Cross-lingual comparison between distributionally determined word similarity networks

Olof Görnerup
Swedish Institute of Computer Science (SICS)
164 29 Kista, Sweden
olofg@sics.se

Jussi Karlgren
Swedish Institute of Computer Science (SICS)
164 29 Kista, Sweden
jussi@sics.se

Abstract

As an initial effort to identify universal and language-specific factors that influence the behavior of distributional models, we have formulated a distributionally determined word similarity network model, implemented it for eleven different languages, and compared the resulting networks. In the model, vertices constitute words and two words are linked if they occur in similar contexts. The model is found to capture clear isomorphisms across languages in terms of syntactic and semantic classes, as well as functional categories of abstract discourse markers. Language specific morphology is found to be a dominating factor for the accuracy of the model.

1 Introduction

This work takes as its point of departure the fact that most studies of the distributional character of terms in language are language specific. A model or technique—either geometric (Deerwester et al., 1990; Finch and Chater, 1992; Lund and Burgess, 1996; Letsche and Berry, 1997; Kanerva et al., 2000) or graph based (i Cancho and Solé, 2001; Widdows and Dorow, 2002; Biemann, 2006)—that works quite well for one language may not be suitable for other languages. A general question of interest is then: What strengths and weaknesses of distributional models are universal and what are language specific?

In this paper we approach this question by formulating a distributionally based network model, apply the model on eleven different languages, and then compare the resulting networks. We compare the networks both in terms of global statistical properties and local structures of word-to-word relations of linguistic relevance. More specifically, the generated networks constitute words (vertices) that are connected with edges if they are observed to occur in similar contexts. The networks are derived from the Europarl corpus (Koehn, 2005)—the annotated proceedings of the European parliament during 1996-2006. This is a parallel corpus that covers Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish.

The objective of this paper is not to provide a extensive comparison of how distributional network models perform in specific applications for specific languages, for instance in terms of benchmark performance, but rather to, firstly, demonstrate the expressive strength of distributionally based network models and, secondly, to highlight fundamental similarities and differences between languages that these models are capable of capturing (and fail in capturing).

2 Methods

We consider a base case where a context is defined as the preceding and subsequent words of a focus word. Word order matters and so a context forms a word pair. Consider for instance the following sentence[1]:

> *Ladies and gentlemen, once again, we see it is essential for Members to bring their voting cards along on a Monday.*

Here the focus word *essential* occurs in the context *is* $*$ *for*, the word *bring* in the context *to* $*$ *their* etcetera (the asterisk $*$ denotes an intermediate focus word). Since a context occurs with a word with a certain probability, each word w_i is associated with a probability distribution of contexts:

$$P_i = \{\Pr[w_p w_i w_s | w_i]\}_{w_p, w_s \in \mathcal{W}}, \qquad (1)$$

[1] Quoting Nicole Fontaine, president of the European Parliament 1999-2001, from the first session of year 2000.

Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 48–54,
Uppsala, Sweden, 16 July 2010.

where $\mathcal{W}$ denotes the set of all words and $\Pr[w_p w_i w_s | w_i]$ is the conditional probability that context $w_p * w_s$ occurrs, given that the focus word is w_i. In practice, we estimate P_i by counting the occurring contexts of w_i and then normalizing the counts. Context counts, in turn, were derived from trigram counts. No pre-processing, such as stemming, was performed prior to collecting the trigrams.

2.1 Similarity measure

If two words have similar context distributions, they are assumed to have a similar function in the language. For instance, it is reasonable to assume that the word "salt" to a higher degree occurs in similar contexts as "pepper" compared to, say, "friendly". One could imagine that a narrow 1+1 neighborhood only captures fundamental syntactic agreement between words, which has also been argued in the literature (Sahlgren, 2006). However, as we will see below, the intermediate two-word context also captures richer word relationships.

We measure the degree of similarity by comparing the respective context distributions. This can be done in a number of ways. For example, as the Euclidian distance (also known as L_2 divergence), the Harmonic mean, Spearman's rank correlation coefficient and the Jensen-Shannon divergence (information radius). Here we quantify the difference between two words w_i and w_j, denoted d_{ij}, by the variational distance (or L_1 divergence) between their corresponding context distributions P_i and P_j:

$$d_{ij} = \sum_{c \in \mathcal{C}} |P_i(X = c) - P_j(X = c)|, \quad (2)$$

where X is a stochastic variable drawn from $\mathcal{C}$, which is the set of contexts that either w_i or w_j occur in. $0 \leq d_{ij} \leq 2$, where $d_{ij} = 0$ if the two distributions are identical and $d_{ij} = 2$ if the words do not share any contexts at all. It is not obvious that the variational distance is the best choice of measure. However, we chose to employ it since it is a well-established and well-understood statistical measure; since it is straightforward and fast to calculate; and since it appears to be robust. To compare, we have also tested to employ the Jensen-Shannon divergence (a symmetrized and smoothed version of Kullback information) and acquire very similar results as those presented here. In fact, this is expected since the two measures are found to be approximately linearly related in this context. However, for the two first reasons listed above, the variational distance is our divergence measure of choice in this study.

2.2 Network representation

A set of words and their similarity relations are naturally interpreted as a weighted and undirected network. The vertices then constitute words and two vertices are linked by an edge if their corresponding words w_i and w_j have overlapping context sets. The strength of the links vary depending on the respective degrees of word similarities. Here the edge between two words w_i and w_j's is weighted with $w_{ij} = 2 - d_{ij}$ (note again that $\max_{ij} d_{ij} = 2$) since a large word difference implies a weak link and vice versa.

In our experiment we consider the 3000 most common words, excluding the 19 first ones, in each language. To keep the data more manageable during analysis we employ various thresholds. Firstly, we only consider context words that occur five times or more. As formed by the remaining context words, we then only consider trigrams that occur three times or more. This allows us to cut away a large chunk of the data. We have tested to vary these thresholds and the resulting networks are found to have very similar statistical properties, even though the networks differ by a large number of very weak edges.

3 Results

3.1 Degree distributions

The degree g_i of a vertex i is defined as the sum of weights of the edges of the vertex: $g_i = \sum w_{ij}$. The degree distribution of a network may provide valuable statistical information about the networks structure. For the word networks, Figure 1, the degree distributions are all found to be highly right-skewed and have longer tails than expected from random graphs (Erdős and Rényi, 1959). This characteristics is often observed in complex networks, which typically also are scale-free (Newman, 2003). Interestingly, the word similarity networks are not scale-free as their degree distributions do no obey power-laws: $\Pr(g) \sim g^{-\alpha}$ for some exponent α. Instead, the degree distributions of each word network appears to lay somewhere between a power-law distribution and an exponential distribution ($\Pr(g) \sim e^{-g/\kappa}$). However, due to quite noisy statistics it is difficult to reliably

measure and characterize the tails in the word networks. Note that there appears to be a bump in the distributions for some languages at around degree 60, but again, this may be due to noise and more data is required before we can draw any conclusions. Note also that the degree distribution of Finnish stands out: Finnish words typically have less or weaker links than words in the other languages. This is reasonably in view of the special morphological character of Finnish compared to Indo-European languages (see below).

3.2 Community structures

The acquired networks display interesting global structures that emerge from the local and pairwise word to word relations. Each network form a single strongly connected component. In other words, any vertex can be reached by any other vertex and so there is always a path of "associations" between any two words. Furthermore, all word networks have significant community structures; vertices are organized into groups, where there are higher densities of edges within groups than between them. The strength of community structure can be quantified as follows (Newman and Girvan, 2004): Let $\{v_i\}_{i=1}^{n}$ be a partition of the set of vertices into n groups, r_i the fraction of edge weights that are internal to v_i (i.e. the sum of internal weights over the sum of all weights in the network), and s_i the fraction of edge weights of the edges starting in v_i. The modularity strength is then defined as

$$Q = \sum_{i=1}^{n}(r_i - s_i^2). \quad (3)$$

Q constitutes the fraction of edge weights given by edges in the network that link vertices within the same communities, minus the expected value of the same quantity in a random network with the same community assignments (i.e. the same vertex set partition). There are several algorithms that aim to find the community structure of a network by maximizing Q. Here we use an agglomerative clustering method by Clauset (2005), which works as follows: Initialize by assigning each vertex to its own cluster. Then successively merge clusters such that the positive change of Q is maximized. The procedure is repeated as long as Q increases.

Typically Q is close to 0 for random partitions and indicates strong community structure when approaching its maximum 1. In practice Q is typically within the range 0.3 to 0.7, also for highly modular networks (Newman and Girvan, 2004). As can be seen in Table 1, all networks are highly modular, although the degree of modularity varies between languages. Greek in particular stands out. However, the reason for this remains an open question that requires further investigations.

Dutch	0.43	Swedish	0.58
German	0.43	French	0.63
Spanish	0.48	Finnish	0.68
Portuguese	0.51	Italian	0.68
English	0.53	Greek	0.78
Danish	0.55		

Table 1: Community modularity.

Communities become more apparent when edges are pruned by a threshold as they crystalize into isolated subgraphs. This is exemplified for English in Figure 2.

4 Discussion

We examine the resulting graphs and show in this section through some example subgraphs how features of human language emerge as charactersitics of the model.

4.1 Morphology matters

Morphology is a determining and observable characteristic of several languages. For the purposes of distributional study of linguistic items, morphological variation is problematic, since it splits one lexical item into several surface realisations, requiring more data to perform reliable and robust statistical analysis. Of the languages studied in this experiment, Finnish stands out atypical through its morphological characteristics. In theory, Finnish nouns can take more than 2 000 surface forms, through more than 12 cases in singular and plural as well as possessive suffixes and clitic particles (Linden and Pirinen, 2009), and while in practice something between six and twelve forms suffice to cover about 80 per cent of the variation (Kettunen, 2007) this is still an order of magnitude more variation than in typical Indo-European languages such as the others in this sample. This variation is evident in Figure 1—Finnish behaves differently than the Indo-European languages in the sample: as each word is split in several other surface forms, its links to other forms will be weaker. Morphological analysis, transforming surface forms to base forms

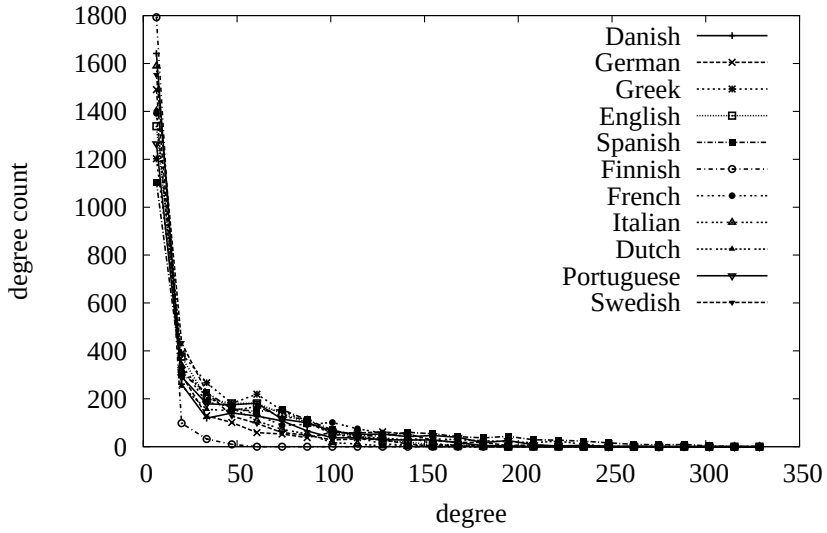

Figure 1: Degree histograms of word similarity networks.

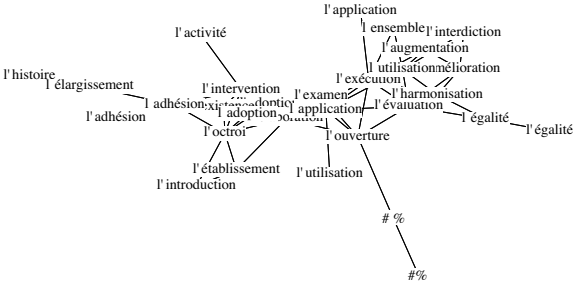

Figure 3: French definite nouns clustered.

would strengthen those links.

In practice, the data sparsity caused by morphological variation causes semantically homogenous classes to be split. Even for languages such as English and French, with very little data variation we find examples where morphological variation causes divergence as seen in Figure 3, where French nouns in definite form are clustered. It is not surprising that certain nouns in definite form assume similar roles in text, but the neatness of the graph is a striking exposition of this fact.

These problems could have been avoided with better preprocessing—simple such processing in the case of English and French, and considerably more complex but feasible in the case of Finnish—but are retained in the present example as proxies for the difficulties typical of processing unknown languages. Our methodology is robust even in face of shoddy preprocessing and no knowledge of the morphological basis of the target language. In general, as a typological fact, it is reasonable to assume that morphological variation is offset for the language user in a greater freedom in choice of word order. This would seem to cause a great deal of problems for an approach such as the present one, since it relies on the sequential organisation of symbols in the signal. However, it is observable that languages with free word order have preferred unmarked arrangements for their sentence structure, and thus we find stable relationships in the data even for Finnish, although weaker than for the other languages examined.

4.2 Syntactic classes

Previous studies have shown that a narrow context window of one neighour to the left and one neighbour to the right such as the one used in the present experiments retrieves syntactic relationships (Sahlgren, 2006). We find several such examples in the graphs. In Figure 2 we can see subgraphs with past participles, auxiliary verbs, progressive verbs, person names.

4.3 Semantic classes

Some of the subgraphs we find are models of clear semantic family resemblance as shown in Figure 4. This provides us with a good argument for blurring the artificial distinction between syntax and semantics. Word classes are defined by their meaning and usage alike; the *a priori* distinction between classification by function such as auxiliary verbs given above and classification by meaning such months and places given here is not fruitful. We expect to be able to provide much more in-

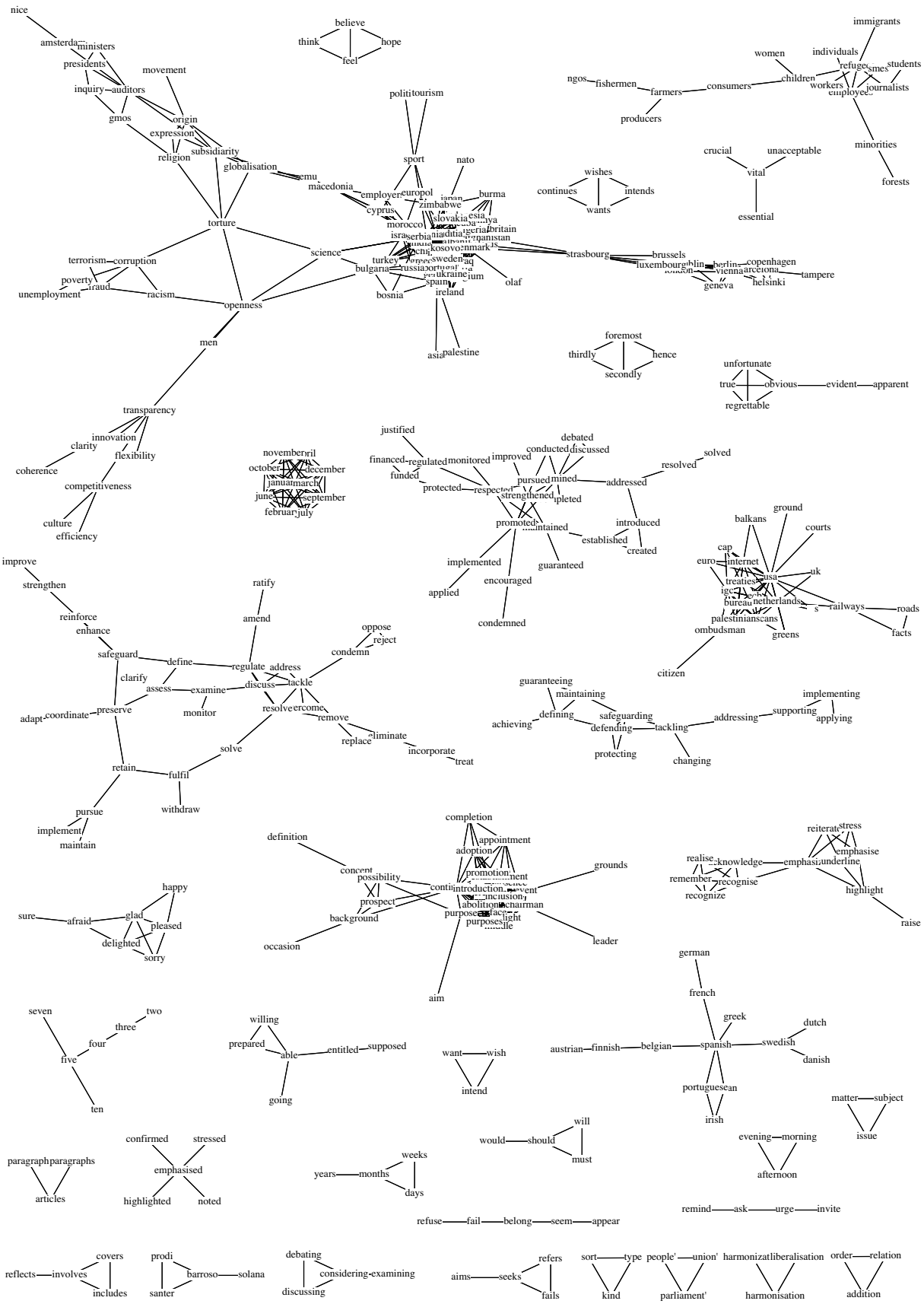

Figure 2: English. Network involving edges with weights $w \geq 0.85$. For sake of clarity, only subgraphs with three or more words are shown. Note that the threshold 0.85 is used only for the visualization. The full network consists of the 3000 most common words in English, excluding the 19 most common ones.

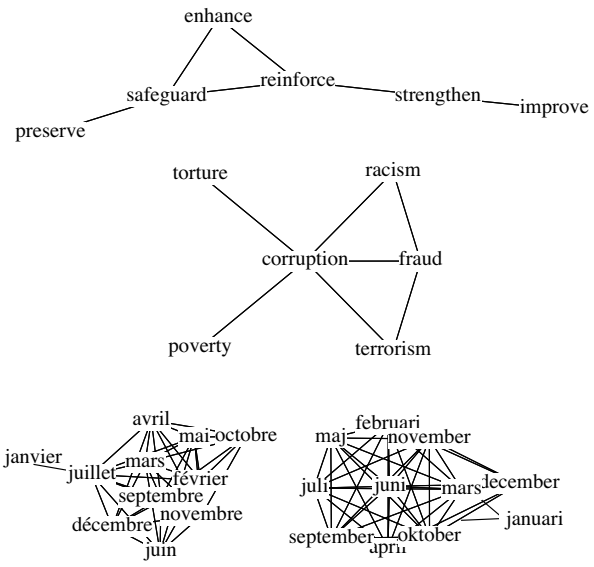

Figure 4: Examples of semantically homogenous classes in English, French and Swedish.

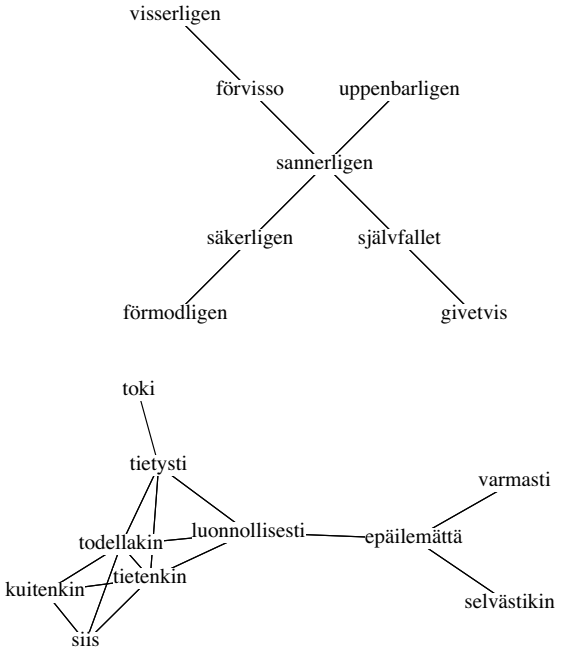

Figure 5: Examples of discourse functional classes in Swedish and Finnish. The terms in the two subgraphs are discourse markers and correspond to English “certainly”, “possibly”, “evidently”, “naturally”, “absolutely”, “hence” and similar terms.

formed classification schemes than the traditional “parts of speech” if we define classes by their distributional qualities rather than by the “content” they “represent”, schemes which will cut across the function-topic distinction.

4.4 Abstract discourse markers are a functional category

Further, several subgraphs have clear collections of discourse markers of various types where the terms are markers of informational organisation in the text, as exemplified in Figure 5.

5 Conclusions

This preliminary experiment supports future studies to build knowledge structures across languages, using distributional isomorphism between linguistic material in translated or even comparable corpora, on several levels of abstraction, from function words, to semantic classes, to discourse markers. The isomorphism across the languages is clear and incontrovertible; this will allow us to continue experiments using collections of multilingual materials, even for languages with relatively little technological support. Previous studies show that knowledge structures of this type that are created in one language show considerable isomorphism to knowledge structures created in another language if the corpora are comparable (Holmlund et al., 2005). Holmlund et al show how translation equivalences can be established using two semantic networks automatically created in two languages by providing a relatively limited set of equivalence relations in a translation lexicon. This study supports those findings.

The results presented here display the potential of distributionally derived network representations of word similarities. Although geometric (vector based) and probabilistic models have proven viable in various applications, they are limited by the fact that word or term relations are constrained by the geometric (often Euclidian) space in which they live. Network representations are richer in the sense that they are not bound by the same constraints. For instance, a polyseme word (“may” for example) can have strong links to two other words (“might” and ”September” for example), where the two other words are completely unrelated. In an Euclidean space this relation is not possible due to the triangle inequality. It is possible to embed a network in a geometric space, but this requires a very high dimensionality which makes the representation both cumbersome and inefficient in terms of computation and memory. This has been addressed by coarse graining or dimension reduction, for example by means of singular value de-

composition (Deerwester et al., 1990; Letsche and Berry, 1997; Kanerva et al., 2000), which results in information loss. This can be problematic, in particular since distributional models often face data sparsity due to the curse of dimensionality. In a network representation, such dimension reduction is not necessary and so potentially important information about word or term relations is retained.

The experiments presented here also show the potential of moving from a purely probabilistic model of term occurrence, or a bare distributional model such as those typically presented using a geometric metaphor, in that it affords the possibility of abstract categories inferred from the primary distributional data. This will give the possibility of further utilising the results in studies, e.g. for learning syntactic or functional categories in more complex constructional models of linguistic form. Automatically establishing lexically and functionally coherent classes in this manner will have bearing on future project goals of automatically learning syntactic and semantic roles of words in language. This target is today typically pursued relying on traditional lexical categories which are not necessarily the most salient ones in view of actual distributional characteristics of words.

Acknowledgments: OG was supported by Johan and Jacob Söderberg's Foundation. JK was supported by the Swedish Research Council.

References

Chris. Biemann. 2006. Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, New York, USA.

Aaron Clauset. 2005. Finding local community structure in networks. *Physical Review E*, 72:026132.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Pal Erdős and Alfréd Rényi. 1959. On random graphs. *Publications Mathematicae*, 6:290.

Steven Finch and Nick Chater. 1992. Bootstrapping syntactic categories. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 820–825, Bloomington, IN. Lawrence Erlbaum.

Jon Holmlund, Magnus Sahlgren, and Jussi Karlgren. 2005. Creating bilingual lexica using reference wordlists for alignment of monolingual semantic vector spaces. In *Proceedings of 15th Nordic Conference of Computational Linguistics*.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The small world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268:2261–2266.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–6.

Kimmo Kettunen. 2007. Management of keyword variation with frequency based generation of word forms in ir. In *Proceedings of SIGIR 2007*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Todd Letsche and Michael Berry. 1997. Large-scale information retrieval with latent semantic indexing. *Information Sciences*, 100(1-4):105–137.

Krister Linden and Tommi Pirinen. 2009. Weighting finite-state morphological analyzers using hfst tools. In *Proceedings of the Finite-State Methods and Natural Language Processing*. Pretoria, South Africa.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Mark Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2).

Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD Dissertation, Department of Linguistics, Stockholm University.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *In 19th International Conference on Computational Linguistics*, pages 1093–1099.