $\mathbb{X} = \mathbb{R}^n$ for some $n > 0, n \in \mathbb{N}$, and $(\mathbb{X}, d)$ be a metric space. A semantic space is a quintuple $(L, S, \alpha, \beta, (\mathbb{X}, d))$ where $\alpha : L \to \mathbb{X}$ is a function mapping types $a \in L$ onto representations of the contexts of their tokens $\boldsymbol{a} \in L$ in segments $x \in S$. Further, $\beta\colon S \to \mathbb{X}$ is a function mapping segments $x \in S$ onto $\mathbb{X}$ by operating on the context representations of their components according to $\mathbb{S}$ down to the level of tokens $\boldsymbol{a} \in T(x)$ as instances of types $a \in L(x)$.*

This definition relates to completely connected weighted graphs as follows:

Definition 3 *Let $(L, S, \alpha, \beta, (\mathbb{X}, d))$ be a semantic space according to definition (2). Let further $Z = \{z \in \mathbb{X} \,|\, \exists a \in L : \alpha(a) = z \vee \exists x \in S : \beta(x) = z\}$ be the set of all meaning points assigned to signs in $L \cup S$. The completely connected, weighted, undirected graph $G = \langle V, E, \omega \rangle$ induced by this semantic space is a graph with the set of vertices $V = Z$, the set of edges $E = \{\{x, y\} \,|\, x, y \in Z, x \neq y\}$, where $|E| = \frac{n(n-1)}{2}$. Further, $\omega : E \to [0, 1]$ is a weighting function with $\omega(\{x, y\}) = 1 - d(x, y)/Max(d)$ for all $\{x, y\} \in E$. $Max(d)$ is the maximum value assumed by d.*

These definitions motivate that semantic spaces naturally induce graph models in terms of completely connected graphs in which every vertex is directly connected with every other vertex of the same graph where the weight of the corresponding edge is determined by the metric d. Obviously, all vertices of such a graph trivially have the same

maximum cluster value as the graph as a whole has a minimum average geodesic distance with zero variance.

If Steyvers and Tenenbaum (2005) are right, LSA and related models of semantic spaces are, at least in its present form, cognitively implausible. In spite of its grounding in terms of a two-stage process of inductive learning, LSA outputs semantic spaces which – in the form of completely weighted graphs – lack the SW property. In other words, although LSA operates in a corpus analytic, unsupervised fashion, it is inadequate in terms of the space complexity of its representation format. A simple way to avoid complete linkage in semantic space would be to introduce a threshold ϵ in order to delete edges for which $\omega(\{x, y\}) < \epsilon$. Obviously, this does not automatically guarantee the SW property of the resulting graph so that a more sophisticated procedure is needed to generate semantic spaces in accordance with the SW-model.

2 Toward Reconstructing the Semantic Space Approach

Following this line of argumentation, Steyvers and Tenenbaum have demonstrated how to utilize complex network theory in order to derive criteria for judging the cognitive plausibility of corpus-based models of lexical association. This raises the question about a computational linguistic model *which combines the corpus-analytic stance of LSA with a cognitively more adequate representation format as an alternative to the predominant model of semantic spaces*. The present paper focuses on this question. Starting from Steyvers and Tenenbaum's findings it *additionally* asks for the SW-like networking of the underlying text corpora. From a corpus linguistic point of view, the SW property of text networks can be seen as an argument in favor of representative samples as input to computing cognitively plausible models of lexical association. Although it is known from quantitative linguistics that such samples are hardly possible in linguistic statistics – cf., for example, Orlov (1982) – this property can at least be utilized as a necessary condition which has to be fulfilled by corpora in order to be judged as reliable data bases for computing lexical memory models showing the SW property on their own.

Consequently, we propose to reconstruct LSA in terms of SW-like networking of lexical *and* of

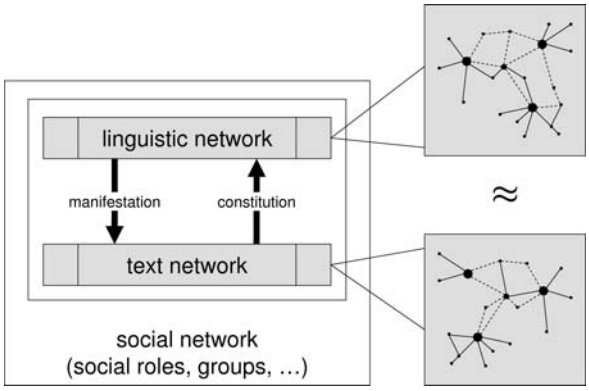

Figure 1: A three-level model of linguistic systematic, textual and social networking.

textual units. That is, we propose a model of LSA which obeys SW characteristics on the level of lexical and textual semantics. This can be done by means of semantic spaces which are no longer based on completely weighted, but on SW graphs. The basic idea of this model is that the SW property of lexical association networks is an epiphenomenon of the SW property of networking within corpora of natural language texts whose linkage is due to genre- and topic-based intertextual relations (Mehler, 2006c).

As illustrated by the three-layer model in Figure (1), our starting point is, so to speak, a *correlation* of SW-like networking on the lexical-semantic level as well as on the level of the textual network manifesting the latter network. In terms of complex network theory (Watts, 2003), this is tantamount to a bipartite graph model in which the top-mode as well as the bottom-mode are small world-networks on their own.

3 Conclusion

In summary, this paper pleads for a model of semantic spaces which combines restrictions on corpus internal networking with constraints on the networking of the relations of cognitive units derived from these corpora. It focuses on the validity of lexical association models from the point of view of their time and space complexity and, vice versa, sheds light on the validity of computational models from the point of view of their SW characteristics. A central outcome of the paper is that complex network theory allows to derive *necessary conditions* which have to be fulfilled by corpora of natural language texts in order to be a reliable input to computing lexical associations

thereof.

References

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

Barwise, J. and Seligman, J. (1997). *Information Flow. The Logic of Distributed Systems*. University Press, Cambridge.

Bollobás, B. (1985). *Random Graphs*. Academic Press, London.

Burgess, C., Livesay, K., and Lund, K. (1999). Exploration in context space: Words, sentences, discourse. *Discourse Processes*, 25(2&3):211–257.

Foltz, P. W., Kintsch, W., and Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

Kintsch, W. (1998). *Comprehension. A Paradigm for Cognition*. Cambridge University Press, Cambridge.

Kintsch, W. (2001). Predication. *Cognitive Science*, 25:173–202.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Leopold, E. (2005). On semantic spaces. *LDV Forum*, 20(1):63–86.

Mehler, A. (2006a). Compositionality in quantitative semantics. a theoretical perspective on text mining. In Mehler, A. and Köhler, R., editors, *Aspects of Automatic Text Analysis*, Studies in Fuzziness and Soft Computing, Berlin. Springer.

Mehler, A. (2006b). Stratified constraint satisfaction networks in synergetic multi-agent simulations of language evolution. In Loula, A., Gudwin, R., and Queiroz, J., editors, *Artificial Cognition Systems*, pages 140–174. Idea Group Inc., Hershey.

Mehler, A. (2006c). Text linkage in the wiki medium – a comparative study. In *Proceedings of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources, Trento, Italy, April 3-7*.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C., and Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E*, 65(065102).

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

Orlov, J. K. (1982). Dynamik der Häufigkeitsstrukturen. In Orlov, J. K., Boroda, M. G., and Š. Nadarejšvili, I., editors, *Sprache, Text, Kunst. Quantitative Analysen*, pages 82–117. Brockmeyer, Bochum.

Renear, A., Mylonas, E., and Durand, D. (1996). Refining our notion of what text really is: The problem of overlapping hierarchies. In Ide, N. and Hockey, S., editors, *Research in Humanities Computing*, pages 263–280. Oxford University Press, Oxford.

Rieger, B. B. (1989). *Unscharfe Semantik: Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*. Peter Lang, Frankfurt a.M.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*, volume 71 of *CSLI Lecture Notes*. CSLI Publications, Stanford.

Sigman, M. and Cecchi, G. (2002). Global organization of the WordNet lexicon. *Proceedings of the National Academy of Sciences*, 99(3):1742–1747.

Steyvers, M. and Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Watts, D. J. (2003). *Six Degrees. The Science of a Connected Age*. W. W. Norton & Company, New York/London.

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