Lindh-Knuutila and Honkela *Computational Cognitive Science* (2015) 1:2
DOI 10.1186/s40469-015-0001-1

Computational Cognitive Science
a SpringerOpen Journal

RESEARCH Open Access

Exploratory analysis of semantic categories: comparing data-driven and human similarity judgments

Tiina Lindh-Knuutila[1,3*] and Timo Honkela[2,4,3]

Abstract

Background: In this article, automatically generated and manually crafted semantic representations are compared. The comparison takes place under the assumption that neither of these has a primary status over the other. While linguistic resources can be used to evaluate the results of automated processes, data-driven methods are useful in assessing the quality or improving the coverage of hand-created semantic resources.

Methods: We apply two unsupervised learning methods, Independent Component Analysis (ICA), and probabilistic topic model at word level using Latent Dirichlet Allocation (LDA) to create semantic representations from a large text corpus. We further compare the obtained results to two semantically labeled dictionaries. In addition, we use the Self-Organizing Map to visualize the obtained representations.

Results: We show that both methods find a considerable amount of category information in an unsupervised way. Rather than only finding groups of similar words, they can automatically find a number of features that characterize words. The unsupervised methods are also used in exploration. They provide findings which go beyond the manually predefined label sets. In addition, we demonstrate how the Self-Organizing Map visualization can be used in exploration and further analysis.

Conclusion: This article compares unsupervised learning methods and semantically labeled dictionaries. We show that these methods are able to find categorical information. In addition, they can further be used in an exploratory analysis. In general, information theoretically motivated and probabilistic methods provide results that are at a comparable level. Moveover, the automatic methods and human classifications give an access to semantic categorization that complement each other. Data-driven methods can furthermore be cost effective and adapt to a particular domain through appropriate choice of data sets.

Keywords: Text mining; Semantic modeling; Machine learning; Lexical meaning; Semantic similarity; Independent component analysis; Latent Dirichlet Allocation

Background

In this article, we explore the relationship between human and data-driven semantic similarity judgments. The general architecture of this work is presented in Figure 1. We aim to see a) whether the representations that are automatically generated in a data-driven manner coincide with manually constructed semantic categories, and b) critically assess manually constructed semantic categories and semantically annotated data using statistical machine learning and visualization methods.

Challenge of semantics

Semantics is an intriguing but also a challenging area of linguistics. Linguists and researchers in nearby disciplines have created a number of theories related to semantics. These theories have been used as frameworks for semantic description or for labeling of lexica and corpora (Cruse 1986). On the other hand, availability of large

*Correspondence: tiina.lindh-knuutila@aalto.fi
[1] Aalto University, Department of Neuroscience and Biomedical Engineering, P.O. Box 15100, 00076 AALTO, Espoo, Finland
[3] Aalto University, Department of Information and Computer Science, P.O. Box 15400, 00076 AALTO, Espoo, Finland
Full list of author information is available at the end of the article

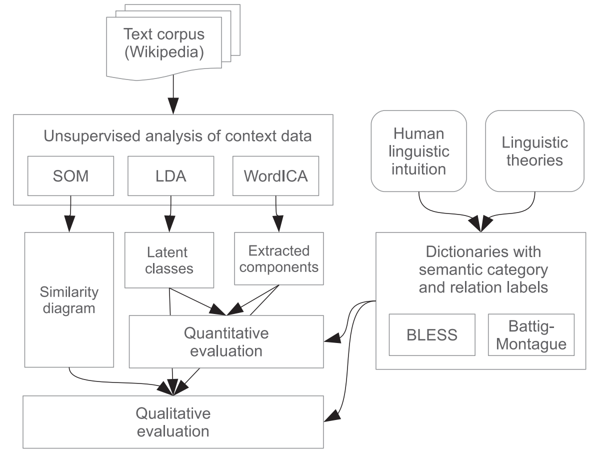


Figure 1 A process architecture for comparing data driven and human semantic judgments.

text corpora and sophisticated statistical machine learning algorithms has made it possible to automatically conduct semantically oriented analysis of corpora and lexical items (Manning and Schütze 1999).

When the objective is to create linguistic models and theories, a traditional approach is to rely on linguists' intuition and knowledge building in a community of professional linguists. In corpus linguistics, linguistic theories are usually the starting point and statistical analyses on corpus data are used to confirm, reject and refine these theories (McEnery 2001). In such a paradigm, basic linguistic categories like noun and verb are taken as given and may even be assumed to have an objective status. Similarly, when computer scientists work on some linguistic data, they very often use human-constructed categories and labels as a ground truth to evaluate the performance of the computational apparatus. For syntax, there is a large number of competing and mostly mutually incompatible theories and category systems (Rauh 2010). In recent years, some linguists have pointed out that there is no good evidence for pre-established syntactic categories that would be shared by all or a large number of languages (Haspelmath 2007).

A unidirectional view on knowledge formation within computational linguistics is problematic. There is no generally accepted theory of semantics at the level of semantic categories or primitives, even though the quest for universal primitives has been active (Goddard and Wierzbicka 2002). In general, any classification system is prone to subjective variation even among experts in the field (Johnston 1968). Some research has been conducted on modeling this subjective variation (Caramazza et al. 1976; Honkela et al. 2012). In information retrieval, it has been known for a long time that indexers are inconsistent from one to another or from one time to another (Bates 1986) and that two individuals often use different expressions to describe the same thing (Chen 1994). This kind of inherent human subjectivity should also influence semantic theories in linguistics. It is useful to view language as a complex adaptive socio-cognitive system, rather than a static system of abstract grammatical principles (Beckner et al. 2009).

Unsupervised learning of linguistic models

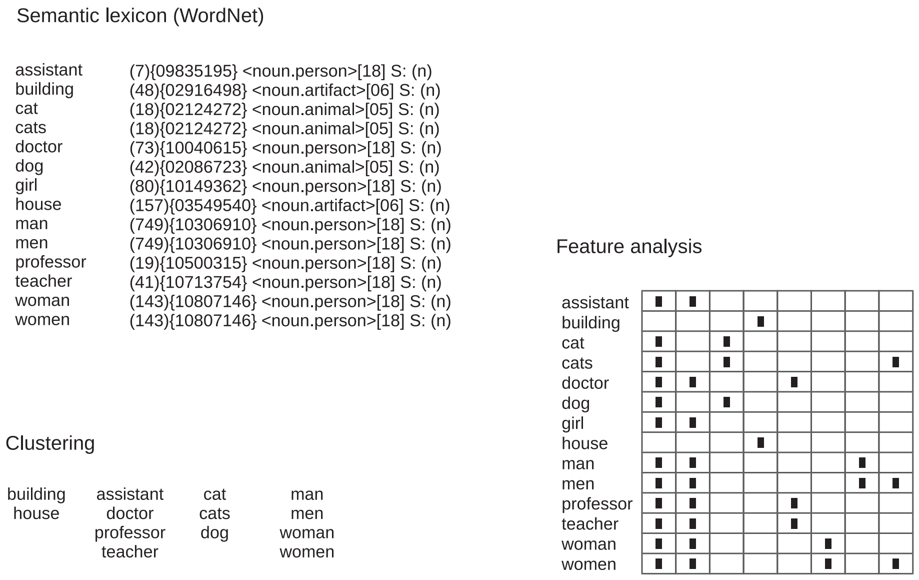
Based on what was discussed above, we must consider any semantic category system or a semantically labeled corpus as a representation which may have well motivated alternatives. Based on the availability of text and speech corpora as well as sophisticated computational tools, an increasingly popular approach is data-driven: linguistic models are created using statistical and machine learning methods.

We are particularly interested in methods that are applicable without strong linguistic assumptions. Therefore, we focus on the unsupervised learning approach rather than any supervised learning (classification) methods. More specifically, we first compare the use of Independent Component Analysis (ICA) (Hyvärinen et al. 2001) and generative topic models, in particular Latent Dirichlet Allocation (LDA) (Blei et al. 2003) in extracting automatically linguistic features in a data-driven manner. In comparison with clustering methods that also belong to the unsupervised learning methods, ICA and LDA provide an important additional advantage. Namely, they find feature representations for words, i.e., they do not simply position words to different clusters but represent words through a collection of features. In the ICA method, these emergent features are called components, whereas in the LDA model they are called topics. For example, the word 'women' could be associated with emergent categories of living things, humans and females. Furthermore, the methods can also come up with a representation where the syntactic category plural is also associated with the word 'women'. In this, like in many other cases, syntactic categories are actually related to an abstract level of meaning. The difference between clustering and feature analysis is illustrated in Figure 2.

We further analyze and visualize the data using the Self-Organizing Map (SOM) (Kohonen 2001). The SOM is widely used as a visualization method and has proven to be a viable alternative even when compared with more recent developments (Venna and Kaski 2006). We use the SOM for an analysis of special cases highlighted by the ICA and LDA analysis to reveal additional structure and to consider potential problems and ambiguities related to manually constructed semantic models.

Earlier and related work

Here the basic building blocks for this research are described including methods for vector space modeling,



assistant	(7){09835195} <noun.person>[18] S: (n)
building	(48){02916498} <noun.artifact>[06] S: (n)
cat	(18){02124272} <noun.animal>[05] S: (n)
cats	(18){02124272} <noun.animal>[05] S: (n)
doctor	(73){10040615} <noun.person>[18] S: (n)
dog	(42){02086723} <noun.animal>[05] S: (n)
girl	(80){10149362} <noun.person>[18] S: (n)
house	(157){03549540} <noun.artifact>[06] S: (n)
man	(749){10306910} <noun.person>[18] S: (n)
men	(749){10306910} <noun.person>[18] S: (n)
professor	(19){10500315} <noun.person>[18] S: (n)
teacher	(41){10713754} <noun.person>[18] S: (n)
woman	(143){10807146} <noun.person>[18] S: (n)
women	(143){10807146} <noun.person>[18] S: (n)

building	assistant	cat	man
house	doctor	cats	men
	professor	dog	woman
	teacher		women

assistant	■	■						
building				■				
cat	■		■					
cats	■		■					■
doctor	■	■			■			
dog	■		■					
girl	■	■						
house				■				
man	■	■					■	
men	■	■					■	■
professor	■	■			■			
teacher	■	■			■			
woman	■	■				■		
women	■	■				■		■

Figure 2 The difference between clusters and feature representations illustrated.

semantic similarity calculations, and unsupervised learning algorithms for linguistic processing. Earlier work in these areas is also discussed.

Word vector space model

Word vector space models (VSM) are based on (Miller and Charles 1991) a well-known hypothesis on the relationship between semantic similarity and context data: “two words are semantically similar to the extent that their contextual representations are similar” (Miller and Charles 1991). They capture meaning through word usage and are widely used in computational linguistics (Honkela et al. 2010; Landauer and Dumais 1997; Sahlgren 2006; Schütze 1993; Turney and Pantel 2000). For example, Turney and Pantel (2000) and Erk (2012) provide extensive reviews on the current state-of-the-art of vector space models. In a vector space model, it is assumed that semantic relatedness equals proximity in the vector space: related words are close, and unrelated words are distant (Schütze 1993).

The model construction takes place in several steps. First, the text data is pre-processed and feature selection can be applied. The context word frequencies are calculated, and raw frequency counts are transformed by weighting. Dimensionality reduction can be applied to smooth the space. Finally, the similarities between word vectors are calculated by using a vector distance measure (Turney and Pantel 2000).

To obtain the raw word co-occurrence count representation for N target words, the number of context words C occurring inside a window of size l positioned around each occurrence of the target word is counted. The accumulation of the occurrences of the context word in the window creates a word-co-occurrence matrix $X_{C\times N}$. The size of context around the target word affects the results. The context used can be a document, or a more immediate context around the target word. Bullinaria and Levy (2007) provide a systematic analysis on different context sizes. Sahlgren (2006) concludes that a small context around a target word gives rise to paradigmatic relations between words, whereas larger context allows syntagmatic relations to be more prominent. See also Rapp (2002) for comparisons of paradigmatic and syntagmatic relations. As the concepts in the categories are mostly in paradigmatic relationship, we use a bag-of-words representation with a window of size $l = 1 + 1$, that is, one word left and one word to the right around the target word.

Semantic similarity judgments

Similarity judgment is considered to be one of the most central functions in human cognition (Goldstone 1994). Humans use similarity to store and retrieve information, and to compare new situations to similar experiences in the past. Category learning and concept formation also depend on similarity judgment (Schwering 2008). Research has been carried out to obtain information on human similarity judgments and different types of similarity have been identified, such as synonymy (*automobile:car*), antonymy (*good:bad*), hypernymy (*vehicle:car*) and meronymy (*car:wheel*) (Cruse 1986). A special case is family resemblance, in which the members of a category are perceived as possessing some similar characteristics (VEHICLE: *car, bicycle*). Based on similarity judgment research in psychology and related fields, data sets that list words that are judged to be similar have been used to evaluate vector space models, explored for example in Baroni and Lenci (2011) and Lindh-Knuutila et al. (2012), with an intuition that the similarity perceived by humans should be translated as proximity in a word vector space. Another approach is to use a taxonomy or ontology as a basis for the similarity calculations (Seco et al. 2004). A new prominent evaluation direction is comparing corpus-derived vector representations to brain imaging results obtained with functional Magnetic Resonance Imaging (fMRI) (Mitchell et al. 2008; Murphy et al. 2012) or magnetoencephalography (MEG) (Sudre et al. 2012).

Direct vector space model evaluation concentrates on VSM performance, and measures the similarities of given words in the VSM model, and require human-annotated sources. For English, there are several such evaluation sets for analyzing the semantic similarity of the vector space models, that use synonym or antonym pairs, categories and association data (Sahlgren 2006) or separating a correct answer from the incorrect ones such as the TOEFL test set (Landauer and Dumais 1997).

General purpose algorithms for linguistic processing

In this article, we compare two methods, Independent Component Analysis (ICA) (Hyvärinen et al. 2001) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003) in the analysis of vector spaces and contextual information. In particular, we are interested in how well these methods are able to extract meaningful linguistic information in an automated fashion. Latent semantic analysis (LSA) is a very popular method that is used to analyze linguistic vector spaces (Deerwester et al. 1990; Landauer and Dumais 1997). It has been shown, however, that even though LSA is useful in applications, it fails to provide explicit representations that would be comparable to linguists' intuitions. In this task ICA is successful (Honkela et al. 2010). Now we wish to find out how the information-theoretically motivated ICA and the probabilistically motivated LDA succeed in this task. In other words, do these methods automatically find categorizations that would coincide with manually constructed semantic resources? Moreover, do these corpus based methods detect semantic similarities that have been neglected by linguists?

Terms that have been used to describe semantically related words or semantic categories that have been found using unsupervised learning methods include 'emergent category' (Honkela 1998), 'latent class' (Hofmann 1999), 'topic' (Blei et al. 2003; Steyvers and Griffiths 2007) and 'sense' (Brody and Lapata 2009). The first three can be considered to be synonymous. The term 'sense' is often used when multiple meanings of words are considered. Essentially, the phenomenon is still the same: What are the semantic distinctions that are made?

Methods

In this section, the corpus and evaluation data sets and the computational metodology are described in more detail. We begin by describing the evaluation data sets, and continue with the details of the corpus and methodological choices for building a vector space model. We then further describe the unsupervised learning methods that are used in the analysis.

Data and pre-processing

In this article, we use evaluation sets that contain information on semantic categories, that is, groups of words that are judged similar in some sense. The two test sets used in this article, the Battig set (Bullinaria 2012), based on 56 categories collected by (Battig and Montague 1969), and BLESS (Baroni and Lenci 2011) are introduced in more detail in the following sections. Other category-based evaluation sets not used in this article include the ESSLLI 2008 set (Baroni et al. 2008), which contains 44 concrete nouns that belong to six classes, and 45 verbs that belong to nine semantic classes; Baroni's category list of 83 concepts in 10 categories (Baroni et al. 2010) based on an updated version of the Battig-Montague list (Van Overschelde et al. 2004); and the Almuhareb list (Almuhareb 2006), which contains 402 concepts.

Battig set

The Battig evaluation set (Bullinaria 2012) has earlier been used, for example, in formulating and validating representations of word meanings from word co-occurrence statistics (Bullinaria and Levy 2007, 2012). The test set contains 53 categories with 10 words in each category. The total evaluation set size is 530 words, out of which 528 words are unique. The categories are listed in Table 1. The set contains the words in each category in the frequency order they are listed in (Battig and Montague 1969). All words in the set are nouns, and only two word forms have more than one label: 'orange' is labeled with FRUIT and in COLOR, and 'bicycle' with TOY, and VEHICLE. For this article, the British English spelling of some words was changed back into American English (e.g., 'millimetre'–'millimeter') to better conform to the English used in the Wikipedia corpus used in this article.

BLESS set

The second annotated vocabulary used in this article is the BLESS (Baroni-Lenci Evaluation of Semantic Spaces) (Baroni and Lenci 2011) test set, which is based on a body of earlier work on human similarity judgments. The data set contains 200 concepts in 17 broader classes or categories with 5-21 words per class. Each concept is further linked with other words that are in a certain defined relation with the concept. Attributive (ATTR)

Table 1 The Battig categories used in this article

Precious stone	Furniture	Sport	Vegetable
Unit of time	Fruit	Dance	Type of footgear
Relative	Weapon	Article of clothing	Insect
Unit of distance	Elective office	Part of a building	Girl's first name
Metal	Human dwelling	Chemical element	Male's first name
Reading material	Toy	Science	Flower
Military title	Country	Kind of money	Disease
City	Crime	Type of music	Tree
Kind of cloth	Carpenter's tool	Bird	Ship
Color	Type of fuel	Kitchen utensil	Fish
Four-footed animal	Vehicle	Part of human body	Alcoholic beverage
Nonalcoholic beverage	Substance for flavoring food	Weather phenomenon	Natural earth formation
Building for religious services	Member of the clergy	Occupation or profession	
Musical instrument	Part of speech		

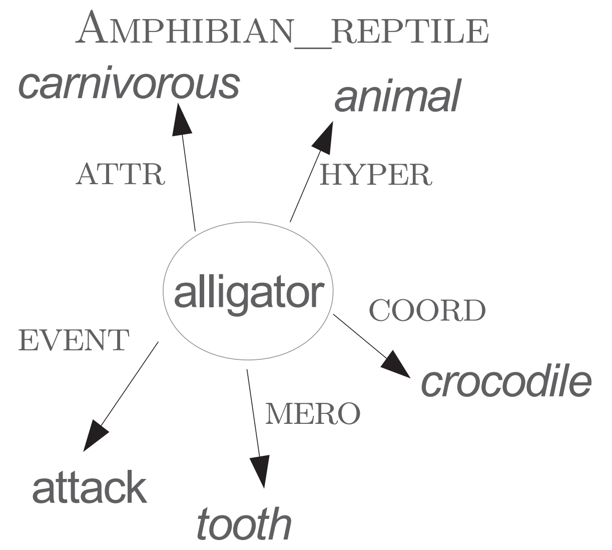


Figure 3 Example of a sample concept of 'alligator' in BLESS with words in different relations to the concept.

relation describes a property of the concept, and belongs to the class of adjectives. Coordinating concept (COORD) belongs to the same category as the given concept and is a noun. An event (EVENT) is a verb related to the concept. A word that is in a hypernymous relation (HYPER) is a superordinate concept for the word, and a meronymous relation (MERO) is in a part-whole relation with the concept. Both hypernyms and meronyms are nouns. Figure 3 gives an example of such a concept and its relations. In total, there are 14 400 word-relation pairs in the data set.

Each word in the vocabulary is labeled with a combination of the relation and the category, and multiple labels per word are allowed. For example, a word 'aeroplane' is labeled with VEHICLE-COORD and VEHICLE-HYPER, and 'back' with CLOTHING-MERO, FURNITURE-MERO, MUSICAL_INSTRUMENT-MERO, and VEHICLE-MERO. Table 2 shows the BLESS categories, the relations and number of words with each label. The sum of the words is larger than the size of vocabulary as words can have multiple labels as explained above.

Wikipedia corpus

Our corpus was built from the documents in the English Wikipedia (Wikimedia Project 2008), using the October 2008 edition, which is no longer available at the Wikipedia dump download site. A size threshold of 2 kB was used when selecting the documents to reduce the effect of empty or very short documents. In pre-processing, all non-text markup was removed, the words were lowercased and punctuation was removed except for word-internal hyphens and apostrophes. The VSM representations used are often very high-dimensional, for example 100 000 features. To reduce computational load in the ICA calculation, we opted to keep the matrix size reasonable, and used a smaller feature space: the 5 000

Table 2 The categories and relation types of the BLESS set with number of words that belong to each class

The category	Attr	Coord	Event	Hyper	Mero	Total
AMPHIBIAN_REPTILE	42	14	41	11	22	130
APPLIANCE	37	14	64	10	80	205
BIRD	39	23	42	14	14	132
BUILDING	55	21	78	18	125	297
CLOTHING	35	42	38	14	44	173
CONTAINER	33	22	41	10	50	156
FRUIT	33	22	20	5	20	100
FURNITURE	27	10	59	5	60	161
GROUND_MAMMAL	85	57	98	20	50	310
INSECT	33	16	34	7	10	100
MUSICAL_INSTRUMENT	23	23	24	7	41	118
TOOL	34	35	93	14	23	199
TREE	19	10	7	6	17	59
VEGETABLE	29	26	31	11	24	121
VEHICLE	54	31	91	16	118	310
WATER_ANIMAL	36	34	30	10	17	127
WEAPON	34	20	61	12	57	184
Total	648	420	852	190	772	

most frequent words as features. This choice is intentional, a larger feature space might improve the similarity calculation results slightly, but make the ICA and LDA calculation very long. The co-occurrence count representations were calculated for the vocabulary of the 200 000 most frequent words. This vector space model performance has been evaluated previously using several syntactic and semantic test sets (Lindh-Knuutila et al. 2012), and it is found comparable to other VSM evaluation results such as (Bullinaria and Levy 2007).

We carry out an ICA experiment with the full vocabulary of 200 000 words, but for most of the analysis, a subset of the full vector space was used that corresponds to the test set vocabulary. In the experiments with the Battig set, the 528 word vectors of the Battig vocabulary were used. In the case of the BLESS set, we use the 1 673 unique words that appear within the 200 000 most frequent words of the Wikipedia corpus. In the last visualization experiment with the SOM, a joint vocabulary of 1 997 words consisting of both the Battig and the BLESS sets was used in training.

Term weighting

A weighting scheme is often used to decrease the effect of the most frequent words in the representation. In this article, we use the positive pointwise mutual information (PPMI) (Niwa and Nitta 1994) weighting scheme (Eq. 1), which was reported to give best results in Bullinaria and Levy (2007). The positive pointwise mutual information is given by using only the non-negative values of the pointwise mutual information:

$$ppmi_{ij} = \begin{cases} pmi_{ij} = \log \frac{p_{ij}}{p_{i*}p_{*j}}, & \text{if } pmi_{ij} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}$, $p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}$, $p_{*j} = \frac{\sum_{i=1}^{n_r} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}}$, and f_{ij} is the frequency of the i^{th} word in the context of the j^{th} context word.

The similarity of vectors in a space is measured using a similarity metric. In the word space models, most commonly used similarity measure (Turney and Pantel 2000) is the cosine similarity (Landauer and Dumais 1997), which is also used throughout this article.

$$d_{cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{2}$$

Unsupervised learning methods

In this work, three unsupervised learning methods are used. They are Independent Component Analysis, Latent Dirichlet Allocation, and the Self-Organizing Map. The following sections detail their use in obtaining corpus-based representations.

Independent component analysis

Independent component analysis (ICA) (Comon 1994; Hyvärinen et al. 2001) is a blind-source separation method that can be used to extract components that correspond to different categories, either syntactic (Honkela et al. 2010) or semantic (Lindh-Knuutila et al. 2012). In this context, the automatically extracted independent components can also be called emergent features. ICA represents a matrix of observed signals $X_{C\times N}$ as

$$X_{C\times N} = AS, \tag{3}$$

where $A_{C\times d}$ is a mixing matrix, and $S_{d\times N}$ contains the independent components. The columns for the matrix $S_{d\times N}$ give a d-dimensional representation for the target words. We use the FastICA algorithm (Hyvärinen and Oja 1997), which estimates the model by first using dimensionality reduction and whitening and then finding a rotation that maximizes the statistical independence of the components. The dimensionality reduction and de-correlation step can be computed, for instance, with principal component analysis. Earlier, ICA has been used to find components that match the syntactic categories (Honkela et al. 2010) and semantic categories in the Battig category set (Lindh-Knuutila et al. 2012) and BLESS set (Lindh-Knuutila and Honkela 2013). The premise of the ICA method is that the components can be interpreted, compared to for example the components of LSA (Honkela et al. 2010). Often the words for which the values are high in a given component are similar, which can be evaluated using known category labels.

Probabilistic topic modeling

Generative topic models (Latent Dirichlet Allocation, LDA, and derivations) have gained popularity (Blei et al. 2003). They are probabilistic models that have been explicitly developed to model count data, and make assumptions about the distributions in different levels. The models are mostly based on document representations, but some experiments have been also carried out with a short context around the target word (Brody and Lapata 2009; Chrupała 2011; Dinu and Lapata 2010).

Latent Dirichlet Allocation (LDA) (Blei et al. 2003) is a generative probabilistic model designed explicitly for discrete data such as text corpora. It is based on the idea that documents are represented as random mixtures over latent (hidden) topics. Each topic then has a distribution over words.

The parameterized model assumes a generative process for each document d in a corpus. The length of the document N is generated from a Poisson distribution, and the topic θ from a Dirichlet distribution $\mathrm{Dir}(\alpha)$. For each of the N words in the document, the topic z_k, where $k \in [1, T]$ is chosen from a Multinomial distribution with a parameter θ and finally the word w_n is chosen from $p(w_n|z_k, \beta)$ (Blei et al. 2003). As a simplifying assumption, the dimensionality of the Dirichlet distribution, that is, the dimensionality of the topic variable z, is fixed.

Brody and Lapata (2009) use a method related to LDA in a sense induction task. Their model operates on what they call a local context, a small context around the target word, instead of a global topic around a document. They use a $P(s)$ as a distribution over senses of an ambiguous target word in a context window, and $P(w|s)$ for the probability distribution over context words w given a sense s. Their model generates each word w_i in the context window by first sampling a sense from the sense distribution, and then choosing a word from the sense-context distribution. All the other variables except for the word itself are hidden. Their model specifies a distribution over words within a context window:

$$P(w_i) = \sum_{j=1}^{S} P(w_i|s_i = j)P(s_i = j), \tag{4}$$

where S is the number of senses. It is assumed that each target word has C contexts and each context c consists of N_c word tokens.

In another related work, a similar model is used with a $1 + 1$ context for several statistical NLP tasks: named entity recognition, morphological analysis and classification of semantic relations (Chrupała 2011). In the Chrupała model, a word type corresponds to a document in the LDA model, a word is replaced by a context feature, and topic by a word class. In the generative model, the K from the LDA model corresponds to the number of latent classes, D is the vocabulary size, N_d the number of left and right contexts in which word type d appears, z_{nd} is the class of the word type d in the n_d^{th} context and f_{nd} is the n_d^{th} context feature of word type d. The model provides two types of word representations once trained: Each θ_d gives the latent class probability distribution given a word type and each ϕ_k gives a feature distribution given a latent class (Chrupała 2011).

Visualization with the self-organizing map

The Self-Organizing Map (SOM) is an unsupervised learning method which produces a low-dimensional discretized representation of the input space (Kohonen 2001). It preserves the topological properties of the input space, which makes it a useful tool for visualizing high-dimensional data. The vector space model representations are usually very high-dimensional. This makes dimensionality reduction methods such as the SOM practical tools for the text data exploration. Honkela et al. (1995) and Ritter and Kohonen (1989) are early examples of this kind of exploration. In the SOM experiments, the high-dimensional contexts have often been approximated with the random projection model (Honkela et al. 1995; Ritter and Kohonen 1989). In this article, we use only a small number of the most frequent words as our context words, and thus we do not need to apply random projection. The experiments have been carried out with SOM Toolbox for Matlab (Vesanto et al. 1999).

The SOM has earlier been carefully compared with several other methods, including principal component analysis (PCA), Isomap, curvilinear component analysis (CCA), locally linear embedding (LLE) regarding their trustworthiness and continuity of the visualization. It was found our that only SOM and CCA can be recommended for general visualization tasks where high trustworthiness is required (Venna and Kaski 2006). A projection from a high-dimensional space onto a low-dimensional display is considered trustworthy if the k closest neighbors of a point on the display are also neighbors in the original high-dimensional space (Venna and Kaski 2006).

The SOM method is well suited for analyzing and visualizing high-dimensional data. It can show in an intuitive manner the relationships between prototypical representations of the original data points. In our application, this means that the method can visualize the relationships between different linguistic phenomena, and more specifically, between different semantic categories.

A Self-Organizing Map was created based on the Wikipedia data described earlier. The data was chosen to include the combined vocabulary of the Battig and BLESS sets (1 997 words) to enable a comparison between emergent structures and linguistic category labels. For the SOM creation, we use the SOM Toolbox (Alhoniemi et al. 2005) with default parameters and batch training. The initialization of the map was based on the largest variance of the data, according to current best practices (Kohonen and Honkela 2007). Training the SOM with the full vocabulary and then inspecting the relations between words is easy on a computer, when the map can be examined interactively, but it is poorly suited to be presented on paper. Hence, we present only an illustrative set of sample visualizations.

The category labels in the Battig and BLESS data sets can be visualized on the map to gain further insight of the relations and distance of the words in a given category or several categories.

Finding category information

We consider two different approaches for finding linguistic category information in a data-driven manner. Experiments are run both with the ICA and the LDA method following Lindh-Knuutila et al. (2012) and Lindh-Knuutila and Honkela (2013).

The basic comparison made is the same as in our earlier work with ICA (Lindh-Knuutila et al. 2012; Lindh-Knuutila and Honkela 2013), but in this work we use a larger set of model sizes and and both Battig and BLESS sets. In the analysis, we study the words with highest values for each independent component or topic. In the case of ICA, the component values are usually skewed in one direction, and it thus suffices to restrict the analysis to the maximum values in the direction of the skewness of each component.

In the ICA experiment, we use the PPMI weighting scheme explained earlier. In the case of the LDA models, the question of weighting is slightly more complicated. As the LDA model expects discrete vectors, the same vector representations cannot be used for both ICA and the topic modeling task. Wilson and Chew (2010) point out that term weighting has not usually been used in conjunction with the LDA, but the models should benefit from introducing a weighting. In Dinu and Lapata (2010), a simple scaling of the counts by a factor of 1/70 was used. In this article, two different setups were used: (1) using a heuristic, where the ppmi-weighted vectors were changed to only contain integers by rounding the values of the vectors *up* to the next integer (ceil-ppmi), and (2) using the word vectors without any weighting (noweight).

When applying the model for the word vectors in the LDA case, we have used Chrupała's approach with the Matlab Topic Modeling Toolkit (Steyvers and Griffiths 2007). We treated word types as documents and context words in the place of documents, see Table 3.

The LDA model was run with the parameter values $\alpha = 50/T$, where T is the size of the model (i.e. number of topics) and $\beta = 200/W$, where W is the length of the vocabulary, suggested by the authors of the topic modeling toolkit. Initial runs were ran with 500 iterations. After the initial experiments, we also ran the LDA ceil-ppmi experiments again with a longer training time to obtain more stable results. The training length was 2000 iterations in this case.

Another difference in performance is the length of computation. The LDA ppmi-ceil weighted model calculation takes 30-40 minutes regardless of the model size, but with the unweighted model, small model sizes already take 2 hours of computation per run with the topic modeling toolbox and with the largest model size, the computation of one run lasts almost 11 hours. Hence, long training used with the ppmi-ceil was not attempted for the unweighted case. It may well be that this problem is solvable by different programming, but at least the current results support using the weighting heuristic.

Table 3 The correspondence of the LDA model and the Chrupała model of the word classes

LDA	Chrupała model
Topics	word classes
documents	word types
words	context features

The evaluation setup is the following: For each component or topic, the words were sorted in the order of the value of the component (ICA) or topic (LDA), and $N = 10$ words with highest value were chosen for analysis. In ICA, these values are taken from the matrix S, and in the case of LDA, they are the word-topic co-occurrence counts from the Gibbs sampler. The limit of analysis $N = 10$ corresponds to the number of words in each category in the Battig set. As the labels of each word are known, an automatic check was performed to see how many of them belong to the same Battig or BLESS category.

Two analysis thresholds, *strict* and *lax* were defined similarly as in Lindh-Knuutila et al. (2012) and Lindh-Knuutila and Honkela (2013). These correspond to a minimum of $P_{strict} = \frac{9}{10}$ and a minimum of $P_{lax} = \frac{6}{10}$ of words belonging to the same category or relation group, respectively. All experiments were run for ten separate iteration runs for each model size. The model sizes, i.e. the number of topics or independent components used were $T = [10, 20, 30, 40, 50, 60, 80, 100]$. In these experiments, the model was trained with only the data we have labels for, i.e. the word vectors representing the Battig vocabulary in the first experiment and the word vectors that represent the BLESS vocabulary in the second experiment. It is worth noting that the Battig set contains only category information, whereas BLESS set contains category, relation, and category-relation class labels and words can have multiple labels. In addition to these experiments with the subset of vectors, we also carried out an ICA experiment using the vector space of 200 000 word vectors. This experiment was only carried out for $T = 100$ and for 10 different ICA runs, as it is more computationally intensive. Unfortunately, the current setup of the LDA model did not allow us to experiment with the large vocabulary, and thus such an experiment remains as a future work.

Results and discussion

Battig

Method performance

In the first experiment, the ICA and LDA models were trained with the vector representations that correspond to the Battig vocabulary. We used all 53 categories and checked how many categories were found in any topic or independent component using both strict and lax criteria. We include both results in which multiple components can cover the same criteria, and results with the number of unique Battig categories found. The latter are indicated with the '-uniq' ending. The results, given in Table 4,

Table 4 Number of Battig categories found: results for strict (S) and lax (L) condition for different model sizes for ICA, LDA 1 (ppmi-ceil), LDA 2 (ppmi-ceil long), and LDA 3 (no weighting)

		Model size							
	Model type	10	20	30	40	50	60	80	100
S	ICA	**1.0**	**8.3**	**10.7**	**16.2**	**17.0**	16.6	17.1	14.9
	LDA 1	0	1.2	3.9	10.3	14.8	16.3	18.4	16.2
	LDA 2	0	1.2	3.6	10.8	14.8	17.8	**19.9**	**16.8**
	LDA 3	0	1.4	3.7	9.7	14.0	**18.7**	18.8	16.5
	ICA-uniq	**1.0**	**8.3**	**10.7**	**16.2**	**17.0**	16.6	17.1	14.9
	LDA 1-uniq	0	1.2	3.9	10.3	14.8	16.2	18.3	16.0
	LDA 2-uniq	0	1.2	3.6	10.8	14.8	17.7	**19.6**	**16.5**
	LDA 3-uniq	0	1.4	3.7	9.7	14.0	**18.7**	18.7	16.3
L	ICA	**7.7**	**14.3**	**23.0**	**34.0**	**39.2**	36.8	37.9	40.0
	LDA 1	4.8	12.0	22.7	28.9	36.2	38.1	41.3	**43.1**
	LDA 2	4.1	11.9	21.5	29.6	36.4	**39.7**	**42.9**	43.0
	LDA 3	4.1	12.3	21.8	30.9	36	38.9	41.8	**43.1**
	ICA-uniq	**7.7**	**14.3**	**23.0**	**33.0**	**37.3**	36.4	37.4	**38.4**
	LDA 1-uniq	4.8	12.0	22.7	28.8	35.0	36.3	37.0	36.7
	LDA 2-uniq	4.1	11.9	21.5	29.5	35.8	**38.2**	**39.0**	35.9
	LDA 3-uniq	4.1	12.3	21.8	30.9	35.2	37.6	37.8	36.1

The highest value for each model size and condition is marked in boldface. The number of different category types is 53.

indicate the number of categories found. They are averaged over the 10 iteration runs for each model size and experiment type for both strict (S) and lax (L) condition. In this and the following tables, LDA 1 corresponds to the LDA model with the ppmi-ceil weighting with the training length of 500 iterations; LDA 2 to the ppmi-ceil with a longer training of 2000 iterations; and LDA 3 to no weighting with the 500 iterations.

It can be seen that with a smaller model size, the ICA is able to find more categories than the LDA models, but as the model size approaches the number of categories in the Battig set, the performance difference evens out. When we compare results with the model size $T = 50$, we can see that the ICA model is able to find 37 out of the 53 unique categories with the lax condition and 17 categories with the strict condition. With the different LDA versions, the results were slightly worse: the models found approximately 35 unique categories with the lax condition, and 14 categories with the strict condition. Using $T = 60$, the LDA models are slightly better with both the strict and the lax condition.

When the model size is considerably larger than the number of categories, the ICA performance declines, probably due to splitting of the categories into several components. The LDA models suffer less from this phenomenon. There is only a small difference between the averaged results of the different variations of the LDA. To make sure the results were not due to chance, a comparison to randomly assigned categories for the words was also made. The randomly assigned categories were never found with the methods, which verifies that the results are not due to chance.

Analysis of the categories

Next, we visualized all the categories found with ICA and LDA 2 for all model sizes. The visualizations are shown in Figure 4 for the strict condition and Figure 5 for the lax condition. The Battig category names are given in the middle and the shades of gray indicate a found category according to either strict or lax criterion. Note that the direction of the x-axis is from right to left in the left hand figure. A black rectangle indicates that the category was found with every iteration and the lighter the shade of gray, the less frequently it was found.

It is clear that not all the categories are equal. Some categories are found early on with the strict condition, for example Sport, Male’s first name, Science, and Musical Instrument and Color are found early on, whereas some categories such as Kind of cloth, Kitchen utensil, Furniture or Carpenter’s tool are rarely if ever found. The effect of the large model size is evident in the strict condition as well: When the model size exceeds the number categories in the set, some of the categories are lost again, possibly due to words in categories splitting into subcategories in different topics or components. The more graded shades in the LDA results are due to the random initialization in the model calculation, whereas in ICA, the principal component analysis step is always the same and thus variation between iteration runs is not as large.

BLESS

Method performance

The BLESS data set is more complex than the Battig set as it includes both category and relation information. There are 17 categories, 5 relations and 85 joint category-relation classes with considerably larger number of words in each, and setting a cap on analyzing only 10 words per each topic or independent component will not allow every word to be present. Thus, a larger number of components or topics allows several of them to represent different parts of the same category or relation.

For the evaluation, the methodology of the Battig experiments was used, but three separate evaluations were carried out for the BLESS categories, relations and joint category-relation labels. We again applied the strict and lax evaluation criteria, and the LDA experiment variants are the same as in the Battig experiment: LDA 1 corresponds to ppmi-ceil weighting and training length of 500 iterations. LDA 2 corresponds to ppmi-ceil weighting and

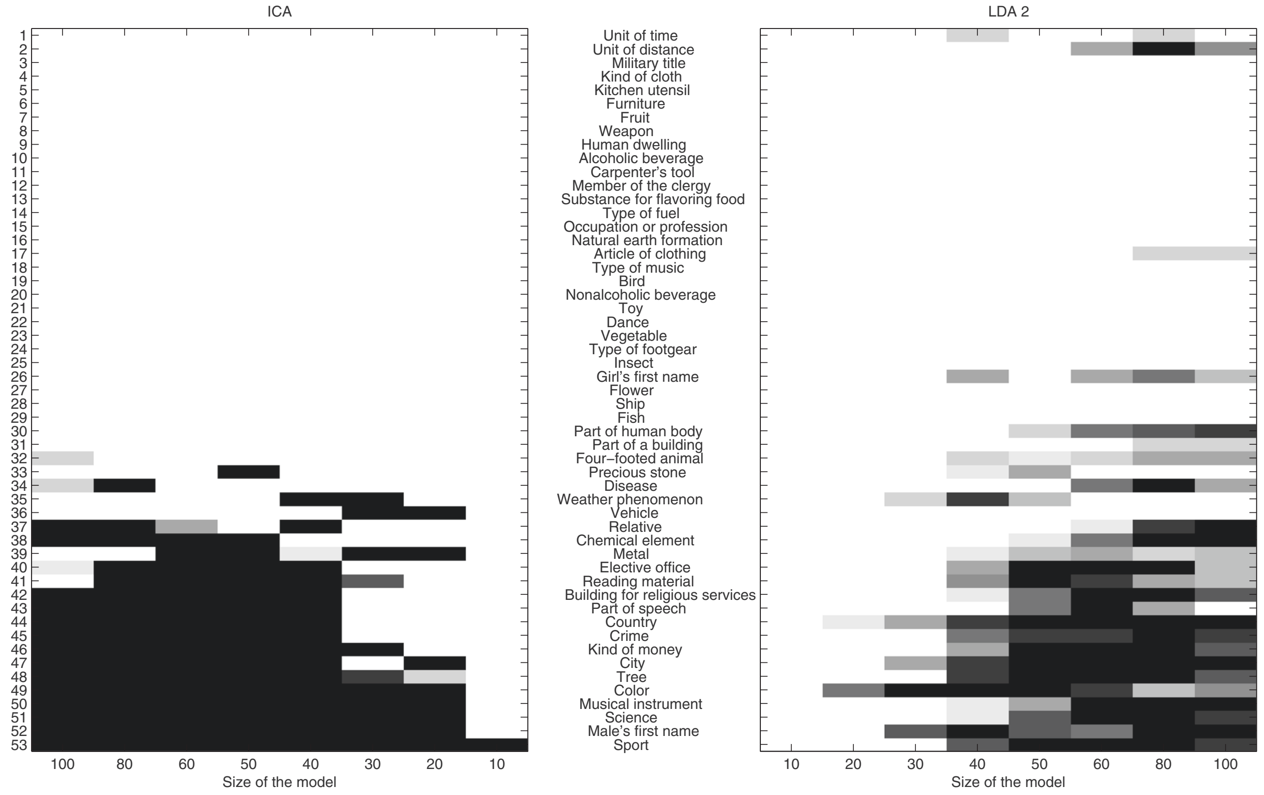


Figure 4 Battig categories found with ICA and the LDA 2 with the strict criterion for different model sizes (number of independent components or topics). The darker the square is, more often it was found in different iteration rounds.

longer training length of 2000 iterations and LDA 3 to the unweighted case with the training length of 500 iterations. The long training length was not used with the unweighted case, as the computation took a very long time even with the shorter training length. As earlier, we also include an evaluation on how many unique categories are found.

The results are given in Tables 5, 6 and 7. The results are similar to those results with the Battig set: the ICA model performs better with a small model size, but the effect evens out when the model size grows. For model size $T = 20$, which corresponds to the number of categories, it can be seen that 10 out of 17 unique categories are found with ICA in the lax case, and only 4 unique categories in the strict case. On the other hand, there are 15 components in the non-unique lax case that correspond to some of the categories, which means that several components represent the same category.

The numbers for LDA models are again slightly worse: The LDA 1 and LDA 2 perform fairly well, finding approximately 7 categories in the lax case, and 3.6 and 3.9 categories in the strict case, but LDA 3 performs very poorly. This might be due to too short training which does not reach convergence. We can also notice that with the largest model size ($T = 100$), only 11 out of 17 possible categories are found with the ICA model according to the strict criterion, and only 8 categories with the LDA 1 and LDA 2 models. Looking at the results in which categories can be represented with several components, we see that this is indeed the case: Categories are split into different subcategories, which still can fulfill especially the lax criterion.

In the case of relations (Table 6), there are only five different relation types and with the smallest model size $T = 10$, 3 out of 5 unique categories can be found with the strict condition, and 4 out of five with the lax condition. The LDA 3 model again performs worse than any of the other models. With the largest model size, all of the different relation types are represented by at least one component with the lax condition. The ICA method almost always also finds the fifth category, whereas the LDA 1 and LDA 2 again perform slightly worse, but LDA 3 only finds three out of five categories.

The difference between the number of unique relation types found and the total number of components that represent some relation type well is very large. With larger model sizes, the methods are able to find a large number of topics or independent components that correspond to

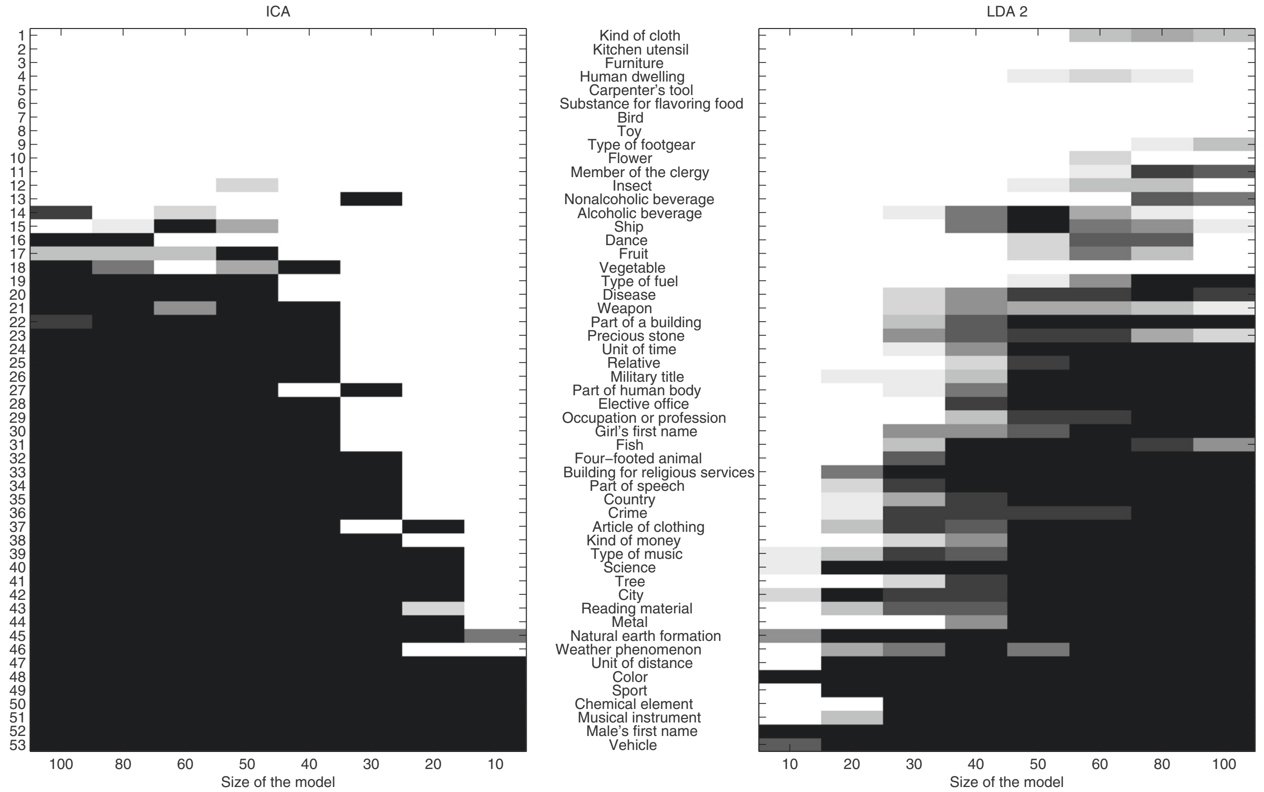


Figure 5 Battig categories found with ICA and the LDA 2 with the lax criterion for different model sizes. The darker the square is, more often it was found in different iteration rounds.

one of them. With $T = 100$ strict condition, the numbers are 38 for ICA and LDA 2 and 36 for LDA 1 in the strict case and 73 for ICA and approximately 75 topics or components in the lax case. This is explained by the fact that the relation classes are very large, and these methods are actually able to further divide the relation types into subclasses. This phenomenon will be discussed in more detail later.

The performance of the unweighted LDA (LDA 3) is worse than the ppmi-ceil weighted model (LDA 1 and LDA 2) except with the smallest model size with the lax condition. With the strict condition, and in the relation (Table 6) and the category-relation test (Table 7), the unweighted model performs very poorly. This may of course be due to the fact that the training length is too short and the model does not converge.

Analysis of the categories

The BLESS classes are visualized with a similar gray scale visualization as the Battig set to see which categories the methods are able to find. Figure 6 shows the categories found with the lax condition. Again we can see that the categories are not equal. Some categories can be found early on, but others, such as TOOL only with a larger model size. In this visualization, all of the possible relations are grouped together within a category. Both methods can find different relation types (not visualized). The only exception is least frequent relation, hypernym, which cannot be found with the two smallest model sizes. The category-relation groups are problematic. The found category-relation groups are visualized on Figures 7 and 8 for strict and lax condition respectively. The classes that were not found are left out of the visualization.

With the strict condition, MUSICAL-INSTRUMENT_COORD can be found even with the model size $T = 20$. The second best category found is VEHICLE-MERO which contains words for different parts of vehicles. The LDA 2 model also finds a similar category for parts of building, but this category is not found by the ICA model at all, even though the models give otherwise fairly similar results.

The number of classes found grows considerably when the lax condition is applied. With a large model size, both methods find most of the coordinating concept classes, except FURNITURE-COORD and AMPHIBIAN_REPTILE-COORD. Both methods find several meronym, coordinating concept or attribute classes, but there are only a few separated event classes, VEHICLE-EVENT, BUILDING-EVENT and WEAPON-EVENT that are found reliably. This may be caused by the fact that many events (or verbs), especially in animal categories, are not category specific.

Table 5 Number of BLESS categories (cat) found with ICA and the different LDA variations: LDA 1 (ppmi-ceil), LDA 2 (ppmi-ceil long) and LDA 3 (no weighting), strict (S) and lax (L) condition, and different model sizes

		Model size							
	Model type	**10**	**20**	**30**	**40**	**50**	**60**	**80**	**100**
	ICA	1.0	**5.0**	**9.0**	**10.6**	**12.1**	**13.8**	**16.4**	**18.7**
	LDA 1	1.1	4.3	6.0	8.4	10.9	12.8	14.6	17.4
S	LDA 2	**1.2**	4.8	5.9	9.2	11.8	13.4	14.1	17.2
	LDA 3	0.1	0.8	1.9	3.0	3.3	4.2	5.1	5.1
	ICA-uniq	1.0	**4.0**	**7.0**	**7.7**	**7.8**	**7.9**	**9.6**	**11.0**
	LDA 1-uniq	1.1	3.6	4.5	6.0	6.7	7.1	7.9	8.3
	LDA 2-uniq	**1.2**	3.9	4.4	6.2	7.2	7.3	7.9	8.2
	LDA 3-uniq	0.1	0.8	1.7	2.2	2.4	2.8	3.1	2.9
	ICA	**6.3**	**15.6**	**24.0**	**27.2**	**35.2**	**40.2**	**53.3**	**64.4**
	LDA 1	4.7	11.0	18.5	23.3	31.3	35.7	43.4	53.1
L	LDA 2	4.7	10.4	18.2	23.3	30.0	34.8	45.5	53.5
	LDA 3	6.3	9.3	16	18.9	24.3	29.2	40.7	46.9
	ICA-uniq	**6.0**	**9.6**	**12.3**	**12.7**	12.8	13.3	**14.5**	**15.3**
	LDA 1-uniq	4.3	7.6	9.7	10.9	**13.1**	**13.9**	13.8	15.0
	LDA 2-uniq	4.3	7.2	9.9	10.6	12.8	13.5	14.0	14.7
	LDA 3-uniq	4.8	5.0	7.5	7.6	8.3	9.4	11.3	11.9

The highest value for each model size and condition is marked in boldface. The number of different category types is 17.

Table 6 Number of BLESS relations (rel) found with ICA and the different LDA variations: LDA 1 (ppmi-ceil), LDA 2 (ppmi-ceil long) and LDA 3 (no weighting), strict (S) and lax (L) condition, and different model sizes

		Model size							
	Model type	**10**	**20**	**30**	**40**	**50**	**60**	**80**	**100**
	ICA	**4.0**	**8.0**	**13.0**	**16.0**	18.8	23.8	27.9	**38.7**
	LDA 1	3.5	6.3	11.4	15.9	21.2	**25.0**	32.6	36.1
S	LDA 2	3.3	6.2	12	15.4	**21.6**	**25.0**	**35.2**	38.4
	LDA 3	2.3	3.2	5.6	6.1	8.0	8.3	12.9	14.8
	ICA-uniq	3.0	**4.0**	4.0	4.0	4.0	**4.0**	**4.2**	**4.9**
	LDA 1-uniq	**3.4**	3.8	4.1	4.5	4.3	**4.0**	4.0	4.5
	LDA 2-uniq	3.2	**4.0**	**4.1**	**4.5**	**4.5**	**4.0**	**4.2**	4.3
	LDA 3-uniq	2.0	2.0	2.3	2.5	3.0	2.9	3.0	3.0
	ICA	**8.0**	13.0	**23.1**	31	38.5	45.3	56.6	73.4
	LDA 1	6.4	**14.2**	22.1	**31.5**	**39.1**	**47.2**	61.4	74.5
L	LDA 2	6.4	14	22.1	30.9	**39.1**	46.4	**61.5**	**74.7**
	LDA 3	7.2	11.5	16.8	22.9	27.5	32.6	46.3	58.6
	ICA-uniq	**4.0**	4.0	4.6	4.8	**5.0**	**5.0**	**5.0**	**5.0**
	LDA 1-uniq	**4.0**	**4.3**	**5.0**	**5.0**	**5.0**	**5.0**	**5.0**	**5.0**
	LDA 2-uniq	**4.0**	4.1	**5.0**	**5.0**	**5.0**	**5.0**	**5.0**	**5.0**
	LDA 3-uniq	3	3	3	3.3	3.3	3.5	4.1	4.7

The highest value for each model size and condition is marked in boldface. The number of different relation types is 5.

Table 7 Number of BLESS joint category-relation (cat-rel) classes found with ICA and the different LDA variations: LDA 1 (ppmi-ceil), LDA 2 (ppmi-ceil long) and LDA 3 (no weighting), strict (S) and lax (L) condition, and different model sizes

		Model size							
	Model type	**10**	**20**	**30**	**40**	**50**	**60**	**80**	**100**
	ICA	0	1.0	**1.6**	**2.4**	**3.7**	3.3	4.1	**7.1**
	LDA 1	0	0.8	1.2	2.0	3.3	3.6	**4.5**	5.2
S	LDA 2	0	**1.1**	1.2	2.0	3.6	**4.2**	3.7	5.7
	LDA 3	0	0	0	0	0	0.1	0.3	0.1
	ICA-uniq	0	1.0	**1.6**	**2.4**	**3.7**	3.3	4.1	**7.1**
	LDA 1-uniq	0	0.8	1.2	2.0	3.3	3.5	**4.3**	5.2
	LDA 2-uniq	0	**1.1**	1.2	2.0	3.6	**4.1**	3.6	5.2
	LDA 3-uniq	0	0	0	0	0	0.1	0.3	0.1
	ICA	**3.3**	**8.0**	**17.5**	**18.3**	**23.9**	**25.6**	**35.7**	**42.6**
	LDA 1	1.8	5.5	12.4	17.0	22.7	25.1	28.4	33.0
L	LDA 2	1.8	5.5	13.5	16.9	21.7	23.3	30.7	33.6
	LDA 3	2.6	3.3	6.6	7.1	8.5	10	14.2	16.2
	ICA-uniq	**3.3**	**7.0**	**15.1**	14.2	16.5	18.6	**23.5**	**26.8**
	LDA 1-uniq	1.8	5.4	11.7	14.7	**18.9**	**19.7**	20.3	21.1
	LDA 2-uniq	1.8	5.5	12.5	**15.0**	17.8	18.5	21.2	20.9
	LDA 3-uniq	2.4	3	5.5	6.3	7	8	11.2	11.3

The highest value for each model size and condition is marked in boldface. The number of different category-relation types is 85.

This is further confirmed by comparing to the relation results without specifying category label, where several event categories can be found. Again, the performance of the two methods is fairly similar, except that the hypernym class GROUND-MAMMAL-HYPER was only found by the LDA model.

ICA on large vocabulary

All of the previous analyses were carried out with a set of word vectors that all belong to some of the test categories. We also carried out an additional analysis for the 200 000 most frequent words in the Wikipedia data. Due to extended computation time for such a large vocabulary, the analysis was limited to only the ICA method with 10 iteration runs, using $T = 100$ components. The LDA model calculation was not carried out as it was not feasible with our current computation setup.

The analysis step was similar to previous experiments. From the $200,000 \times 100$-dimensional matrix produced by the ICA, we extract the 100-dimensional vector representations for all of the words in either the Battig or the BLESS subset, and look at the maximum values within this subset for each component as before. Thus, for each component we look at the 10 words that have the highest value, and

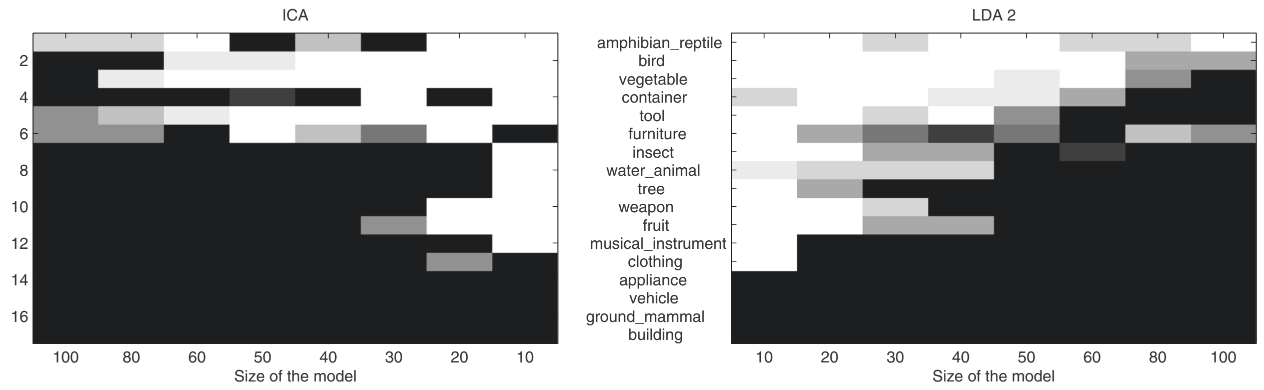


Figure 6 The categories of BLESS found by the ICA and LDA 2 with the lax condition for different model sizes.

see whether they have the same category label, again with the strict (9/10 words belong to the same category) or the lax (6/10 words belong to the same category) condition. These results are reported in Table 8. We can see that ICA is able to find 4.8 unique categories with the strict condition, and 18.9 categories with the lax condition. This can be contrasted to the results in Table 4 with the same size, in which 14.9 unique categories were found with the strict condition, and 40 categories with lax condition. With the BLESS results, ICA finds 3 unique categories in the strict case, and 54 in the lax case; 2 relations in the strict, and 4 relations in lax case, and 3 unique category-relation classes in the lax case. This analysis only reveals a partial truth as the highest value for our labeled Battig example can have the highest value of all Battig vocabulary words, but at the same time it can have a mid- or even low rank when all words in the 200,000 word vocabulary are covered.

To analyze this effect we carried out a further analysis in which we checked the components which fulfilled either the strict or the lax condition of the previous analysis. We did not require the uniqueness of the category, but all components that passed the criterion were included. As a confidence score on how far from the top values for each component, we calculated the ratio of the mean value of the five top Battig or BLESS words and highest value of each component.

The example components obtained with the strict criterion on one ICA run are given in Table 9, sorted according to the confidence score. Shown are the 15 words with the highest value for these components from the full vocabulary of 200 000 words. As we can see, with a confidence score of over 0.8, the Battig category labels are perfect matches to the word sets. The component labeled with COUNTRY seems to have geographical information such as 'north-eastern', 'bordering' etc. mixed with country names. Hence, the labeling is fairly good. The fifth component, labeled with ELECTIVE OFFICE, seems to contain names of world leaders, which are not included in the Battig word set. Nevertheless, this result is also fairly good and indicates a certain semantic relatedness between the label and the word set.

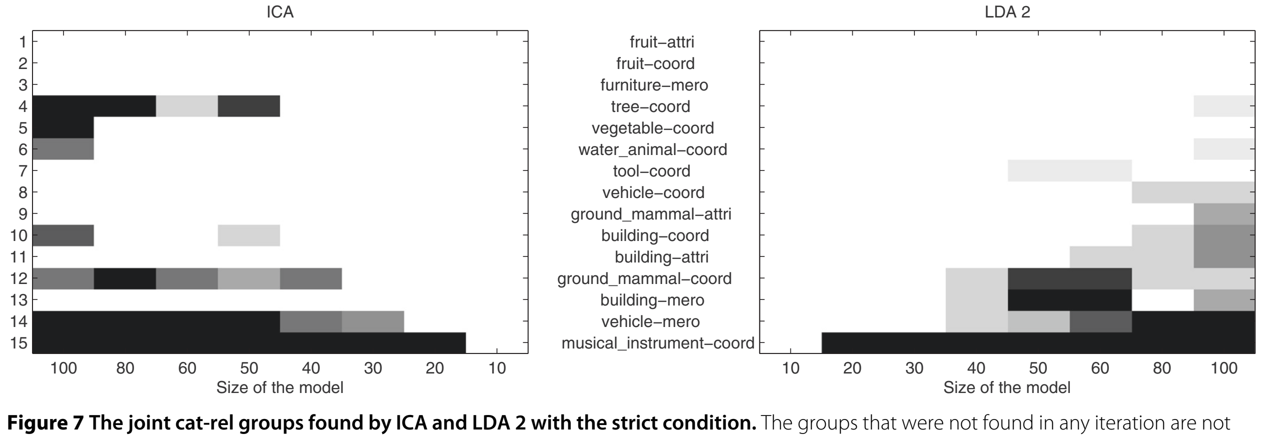


Figure 7 The joint cat-rel groups found by ICA and LDA 2 with the strict condition. The groups that were not found in any iteration are not visualized.

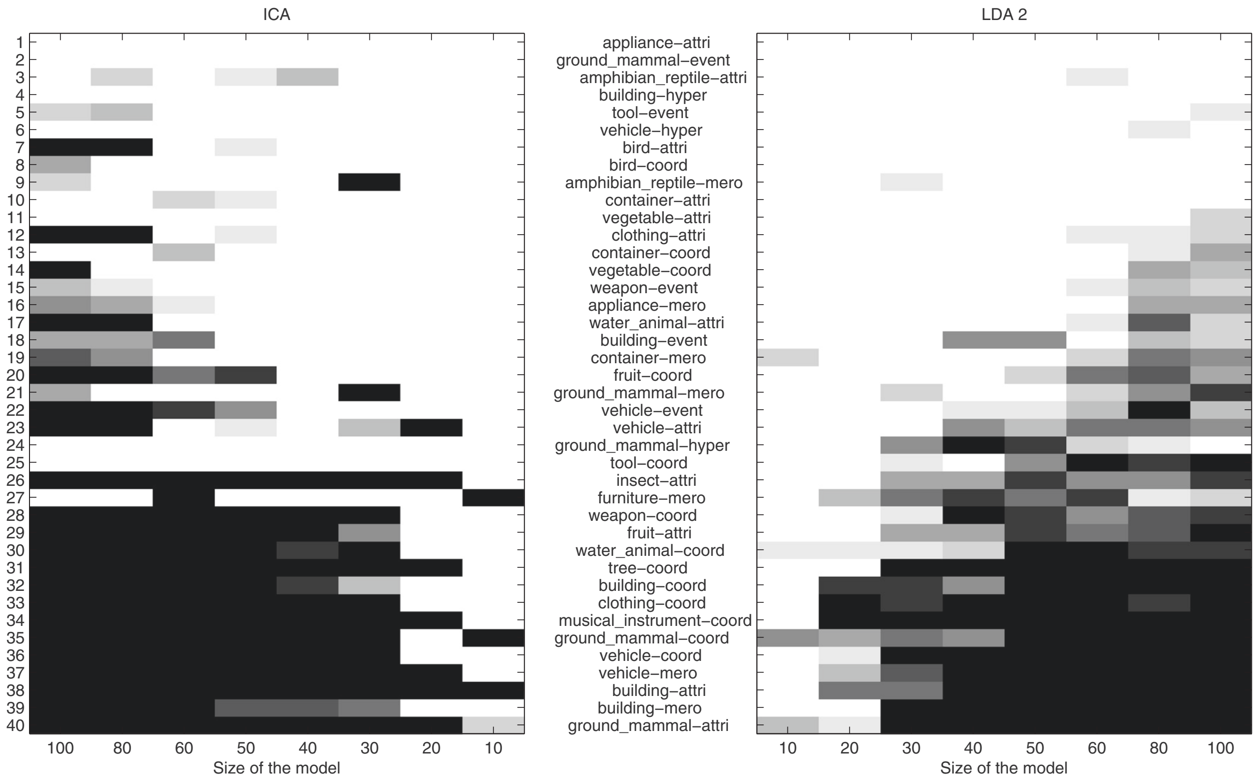


Figure 8 The joint cat-rel groups found by ICA and LDA 2 with the lax condition. The groups that were not found in any iteration are not visualized.

A review on each iteration run shows very similar results: The SPORT category appears in every run, with an average confidence score of 0.94, MALE'S FIRST NAME also appears in each of the runs with an average confidence score of 0.92. MUSICAL INSTRUMENT appears in 6/10 runs with an average confidence score of 0.89. COUNTRY and ELECTIVE OFFICE are also found in each of the ICA runs, with confidence scores of 0.75 and 0.50 respectively. In addition, the category RELATION is found once with confidence score of 0.63.

We can also look at the BLESS labelings in category and relation type in a similar way. Table 10 gives the results for strict category and relation results for the same ICA run as those analysed above with the Battig test set. We notice that the MUSICAL INSTRUMENT category is the same in both sets and is found with BLESS category test as well in each ICA run with an average confidence of 0.94. In addition, the BLESS category test finds a word set labeled with APPLIANCE in 3/10 ICA runs with a confidence of 0.86. In addition, there are two distinct sets labeled with VEHICLE, which are good examples. The first of them, which seems to describe military vehicles is found on average confidence of 0.82 in 9/10 of the ICA runs, and the second in 9/10 runs with an average confidence of 0.75.

Table 8 Results with ICA with 100 components on large vocabulary of 200 000 words, for Battig category results and BLESS category, relation and category-relation analysis for strict and lax condition

		BLESS		
	Battig	Cat	Rel	Cat-Rel
strict	4.8	5.0	6.3	0
strict+uniq	4.8	3.3	2.2	0
lax	23.9	20.5	39.3	3.4
lax+uniq	18.9	5.4	4.0	3.2

The fifth component found with the strict condition, BUILDING is not as clear, which is also highlighted by the lower confidence score, on average 0.56 but found on every ICA run. This result is understandable, though, when one inspects the words in that particular category: words such as 'student' are listed in being meronymous relationship in this category, hence they are included in the analysis. There is another similar component which is labeled with BUILDING. This component contains religion words such as 'orthodox' 'catholic' and 'hindu'. Different words related to religion are also listed as meronyms in the BUILDING category. This word group found in 7/10 runs with an average confidence score of 0.70.

Table 9 The top 15 words in the five components found with ICA on 200,000 word vocabulary using the strict criterion on Battig set, sorted by the confidence score

Score	**0.93989**	**0.91928**	**0.88853**	**0.7655**	**0.51271**
Category	**Sport**	**Male's first name**	**Musical instrument**	**Country**	**Elective office**
	soccer	jim	piano	north-eastern	deputy
	football	steve	percussion	bordering	putin
	tennis	chris	synthesizer	north-western	chirac
	volleyball	mike	synth	present-day	yeltsin
	championship	greg	guitar	southern	incumbent
	basketball	jeff	acoustic	indonesia	saakashvili
	handball	gary	harmonica	northern	tory
	rowing	dave	keyboard	kazakhstan	musharraf
	boxing	doug	orchestral	slovakia	féin
	lacrosse	tim	bass	thailand	republican
	hockey	david	flute	mongolia	mulroney
	netball	tom	keyboards	northeastern	chrétien
	league	brian	vocal	bulgaria	ndp
	softball	kevin	accordion	slovenia	chaudhry
	badminton	ron	mandolin	turkey	whitlam

The relation classes are not as clear, as the classes each contain large number of words. The two components with highest scores, labeled EVENT (present on all ICA runs, with an average score of 0.85 and ATTRI (in 10/10 runs with average score of 0.84) are good examples of those relations. The three other attributes (each found in 10/10 runs) labeled with ATTRI are less clear. The second of them lists different nationalities with an average confidence score of 0.55, but the two others just seem to be mis-classified. Thus the BLESS relation labels do not seem to be very useful in this task.

Hence we only look at the lax criterion on the Battig set to see whether we can find more components that are interesting but do not fulfill the strict criterion. In Table 11, in the first set of component, the confidence scores are over 0.7, and the labels are very good or fairly good matches of the word sets. These sets also appear in all ten of the separate ICA runs: CHEMICAL ELEMENT, with an average score of 0.87, COLOR, with an average score of 0.84, TYPE OF MUSIC with an average score of 0.81, GIRL'S NAME, with average score of 0.79, and VEHICLE, with average score of 0.72. All of these category labels describe the word set well. The VEHICLE label corresponds to the component also found with the BLESS category test.

On the second set, word sets at the mid range confidence score (between 0.7 and 0.5) give mixed results. Some, such as the first component on that set, labeled PART OF SPEECH lists different languages and language-related words (appearing on 5/10 ICA runs with an average confidence score of 0.73; the second component is labeled RELATIVE (in 10/10 runs, with avg. score of 0.62) which also contains words of the category; and the fourth component on that set, labeled with SCIENCE, contains words related to science and research (10/10 runs, with avg. score 0.60). On the other hand, the word set in the middle, labeled with ELECTIVE OFFICE, and with a higher score than the one labeled with SCIENCE contains words mostly in past participle form, which is an erroneous labeling. Still, it is a word set that persists through iterations appearing in 10/10 runs with an avg. score of 0.61, which shows that the ICA finds meaningful structure in the data. The same can be said of the fifth component in this set that contains abbreviations, labeled with UNIT OF DISTANCE (present in 10/10 runs with avg. score of 0.58).

Even at the lowest scores there may be some information: The two components labeled as TYPE OF READING MATERIAL do contain words that are related to reading and books: authors and text type. The former appears in 10/10 runs with an avg. score of 0.42 and the latter on 8/10 runs with an avg. score of 0.35. On the other hand, the lowest score of all contains words in genitive form, again the label is all wrong. Also this kind of component is present in each of the ICA runs, with an average score of 0.22. Based on this analysis, one can deduce that an additional measure such as the confidence score is needed to complement the strict/lax analysis, and the score needs to be adjusted according to the test set. In the lax Battig case, approximately 9.5 out of 24 of the found components have a confidence score $>$ 0.7, and thus can be

Table 10 The top 15 words on example components found with ICA on 200,000 word vocabulary using the strict criterion on BLESS category and relation tests, sorted by the confidence score

Score	**0.94328**	**0.8625**	**0.82388**	**0.76589**	**0.54407**
Category	**Musical instrument**	**Appliance**	**Vehicle**	**Vehicle**	**Building**
	piano	optical	reconnaissance	v8	undergraduate
	percussion	electrical	amphibious	engined	graduate
	synthesizer	manufacturing	naval	turbocharged	postgraduate
	synth	pneumatic	tactical	v6	nursing
	guitar	mechanical	bomber	diesel	vocational
	acoustic	electronics	anti-submarine	four-cylinder	post-graduate
	harmonica	laser	long-range	engine	education
	keyboard	hydraulic	combat	sedan	medical
	orchestral	microwave	submarine	v12	university’s
	bass	analog	helicopter	turbo	engineering
	flute	electronic	squadron	high-performance	humanities
	keyboards	hand-held	airborne	air-cooled	courses
	vocal	portable	patrol	bmw	academic
	accordion	welding	aerial	litre	college
	mandolin	imaging	land-based	gt	interdisciplinary

Score	**0.84948**	**0.84304**	**0.55771**	**0.54577**	**0.38759**
Relation	**Event**	**Attri**	**Attri**	**Attri**	**Attri**
	expect	unpleasant	person’s	slavic	konstantin
	perceive	aggressive	one’s	berber	nikolai
	contain	awkward	spiritual	turkic	nikolay
	understand	arrogant	individual’s	germanic	pavel
	hear	cynical	personal	kurdish	mikhail
	give	risky	character’s	albanian	józef
	find	optimistic	whose	tatar	giuseppe
	happen	sarcastic	your	iranian	aleksandr
	ask	eccentric	our	aboriginal	josef
	lose	ambiguous	mystical	chinese	stanislaw
	ignore	realistic	emotional	lithuanian	viktor
	deny	annoying	inherent	indigenous	andrzej
	affect	cautious	cultural	inuit	františek
	choose	dangerous	tremendous	semitic	vladimir
	appreciate	unstable	man’s	somali	sergei

thought as reliable labelings. Thus it means that with the Battig category labels, we can correctly label 10% of all of the components correctly as the findings with the lax condition also include the findings with the strict condition. The remaining components are thus unaccounted for. In addition, the BLESS strict labeling allows us to correctly label two extra categories with the category labels, and three categories with the relation label, giving a total 15 of 100 components. This still leaves 85 components unaccounted for.

The results are nevertheless promising. Considering that the ICA method attempts to describe any kind of structure in the data, and the Battig vocabulary covers only 0.2% and the BLESS vocabulary only 0.8% of the vocabulary of 200 000 words, these results show that even partial labelings can be very useful when studying such a large dataset. The results would probably improve, if some kind of pre-selection of words was carried out, for example removing numbers, abbreviations, and foreign words from the data, as now the ICA method also attempts

Table 11 Some word sets and category labels with lax condition on Battig set for ICA with 100 components and a vocabulary of 200 000 words

Score	0.92066	0.8262	0.8179	0.7951	0.72163
Category	Chemical element	Color	Type of music	Girl's first name	Vehicle
	liquid	colored	reggae	louise	v8
	nitrogen	coloured	rap	louisa	engined
	sulfur	yellow	pop	frances	turbocharged
	sodium	stripe	hip-hop	margaret	v6
	hydrogen	translucent	jazz	katherine	diesel
	compounds	blue	disco	elisabeth	four-cylinder
	carbon	pale	motown	anne	engine
	organic	metallic	rock	josephine	sedan
	chlorine	pink	punk	elizabeth	v12
	ammonia	red	solo	christina	turbo
	acid	striped	blues	sophie	high-performance
	gaseous	reddish	rockabilly	agnes	air-cooled
	magnesium	hair	funk	catherine	bmw
	oxide	purple	bluegrass	anna	litre
	toxic	glossy	indie	mary	gt
Score	**0.66905**	**0.62344**	**0.6083**	**0.59476**	**0.58315**
Category	**Part of speech**	**Relative**	**Elective office**	**Science**	**Unit of distance**
	sanskrit	ancestors	increasingly	undergraduate	ex-pg
	literally	courtiers	obsessed	graduate	ddg
	hebrew	forebears	reacquainted	postgraduate	cve
	colloquial	lineage	addicted	nursing	apd
	aramaic	servant	enamored	vocational	unterseeboot
	adonai	forefathers	infatuated	post-graduate	aog
	derogatory	members	entangled	education	op
	slang	progeny	enmeshed	medical	pf
	gnosis	grandfathers	embroiled	university's	pgm
	archaic	piety	acquainted	engineering	year-old
	arabic	mutant	disillusioned	humanities	hp
	elohim	sole	romantically	courses	seibel
	bushi	ancestral	enamoured	academic	percent
	dharma	heirs	extratropical	college	kg
	euphemistic	husbands	disenchanted	interdisciplinary	yfd
Score	**0.47718**	**0.41936**	**0.36343**	**0.30244**	**0.23024**
Category	**Science**	**Reading material**	**Reading material**	**Science**	**Unit of time**
	perspective	tolkien	philosophical	mathematical	country's
	considerations	kant	satirical	geometric	japan's
	aspects	eliade	biographical	algebraic	germany's
	concepts	jokingly	scholarly	geometrical	china's
	purely	plutarch	biblical	logical	france's
	problems	rowling	prose	analytic	india's

Table 11 Some word sets and category labels with lax condition on Battig set for ICA with 100 components and a vocabulary of 200 000 words *(Continued)*

contexts	nietzsche	autobiographical	linguistic	russia's
attitudes	asimov	literary	computational	pakistan's
topics	allmusic	unpublished	empirical	egypt's
studies	wittgenstein	historical	evolutionary	poland's
theories	aristotle	humorous	fundamental	israel's
phenomena	freud	modernist	combinatorial	team's
themes	herodotus	english-language	qualitative	korea's
perspectives	maimonides	poetic	theoretical	ireland's
standpoint	strabo	contemporary	quantitative	nation's

to cover them. Additional labeled word sets or manual evaluation by experts could also be added for further analysis.

Exploration

An important application of the methods based on unsupervised learning is exploration. In addition to simply checking the overlap between the provided category labels and the retrieved word sets found with ICA or LDA, it is useful to explore the data with the help of the unsupervised methods to extract other meaningful structure from the data. In this article, we search for stable sets of words that occur frequently in multiple iteration runs. For this purpose, a simple search algorithm that finds frequent word sets is used.

We start with a set of word lists obtained by taking the ten words with largest value for each component or topic for each experiment run with the BLESS set of vectors. That is, for model size $T = 60$, we obtained 60 word sets of ten words each from each of the ten iteration runs, and for the model size $T = 100$, we would have 100 word sets of ten words each from each of the ten iteration runs. However, we do not expect to gain exactly the same set of ten words in every run. Therefore, we also study the subsets to extract partial matches.

We search for word sets that appear in multiple iteration runs and retain those sets that are at least of size N_{set} and are found in at least M iteration runs out of 10. Only the largest subset that exceeds the limit is counted: sets that are a subset of a larger frequent set also exceeding the limit are not counted again. The *strict* and *lax* limit for M_{limit} are defined. The set size is limited to $10 \geq N_{set} \geq 7$. To be included in the results, a word set must be found in $M_{strict} \geq \frac{9}{10}$ iterations and $M_{lax} \geq \frac{6}{10}$. We report results for $T = 60$ and $T = 100$ with both criteria for ICA and the LDA ceil-ppmi weighted model with both normal (LDA 1) and long (LDA 2) training length. Models were trained with the BLESS vocabulary. The details of the found frequent word sets per set size, and model type and size are shown in Table 12.

In exploration, no specific labels can be attached to the found structure, but human evaluators must be used instead. To obtain insight on the quality of the retrieved word set, a simple evaluation criterion was devised. Each word set was checked against all existing BLESS labels, and majority labels for category, relation and category-relation were calculated for each word set. Four different qualitative classes or types were then defined. The types describe how well the majority category label describes the words of the set. The types are listed in Table 13. These types are a) Descriptive: majority label exists and it describes the word set well; b) Partial: there is a majority label but a more specific description can be easily found; c) Meaningful: no descriptive majority label exists, but words are clearly related; and d) Nonsense: there is no descriptive majority label, nor any clear semantic relation between the words.

Examples of these types are shown on Table 14. In the first column, all words belong to the BLESS category TREE, or more specifically TREE-COORD, and one is

Table 12 The number of frequent sets per each set size analyzed

			Set size				
Model size	**Condition**	**Method**	**10**	**9**	**8**	**7**	**Total**
60	S	ICA	9	12	6	11	38
		LDA 1	0	3	7	2	12
		LDA 2	0	3	3	7	13
	L	ICA	17	23	21	8	69
		LDA 1	1	12	16	12	41
		LDA 2	3	14	15	18	50
100	S	ICA	16	24	8	13	61
		LDA 1	0	1	3	9	13
		LDA 2	0	3	10	6	19
	L	ICA	33	34	20	25	112
		LDA 1	2	10	22	28	62
		LDA 2	2	20	28	26	76

Table 13 Types of qualitative classes the word sets found were classified into

Type	Description of the qualitative class
Descriptive	Words are related in some way and the majority label given is as descriptive as possible of the words in the set.
Partial	Words are related in some way and the majority label is somewhat descriptive, but a more descriptive account can be easily given.
Meaningful	Words are related in some way, but there is no majority label that describes the words
Nonsense	There is no majority label, nor is there any perceived relation between the words in the set.

hard-pressed to find a more specific description. In the second column, all words, except 'pilot' are *flying* vehicles, and separate word sets for water and land vehicles also exist. In this case, the word 'pilot' is considered an outlier, even though it is related to flying vehicles. In the third column, one can point out that all of the words are sports-related, but no majority BLESS label exists. The fourth column then has a more ambiguous word set, with no majority label, nor clear description, and the word set is classified as Nonsense.

The word sets that belong to any of the other relation class except COORD are mostly classified to the Partial class, as a more specific description beyond the relation class often exists. An example of different attribute sets found with the ICA method (100 ICs, strict condition) is given in Table 15, each with a characterizing title given by the researchers for that attribute. One can notice that very different attribute groups are found. These range from cultural attributes, to general attributes that describe like or dislike, to colors, and attributes related to taste. Category judgments for nouns seem to be easier than characterizing adjectives or attributes, but these results show that such groupings can be found from corpus data in an unsupervised manner. Similar results can also be found with the LDA model, although the number of frequent sets found is smaller. Attribute word sets found with LDA 2 (100 topics, strict condition) are shown in Table 16.

Table 17 presents the number of the word sets classified to the different groups by the researchers. The coverage of the existing BLESS labels that describe the groups well (Descriptive) ranges from 0 to 34% for different setups, but it is worth noting that the partially descriptive class (Partial) covers from 25% to 70% of the found frequent

Table 14 Examples of the word sets of different qualitative types taken from an analysis of ICA with 100 components and the strict condition

Descriptive	Partial	Meaningful	Nonsense
acacia	aeroplane	bowl	american
birch	aircraft	coach	anchor
cedar	airplane	cricket	dusty
cypress	bomber	cup	herb
evergreen	fighter	game	miss
fir	glider	hall	robin
oak	helicopter	national	rosemary
pine	jet	player	rusty
poplar	pilot	pool	sly
willow	plane	squash	spike

Table 15 Different attribute word sets with an with a characterization given by the authors, from the ICA with 100 components and strict condition

'Dangerous'	'General'	'Color'	'Cultural'
aggressive	bad	black	african
armed	clever	blue	american
bitter	cute	gray	ancient
deadly	dirty	green	asian
destructive	funny	grey	christian
ferocious	nice	pink	indian
fierce	pretty	purple	medieval
heavy	scary	red	modern
strong	stupid	white	roman
stubborn	ugly	yellow	
'Temperature'	**'Animal'**	**'Taste'**	**'Shape'**
antarctic	aquatic	bitter	circular
clean	arboreal	delicious	curved
cold	carnivorous	juicy	cylindrical
cool	endangered	oily	flat
dry	gigantic	sour	narrow
hot	herbivorous	spicy	oval
soft	nocturnal	sweet	rectangular
tropical	solitary	tart	rounded
warm	wild	tasty	spiral

Table 16 Different attribute word sets with a characterization given by the authors, from the LDA 2 with 100 topics and strict condition

'General'	'Temperature'	'Size'	'Dangerous'
bad	cold	gigantic	aggressive
dirty	cool	heavy	bitter
funny	dry	huge	deadly
nice	frozen	immense	destructive
pretty	hot	large	ferocious
scary	muddy	little	fierce
stupid	tropical	small	lethal
ugly	warm	tiny	

Table 17 The number of all frequent sets classified into the qualitative classes by the authors

Model size	Condition	Method	Descriptive	Partial	Meaningful	Nonsense	Total
60	S	ICA	11	21	4	2	38
		LDA1	5	3	2	2	12
		LDA2	5	5	1	2	13
	L	ICA	14	47	3	5	69
		LDA1	14	20	5	2	41
		LDA2	14	30	4	2	50
100	S	ICA	16	34	5	6	61
		LDA1	0	10	2	1	13
		LDA2	0	13	3	3	19
	L	ICA	27	72	2	11	112
		LDA1	8	39	7	8	62
		LDA2	13	51	3	9	76

sets - i.e. the methods are able to find meaningful sets that are more fine grained than the BLESS labels allow. The fraction of the meaningful sets for which no labels exist are not negligible, either, and they cover from 7% to 20% of occurrences, depending on the case, whereas nonsense word sets cover about the same amount of data: from 5% to 17%. In cases where there is a higher number of nonsense cases, there are actually different word sets that are mostly same and only vary by one word, which indicates that the method finds a reasonable amount of meaningful word sets. ICA finds more stable word sets than the LDA. This is partly due to the underlying PCA component of the ICA method, which is the same in every run, and only the solution for the rotation search differs, whereas LDA has a truly random starting point. The longer training in the LDA 2 case improves results considerably, with a quarter of more sets found in lax case.

Visualization of categories and relations

Analysis based on BLESS set

The part of speech and relation information of the BLESS data set was visualized on the self-organizing map with hit histograms (Figures 9 and 10). In a hit histogram, each word vector is mapped to the best matching map unit, which is the *hit* for that vector. Instead of visualizing every individual hit, the number of hits for each map node is visualized. The size of the black dot on a map unit is proportional to the number of hits on that node. A completely filled node contains five or more hits.

Figure 9 shows hit histograms on the map for the three part-of-speech categories: ADJECTIVE (a), NOUN (b) and VERB (c) for the words with the BLESS labels for which the part-of-speech label is given. The verb category is on the left side of the map, markedly on the top-left corner, the adjectives have a prominent area next to the verbs, and the nouns, the largest group, occupy most of the right side of the map. The relation categories of the BLESS data are shown in Figure 10. In the case of the relations, the ATTR (a) category corresponds to ADJECTIVE category and the EVENT (d) to VERB. The NOUN category is divided into COORD (b), HYPER (d), and MERO (e). The division of COORD and MERO is visible on the map with most of the coordinating concepts on the bottom right and the meronyms on the top of the map, whereas the

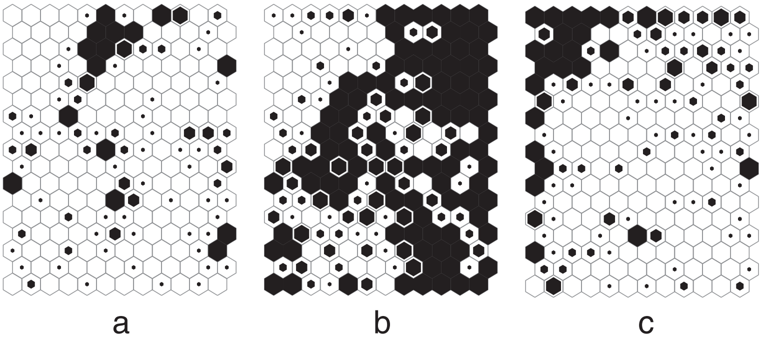


Figure 9 Hit histograms for three part of speech classes. a) ADJECTIVE, **b)** NOUN, and **c)** VERB.

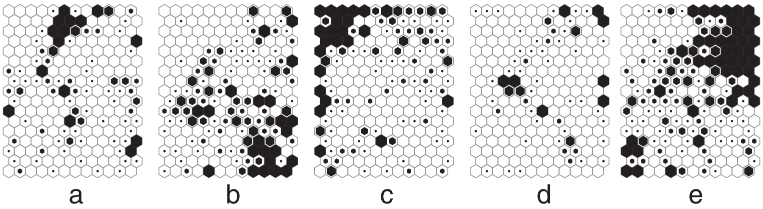


Figure 10 Hit histograms for the five relation types of BLESS. a) ATTR, **b)** EVENT, **c)** COORD, **d)** HYPER, and **e)** MERO.

HYPER relation is scattered all over the map. A statistical analysis of co-occurrence data is thus able to find also this semantic distinction among the words.

Analysis based on Battig categories

We can also examine the semantic categories and their relations. Earlier, we examined the categories of the BLESS set that were not found by the ICA method (Lindh-Knuutila and Honkela 2013) and concluded that the categories that are not found were either too spread out on the map or overlapping with another category.

We repeated the experiment on the semantic categories of the Battig data set to find explanations why some of the categories can be found with the ICA or LDA model, but others are not found. For this purpose, we plotted the hit histograms of the 10 best categories found with ICA or LDA with the strict condition, and 10 worst that were not found even with the lax condition. Figure 11 shows the worst categories. All of the hit histograms are fairly scattered. The ones with most concise clusterings are a) KIND OF CLOTH, f), SUBSTANCE FOR FLAVORING FOOD, g) BIRD and i), INSECT. All of these categories have one thing in common: While there is some spread all over the map, they also overlap with other categories. BIRD and INSECT cannot be separated from other the animal categories, KIND OF CLOTH overlaps with ARTICLE OF CLOTHING, and SUBSTANCE FOR FLAVORING FOOD overlaps with both VEGETABLE and FRUIT, and indeed the category contains words that could be classified in the VEGETABLE class, such as 'onions' and 'garlic'.

Also polysemous words influence the results. In the INSECT category, one hit in the top left corner is away from rest of the hits. This hit corresponds to the word 'fly' with an obvious polysemy. Comparing to the map of the relations in Figure 9, it can be seen that this is the area where the majority of verbs are mapped to. Similarly, one of the color items, 'brown', is clustered in the corner, close to nodes where all human-related categories are located. 'Robin' from the BIRD category exhibits similar behavior.

If we now compare these worst cases with the best separable categories in Figure 12, we see that most of them form a concise cluster in either one or several neighboring nodes. The only category that is different in this sense is the category KIND OF MONEY.

We also explored a case where the categories belong to a shared higher level category. Figure 13 shows the hit histogram mapping all the words from the Battig set that belong to a higher level category HUMAN: MALE NAME, GIRL'S NAME, MEMBER OF THE CLERGY, ELECTIVE OFFICE, RELATIVE, OCCUPATION OR PROFESSION, and MILITARY POSITION.

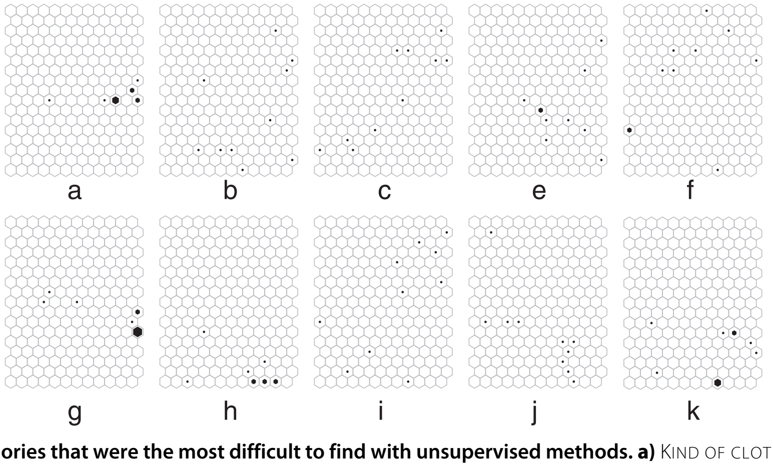


Figure 11 The Battig categories that were the most difficult to find with unsupervised methods. a) KIND OF CLOTH, **b)** KITCHEN UTENSIL, **c)** FURNITURE, **d)** HUMAN DWELLING, **e)** CARPENTER'S TOOL, **f)** SUBSTANCE FOR FLAVORING FOOD, **g)** BIRD, **h)** TOY, **i)** INSECT, and **j)** FLOWER.

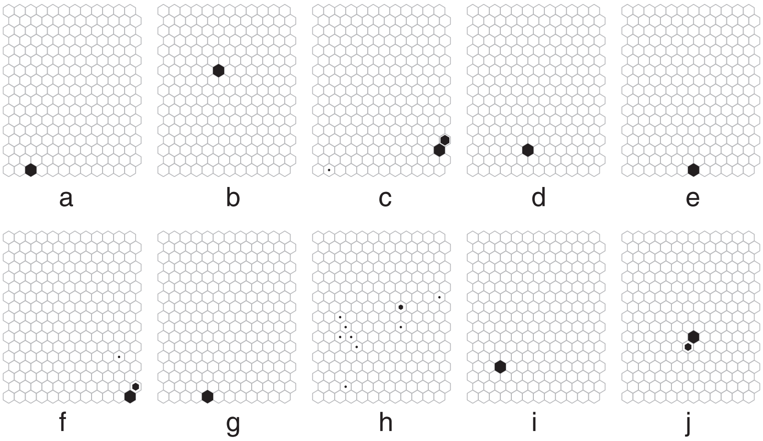


Figure 12 The Battig categories that were found best with unsupervised methods. a) Male's first name, **b)** Science, **c)** Color, **d)** Sport, **e)** City, **f)** Tree, **g)** Musical instrument, **h)** Kind of money, **i)** Country, and **j)** Building for religious services.

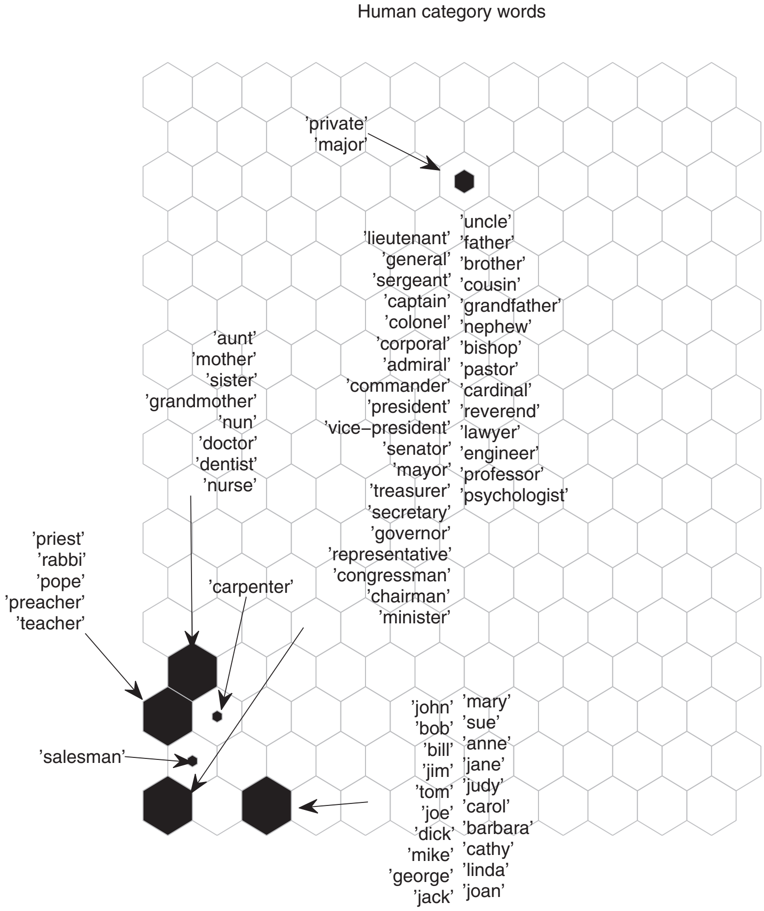


Figure 13 All the words of the categories in the Battig set that are part of a higher level category HUMAN. These categories are Male name, Girl's name, Member of the Clergy, Elective office, Relative, Occupation or Profession, and Military Position.

These words are mapped fairly closely together in the lower left corner of the map, except for two outliers on the top. Analyzing this figure further some remarks can be made. The names separate on their own on a map unit at the bottom of the map. The category RELATIVE is split into two: The words indicating a female relative are on a separate node from the rest. Words meaning a female occupation such as 'nun' and 'nurse' are also in that node, with 'doctor' and 'dentist', which are not as gender dependent.

The most populated map unit in this visualization is the one that contains the male relatives, most military titles, all words from the elective office category and the rest of the professions or occupations, and the rest of the members of the clergy. The words 'carpenter' and 'salesman' are slightly different from the rest.

It is also worth noting that there are two outliers away from the general cluster of hits on the top of the map. These words are 'private' and 'major' from the MILITARY TITLE category. Comparing to Figure 9, we notice that these words are mapped close to the cluster of adjectives. Indeed, both of these word forms are polysemous, and have another prominent sense in the adjective class.

Conclusions

In this article, we have compared the word sets found with ICA and LDA with different model sizes to semantic labels of two semantic dictionaries: Battig and BLESS. We can verify that both of these unsupervised methods are able to find components or topics that have a semantic interpretation. We also found that not all categories are equally easy to find. Some of the categories listed in these dictionaries are more concise, and the similarity based comparison of these vector representations finds them, whereas other categories cannot be found with these methods. One can question whether there is a vector space model that is able to represent the difficult categories such as TOOL, where many words are more general and fairly polysemous. Perhaps this could be possible with a larger context and differentiating between different context types or senses (Erk and Padó 2008).

We also carried out an analysis on ICA using a large vocabulary of the 200,000 most frequent words in the Wikipedia corpus. These results show that ICA is able to extract meaningful semantic information, and partial labeling of the corpus can be used in explorative analysis of the word sets: a confidence score which tells how close to the highest value for each component seems to be an useful heuristic. This work needs to be repeated with different semantic test sets and model sizes though, carried out in the future. The current setup of the LDA model could not be extended to include such a large vocabulary—to better compare with the ICA, a computationally more efficient approach will need to be devised.

Not all manually defined class labels can be found in an unsupervised way: Instead, structure that may or may not correspond to class labels can be found. Exploration was the second important topic of this article. We devised a search algorithm to find frequent sets of words, and made a preliminary qualitative analysis of the magnitude of how well BLESS labels describe the retrieved word sets. We found that often the models are able to divide the classes into meaningful subsets, for example dividing attributes into more fine grained and meaningful sets. In addition, some meaningful word sets with no labeling were found. On the other hand, the number of the nonsense word groups was fairly low.

Comparison of the automatically generated structures and manually defined classes provides useful information. In order to explore this relationship in more detail, we have demonstrated how the SOM can be used for this purpose. It serves as a visualization tool for category information, which can yield information on the conciseness of the categories or relations between different categories. This work can be further extended by combining different separate data for more labeled data or comparing the ICA and LDA results with other manually built resources such as ontologies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TH and TL-K defined the basic problem setting jointly. TL-K defined the details of the evaluation and analyses discussing them with TH, and carried out the experiments. TL-K and TH wrote the article jointly. Both authors read and approved the final manuscript.

Acknowledgments

The authors are grateful to the Department of Information and Computer Science at Aalto University School of Science that has supported the research in various ways, and to the Academy of Finland whose funding has significantly helped the research within the Computational Cognitive Systems research group. For support at later stages of this work, TL-K wishes to thank O.V.Lounasmaa Laboratory at Aalto University. Similarly, TH thanks the support of the Dept. of Modern Languages at University of Helsinki. We specifically thank Jaakko J. Väyrynen for his valuable help. TL-K has been supported by the Finnish Cultural Foundation. TH wishes to acknowledge collaborators in the META-NET Network of Excellence for useful exchange of ideas. We also would like to thank the anonymous reviewers for their helpful comments on different stages of this manuscript.

Author details

[1] Aalto University, Department of Neuroscience and Biomedical Engineering, P.O. Box 15100, 00076 AALTO, Espoo, Finland. [2] University of Helsinki, Department of Modern Languages, P.O. Box 24, 00014, University of Helsinki, Helsinki, Finland. [3] Aalto University, Department of Information and Computer Science, P.O. Box 15400, 00076 AALTO, Espoo, Finland. [4] Center for Preservation and Digitisation, National Library of Finland, Saimaankatu 6 50100, Mikkeli, Finland.

Received: 17 January 2014 Accepted: 15 January 2015
Published online: 07 July 2015

References

Alhoniemi, E, Himberg, J, Parhankangas, J, Vesanto, J (2005). SOM toolbox for matlab. http://www.cis.hut.fi/projects/somtoolbox/. Accessed 1.8.2013.

Almuhareb, A (2006). *Attributes in Lexical Acquisition*. PhD thesis, University of Essex.

Baroni, M, & Lenci, A (2011). How we BLESSed distributional semantic evaluation. In S Pado & Y Peirsman (Eds.), *Proc. of EMNLP 2012, Geometrical Models for Natural Language Semantics (GEMS 2011) Workshop* (pp. 1–10). Stroudsburg, PA: Association for Computational Linguistics, (ACL).

Baroni, M, Evert, S, Lenci, A (2008). *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*. Hamburg: Association of Logic, Language and Information (FoLLI).

Baroni, M, Barbu, E, Murphy, B, Poesio, M (2010). Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

Bates, MJ (1986). Subject access in online catalogs: A design model. *Journal of the American society for information science*, *37*(6), 357–376.

Battig, WF, & Montague, WE (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, *80*(3, part 2.), 1–45.

Beckner, C, Blythe, R, Bybee, J, Christiansen, MH, Croft, W, Ellis, NC, Holland, J, Ke, J, Larsen-Freeman, D, Schoenemann, T (2009). Language is a complex adaptive system. *Language learning*, *59*(s1), 1–26.

Blei, DM, Ng, AY, Jordan, MI (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brody, S, & Lapata, M (2009). Bayesian word sense induction, In *Proceedings of the 12th conference of the European Chapter of the ACL* (pp. 103–111). Stroudsburg, PA: Association for Computational Linguistics.

Bullinaria, JA (2012). Semantic category set. http://www.cs.bham.ac.uk/~jxb/Corpus/semcat.txt. Accessed March 8, 2012.

Bullinaria, JA, & Levy, JP (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Bullinaria, JA, & Levy, JP (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, *44*, 890–907.

Caramazza, A, Hersh, H, Torgerson, WS (1976). Subjective structures and operations in semantic memory. *Journal of verbal learning and verbal behavior*, *15*(1), 103–117.

Chen, H (1994). Collaborative systems: solving the vocabulary problem. *Computer*, *27*(5), 58–66.

Chrupała, G (2011). Efficient induction of probabilistic word classes with LDA, In *Proceedings of 5th International Joint Conference of Natural Language Processing* (pp. 363–372). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.

Comon, P (1994). Independent component analysis—a new concept? *Signal Processing*, *36*, 287–314.

Cruse, DA (1986). *Lexical semantics*. Cambridge, UK: Cambridge University Press.

Deerwester, S, Dumais, ST, Furnas, GW, Landauer, TK, Harshman, R (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dinu, G, & Lapata, M (2010). Measuring distributional similarity in context, In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 162–1172). Stroudsburg, PA: MIT, Mass, Association for Computational Linguistics.

Erk, K (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635–653.

Erk, K, & Padó, S (2008). A structured vector space model for word meaning in context, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 897–906). Stroudsburg, PA: Association for Computational Linguistics.

Goddard, C, & Wierzbicka, A (2002). *Meaning and universal grammar: Theory and empirical findings, volume 1*. Philadelphia, PA: John Benjamins Publishing.

Goldstone, RL (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*, 125–157.

Haspelmath, M (2007). Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, *11*(1), 119–132.

Hofmann, T (1999). Probabilistic latent semantic indexing, In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 50–57). New York, NY: ACM.

Honkela, T (1998). Learning to understand - general aspects of using self-organizing maps in natural language processing, In *AIP Conference Proceedings, volume 437* (pp. 563–576). Liege, Belgium: American Institute of Physics (AIP).

Honkela, T, Pulkki, V, Kohonen, T (1995). Contextual relations of words in Grimm tales, analyzed by self-organizing map, In *Proc. of ICANN'95,* volume II, (pp. 3–7). Paris, France: EC2 & Cie.

Honkela, T, Hyvärinen, A, Väyrynen, JJ (2010). WordICA — emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, *16*, 277–308.

Honkela, T, Raitio, J, Lagus, K, Nieminen, IT, Honkela, N, Pantzar, M (2012). Proceedings of IJCNN 2012 International Joint Conference on Neural Networks (pp. 2875–2883): IEEE, (Institute of Electrical and Electronics Engineers).

Hyvärinen, A, & Oja, E (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483–1492. ISSN 0899-7667.

Hyvärinen, A, Karhunen, J, Oja, E (2001). *Independent component analysis*. New York, NY: John Wiley & Sons.

Johnston, RJ (1968). Choice in classification: the subjectivity of objective methods. *Annals of the Association of American Geographers*, *58*(3), 575–589.

Kohonen, T (2001). *Self-Organizing maps*. Heidelberg: Springer.

Kohonen, T, & Honkela, T (2007). Kohonen network. *Scholarpedia*, *2*(1), 1568.

Landauer, TK, & Dumais, ST (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.

Lindh-Knuutila, T, & Honkela, T (2013). Exploratory text analysis: Data-driven versus human semantic similarity judgments, In *Adaptive and Natural Computing Algorithms* (pp. 428–437). Berlin Heidelberg, Germany: Springer.

Lindh-Knuutila, T, Väyrynen, J, Honkela, T (2012). Semantic analysis in word vector spaces with ICA and feature selection, In *Proc. of The 11th Conference on Natural Language Processing (KONVENS)* (pp. 98–107). Vienna, Austria: ÖGAI.

Manning, CD, & Schütze, H (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT press.

McEnery, T (2001). *Corpus linguistics: An introduction*. Edinburgh, UK: Edinburgh University Press.

Miller, GA, & Charles, WG (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Mitchell, TM, Shinkareva, SV, Carlson, A, Chang, K-M, Malave, VL, Mason, RA, Just, MA (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*, 1191.

Murphy, B, Talukdar, P, Mitchell, T (2012). Selecting corpus-semantic models for neurolinguistic decoding, In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM),* SemEval '12, (pp. 114–123). Montréal, Canada: Association for Computational Linguistics.

Niwa, Y, & Nitta, Y (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries, In *Proc. of COLING 1994* (pp. 304–309). Stroudsburg, PA: Association for Computational Linguistics.

Rapp, R (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches, In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1–7). Stroudsburg, PA: Association for Computational Linguistics.

Rauh, G (2010). *Syntactic categories: Their identification and description in linguistic theories*. New York, NY: Oxford University Press.

Ritter, H, & Kohonen, T (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*, 241–254.

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

Schütze, H (1993). Word space, In *Advances in Neural Information Processing Systems 5* (pp. 895–902). San Francisco, CA: Morgan Kaufmann.

Schwering, A (2008). Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS*, *12*(1), 5–29.

Seco, N, Veale, T, Hayes, J (2004). An intrinsic information content metric for semantic similarity in WordNet, In *Proceedings of ECAI 2004* (pp. 1089–1090). Amsterdam, the Netherlands: IOS Press.

Steyvers, M, & Griffiths, T (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Sudre, G, Pomerleau, D, Palatucci, M, Wehbe, L, Fyshe, A, Salmelin, R, Mitchell, T (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, *62*(1), 451–463.

Turney, PD, & Pantel, P (2000). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van Overschelde, JP, Rawson, KA, Dunlosky, J (2004). Category norms: An update and expanded version of the Battig and Montague 1969 norms. *Journal of Memory and Language*, *50*, 289–335.

Venna, J, & Kaski, S (2006). Local multidimensional scaling. *Neural Networks*, *19*(6), 889–899.

Vesanto, J, Himberg, J, Alhoniemi, E, Parhankangas, J (1999). Self-organizing map in Matlab: The SOM toolbox, In *Proceedings of the Matlab DSP conference, volume 99* (pp. 16–17).

Wikimedia Project (2008). The English Wikipedia. http://dumps.wikimedia.org/enwiki. Accessed December 11, 2008. The October 2008 edition used to build the corpus is no longer available for download.

Wilson, AT, & Chew, PA (2010). Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (pp. 465–473). Los Angeles, California.